UiT The Arctic University of Norway

Faculty of Science and Technology
Department of Computer Science

Visual Realism in AI-Driven Virtual Training Environments for Child Investigative Interviews

Pegah Salehi

A dissertation for the degree of Philosophiae Doctor — October 2025



Preface

This doctoral dissertation, entitled "Visual Realism in AI-Driven Virtual Training Environments for Child Investigative Interviews," presents my research contributions as part of the interdisciplinary project "Interview Training of Child-Welfare and Law-Enforcement Professionals Interviewing Maltreated Children Supported via Artificial Avatars." The work described here was conducted within the PhD program in Computer Science at the Faculty of Science and Technology, UiT The Arctic University of Norway, Tromsø, Norway. My research activities have primarily taken place at the Simula Metropolitan Center for Digital Engineering (SimulaMet), particularly within the Department of Holistic Systems. SimulaMet has generously provided all the necessary infrastructure, technological resources, and collaborative opportunities that significantly facilitated the completion of my research. This work was guided by the supervision and expertise of Dag Johansen, Pål Halvorsen, Michael Alexander Riegler, Saeed Shafiee Sabet, and Vajira Thambawita, whose invaluable feedback and mentorship have shaped every phase of the thesis.

Pegah Salehi
Oslo, Norway

Abstract

Protecting children from abuse requires highly skilled professionals who can conduct sensitive investigative interviews. These interviews are often pivotal in legal and welfare responses, particularly as a child's testimony can be the primary source of evidence. Effective training for such professionals requires solutions that are not only scalable and dynamic but also address the visual dimension of simulated experiences, which plays a foundational, yet complex, role. Traditional training methodologies, however, frequently encounter limitations in providing interactive and visually convincing practice environments, thereby restricting opportunities for comprehensive skill refinement. This doctoral research confronts these pressing challenges by systematically investigating how artificial intelligence (AI) can be employed to develop advanced avatar-based systems. The core of this investigation lies in understanding and optimizing the visual components of these systems to create impactful and accessible training tools for this sensitive field.

Motivated by these existing limitations in interviewer training, this research aims to answer the central research question: How can AI be used to develop a visually realistic child avatar-based system for investigative interview training with children?

To comprehensively answer this question, the thesis delineates five principal research objectives. These include: identifying the key challenges inherent in the development of child avatar systems for investigative training; investigating how varying degrees of visual realism in these avatars impact the learning experience and user acceptance; conducting a comparative analysis of 2D versus 3D virtual environments to determine their effectiveness for training child protective service (CPS) professionals; developing and refining talking-head generation models through advanced Audio Feature Extraction (AFE) to optimize technical performance and ensure scalability for practical applications; and examining the benefits of emotionally expressive AI-driven child avatars, along with the feasibility of implementing dynamic facial emotion expression capabilities.

Employing a Design Science Research Methodology (DSRM), this research iteratively designed, developed, demonstrated, and rigorously evaluated an AI-driven child avatar system. Key contributions include a multimodal AI system that integrates Natural Language Processing (NLP), speech-to-text (STT), text-to-speech (TTS), and advanced visual avatar generation techniques to simulate authentic and interactive child responses with a focus on visual fidelity. Empirical findings revealed that while photorealistic avatars generally enhance perceived visual authenticity, stylized avatars were often preferred in conversational contexts to mitigate the "uncanny valley" effect. Beyond this distinction, users valued perceptual consistency and clarity above style, with smooth, stable avatars perceived as more effective for training. Comparative evaluations demonstrated that 3D virtual reality (VR) environments significantly boost visual immersion and emotional resonance, whereas 2D environments offer superior usability and comfort for extended training. Significant technical advancements in AFE for visually synchronized talking-head models drastically reduced latency and improved audio-visual coherence. Furthermore, the integration of emotionally expressive avatars, enabled through technologies such as NVIDIA Omniverse Audio2Face and Unreal Engine, was explored as a means to improve trainees' ability to interpret subtle psychosocial cues. Preliminary findings indicated promising perceptual effects, while also highlighting the key role of audiovisual congruence and the technical challenges involved in achieving emotionally believable avatar interactions.

Collectively, this dissertation establishes a foundational framework for developing more accessible, visually engaging, and ultimately more effective training tools. By advancing the understanding and application of visual realism and interactive capabilities of AI avatars, this research endeavors to improve the skills and preparedness of professionals, thereby enhancing the quality of child testimonies, minimizing trauma during investigative processes, and contributing positively to child welfare outcomes.

Acknowledgements

This research journey has been shaped by a multitude of individuals, and I offer them my deepest thanks. I was fortunate to have a dedicated team of supervisors, and I begin by thanking Professor Pål Halvorsen. His modest nature and consistent availability created an environment where our interactions were both highly productive and genuinely enjoyable. His probing questions were a constant catalyst, encouraging me to think more expansively and critically. I am deeply grateful to Professor Michael Alexander Riegler, another key member of my supervisory team, who taught me the invaluable lesson of believing in my own capabilities. Working with him and witnessing his open-minded and creative approach to research had a significant impact on my own thinking and development. My thanks also go to Professor Dag Johansen, whose supervisory input, though our engagements were brief, offered deep expertise and sage advice that benefited my research. I am especially grateful to Dr. Saeed Shafiee Sabet for his supervision. His support was invaluable in helping me navigate the early, uncertain phases of the project and in finding a clear direction. His positive outlook and holistic problem-solving strategies have had a lasting impact on my work. Rounding out my supervisory team, Vajira Thambawita provided valued technical support and an encouraging presence in meetings. Moreover, I benefited from learning his cautious and thorough approach to research.

To my husband, Mehran, thank you for being my constant support through every phase of this journey. Your encouragement, serene presence, and unwavering faith in me were essential in keeping me grounded and focused, particularly when challenges seemed insurmountable. My dearest parents, Nahid and Hassan, and my wonderful sisters, Saba and Donya, thank you for your unending love and steadfast support. Your presence, even from afar, was a powerful motivator that kept me moving forward.

Lastly, my sincere thanks are extended to the many colleagues and fellow researchers I've had the pleasure of working alongside. The collaborative spirit, insightful discussions, and shared moments have made this entire experience both deeply enriching and truly memorable.

AI Assistance Declaration

Generative AI has been used to improve the overall quality of this work. As the author of the document, I take full responsibility for its content, claims, and references. Specifically, *OpenAI GPT*, *Writefull for Overleaf*, and *GitHub Copilot* have been used for the following purposes:

- As a writing tool to improve grammar, spelling, and clarity of written content. **Example use case:** A draft version of the text was pasted in its entirety, and the model was prompted with: *"Rewrite this text to make the language more lively."* The resulting output was reviewed, and the text was further iterated using additional prompts to improve structure and flow. The final version was fact-checked and partially rewritten by the author(s).
- To expedite repetitive coding tasks, particularly those related to debugging and visualization.

AI was not used to directly generate any original content for this thesis or to complete any unverified code associated with this work.

List of Articles

Article I G. A. Baugerud, M. S. Johnson, R. K. Røed, M. E. Lamb, M. Powell, V. Thambawita, S. Hicks, *P. Salehi*, S. Z. Hassan, P. Halvorsen and M. Riegler (2021).
Multimodal virtual avatars for investigative interviews with children. Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR) - Taipei, Taiwan.
DOI: https://dl.acm.org/doi/10.1145/3463944.3469269

Article II S. Z. Hassan, *P. Salehi*, R. K. Røed, P. Halvorsen, G. A. Baugerud, M. S. Johnson, P. Lison, M. Riegler, M. E. Lamb, C. Griwodz and S. S. Sabet (2022).
Towards an AI-driven talking avatar in virtual reality for investigative interviews of children. Proceedings of the 2nd Workshop on Games Systems (GameSys) - Athlone, Ireland.
DOI: https://dl.acm.org/doi/10.1145/3534085.3534340

Article III *P. Salehi*, S. Z. Hassan, S. S. Sabet, G. A. Baugerud, M. S. Johnson, P. Halvorsen, and M. A. Riegler (2022).
Is more realistic better? a comparison of game engine and gan-based avatars for investigative interviews of children. In Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR) - Newark, NJ, USA . DOI: https://dl.acm.org/doi/10.1145/3512731.3534209

Article IV *P. Salehi*, S. Z. Hassan, M. Lammerse, S. S. Sabet, I. Riiser, R. K. Røed, M. S. Johnson, V. Thambawita, S. A. Hicks, M. Powell, M. E. Lamb, G. A. Baugerud, P. Halvorsen, and M. A. Riegler (2022).
Synthesizing a talking child avatar to train interviewers working with maltreated children. Big data and cognitive computing, vol. 6, no. 2, pp. 62–84.
DOI: https://www.mdpi.com/2504-2289/6/2/62

Article V S. Z. Hassan, S. S. Sabet, M. Riegler, G. A. Baugerud, H. Ko, *P. Salehi*, R. K. Røed, M. S. Johnson and P. Halvorsen (2023).
Enhancing investigative interview training using a child avatar system: a comparative study of interactive environments. Scientific Reports 13, 20403 (2023). DOI: https://www.nature.com/articles/s41598-023-47368-2

Article VI *P. Salehi*, S. Z. Hassan, G. A. Baugerud, M. Powell, M. C. López Cano, M. S. Johnson, R. K. Roed, D. Johansen, S. S. Sabet, M. A. Riegler, and P. Halvorsen (2024).
Immersive Virtual Reality in Child Interview Skills Training: A Comparison of 2D and 3D Environments. In Proceedings of the 16th ACM International Workshop on Immersive Mixed and Virtual Environment systems (MMVE) - Bari, Italy. DOI: https://dl.acm.org/doi/10.1145/3652212.3652219

Article VII *P. Salehi*, S. Z. Hassan, G. A. Baugerud, M. Powell, M. S. Johnson, D. Johansen, S. S. Sabet, M. Riegler and P. Halvorsen (2024).
A theoretical and empirical analysis of 2D and 3D Virtual Environments in Training for Child Interview Skills. IEEE Access, vol. 12, pp. 131842-131864. DOI: https://ieeexplore.ieee.org/document/10634162

Article VIII *P. Salehi*, S. A. Sheshkal, V. Thambawita, S. Gautam, S. S. Sabet, D. Johansen, M. A. Riegler, and P. Halvorsen (2024).
Comparative Analysis of Audio Feature Extraction for Real-Time Talking Portrait Synthesis. Big data and cognitive computing, vol. 9, no. 3, pp. 59–80. DOI: https://www.mdpi.com/2504-2289/9/3/59

Article IX *P. Salehi*, S. A. Sheshkal, V. Thambawita, M. A. Riegler, and P. Halvorsen (2025).
Multimodal Integration Challenges in Emotionally Expressive Child Avatars for Training Applications. arXiv preprint arXiv:2506.13477v2.
DOI: https://doi.org/10.48550/arXiv.2506.13477

Contents

AI Declaration **vii**

Acronyms **xv**

I Research Overview **1**

1 Introduction **3**

1.1 Background and Motivation 3

1.1.1 The Importance of Child Testimonies 3

1.1.2 Obstacles and Gaps in Interviewer Training 4

1.1.3 Existing Interview Training Systems 5

1.1.4 Potentials of Technology and AI in Interviewer Training 7

1.1.4.1 Visual Avatars 7

1.1.4.2 Dialogue Generation 10

1.2 Research Scope and Objectives 11

1.3 Research Methodology 15

1.4 Ethical Concerns and Considerations 19

1.4.1 Avoiding Gender Stereotypes 20

1.4.2 Deepfakes Technology 21

1.4.3 VR Technology 22

1.4.4 Practical Use of AI Avatars 22

1.5 Main Contributions 23

1.6 Other contributions 25

1.7 Thesis Outline 26

2 Research Contributions and Discussion **29**

2.1 Challenges in Child Avatar Creation 30

2.2 Visual Realism and User Acceptance 33

2.3 2D vs. 3D: Comparative Effectiveness 39

2.4 Refining Talking-Head Models 42

2.5 Emotional Expressiveness in AI Avatars 45

2.6 AI-Driven Visual Avatars for Effective Training 54

3 Conclusion and Future Work **57**

3.1 Conclusion 57

3.2 Future Work 59

3.3 Final Remarks 61

Bibliography **63**

II Research Papers **83**

Article I: Multimodal Virtual Avatars for Investigative Interviews with Children **85**

Article II: Towards an AI-driven Talking Avatar in Virtual Reality for Investigative Interviews of Children **93**

Article III: Is More Realistic Better? A Comparison of Game Engine and GAN-based Avatars for Investigative Interviews of Children **101**

Article IV: Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children **111**

Article V: Enhancing Investigative Interview Training Using a Child Avatar System: A Comparative Study of Interactive Environments **135**

Article VI: Immersive Virtual Reality in Child Interview Skills Training: A Comparison of 2D and 3D Environments **153**

Article VII: A Theoretical and Empirical Analysis of 2D and 3D Virtual Environments in Training for Child Interview Skills **161**

Article VIII: Comparative Analysis of Audio Feature Extraction for Real-Time Talking Portrait Synthesis **185**

Article IX: Multimodal Integration Challenges in Emotionally Expressive Child Avatars for Training Applications **207**

Acronyms

AFE Audio Feature Extraction iii, iv, xvii, 14, 42–45, 54, 58, 60, 185

AI artificial intelligence iii, iv, vii, 7, 11, 17, 45, 61, 93

CPS child protective service iii, 12, 13, 16, 18, 19, 22, 26, 31, 32, 55, 58, 59, 111

CSA Child sexual abuse 3, 4, 20, 22

DSRM Design Science Research Methodology iv, xvii, 15, 16, 18, 19, 26

GAN Generative Adversarial Network 9, 25, 30, 33, 101

LLM Large Language model xviii, 7, 10, 11, 17, 48

NICHD The National Institute of Child Health and Human Development 4, 5

NLP Natural Language Processing iv, 6, 10, 11, 13, 26, 32, 45, 54

QoE Quality of experience 32, 35, 36, 40, 135

RO research objective xvii, 12, 15, 17, 18, 23, 24, 27, 29, 39, 54, 57

SEM structural equation modeling 18, 37, 40, 46

STT speech-to-text iv, xvii, xviii, 7, 11, 13, 17, 26, 30–32, 43, 44, 48, 54

TTS text-to-speech iv, xvii, xviii, 7, 11, 13, 17, 26, 30–32, 43, 44, 48, 54

VR virtual reality iv, xi, xviii, 6, 7, 22, 25, 38–40, 48, 54, 61, 93, 135

List of Figures

1.1 Architecture of the AI-driven training system for investigative interview training with children. The blue boxes represent the specific tools and methods used for each component. The orange boxes within the "Visual Interactive Environment," which is the main focus of this thesis, highlight key features of the visual component that are systematically explored throughout the work. 12

1.2 Overview of the DSRM process as applied in the development of the AI-driven child avatar training system. 16

1.3 Development pathway of the visual component in the training system. The diagram highlights the specific research objectives (ROs) and associated research articles at each stage, illustrating how each step contributed to the overall research trajectory. Green boxes indicate key challenges or limitations that motivated transitions to subsequent developmental phases. 24

2.1 a) The architecture of the proposed child avatar prototype includes: the Visual Module (frontend built with Unity 3D and Oculus Quest 2), the Auditory Module (IBM Watson APIs for STT and TTS), and the Language Module (backend dialogue system using Rasa). b) Screenshots of the virtual environment demonstrating user interviews with the child avatar. 31

2.2 Representative examples of avatar appearance across different realism levels, ranging from stylized to photorealistic. 34

2.3 System architecture and user interaction with the interactive child avatar. . 43

2.4 Execution time comparison of different AFE models, including Deep-Speech [1], Wav2Vec [2], HuBERT [3], and Whisper [4]. 44

2.5 Architecture for real-time emotionally expressive child avatars. PC1 processes speech input (Google Cloud STT), generates responses (Large Language model (LLM)), synthesizes speech (Amazon Polly TTS), and extracts emotional parameters (NVIDIA Audio2Face) to drive facial animation. PC2 renders a customizable Metahuman avatar (Unreal Engine) driven by this real-time emotional data via Wi-Fi. The Live Link plugin ensures synchronization, and Metahuman's anatomical rigging enables nuanced expressions. Supports desktop and VR environments. 48

2.6 Example frames of the two avatars used in the user study: Amelia (female-presenting) and Emory (male-presenting). The top row displays their neutral facial expressions. The bottom rows show expressions of joy (left) and sadness (right) for each avatar, as used in the study. 50

2.7 Mean ratings ($M \pm SD$) of emotional expressiveness across five emotions (Angry, Sad, Joy, Fear, Disgust) for the avatars *Emory* and *Amelia*, under both audio+visual and visual-only conditions. Bars represent group means; values next to each bar report the mean and standard deviation, while horizontal error bars indicate 95% confidence intervals (CI). Each video clip is labeled on the left, with color-coded bars distinguishing the experimental conditions. Green-highlighted boxes indicate target emotions critical for the study's recognition task. 51

Part I

Research Overview

Chapter 1

Introduction

1.1 Background and Motivation

Imagine a classroom where one in five children carries the invisible scars of physical or sexual abuse. Around the world, more than one billion children are exposed to violence that can irrevocably alter the course of their lives [5, 6]. Child abuse and neglect are not just individual tragedies. They ripple outward, profoundly affecting mental health, increasing the risk of substance abuse, compromising educational success, diminishing productivity, and in the worst cases, leading to premature mortality [7, 8]. This epidemic of maltreatment underscores an urgent need to better understand, prevent, and intervene in the lives of vulnerable children worldwide.

Beyond these impacts, child abuse imposes a staggering economic burden. The costs associated with healthcare, social services, and lost productivity amount to hundreds of billions of dollars annually. In the United States alone, the lifetime economic impact of child maltreatment is estimated at $124 billion for a single year of confirmed fatal and non-fatal cases [9]. Notably, the long-term financial burden for survivors is comparable to that of chronic conditions such as Type 2 diabetes [10]. These figures highlight the need for effective prevention and intervention strategies worldwide to mitigate the prevalence and impact of child abuse.

1.1.1 The Importance of Child Testimonies

Child sexual abuse (CSA) poses unique challenges for both victims and investigators. Unlike other forms of abuse, CSA cases rarely include corroborative physical evidence,

with fewer than 15% of cases supported by such data [11]. Additionally, a significant proportion of CSA cases go unreported, with research indicating that between 55% and 70% of individuals who experienced CSA delay disclosure until adulthood [12]. Moreover, in approximately 70% of CSA cases, the child is the sole witness, making their testimony essential for both investigation and prosecution [13, 14]. As a result, the first interview with a child plays a key role in the investigative process, as errors at this stage are often difficult to correct later. The child's testimony is a fundamental component of criminal investigations in alleged CSA cases. However, conducting interviews is among the most complex tasks faced by forensic psychologists and law enforcement officers. Research has shown that following recommended questioning techniques increases the likelihood of obtaining reliable information from children, whereas ineffective questioning can compromise the accuracy of their statements [15].

1.1.2 Obstacles and Gaps in Interviewer Training

Research indicates that professionals conducting interviews with children in suspected abuse cases often fail to adhere to best-practice guidelines. These guidelines emphasize the use of open-ended questions designed to elicit detailed, accurate, and reliable information from child witnesses [16, 17]. However, in practice, interviewers frequently rely on suggestive, closed, and directive questions, which reveals a significant gap between recommended techniques and actual interviewing practices [18, 19, 20, 21]. In response, various training programs have been developed to improve interviewing techniques. Traditional methods, such as lectures, discussions, and role-playing exercises, have generally not yielded satisfactory improvements in interview quality [22].

Providing feedback on real-life interviews has been shown to lead to increased use of open-ended questions, particularly when combined with the protocol The National Institute of Child Health and Human Development (NICHD) [23]. However, legal and organizational barriers often make reviewing actual case interviews challenging [24]. Additionally, while process feedback focusing on question types is possible, outcome feedback evaluating the accuracy of interview conclusions is limited due to the lack of verifiable ground truth [25].

Mock interviews offer a solution by allowing both process and outcome feedback. Since ethical concerns prevent children from participating, trained adult actors are used to sim-

ulate realistic child responses. Effective training programs for these actors require extensive preparation, such as the 25-hour training over 12 weeks provided to postgraduate students with prior interviewing experience [26, 27]. Research suggests that interviews conducted with well-trained actors lead to improved adherence to structured protocols like NICHD [26, 27, 28, 29]. However, the extent to which actors' responses truly reflect children's natural behavior remains uncertain, raising concerns about the validity of such training [30, 31]. Therefore, given the critical importance of high-quality investigative interviews in child abuse cases and the scalability challenges of resource-intensive actor-based training programs, there is an urgent need for more effective, accessible, and automated training solutions [32].

1.1.3 Existing Interview Training Systems

Technology is rapidly transforming how we learn and train, especially in complex fields like psychology [33, 34, 35]. Consequently, the integration of technology into skill development has revolutionized training methodologies across various disciplines, including investigative interviewing. Traditional approaches often lack scalability, adaptability, and real-time feedback, prompting the exploration of digital alternatives. Among these, computer-based simulations have emerged as a promising solution, offering structured yet flexible environments for trainees to refine their questioning techniques.

Recent advancements in eLearning platforms demonstrate the potential of technology-assisted training. For instance, self-paced digital programs, such as those implemented at the Griffith Institute's Center for Investigative Interviewing, utilize branching decision structures to guide trainees through simulated interviews [36]. These systems assess users' question selections and provide feedback on optimal strategies, improving interview quality over time [37, 38]. However, such platforms remain constrained by their static design, limiting exposure to the dynamic, unpredictable nature of real-world interactions. Additionally, they do not allow trainees to practice verbalizing questions which is a important skill for conducting actual interviews.

Early systems like the one developed by Linnæus University and AvBIT Labs employed prerecorded child responses manually curated by human operators to simulate conversational exchanges during interviews [39]. Though groundbreaking for its time,

this semi-automated approach limited authentic dialogue by relying entirely on static, predefined reply options—a constraint that hindered adaptive responsiveness.

Subsequent research efforts introduced LiveSimulation [40, 41], a system enabling simulated forensic interview practice through interaction with a videotaped developmentally representative five-year-old child avatar. In this paradigm, trainees select from four predefined inquiry options, to which the system responds with video segments meticulously designed to mirror age-typical verbal and non-verbal response patterns observed in cases of alleged sexual abuse. Empirical evidence demonstrates that this scenario-based training produces a statistically significant increase in practitioners' adoption of open-ended questioning techniques, with longitudinal efficacy confirmed through 12-month follow-up assessments showing maintained competency levels [22, 36].

Similarly, the Empowering Interviewer Training (EIT) system [42] employs a virtual child avatar, constructed from animated, morphed images of real children, to simulate dynamic interactions where participants can freely pose verbal questions. While EIT uses a probabilistic rule-based algorithm to generate contextually appropriate responses from predefined video clips, its reliance on manual question categorization introduces scalability limitations [43, 44]. After each session, comprehensive feedback on both interview techniques and outcomes is provided, reinforcing learning effectiveness [43, 42, 45].

Building on these foundations, systems like ViContact [43] use virtual reality (VR) to simulate child interviews in abuse-related scenarios. It allows trainees to conduct simulated verbal interactions with virtual child avatars in realistic abuse scenarios, combining VR interactions with automated, personalized feedback and traditional seminar training. While this integrated approach enhances trainees' use of open-ended questions and socio-emotional support skills, opportunities remain to integrate modern Natural Language Processing (NLP) and AI-driven character generation for create more lifelike, adaptive, and context-aware interactions, potentially refining the responsiveness and adaptability of virtual avatars (VAs) in complex training environments.

While current systems demonstrate progress, they remain constrained by semi-automated architectures requiring human intervention, static visual assets, and limited dynamic interaction capabilities. Reliance on pre-recorded responses and fixed emotional expressions results in reduced realism, diminished trainee immersion, and restricted skill transfer to real-world contexts. To address these gaps, our proposed AI-driven system inte-

grates advanced visual generation technologies to produce photorealistic avatars with real-time, speech-synchronized animations. These are coupled with Large Language models (LLMs), speech-to-text (STT), and text-to-speech (TTS) systems to enable fully automated, context-aware dialogues. By automating non-verbal cues (e.g., micro-expressions, adaptive gaze) within immersive VR environments, the system enhances visual fidelity, scalability, and engagement while minimizing operational reliance on human actors. This innovation establishes a cost-effective, high-fidelity training benchmark, directly addressing the urgent need for dynamic, lifelike simulations in child abuse investigations through authentic, responsive interactions that mirror real-world scenarios.

1.1.4 Potentials of Technology and AI in Interviewer Training

Advances in artificial intelligence (AI) and immersive technologies are revolutionizing educational methodologies, offering scalable, adaptive, and cost-effective solutions for skill development. These innovations hold transformative potential for forensic interview training by enabling realistic simulations of interactions with children, thereby allowing professionals to refine their questioning techniques in a safe, structured setting. Below, we explore the technological landscape through two key dimensions: visual avatars and dialogue generation, with a focus on the former as the central pillar of this PhD research.

1.1.4.1 Visual Avatars

Technological advancements have streamlined the creation of virtual avatars, making them increasingly efficient and widely accessible in computer games, online chats, virtual meetings, and social media platforms. Modern tools, including high-quality rendering engines* and deep-learning-based technologies known as Talking Head Generation models [46], have significantly improved the fidelity of these digital counterparts compared to those developed in previous decades. Today, avatars not only exhibit lifelike visual qualities but also replicate natural human movements and cognitive behaviors, narrowing the gap between virtual and real-world interactions.

High realism (often termed high-fidelity) is widely regarded as a beneficial design approach for virtual avatars [47]. Studies suggest that users prefer interacting with human-like avatars over non-human representations [48, 49]. Such avatars are shown to improve

*https://www.unrealengine.com/en-US/metahuman

user engagement [50] and enjoyment by fostering a sense of familiarity, which can evoke psychological responses akin to real-world interactions. Thus, visually realistic avatars are critical for creating immersive experiences, and their ability to simulate lifelike interactions makes them particularly well-suited for developing automated training systems for conducting forensic interviews with children.

Conceptual Foundations of Human Perception of Avatars The appearance of virtual avatars can portray diverse impressions to audiences, such as cartoon-like emoticons or photorealistic humanoids [51]. The visual fidelity of virtual avatars significantly influences users' perception and interaction experiences within virtual environments. Visual fidelity is defined by three criteria: anthropomorphism, realism, and truthfulness [50, 52].

- Anthropomorphism refers to the degree to which avatars mimic human morphology, directly influencing users' body schema and identity adaptation—a phenomenon termed the Cyborg Dilemma [53]. While highly anthropomorphic avatars can amplify embodiment, incongruent designs risk distorting users' mental models [54]; however, neural adaptability enables interaction even with non-humanoid forms [55].

- Realism, particularly perceptual realism [56], involves balancing photorealistic detail with stylization. Inconsistencies in rendering (e.g., mismatched textures or proportions) may reduce appeal and induce eeriness [57]. Paradoxically, hyperrealistic avatars may not inherently enhance user experience; instead, they may provoke discomfort or eeriness, as excessive fidelity can heighten expectations of human-like behavior. Applications prioritizing user comfort or creativity, for instance, often opt for stylized or cartoon-like avatars, sidestepping the pitfalls of hyperrealism while fostering engagement through abstraction or artistic expression. This balance underscores that technical precision alone cannot dictate effectiveness; context and user needs remain paramount [58].

- Truthfulness is the degree of similarity between users' appearance and virtual characters [59, 60]. Crucially, mismatches between visual fidelity and behavioral realism disrupt co-presence, underscoring the necessity for coherence across design dimensions [61].

This need for coherence connects directly to the concept of the uncanny valley, first described by Mori [62], wherein increased avatar realism paradoxically evokes negative emotional responses if it approaches, but does not fully achieve, lifelike authenticity [62]. While users initially demonstrate greater affinity toward avatars with increased human-likeness [48, 49], a critical threshold exists beyond which small imperfections in appearance or motion suppress engagement and elicit eeriness [63].

Despite ongoing debates about the universality of these effects [64], high-fidelity, human-like avatars remain critically important in domains such as medical training [65] and spatial perception [66] within virtual reality environments, where enhanced realism significantly improves user experiences in interactive scenarios.

Our goal of developing a child avatar for investigative interviewing requires guiding avatar design toward appropriate fidelity levels tailored to specific user needs and application contexts. This involves balancing realism with contextual goals, ensuring behavioral alignment to mitigate perceptual disruptions, and leveraging stylization to enhance user experience.

Visual Design Techniques for Avatar Creation The visual design of virtual avatars can be strategically tailored to application-specific objectives through two primary approaches: stylized or photorealistic representations. Stylized designs employ artistic abstraction to mitigate uncanny valley effects and enhance user comfort, while photorealistic avatars prioritize anatomical accuracy through high-fidelity modeling to maximize realism. Methodologies for avatar creation broadly fall into two categories: AI generative-based techniques (e.g., talking-head generation models for dynamic facial synthesis) and engine-driven pipelines (e.g., game engines for customizable avatars).

Talking-head generation models, such as those leveraging Generative Adversarial Networks (GANs) [67] or Neural Radiance Fields (NeRF) [68], enable adaptive animations with emotional expressiveness. However, they often face challenges in real-time performance and computational efficiency. In contrast, engine-driven platforms like Unity [69] and Unreal Engine's MetaHuman [70] excel in delivering interactive, real-time avatars within immersive VR environments, leveraging tools like NVIDIA Omniverse [71] to automate emotionally responsive animations, though they may lack the adaptive fidelity of AI-generated animations.

The design of child avatars for investigative interview training with children demands a careful balance between technical feasibility and psychological realism. Through objective metrics (e.g., rendering speed, synchronization accuracy) and subjective evaluations (user feedback on realism and engagement), we systematically determine the optimal design strategy, ensuring avatars meet the fidelity and interactivity demands of effective, scalable training.

Operational Definitions of Visual Realism Levels Operational Definitions of Visual Realism Levels To enhance clarity and consistency throughout the thesis, the following definitions are adopted for avatar visual realism: *Stylized*: Avatars that use artistic abstraction and do not attempt to mimic real human appearance (e.g., cartoon-like or painted visuals). *Semi-realistic*: Avatars that balance stylized design with realistic proportions and features, but avoid hyper-detailed textures to prevent uncanny effects. *Photorealistic*: Avatars designed to closely replicate the appearance of real human beings with high fidelity in texture, lighting, and facial detail. *Human-like*: A general term referring to avatars with anthropomorphic features, including all of the above categories, depending on context.

1.1.4.2 Dialogue Generation

NLP plays a foundational role in empowering dynamic, context-aware dialogue generation for realistic child avatars in investigative interview training with children. Modern NLP systems, particularly LLMs, such as GPT-4, offer significant potential to simulate interactive conversations that adapt to trainees' questions, moving beyond the rigid, scripted interactions of traditional training tools [72, 73]. These technologies allow avatars to generate responses in real time, mimicking the unpredictable nature of real-world child interviews while adhering to age-appropriate language patterns.

However, existing NLP models face inherent challenges when applied to child-specific contexts. For instance, systems trained predominantly on adult language corpora may struggle to replicate the lexical simplicity, emotional nuance, and stress-induced disfluencies (e.g., hesitations, repetitions) characteristic of children's speech under trauma [74, 75]. Emotion recognition in conversations (ERC) and trauma-informed dialogue generation also remain active research frontiers, as text-only datasets often fail to capture the psy-

chosocial complexities of child abuse disclosures [76, 77]. Addressing these gaps—such as refining child-centric lexicons, integrating disfluencies for naturalness, and improving real-time feedback mechanisms—requires specialized adaptation of NLP frameworks.

In this project, these challenges are addressed through collaborative efforts, with dedicated focus on optimizing NLP components for child avatar interactions. For example, our team has worked to fine-tune LLMs and implement rule-based filters to better align generated dialogue with developmental stages and emotional states. While advancing core NLP methodologies falls outside the primary scope of this dissertation, the integration of these adapted models demonstrates their viability in creating immersive training environments. By leveraging existing NLP technologies and addressing their limitations through targeted adjustments, the system achieves a balance between flexibility, realism, and practical utility which are key requirements for effective interviewer training.

1.2 Research Scope and Objectives

Traditional approaches and earlier child avatar-based simulations have faced significant challenges related to realism, adaptability, and scalability, revealing a clear gap in current practice. Motivated by these limitations, this research aims to develop an AI-driven training system that uses realistic, interactive child avatars to simulate real-life scenarios. By incorporating visually and emotionally realistic avatars, the system seeks to enhance training quality, helping professionals build the skills and confidence required to conduct sensitive interviews with care and precision. Therefore, this thesis seeks to answer the following central research question:

> ***How can AI be used to develop a visually realistic child avatar-based system for investigative interview training with children?***

As illustrated in Figure 1.1, the proposed system integrates multiple AI components, including STT and TTS processing, NLP, and a visually dynamic child avatar. Developing a realistic AI-driven child avatar training system has been a complex and evolving process. The system's development was guided by continuous evaluation and iteration, with insights from user studies and technical performance reviews informing strategic adjustments in both design and implementation. A key challenge has been achieving a balance between visual and emotional realism, narrative coherence, and natural interaction flow

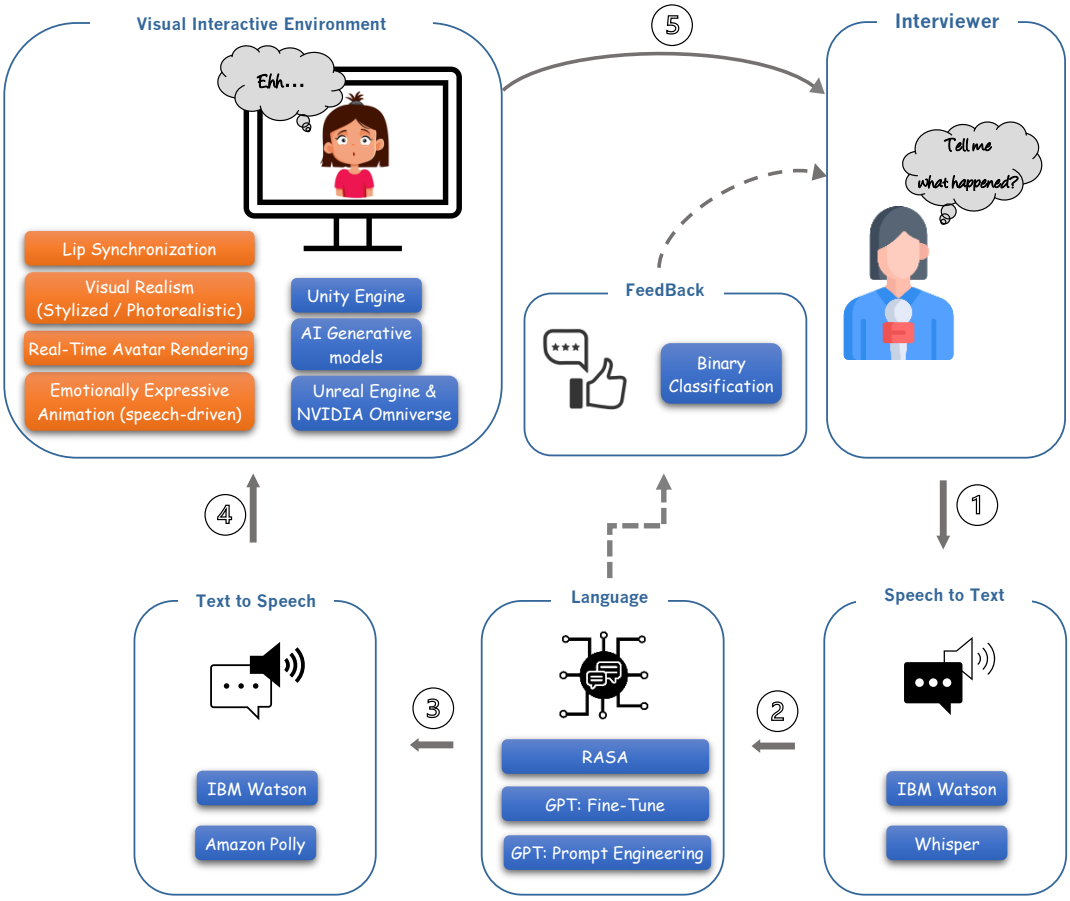


Figure 1.1: Architecture of the AI-driven training system for investigative interview training with children. The blue boxes represent the specific tools and methods used for each component. The orange boxes within the "Visual Interactive Environment," which is the main focus of this thesis, highlight key features of the visual component that are systematically explored throughout the work.

while maintaining an effective learning environment for child protective service (CPS) professionals. Understanding how professionals engage with and respond to the child avatar system is essential for optimizing the learning experience. Comparative studies between different training approaches, such as 2D versus 3D environments, offer valuable insights into the most effective strategies for improving interview skills. These findings have directly shaped the research objectives (ROs) and technological solutions presented in this thesis, ensuring that the system is both technically robust and practically relevant.

The ROs guiding this study are defined as follows:

RO-1 ***Identify the key challenges and considerations in developing the visual components of an AI-driven child avatar system for training professionals to conduct investigative interviews with children, with addi-***

tional attention to multimodal integration. This objective aims to identify and analyze the technical, emotional, and practical hurdles involved in creating a realistic, interactive, and emotionally responsive AI-driven child avatar capable of effectively simulating abused children during forensic interviews. As depicted in Figure 1.1, the study will explore the integration of multimodal technologies (e.g., NLP, TTS, STT, and visual avatar generation), evaluate the feasibility of developing an interactive and adaptive system, and define the key challenges that require further attention to enhance the system's realism, which refers to the visual and behavioral fidelity of the avatar compared to real children, and its responsiveness, meaning how naturally and promptly the system reacts to user input.

RO-2 ***Investigate the impact of varying levels of visual realism of child avatars on the learning experience and user acceptance.*** Varying degrees of visual realism of child avatars can significantly influence how learners perceive and engage with the training environment, as supported by recent research showing that avatar realism affects user perception and immersion in virtual environments [78]. Building on the conceptual foundations of avatar perception outlined in Section 1.1.4.1, this objective empirically evaluates how stylized, semi-realistic, and hyperrealistic avatars influence trainee engagement, confidence, and satisfaction. Theoretical frameworks suggest that overly realistic avatars may evoke discomfort due to subtle imperfections [62], whereas stylized designs can enhance immersion by balancing familiarity with abstraction [47]. This study directly tests these hypotheses in the context of child interview training.

By comparing different avatar configurations, the research quantifies trade-offs between visual fidelity and user acceptance. For example, it examines whether reducing realism to avoid the uncanny valley increases trainees' willingness to adopt the technology.

RO-3 ***Conduct a comparative analysis of the effectiveness, usability, and perception of 2D versus 3D virtual environments for training CPS professionals.*** Comparing 2D and 3D virtual environments provides valuable insights into the ways in which different levels of immersion and interactivity shape the learning experience for CPS professionals. While 2D interfaces typically require

fewer resources and simpler setups, they may reduce the perceived sense of presence and realism. In contrast, 3D environments can simulate more lifelike scenarios, potentially enhancing engagement and skill retention. However, they also demand greater technological support and present additional usability considerations.

By systematically examining these two formats, practitioners can better understand how to balance fidelity, accessibility, and user comfort to optimize training outcomes. This comparison also complements **RO-2** by exploring how different spatial formats (2D vs. 3D) shape users' perception of realism and immersion, thereby contributing to a deeper understanding of the level of visual fidelity required for effective child avatar-based training.

RO-4 ***Develop and refine talking-head generation models to improve performance, responsiveness, and scalability in real-time training environments.*** This objective addresses key bottlenecks in real-time avatar systems, where inefficient audio processing introduces latency, limits scalability, and disrupts synchronization. The research focuses on developing lightweight, adaptive Audio Feature Extraction (AFE) methods to enhance responsiveness and realism in AI-driven avatars. The aim is to integrate these models effectively into resource-limited settings, such as child investigative interview training, while maintaining a balance between computational efficiency and perceptual realism.

RO-5 ***Investigate and implement a technical framework for generating emotionally expressive child avatars by integrating audio-driven animation with real-time rendering, thereby establishing the foundational capability to simulate nuanced psychosocial cues for future training evaluations.*** This objective explores the potential benefits of incorporating emotional expressiveness into AI-driven child avatars used in investigative interview training with children. The ability to simulate emotional states, such as sadness, anxiety, or discomfort, may contribute to a more immersive and psychologically realistic interaction experience. Dynamic facial emotion expression could help trainees better recognize and respond to psychosocial cues, potentially fostering empathy and situational awareness.

The study will examine the feasibility of generating lifelike emotional expressions in real time, considering technical challenges such as facial animation synchronization and the interpretation of affective input signals. By evaluating these aspects, this research aims to assess whether emotional expressiveness can meaningfully enhance the perceived realism and emotional credibility of child avatars in this training context.

Collectively, these ROs establish a structured and interconnected framework for investigating how AI can support the development of realistic and interactive child avatars for investigative interview training with children. **RO-1** identifies the core technical, emotional, and practical challenges involved in avatar development. **RO-2** and **RO-3** examine key perceptual and contextual factors, such as visual realism and the immersive qualities of 2D versus 3D environments, that influence user engagement. **RO-4** focuses on improving the system's responsiveness and scalability through more efficient audio-visual integration. Finally, **RO-5** explores the potential value and feasibility of incorporating emotionally expressive behaviors to enhance the realism and perceived credibility of the avatars.

1.3 Research Methodology

Selecting an appropriate research methodology is essential for effectively achieving the study objectives. Various paradigms exist in computing research, including qualitative and quantitative methods [79], the Hypothetico-Deductive Method [80], experimental prototyping, and artifact-based approaches such as the ACM's theory-abstraction-design framework [81]. While each offers distinct strengths, Design Science Research Methodology (DSRM) [82] was selected for this study due to its structured yet iterative approach to developing and evaluating socio-technical artifacts. Unlike purely theoretical methods or isolated experimental prototyping, DSRM integrates problem-solving, design rigor, and empirical validation within a cohesive framework. This methodology is particularly suited to developing AI-driven solutions for child interview training, where iterative refinement through experimental prototyping and practical relevance are essential to addressing complex socio-technical challenges.

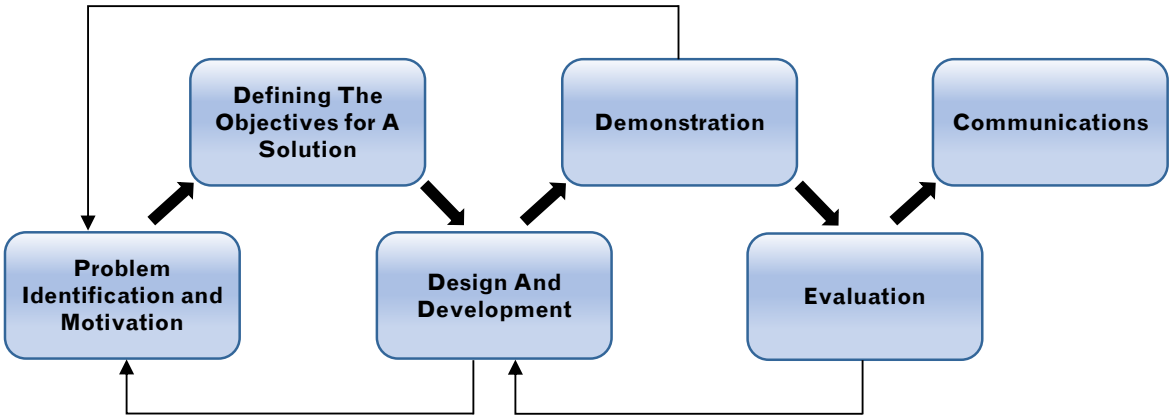


Figure 1.2: Overview of the DSRM process as applied in the development of the AI-driven child avatar training system.

The DSRM framework used in this study, proposed by Peffers et al. [82] and aligned with the principles outlined by Hevner et al. [83], emphasizes the creation and evaluation of artifacts to solve real-world problems and enhance existing practices. It is especially applicable to fields such as information systems, engineering, and management. This methodology is well-suited to the goals of this thesis, which aims to develop innovative solutions for challenges in adopting AI-driven child interview training for CPS professionals.

The research strategy combines prototyping, simulation, and evaluation, enabling iterative refinement and assessment of the artifact. Data collection methods include user studies and interviews to gather feedback, complemented by a systematic literature review to inform the design and validate the findings. These methods were selected for their ability to provide diverse insights into user experiences and their relevance to artifact development and evaluation. Both quantitative and qualitative data analysis techniques were employed to ensure a comprehensive and robust understanding of the results.

The DSRM framework was chosen as the overarching methodology due to its problem-solving focus and emphasis on the creation of practical solutions. It bridges theoretical and practical aspects, offering a holistic approach to addressing research challenges. As illustrated in Figure 1.2, the DSRM process structured this thesis into six phases:

- **Problem Identification and Motivation:** Current training tools for child investigative interviews face significant limitations in terms of scalability, interactivity, and achieving an optimal level of visual realism. Mock interviews with trained

actors, while valuable, are resource-intensive and difficult to scale. Meanwhile, existing virtual training systems often lack immersive environments and fail to convey nuanced nonverbal cues essential for effective communication with children. This study addresses these challenges by developing an AI-driven training platform that integrates visually and emotionally realistic child avatars with advanced dialogue systems. The goal is to enhance the accessibility, interactivity, and emotional depth of interviewer training, ultimately supporting professionals in developing the skills required for sensitive and effective child interviews.

- **Defining Design Objectives:** This study defines five key ROs to guide the development of an effective AI-driven training system for child investigative interviews: (1) to identify the core technical, emotional, and practical challenges involved in designing the visual components of a child avatar system; (2) to investigate how varying levels of visual realism in avatars influence user experience, learning outcomes, and acceptance; (3) to conduct a comparative evaluation of 2D and 3D virtual environments in terms of training effectiveness and usability; (4) to develop and optimize talking-head generation models that improve performance, responsiveness, and scalability in real-time applications; and (5) to examine the benefits of incorporating emotionally expressive avatars and assess the feasibility of implementing dynamic facial emotion expressions. Collectively, these objectives address the overarching challenge of creating realistic, scalable, and emotionally engaging avatar-based simulations to enhance interviewer training.

- **Design and Development:** The system was developed through an iterative design process that combined recent advancements in AI with user-centered design principles. The system architecture integrates several key technical components: (1) avatar animation modules to achieve context-appropriate visual realism; (2) speech processing services for generating natural, childlike vocal responses through STT and TTS; and (3) a fine-tuned LLM designed to produce dynamic, age-appropriate dialogue that reflects the linguistic patterns and emotional states typical of children in investigative interview contexts.

- **Demonstration and Evaluation:** The AI-driven child avatar system was demonstrated to child protection and law enforcement professionals through a combination

of technical benchmarking and structured user studies, enabling practical interaction with the artifact. Evaluation employed a mixed-methods approach. Quantitative analyses included ANOVA for comparative assessments, mixed-effects models to investigate factors such as visual realism and environment type (2D/3D), and structural equation modeling (SEM) to analyze relationships within user experience dimensions. In addition, technical performance was assessed using a range of evaluation metrics addressing speed, visual fidelity, lip-sync accuracy, system efficiency, and perceptual quality. Qualitative data gathered from user studies and semi-structured interviews provided insights into usability, realism, emotional engagement, and training relevance. Feedback from these evaluations informed iterative refinements, enhancing both the effectiveness and real-world applicability of the system in the domain of investigative interview training involving children.

- **Communication:** Research findings and insights were actively disseminated through peer-reviewed publications, academic conferences, workshops, and professional demonstrations. Effective communication facilitated engagement with key stakeholders, including CPS workers, law enforcement agencies, and training institutions. These efforts have supported broader awareness, stakeholder adoption, and the ongoing implementation of the AI-driven child avatar system in Norway. Continuous feedback from stakeholders has guided iterative refinements, fostering sustained interest and helping to ensure the system remains relevant and effective for real-world training in conducting investigative interviews with children.

The DSRM framework was adopted as the overarching methodology to guide the development, evaluation, and refinement of the AI-driven child avatar training system. As illustrated in Figure 1.2, the DSRM process was applied to guide the development of the child avatar system. Each phase was aligned with specific ROs, enabling a systematic and iterative approach to addressing key challenges and evaluating the system's effectiveness. Notably, the application of DSRM was not strictly linear. Thick arrows represent the planned development flow, while thin arrows indicate iterative loops that emerged at various stages. These iterations were often driven by feedback from users and CPS professionals, resulting in continuous refinements and enhancements to the system.

The structured nature of DSRM played a pivotal role in achieving the ROs by providing a coherent framework for both development and evaluation. Its iterative design

allowed for ongoing refinement of the child avatar system, addressing critical challenges in visual and functional design (**RO-1**). Rigorous evaluation phases supported investigations into the impact of visual realism on user acceptance and learning outcomes (**RO-2**), as well as comparative assessments of 2D and 3D virtual environments for CPS training (**RO-3**). The demonstration and evaluation stages were also essential for validating the technical performance and scalability of the talking-head generation models (**RO-4**), and for examining both the benefits and the feasibility of integrating emotional expressiveness into AI-driven avatars (**RO-5**). Lastly, the Communication phase focused on disseminating research findings, supporting broader awareness and adoption of AI-based training tools for investigative interviews. By adhering to the structured phases of DSRM, the research maintained a coherent trajectory while upholding scientific rigor throughout the project.

1.4 Ethical Concerns and Considerations

The development of AI-driven child avatars for training professionals to conduct investigative interviews with children raises complex ethical considerations that demand systematic scrutiny and well-defined mitigation strategies. Our research has comprehensively addressed these ethical dimensions through the establishment of a multi-layered ethical framework, reinforced by stringent protocols and validated through multiple institutional approvals.

The ethical foundation of this work was formally recognized through approvals granted by the Norwegian Agency for Shared Services in Education and Research (SIKT) for two interrelated research initiatives. The primary project (reference number #614272), formally designated as "Interview Training of Child-Welfare and Law-Enforcement Professionals Interviewing Maltreated Children Supported via Artificial Avatars," implemented comprehensive data governance protocols. These measures were specifically designed to protect sensitive child welfare data while maintaining the highest standards of participant confidentiality during user testing scenarios.

Complementing this work, the secondary project (reference number #385370), "Child Speech Analysis," developed an ethically sound methodology for acquiring and processing linguistic data from children across developmental stages (ages 4-16). The research design

incorporated carefully controlled dialogue sessions focusing exclusively on age-appropriate, non-identifiable subject matter. To ensure absolute privacy protection, the protocol mandated immediate verbatim transcription followed by secure destruction of original audio files, with all retained data undergoing thorough anonymization procedures. The ethical framework was further strengthened through collaborative validation, as evidenced by the approval (reference 2023/501) obtained by our research partners at Griffith University, Australia, for their parallel study titled "Evaluation of Child Avatar Interview Simulation Learning Activity."

Together, these approvals provided a structured foundation for addressing four core ethical concerns in our research, ensuring responsible development while prioritizing participant rights and data protection.

1.4.1 Avoiding Gender Stereotypes

The development of training modules utilizing avatars to simulate child victims of sexual abuse necessitates careful consideration of gender representation to avoid perpetuating harmful stereotypes while maintaining statistical accuracy. Research indicates that CSA disproportionately affects girls, with meta-analytic studies estimating prevalence rates of 19% among female children compared to 8% among males [84]. This disparity justifies the inclusion of female avatars in training scenarios to reflect real-world victimization patterns. However, an overrepresentation of female victims risks reinforcing gendered assumptions that may lead professionals to overlook male victims or dismiss their experiences.

To mitigate this concern, our approach integrates a balanced representation of gender in avatar design, ensuring that training scenarios acknowledge the significant prevalence of abuse among boys while maintaining fidelity to empirical data. By incorporating diverse victim profiles, the training modules foster a more inclusive understanding of CSA, equipping child welfare and law enforcement professionals with the sensitivity required to engage with all potential victims effectively. This methodological choice not only aligns with ethical imperatives to avoid stereotype reinforcement but also enhances the practical utility of the training by preparing interviewers for the full spectrum of cases they may encounter.

Furthermore, the design process included rigorous validation to prevent the avatars from exhibiting gendered behavioral biases that could influence interviewer perceptions.

Dialogue scripts and avatar responses were developed in consultation with child psychology experts to ensure neutrality in language and demeanor, irrespective of the avatar's gender presentation. This attention to detail ensures that the training tool advances equitable investigative practices rather than inadvertently replicating societal biases. The ethical framework governing this aspect of the research was further strengthened by institutional oversight, as detailed in the approvals referenced in the broader ethical considerations section.

1.4.2 Deepfakes Technology

Deepfake-derived avatars, powered by generative AI, offer transformative potential for training professionals in child investigative interviewing by simulating realistic interactions with nuanced facial expressions, vocal tone, and emotional cues [85]. However, their application raises critical ethical concerns, particularly regarding consent, bias, and authenticity.

Unlike traditional simulations, deepfake avatars rely on aggregated datasets of child speech, images, or behavioral data, which may unintentionally retain identifiable features or mannerisms. Minors cannot provide explicit consent for their data's use in synthetic media—a practice that contradicts established pediatric research ethics [86]. This risks normalizing non-consensual digital representations and perpetuating systemic biases if training data disproportionately represents specific demographics in abuse scenarios [87]. Legal scholars further caution that the realism of synthetic media could compromise forensic investigations, as courts lack reliable methods to differentiate AI-generated content from authentic evidence [88].

Despite these challenges, a carefully designed ethical framework can unlock the societal benefits of deepfake technology. By anonymizing victim identities, such systems preserve confidentiality and enable immersive training environments without exposing real children to harm or retraumatization. To balance innovation with accountability, our approach incorporates three core safeguards: (1) exclusive use of anonymized, non-identifiable data; (2) algorithmic audits to identify and reduce potential bias; and (3) transparency protocols that ensure trainees are fully aware they are interacting with synthetic avatars. These practices align with broader ethical standards for responsible AI and synthetic

media, which support both the protection of vulnerable populations and the integrity of professional training environments.

1.4.3 VR Technology

The integration of VR technology in child protection training necessitates ethical navigation between immersive realism and psychological safety. While VR's capacity to simulate high-fidelity CSA scenarios enhances investigative preparedness, its intensity risks triggering secondary trauma, necessitating safeguards such as customizable scenario intensity, mandatory debriefings, and session limits (90-second durations in our implementation) to mitigate discomfort [89, 90, 91]. Ethical implementation further requires scrutiny of VR's long-term impacts across diverse demographics, addressing gaps in systemic risk discourse [92]. Our framework prioritizes trainee well-being through practical protocols: structured break intervals, gradual familiarization with VR interfaces, and alternative training modalities (2D models) for accessibility [93]. These measures align with broader ethical standards in healthcare VR, emphasizing standardized scenarios and user safety [94]. Additionally, while VR's empathy-inducing potential through perspective-taking deepens understanding of victim experiences, it demands oversight to prevent biased emotional responses, ensuring decisions remain rooted in professional objectivity [95]. By integrating interdisciplinary expertise in psychology, VR design, and ethics, we establish evidence-based guidelines that balance training efficacy with institutional accountability, ensuring ethical integrity across development and deployment.

1.4.4 Practical Use of AI Avatars

It is imperative that CPS professionals engage with these avatars solely for educational purposes, adhering strictly to professional conduct standards. Interactions must avoid any form of sexualized dialogue or behavior, as such misuse not only disrespects the simulated child's representation but also undermines the gravity of child abuse cases and the integrity of the training objectives. Furthermore, employing AI avatars for casual entertainment or humor in contexts that replicate serious situations is inappropriate and diminishes the seriousness of the subject matter. Clear guidelines ensure that AI avatars are used only as training tools to enhance CPS workers' skills. This maintains focus on their intended educational purpose [96].

1.5 Main Contributions

This thesis consists of eight research articles (presented in Part II) that collectively explore the development, deployment, and rigorous evaluation of an AI-driven child avatar system aimed at improving training for conducting investigative interviews with children. These contributions systematically address the ROs outlined in Section 1.2, which emerge from critical challenges in creating realistic, scalable, and emotionally engaging avatar-based training solutions. As illustrated in Figure 1.3, the research follows a structured development pathway. Each developmental stage is guided by insights from the published articles, showing how identified challenges informed transitions between stages in the system's evolution. Whenever possible, the developed code and data have been made publicly available as open source and open data to support reproducibility and encourage further research.

The research begins with an early prototype developed using the Unity game engine, which highlights the essential role of visual realism in fostering user engagement (see Article I and Article II). Subsequently, comparative analyses of different avatar configurations, including talking-head generation models and Unity-based avatars, revealed key limitations—such as a lack of real-time interaction and the uncanny valley effect in talking-head systems (see Article III and Article IV). These findings led to two distinct development paths: one focused on enhancing the talking-head model into a more interactive system (see Article VIII), and the other aimed at comparing 2D and 3D environments to evaluate their respective effectiveness and impact on user engagement (see Article V–Article VII).

Earlier child avatar-based tools lacked emotional expressiveness, which reduced engagement and training relevance. To address this, a new system was developed that combines the high-fidelity rendering capabilities of Unreal Engine 5 MetaHuman with the intelligent, audio-driven facial animation of NVIDIA Omniverse Audio2Face. A dedicated study evaluated the feasibility of this architecture by assessing user perceptions of dynamic emotional expressions generated from vocal prosody. The study examined emotional clarity, facial realism, and user empathy across both audio-visual and visual-only conditions.

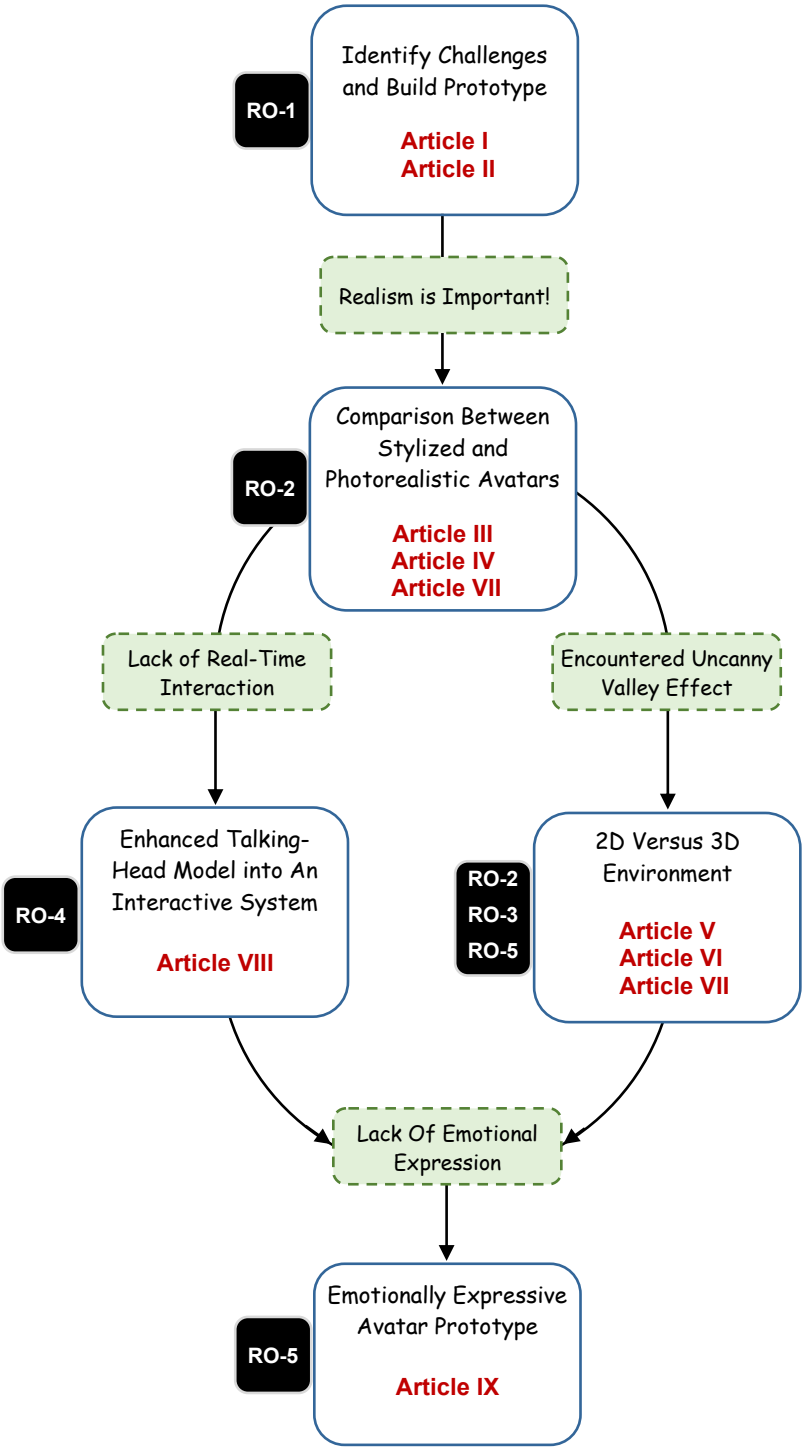


Figure 1.3: Development pathway of the visual component in the training system. The diagram highlights the specific ROs and associated research articles at each stage, illustrating how each step contributed to the overall research trajectory. Green boxes indicate key challenges or limitations that motivated transitions to subsequent developmental phases.

1.6 Other contributions

In addition to the primary research outcomes presented in this thesis, we have made several additional contributions that extend beyond the core research question and research objectives. These contributions include the publication of six peer-reviewed articles, the supervision of two master's students, and the mentoring of one summer intern. While not directly addressing the central focus of the thesis, these contributions are closely connected to the methods applied and the knowledge gained throughout the research process.

Additional Publications: The initial exploration of generative models involved developing a synthetic data pipeline using GANs for medical image segmentation, laying the technical groundwork for realistic avatar creation and data augmentation strategies [97]. Research on the reliability of subjective user studies contributed to understanding how expert and non-expert perceptions influence avatar training effectiveness [98]. Building upon this, we developed a prototype integrating dialogue models, visual avatars, and interactive VR components, establishing the technological foundation for subsequent research [99]. A comparative analysis of interactive environments, VR, 2D desktop, audio-only, and text-based, highlighted VR's advantages for immersive interviewer training [100]. Additionally, an LLM-based chatbot avatar providing immediate feedback was validated for improving questioning skills [101]. Finally, a proof-of-concept study demonstrated the practical effectiveness of dynamic AI-driven avatars in virtual reality at Children's Advocacy Centers, reinforcing the applicability of AI-driven training methods in real-world settings [102]. Collectively, these studies contributed significantly to advancing and validating AI-driven technologies for forensic interviewer training.

Student Supervision: The PhD period also involved supervising master's thesis projects contributing to the broader research agenda. A thesis evaluating AI-generated child avatars for user experience found stylized designs (e.g., cartoony/painted) balanced realism and interaction quality better than photorealistic versions, avoiding the uncanny valley effect [103]. Another thesis assessed the impact of different child avatar interaction types on user Quality of Experience, highlighting the importance of eye contact and graphics realism for effective avatar-mediated communication [104]. Additionally, a sum-

mer internship was supervised, focusing on exploring real-time methods in talking-head generation models.

1.7 Thesis Outline

This dissertation is divided into two main parts. Part I offers a thorough overview of the research and its key contributions, while Part II presents the original research articles that serve as the basis for this work. **Part I** is structured into three chapters, which are outlined as follows:

Chapter 1 Introduction: This chapter highlights the main role of investigative child interviews in abuse cases and identifies limitations in traditional training methods, such as scalability and static practices. It introduces AI-driven virtual avatars as scalable solutions, framed by five research objectives: addressing technical challenges in avatar development, evaluating visual realism's impact, comparing 2D/3D training efficacy, optimizing real-time audio-visual synchronization, and enhancing emotional expressiveness. Guided by the DSRM, the work iteratively develops and evaluates the system. Ethical safeguards address gender representation biases, deepfake risks, and VR safety protocols. Additional contributions include peer-reviewed articles and supervised student projects on avatar interaction.

Chapter 2 Discussion and Research Contributions: This chapter methodically explores the research outcomes, beginning with an identification and analysis of technical, and practical challenges in developing an AI-driven child avatar system include integrating multimodal components (e.g., NLP, TTS, STT, visual generation). The chapter then Investigates the impact of visual realism in avatars on user acceptance and learning outcomes. A Comparative evaluation of 2D versus 3D virtual environments for CPS training. Subsequent sections develops and optimizes real-time talking-head generation models through advanced audio feature extraction. Finally, the chapter examines the role of emotionally expressive avatars in fostering empathy and improving trainees' ability to recognize psychosocial cues, underscoring the importance of dynamic facial animations in realistic simulations.

Chapter 3 Conclusion and Future Work: The last chapter in Part I consolidates the research outcomes and their broader significance. It begins by summarizing the achievements across the defined ROs, including the development of the child avatar system, an exploration of visual realism trade-offs, and an evaluation of comparative 3D training efficacy. The chapter then outlines how these findings can inform the development of more effective training programs for conducting child forensic interviews using adaptive and realistic simulations. It also addresses persistent limitations encountered during the research. Finally, it proposes potential directions for future work to build on the foundations established in this study.

Part II presents our eight research articles that systematically advance the study's objectives: Article I and Article II establish foundational system design, integrating visual avatars and dialogue models while highlighting realism's role in engagement; Article III and Article IV evaluate avatar realism's impact on user experience, identifying key limitations like real-time constraints and the uncanny valley; Article V, Article VI, and Article VII compare 2D and 3D environments, revealing complementary strengths: 2D's usability and accessibility advantages versus 3D's immersive capabilities for professional training, while exploring emotionally expressive avatars to enhance engagement; and Article VIII refines talking-head generation through improved audio processing, adopting Whisper for scalable and realistic interactions. Together, these works provide a cohesive progression from development to evaluation and optimization of the AI-driven interview-training avatar.

Chapter 2

Research Contributions and Discussion

This chapter offers a unified overview of the research contributions resulting from the development of an AI-driven child avatar system designed specifically to support the training of professionals in conducting investigative interviews with children. Rather than duplicating the comprehensive methodological details and complete findings presented in the published articles in Part II, this chapter provides an integrated discussion of key findings and their implications.

Guided by the ROs described in Section 1.2, the following sections elaborate on how each objective was addressed through technical innovations, user evaluations, and comparative analyses. The discussion begins with the identification of challenges and integration considerations in the development of child avatar systems for investigative interviews (**RO-1**), followed by an examination of how visual realism influences the learning experience and user acceptance (**RO-2**), and a comparison of 2D versus 3D virtual environments in terms of their usability and immersive qualities (**RO-3**).

Subsequently, the focus shifts to the challenge of enabling real-time interaction through the development of lightweight talking-head generation models (**RO-4**), aimed at improving system scalability and performance in real-world applications. Finally, **RO-5** investigates the feasibility of incorporating dynamic facial emotion expressions and examines their potential to enhance perceived realism and emotional depth in avatar-based interactions.

The chapter concludes by integrating these findings to revisit the overarching research question: How can AI be used to develop a visually realistic child avatar-based system for investigative interview training with children? This discussion highlights both the potential and the complexities of applying AI in such a sensitive and impactful context.

2.1 Challenges in Child Avatar Creation

In addressing **RO-1**, which focused on identifying the challenges and integration considerations for developing child avatar systems in investigative interview training with children, Article I [105] laid the groundwork by proposing a multimodal AI-driven system designed. The overarching goal is to replace or substantially reduce resource-heavy, actor-based role-plays by creating an immersive and automated environment in which interviewers can practice essential skills repeatedly and at scale. The system architecture integrated advanced technologies such as dialogue models for conversational flow, TTS and STT services for audio processing, and visual manipulation tools to generate lifelike avatars. A critical focus of the research was the visual component, which aimed to achieve age-appropriate realism while ensuring real-time performance and seamless synchronization between speech and facial animations.

Central to achieving these goals in context of a visual avarar was the exploration of deepfake, photorealistic videos using face swaps that leave little trace of manipulation [106], and lip-syncing techniques to synchronize the avatar's lip movements with the spoken content. Initial experiments explored existing tools like Faceswap [107], a GAN-based framework for face-swapping in videos. However, this approach proved unsuitable for child avatars due to its reliance on adult-centric datasets and its inability to adapt to the unique facial proportions of children. Attempts to map adult facial features onto child subjects resulted in uncanny distortions, compromising both realism and user acceptance.

Similarly, ObamaNet [108], a method designed for lip-syncing predefined adult models of presidents such as Barack Obama, faced synchronization challenges and lacked flexibility in replicating child-specific vocal and facial dynamics. These tools highlighted a broader issue: pre-existing models optimized for adult features struggled to accommodate the anatomical and behavioral characteristics of children, necessitating significant modifications. Further exploration turned to Jamaludin et al.'s method [109], which synthesizes

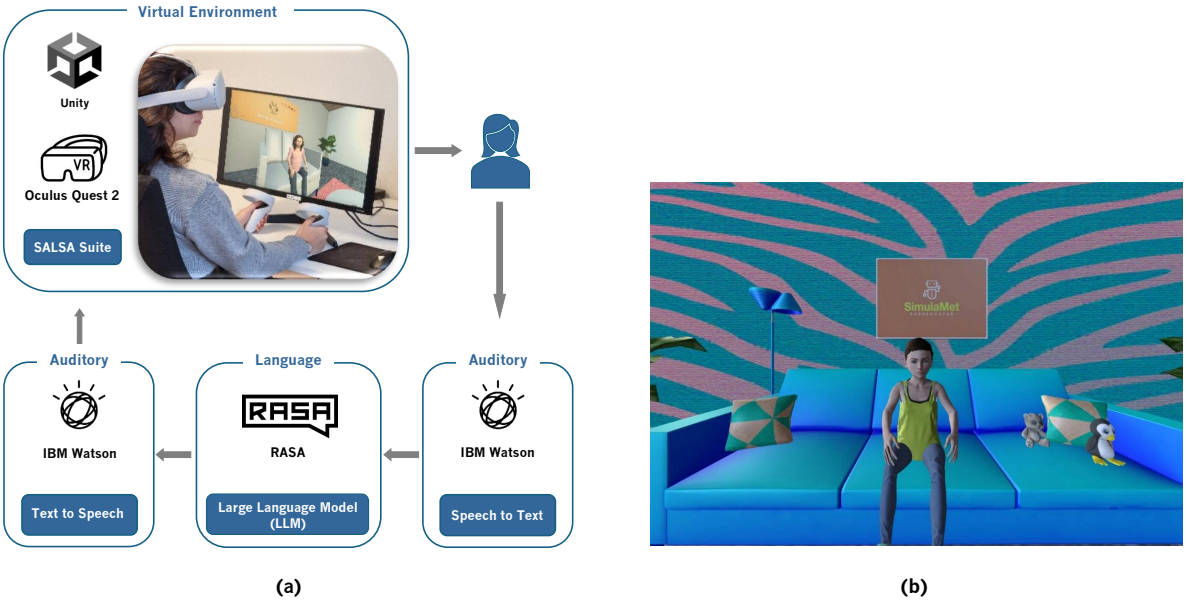


Figure 2.1: a) The architecture of the proposed child avatar prototype includes: the Visual Module (frontend built with Unity 3D and Oculus Quest 2), the Auditory Module (IBM Watson APIs for STT and TTS), and the Language Module (backend dialogue system using Rasa). b) Screenshots of the virtual environment demonstrating user interviews with the child avatar.

lip movements from audio inputs and still images. While promising, re-implementing this approach revealed technical hurdles specific to child avatars. Generating realistic animations required extensive datasets of child facial expressions and computationally intensive tuning to maintain fidelity across diverse emotional states. Challenges included achieving real-time processing speeds, ensuring frame-perfect alignment between audio and visual components, and mitigating the uncanny valley effect, a phenomenon where near-realistic avatars evoke discomfort due to subtle imperfections [62]. Additionally, scalability emerged as a concern, as resource-intensive rendering pipelines posed barriers to deployment in resource-constrained environments such as CPS offices. These findings underscored the necessity for custom solutions tailored to child-specific requirements. Key innovations included the development of adaptive rendering pipelines optimized for real-time performance, the curation of child-centric facial expression datasets, and the integration of lightweight models to reduce computational overhead.

Building on Article I, Article II [110] advanced the child avatar system by creating a fully automated, AI-driven solution for VR-based investigative interviews. We tackled three core challenges outlined in Article I:

- Realistic and responsive avatar behavior: The system employed the open-source Unity Multipurpose Avatar (UMA) framework [111] to create customized child avatars by merging meshes and textures. To synchronize audio with avatar movements, the Unity game engine asset SALSA LipSync Suite [112] was used, enabling coordinated lip-sync, head, and eye motions. These features contributed to a more lifelike and engaging interaction. As illustrated in Figure 2.1, the system's architecture integrated multiple components: dialogue functionality was driven by the Rasa open-source NLP framework [113], integrated with IBM Watson's cloud-based STT and TTS services [114]. While cloud dependencies introduced some latency, the delay was found to be tolerable, as it mirrored the natural hesitations typical of real child responses in investigative contexts.

- Cohesive integration of advanced technologies: Unlike prior systems based on scripted or pre-recorded interactions, this iteration implemented an end-to-end autonomous pipeline. Speech input from the user was transcribed, interpreted using a hybrid rule-based and NLP engine, and then converted into synthesized speech and avatar animations. This eliminated the need for human operators or manually selected video clips, thereby improving scalability and reducing operational complexity.

- Practical feasibility for CPS: The system's practical applicability was evaluated through a user study involving 11 participants, including CPS professionals and child welfare students, as reported in [110]. Each participant conducted two simulated investigative interviews with the VR-based child avatar system. This repeated-interaction design allowed participants to first familiarize themselves with the virtual environment and then demonstrate improved performance and engagement in a second session. Quantitative results showed a notable increase in overall Quality of experience (QoE) between the two sessions, with mean scores rising from 3.62 to 3.92 on a 5-point Absolute Category Rating (ACR) scale (where 1 = "Bad" and 5 = "Excellent"). Participants also reported a statistically significant improvement in perceived system responsiveness (from $M = 2.78$ to $M = 3.18$, $p = .01$), despite the somewhat unnatural requirement of pressing a button to activate dialogue input. Qualitative feedback further indicated that the experience was perceived as immersive, relevant, and pedagogically beneficial, supporting the system's feasibility for integration into real-world CPS training contexts.

By resolving these challenges, Article I and Article II directly addressed **RO-1** by mapping the technical and practical barriers to developing lifelike child avatars for investigative interview training with children. While off-the-shelf tools provided a starting point, our findings showed that achieving the required levels of realism, synchronization, and scalability necessitated tailored, domain-specific innovations. Through iterative development and evaluation, the system matured into a functional prototype capable of delivering immersive and context-aware training. Importantly, qualitative feedback from CPS professionals during the user study consistently emphasized that visual realism was the single most critical factor influencing the perceived effectiveness and authenticity of the system. This directly supports the research objective RO-1, reinforcing that achieving age-appropriate and emotionally coherent realism is not merely a technical aspiration, but a core requirement from the perspective of end-users. The prioritization of realism by domain experts validates the design direction and underscores its centrality in fostering trust, engagement, and successful skill transfer.

2.2 Visual Realism and User Acceptance

In addressing **RO-2**, which focused on investigating how varying the visual realism of child avatars influences the learning experience and user acceptance, four key sources, Article III [115], Article IV [116], the master thesis by Samaneh [103] (which we supervised), and Article VII [117] provide complementary insights.

Building on earlier theoretical frameworks of avatar perception (Section 1.1.4.1), Article III and Article IV examined differences between stylized and photorealistic avatars for training professionals in investigative interviews with children, while the master thesis by Samaneh expanded these findings through a systematic comparison of user experiences across diverse visual styles. Complementing these studies, Article VII explores how visual realism influences user engagement and empathy within both 2D and 3D training contexts.

In particular, we compared stylized avatars developed using a game engine with photorealistic talking-head models generated by recent advancements in AI-driven generative models, including but not limited to GANs. These talking-head systems synthesize dynamic, photorealistic facial animations from static portrait images, using input signals



(a) Example of a photorealistic character generated using StyleGAN [118] (left), which served as input for the talking-head models, and a stylized avatar created using the Unity game engine (right). These two avatar types were directly compared in the user study presented in Article III.



(b) Example of three avatar styles: a photorealistic character generated using StyleGAN [118] (left), and two stylized variants produced using Toonify [119], one in cartoon style (center) and one in painted style (right). Adapted from the master thesis by Samaneh [103].

Figure 2.2: Representative examples of avatar appearance across different realism levels, ranging from stylized to photorealistic.

such as audio, text, or video. Their primary objectives include achieving accurate lip synchronization, naturalistic head movements, and emotionally expressive facial cues. While many approaches employ an audio-driven paradigm, mapping speech signals directly to facial motion, others utilize reference video sequences (video-driven) to extract motion cues or rely on textual input (text-driven) to guide animation.

Our stylized avatars were created using the standard pipeline provided by modern game engines, specifically Unity. This pipeline enables extensive parametric control over facial geometry, surface textures, and physics-based simulations, supporting a high degree of real-time interactivity. These features allow developers to design expressive and customizable avatars with reliable performance across diverse use cases. In our study, this game engine-based approach was contrasted with photorealistic outputs produced by three state-of-the-art AI-driven generative models: PC-AVS [120], MakeItTalk [121], and LSP [122]. Each model adopts a distinct strategy, including implicit modularization

of audiovisual features, landmark-driven animation, and real-time re-targeting of facial movements.

Figure 2.2a illustrates the visual contrast between a stylized avatar generated in Unity and a photorealistic character synthesized using StyleGAN [118], which served as input for the talking-head models. All visual materials were selected and processed in accordance with ethical guidelines, including measures to avoid gender stereotyping and to ensure that all identities remain anonymous and unrecognizable.

Building on established theoretical frameworks of avatar perception, Article III and Article IV report on a user study designed to evaluate how visual realism, facial expressiveness, and synchronization accuracy influence user acceptance, immersion, and perceived authenticity. The study involved 39 participants who were asked to watch and rate 28 short video clips of child avatars—both stylized and photorealistic across four core dimensions: (1) Realism in Talking, (2) Realism in Appearance, (3) QoE, and (4) Accuracy of Lip Sync.

In addition to the quantitative ratings collected after each video, participants completed a post-test questionnaire in which they indicated their preferences across four dimensions: (Q1) most realistic appearance, (Q2) most liked overall, (Q3) most realistic talking, and (Q4) preferred avatar for conversational interaction. As shown in Table 2.1, photorealistic avatars were most frequently selected for appearance, overall preference, and talking realism. However, stylized avatars were slightly preferred for conversational use, despite scoring lower in perceived realism. These findings highlight the nuanced relationship between visual fidelity and user engagement, suggesting that higher realism does not necessarily correspond to increased user preference in interactive contexts.

Table 2.1: Post-test questionnaire items with corresponding participant vote counts comparing photorealistic avatars (generated via AI-driven models) and stylized avatars (animated using Unity).

Question	**photorealistic**	**stylized**	**Winner**
Q1 (Most realistic appearance)	25	14	photorealistic (by 11 votes)
Q2 (Most liked overall)	23	16	photorealistic (by 7 votes)
Q3 (Most realistic talking)	23	16	photorealistic (by 7 votes)
Q4 (Preferred for conversation)	19	20	stylized (by 1 vote)

Although photorealistic avatars received higher ratings for appearance and talking realism, they occasionally evoked discomfort among participants due to subtle visual arti-

facts and imperfect motion. This effect is consistent with the Uncanny Valley hypothesis. In their qualitative feedback, several respondents described a sense of unease when interacting with near-photorealistic child avatars, particularly when facial expressions or movements appeared unnatural or lacked coherence. In contrast, the stylized avatars developed using Unity, while less realistic in visual fidelity, were frequently perceived as more approachable and emotionally comfortable. Post-test responses indicated that nearly half of the participants preferred stylized avatars for conversational use (see Table 2.1). This preference may be attributed to reduced visual distraction and an increased ability to concentrate on the dialogue content, as suggested by several qualitative comments.

From a technical perspective, while talking-head models occasionally achieved highly realistic single-frame outputs, their performance in real-time interaction was limited by issues such as imperfect lip synchronization, unnatural head movements, and inconsistent emotional expressiveness. Even minor synchronization errors were noted as disruptive to the overall experience, underscoring the importance of behavioral realism and responsiveness alongside visual fidelity.

The study also compared synthetic voices (e.g., IBM Watson TTS) with natural human recordings to assess perceived authenticity. Results showed no significant difference in user ratings, suggesting that high-quality synthetic voices can be both realistic and operationally efficient for training applications. Furthermore, matching the avatar's gender with the voice type was not deemed essential for maintaining immersion. In particular, several participants found that female-sounding synthetic voices were acceptable—even for male child avatars—so long as clarity and lip synchronization were preserved.

The master's thesis by Samaneh Taghizadeh (2023) [103] presents a detailed evaluation of various state-of-the-art generative models for talking-head synthesis, including MakeItTalk [121], First Order Motion [123], and PC-AVS [120]. These models were used to generate child avatars in three distinct visual styles: original (photorealistic), cartoony, and painted, as displayed in Figure 2.2b. They were assessed through a user study examining aspects such as realism, lip-sync accuracy, and overall user experience. Contrary to expectations based on the uncanny valley hypothesis, the study found no statistically significant differences in user preference, comfort, or perceived QoE across the different styles. While original-style avatars received slightly higher ratings in lip-sync accuracy and realism, statistical analyses such as repeated-measures ANOVA confirmed that visual

style did not significantly influence overall user perception. Instead, the quality of the underlying generative model, particularly in terms of synchronization accuracy and natural head motion, was the most decisive factor in shaping the user experience. Notably, the PC-AVS model consistently outperformed the others across all evaluated features. The thesis concludes that the ability of a model to preserve identity and align effectively with vocal cues is more important for positive learning outcomes than the level of visual realism alone.

Moreover, while Article VII primarily focuses on comparing 2D and 3D training platforms (as discussed in Section 2.3), its empirical analysis also yields important insights into how the visual realism of child avatars affects user acceptance and the overall learning experience. A central finding reveals that empathy elicitation was not predominantly driven by visual fidelity but rather by the quality of dialogue interactions. SEM demonstrated that response relevance ($\beta = 0.302$, $p = 0.012$) and detailed responses ($\beta = 0.235$, $p = 0.015$) significantly outweighed visual realism in fostering empathetic engagement. This indicates that natural language interaction and contextual realism may significantly outweigh purely photorealistic visual elements in eliciting emotional responses, or possibly that the current level of visual fidelity was insufficiently developed to reliably evoke empathy through visual means alone.

Despite the secondary role of visual fidelity in empathy, its contribution to perceived training effectiveness remains significant. SEM analysis revealed a strong positive relationship between visual realism and training comparability ($\beta = 0.311$, $p = 0.038$), indicating that realistic avatar gestures, lip-syncing, and environmental details reinforced participants' perception of the simulation's authenticity relative to human-led training. However, qualitative feedback exposed critical limitations: disproportionate avatar features (e.g., mature bodily proportions) and unnatural hand movements (e.g., *"floating awkwardly"*) disrupted immersion. This suggests that photorealistic designs risk amplifying subtle imperfections, potentially triggering uncanny valley effects. Conversely, semi-realistic avatars—retaining key expressive features while abstracting non-essential details—could mitigate discomfort while preserving functional realism.

The study also identified scenario-dependent variations in non-verbal cue perception. Hand movement realism differed across scenarios (sexual abuse versus physical abuse) at the conventional threshold for statistical significance ($p = 0.050$). This borderline signifi-

cance suggests a trend-level effect that aligns with qualitative observations but warrants replication in larger samples. Participants interpreted repetitive or incongruent gestures as indicators of the avatar's emotional state, such as anxiety or reluctance, even when animations were imperfect. This implies that visual realism requirements may vary with training objectives: photorealistic avatars could enhance sensitivity training for high-stakes scenarios, while stylized designs might suffice for routine skill practice.

Moreover, visual fidelity further correlated strongly with flow ($\beta = 0.464$, $p = 0.004$), with high levels of visual realism significantly contributing to increased immersion, engagement, and sustained attention. This reinforces the key role of immersive visuals in maintaining a sense of flow, a factor central to effective learning in virtual training scenarios.

Finally, familiarity with VR technologies influenced participants' perceptions significantly. Users who were more accustomed to VR reported higher comfort levels with avatar interactions and system interfaces ($p = 0.046$), suggesting a distinct learning curve for users unfamiliar with immersive technologies. Thus, when implementing training systems with varying realism, user familiarity and comfort with VR may mediate their acceptance and overall effectiveness.

In sum, the findings presented across multiple studies demonstrate that visual realism plays a multifaceted role in shaping user acceptance and the effectiveness of avatar-based training. While photorealistic avatars may enhance perceived authenticity and visual engagement, their benefits are contingent on technical precision, particularly in synchronization and expressiveness. Minor imperfections often led to discomfort, reinforcing concerns around the Uncanny Valley effect. Conversely, stylized avatars, though less realistic, were frequently perceived as more approachable and emotionally comfortable, especially for conversational tasks. Beyond this distinction, however, users consistently emphasized the importance of quality and stability. Clear, smooth avatars without visual glitches were rated as more effective than those with higher realism but lower perceptual consistency. Importantly, user feedback and empirical results highlight that behavioral realism, system responsiveness, and emotionally coherent interactions outweigh mere visual fidelity in determining training efficacy. This suggests that an optimal design strategy balances realism with usability, tailoring avatar appearance to training objectives and user expectations. Therefore, for immersive training environments targeting sensitive contexts

such as investigative child interviews, a hybrid approach that emphasizes emotional authenticity, interaction quality, and moderate realism appears most conducive to fostering engagement and supporting skill acquisition, effectively addressing the goals of **RO-2**.

2.3 2D vs. 3D: Comparative Effectiveness

VR technology has increasingly been adopted across diverse fields including education, training, and psychology, primarily due to its capacity to deliver a heightened sense of immersion and engagement compared to traditional two-dimensional (2D) media. Previous studies have identified substantial advantages of VR environments, highlighting enhanced spatial representation, deeper user engagement, and greater emotional realism, which can facilitate effective learning outcomes [124, 125, 126, 127]. Specifically, VR's immersive nature has been shown to improve learners' understanding and recall of complex 3D structures by providing realistic, dynamic visual cues and spatial interactions not achievable in conventional 2D environments [127]. However, despite these advantages, empirical evidence suggests that increased immersion provided by VR can also lead to heightened physiological arousal independent of content type, potentially introducing discomfort or distraction during prolonged usage [125]. Therefore, a comprehensive evaluation of VR compared to traditional 2D delivery methods remains essential for identifying contexts where VR significantly enhances educational effectiveness and engagement without compromising usability or comfort, alongside evaluations of other critical aspects such as empathy elicitation, visual fidelity, nonverbal cue interpretation, and scenario-specific realism across both environments. Therefore, to address **RO-3**, which aims to evaluate the comparative effectiveness of 2D and 3D training environments, the empirical studies conducted in Article V [128], Article VI [129], and Article VII [117] serve as the primary analytical foundation for achieving this RO.

In Article V, the system under investigation offered four distinct modes of interaction: text-only, audio-based, 2D desktop, and 3D VR conditions. The main focus of the study was to examine how visual fidelity and interactivity impacted training, with particular emphasis on the comparison between 2D desktop and 3D VR environments. In this preliminary study, twenty-one participants, primarily professionals in child welfare or investigative interviewing, interacted with a simulated child avatar in sessions lasting

approximately ninety seconds across the different modes. The study measured factors such as presence, flow, realism, responsiveness, and perceived learning engagement using standardized questionnaires, while short semi-structured interviews provided additional insights into participant impressions. Although participants generally rated the 3D VR environment higher for presence and flow, there was no significant difference between the two settings in terms of overall QoE or the quality of the questions asked. Qualitative feedback highlighted that while the stronger sense of immersion in VR was appealing to many participants, the comfort and simplicity of tasks like note-taking were more manageable in the 2D environment.

Building upon these initial findings, Article VI and Article VII, as extended versions of the same study, conducted an in-depth comparative analysis of 2D desktop and 3D VR environments to evaluate their effectiveness in training professionals for child investigative interviews. 36 participants, including child welfare professionals and individuals with academic backgrounds in psychology, criminology, or related fields, engaged with an AI-driven child avatar across both environments using a repeated-measures design. The study employed psychometrically validated questionnaires, mixed-effects models, and SEM to assess various aspects such as immersion, flow, usability, emotional engagement, and training comparability. Results revealed distinct advantages and limitations for each modality. The mixed-effects analysis, the detailed results of which are presented in Table 2.2, showed that the 3D VR environment significantly enhanced immersion and emotional resonance, with participants reporting stronger feelings of presence and spatial realism. Realistic hand and body movements in VR contributed to a heightened sense of interacting with a child in a shared physical space, fostering cognitive absorption and temporary disconnection from the real-world environment. These immersive qualities were particularly effective in eliciting emotional connections to the avatar, suggesting potential benefits for simulating the stress and sensitivity of real-life interviews.

However, the VR environment faced notable usability challenges. Participants frequently reported physical discomfort, including eyestrain, headaches, and motion sickness, especially during prolonged use. In contrast, the 2D desktop setup was consistently rated higher for ergonomic comfort and practicality, particularly for note-taking and extended training sessions. Despite its lower immersion, the desktop environment's accessibility and ease of use made it preferable for routine practice. Advanced statistical analyses, in-

Table 2.2: Results of Mixed Effects Analysis Examining the Impact of Environment (2D vs. 3D VR), Story Variation, Expertise as investigative interviewers, and Familiarity with VR on User Experience. Cells with blue highlights indicate statistically significant results (p-value ≤ 0.05).

	Environment		**Story**		**Expertise**		**VR Familiarity**	
Questions	Coef	P-Value	Coef	P-Value	Coef	P-Value	Coef	P-Value
Engagement	0.574	0.008	-0.049	0.837	-0.075	0.641	0.219	0.413
Immersion	1.142	0.000	0.018	0.938	-0.047	0.773	0.288	0.288
Concentration	0.387	0.050	-0.01	0.962	-0.162	0.197	0.117	0.571
World Forgetfulness	0.586	0.017	-0.272	0.317	-0.314	0.078	-0.085	0.772
Equipment Comfort	-0.912	0.000	-0.117	0.612	0.066	0.645	0.193	0.414
Avatar Interaction Comfort	0.121	0.473	-0.233	0.213	0.089	0.482	0.416	0.046
Interface Usability	-0.200	0.223	-0.315	0.099	-0.138	0.322	-0.051	0.826
Technical Difficulty	-0.003	0.992	-0.014	0.963	-0.196	0.300	0.482	0.123
Ease of Future Use	-0.466	0.022	-0.11	0.641	-0.139	0.419	0.253	0.374
Appearance Fidelity	0.206	0.225	0.058	0.786	-0.109	0.556	0.267	0.386
Environment Fidelity	0.685	0.000	0.092	0.661	0.190	0.242	-0.136	0.615
Hand Movement Perception	0.416	0.022	-0.431	0.050	-0.135	0.429	-0.104	0.714
Movement Quality	0.238	0.057	0.081	0.618	-0.024	0.896	0.220	0.478
Lip Sync Accuracy	0.06	0.654	-0.265	0.106	0.18	0.212	-0.096	0.689
Facial Expression Fidelity	0.272	0.065	-0.175	0.334	-0.066	0.680	0.051	0.846
Overall Realism Perception	0.263	0.128	-0.078	0.681	0.054	0.671	0.247	0.241
Emotional Response Perception	0.230	0.256	0.185	0.412	-0.027	0.850	0.067	0.779
Emotional Reaction Realism	0.161	0.208	-0.028	0.862	0.048	0.760	0.337	0.199
Emotion-Content Match	0.106	0.576	0.119	0.587	0.067	0.672	-0.084	0.747
Empathy Elicitation	0.104	0.596	-0.034	0.883	-0.218	0.222	-0.337	0.254
Training Comparability	0.687	0.001	-0.038	0.879	-0.211	0.247	0.005	0.988
Tool Inclusion Recommendation	0.227	0.135	-0.118	0.508	-0.015	0.910	0.070	0.749

cluding SEM, further clarified that empathy toward the child avatar was more strongly influenced by the quality of dialogue, such as relevance, age-appropriateness, and emotional authenticity, than by environmental immersion alone. This underscored the importance of refining AI-driven dialogue systems to ensure coherent and contextually appropriate interactions, regardless of the modality used.

Professional expertise emerged as a key moderating factor. Experienced interviewers, familiar with real-world child behaviors and best practices, derived comparable training value from both environments, focusing on substantive interaction quality rather than immersive realism. Novices, however, benefited more from VR's ability to replicate the immediacy of face-to-face interviews, highlighting its value for situational awareness training. Notably, neither environment significantly influenced adherence to evidence-based questioning techniques, such as the use of open-ended questions, indicating that interview quality depended more on user expertise and dialogue coherence than on modality.

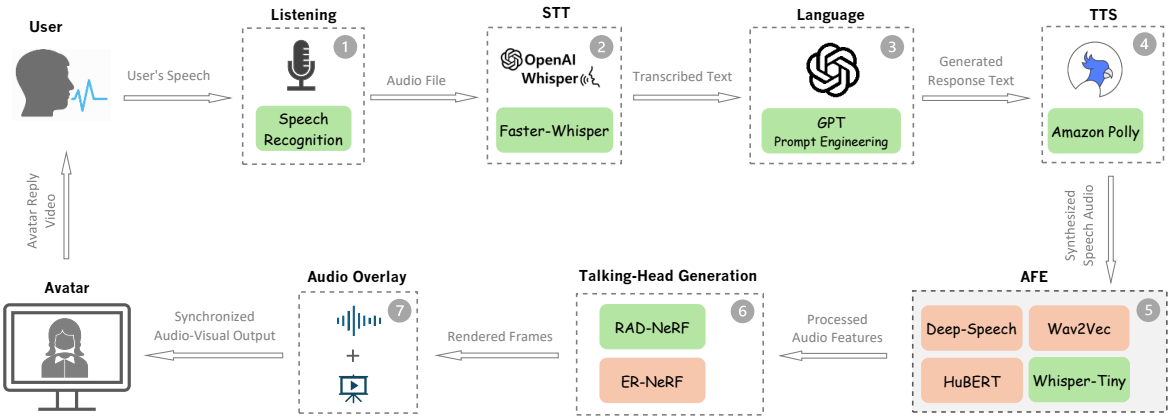
Overall, the comparative analysis highlights the complementary strengths and limitations of each environment. The 3D VR environment excels at fostering immersion, engagement, and emotional realism, making it particularly suitable for training sessions that require emotional intensity and the replication of real-world complexities. Conversely, the 2D desktop environment offers key advantages in terms of usability, comfort, and ease

of interaction, making it ideal for longer, routine training sessions focused on mastering interview techniques. These findings, which directly inform the conclusions for **RO-3**, suggest that an optimal approach for training child welfare professionals and law enforcement officers would involve a hybrid or adaptive system that integrates both VR and 2D modalities, using VR for short, high-impact training sessions and 2D for prolonged practice and mastery of core skills. By strategically combining these environments based on the specific goals and context of the training, programs can maximize learning outcomes and enhance the effectiveness of investigative interview training with children.

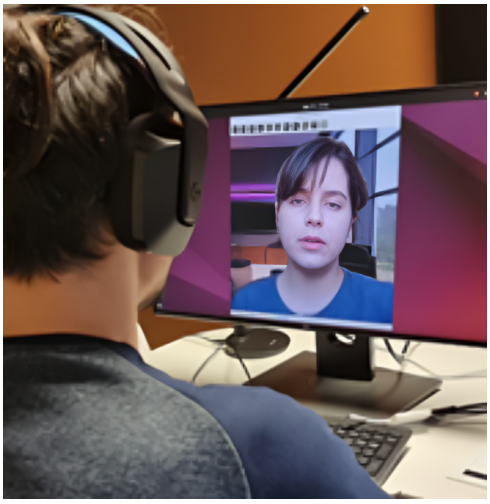
2.4 Refining Talking-Head Models

The **RO-4** addresses the refinement of talking-head generation models through advancements in AFE techniques to optimize responsiveness, realism, and technical scalability for practical deployment in real-world training applications, such as investigative interview simulations. A primary motivation behind this objective is the recognition that existing talking-head systems typically suffer from substantial latency and synchronization issues, critically undermining user immersion and realism during interactive avatar interactions. Furthermore, achieving high perceptual fidelity requires photorealistic avatar rendering methods, which are best realized through AI-driven generative frameworks, given the limitations of conventional game-engine avatars.

Recent advancements in audio-driven portrait animation have remarkable visual fidelity through various methodologies. For instance, the Hallo2 [130] portrait image animation achieves 4K resolution and hour-long animations by integrating vector quantization and temporal alignment techniques. Similarly, EMO [131] employs a direct audio-to-video synthesis approach, bypassing intermediate 3D models to produce expressive and lifelike talking head videos. EchoMimic [132] introduces editable landmark conditioning, enabling the generation of realistic animations driven by audio inputs. Additionally, TANGO [133] utilizes a motion graph-based retrieval framework combined with diffusion-based interpolation to generate co-speech body-gesture videos synchronized with target speech audio. While these models achieve remarkable visual quality, they often require substantial computational resources, posing challenges for real-time applications.



(a) System architecture of the interactive child avatar, detailing the integration of key modules: (1) Listening, (2) STT, (3) Language Processing, (4) TTS, (5) AFE, (6) Frames Rendering, and (7) Audio Overlay. This setup simulates natural conversation, allowing the user to interact with the avatar as if communicating with a real person.



(b) User interaction with the child avatar system.

Figure 2.3: System architecture and user interaction with the interactive child avatar.

Recent approaches have attempted to reconcile the trade-off between quality and speed by exploring hybrid solutions and architecture-level optimizations. For example, some lightweight neural vocoders and frame interpolation techniques have shown potential in accelerating lip-sync generation without heavily compromising fidelity. Yet, empirical evidence indicates that transformer-heavy models such as Wav2Vec 2.0 [2] and HuBERT [3], while highly expressive, introduce substantial processing overheads due to their deep architectures and attention mechanisms. As a result, their deployment in latency-sensitive applications like live avatar simulations remains limited. Conversely, streamlined models such as Whisper [4] offer a compelling alternative by balancing performance and computational efficiency. Notably, Whisper's encoder design, optimized for multilingual and

multitask transcription, enables it to extract high-fidelity audio features at significantly reduced latency. This positions Whisper as an ideal candidate for real-time AFE in applications where rapid response and scalability are critical, such as investigative training with photorealistic avatars.

To address this, in Article VIII we developed and empirically evaluated a fully integrated interactive child avatar system incorporating several key improvements (see Figure 2.3). Notably, this architecture includes optimized modules for STT, TTS, and a more effective Language Processing model leveraging GPT prompt engineering. Compared to earlier versions detailed in prior developed system, these improvements significantly improve the speed and effectiveness of real-time interactions. Moreover, we integrated two leading real-time avatar rendering frameworks (RAD-NeRF [134] and ER-NeRF [135]) with a range of AFE models—including Deep-Speech 2 [1], Wav2Vec 2.0 [2], HuBERT [3], and Whisper [4] in a comprehensive evaluation (Figure 2.4).

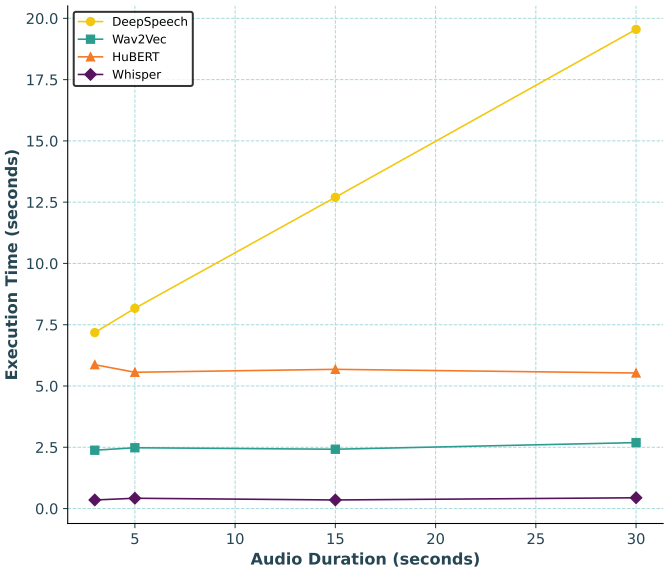


Figure 2.4: Execution time comparison of different AFE models, including Deep-Speech [1], Wav2Vec [2], HuBERT [3], and Whisper [4].

Our experimental analyses clearly highlighted Whisper's superior performance. Whisper's streamlined encoder-decoder architecture demonstrated remarkably lower inference times compared to competing methods (as illustrated in Figure 2.4), substantially improving the overall responsiveness of the avatar interactions. Specifically, Whisper achieved an approximately 80–90% reduction in processing latency relative to Deep-Speech, Wav2Vec, and HuBERT, without compromising rendering quality or synchronization accuracy. These

reductions in inference time directly translated into smoother, more naturalistic lip synchronization and overall more realistic visual portrayals, reinforcing Whisper's practical utility for scalable deployment in real-time interactive environments.

An additional significant finding from our analysis was that Whisper's efficiency made it uniquely suitable for real-world deployment scenarios, including resource-constrained hardware typical of virtual training environments. These results validate the necessity and effectiveness of our optimized audio processing pipeline in supporting photorealistic and interactive avatar representations.

Ultimately, the integration of Whisper within our refined interactive avatar system effectively resolved key limitations outlined in **RO-4**, providing a scalable, responsive, and realistic solution for practical deployment. By enhancing AFE, we significantly improved avatar realism and interaction fluidity, demonstrating the feasibility of deploying sophisticated, AI-generated photorealistic avatars for realistic and effective investigative interview training with children.

2.5 Emotional Expressiveness in AI Avatars

Emotional expressiveness in AI avatars represents a pivotal frontier in affective computing. This field bridges advancements in NLP and multimodal systems to enable machines to approximate human-like emotional intelligence [136, 137, 138, 139]. While early research in this domain concentrated on detecting and modeling emotional cues, recent efforts have shifted towards the dynamic expression and adaptation of these emotions in avatars. The significance of such capabilities is particularly pronounced in high-stakes interactions, such as child forensic interviews, where context-sensitive emotional displays can substantially influence user perceptions and overall system effectiveness. Within this dissertation, **RO-5** extends the investigation into emotionally expressive AI-driven child avatars by not only highlighting their importance for training realism and empathy, but also addressing the technical and perceptual challenges of real-time multimodal emotion integration.

Initial findings from Article IV and Article VII highlighted the importance of emotional expressiveness in child avatars, particularly from a user experience perspective. Building on these findings, Article IX provided a comprehensive investigation into the technical and perceptual challenges of achieving real-time, multimodal emotional integration.

The emotion classification component developed in Article IV contributes to **RO-5** by integrating dynamic emotional modeling into AI-driven child avatars. This research leveraged BART [140] and GPT-3 [141] to analyze textual transcripts from mock interviews. A sliding window approach (3–15 sentences) was employed to balance localized and contextual emotion detection. A key finding was that reducing the emotion taxonomy from seven to four categories (joy, sadness, anger, fear) significantly improved reliability, elevating the Intraclass Correlation Coefficient (ICC) from 0.537 (moderate) to 0.788 (substantial agreement). While GPT-3 demonstrated superiority in contextual reasoning (e.g., detecting fear in multi-turn dialogues), BART performed better on standalone sentences. However, the study also found that threshold-based window resets could disrupt accuracy by overemphasizing isolated cues, underscoring the complexity of balancing abrupt emotional shifts with contextual coherence.

This study (Article IV) highlights the challenges in achieving fluid emotional adaptability, especially in sensitive interactions where limited contextual awareness of interviewer prompts can lead to misclassifications. Addressing these gaps necessitates the development of multimodal datasets that combine text, audio, and visual cues to enhance affective modeling. This work establishes foundational insights for creating emotionally responsive avatars, advocating for simplified taxonomies and context-aware frameworks, while also emphasizing the need to overcome data scarcity and the computational trade-offs inherent in replicating nuanced human-like emotions.

From a broader human-computer interaction (HCI) perspective, anthropomorphism in digital avatars is linked to enhanced user engagement, credibility, and satisfaction [142, 143, 144, 145]. Emotion itself is conceptualized as both a mental feeling and a physiological arousal [146, 147]. Congruently, prior research in virtual therapy and education indicates that insufficiently expressive avatars may lead to misunderstandings or diminished rapport [148, 149].

Further supporting these points, Article VII provides evidence that limited emotional expressiveness remains a substantive issue for AI-driven child avatars, even with improved immersion and visual fidelity. Quantitative results from this study demonstrate that the perceived realism of facial reactions and body language directly affects user experience. SEM revealed that poorly executed emotional cues could negatively shape participants' recommendations for tool adoption ($\beta = 0.254$, $p = 0.071$). Conversely, empathy

elicitation appeared more strongly driven by the dialogue system's response relevance ($\beta = 0.302$, $p = 0.012$) and response detail ($\beta = 0.235$, $p = 0.015$). This suggests that high-quality verbal interaction can, at times, outweigh emotional expressiveness in fostering user empathy. Qualitative feedback corroborated these findings: many participants perceived the avatar's affective displays as too muted or disconnected from distressing interview content, thereby weakening authenticity and limiting the avatar's practical utility as a training tool. Moreover, technical barriers, such as a reliance on static animations, were found to hinder the system's responsiveness to rapidly shifting emotional dynamics. When an avatar fails to accurately mirror micro-expressions or contextually relevant emotional changes, particularly in delicate trauma-related scenarios, user skepticism regarding the avatar's authenticity and effectiveness increases.

Collectively, these findings underscore the necessity of integrating adaptive emotion modeling, more granular affective taxonomies, and context-aware feedback into AI avatars designed to simulate children in forensic interview training and related child-focused domains. Refining emotional realism, whether through optimized emotional frameworks, advanced speech synthesis, or reactive visual displays, can enable future avatar-based platforms to better replicate the psychosocial complexity of child abuse cases, thereby enhancing empathy, engagement, and training outcomes. The advancements detailed in Article IV and Article VII highlight the paramount importance of emotional authenticity in AI-driven child avatars. By aligning a targeted emotional taxonomy, focused on core states like fear, sadness, and anxiety, with high-fidelity, context-aware animations, these studies demonstrate that emotionally resonant design is essential for achieving realism and trust. Emotional precision allows avatars to reflect the nuanced cues of real children, which can deepen trainees' empathic engagement and improve their ability to respond to subtle emotional shifts during abuse-related interviews. Ultimately, dynamic emotional frameworks reveal that effective training hinges not merely on visual accuracy but on the capacity to evoke genuine emotional experiences.

Historically, early efforts in avatar development primarily revolved around 2D video-based avatars [150, 151, 152]. These systems often produced rigid emotional outputs via direct mappings between textual or speech inputs and animated expressions. While more recent 3D parametric models, such as FLAME [153] , have enabled finer-grained control

over facial shape and pose, even sophisticated systems like EMOTE [154] or EmoTalk [155] continue to grapple with achieving high-fidelity emotional realism.

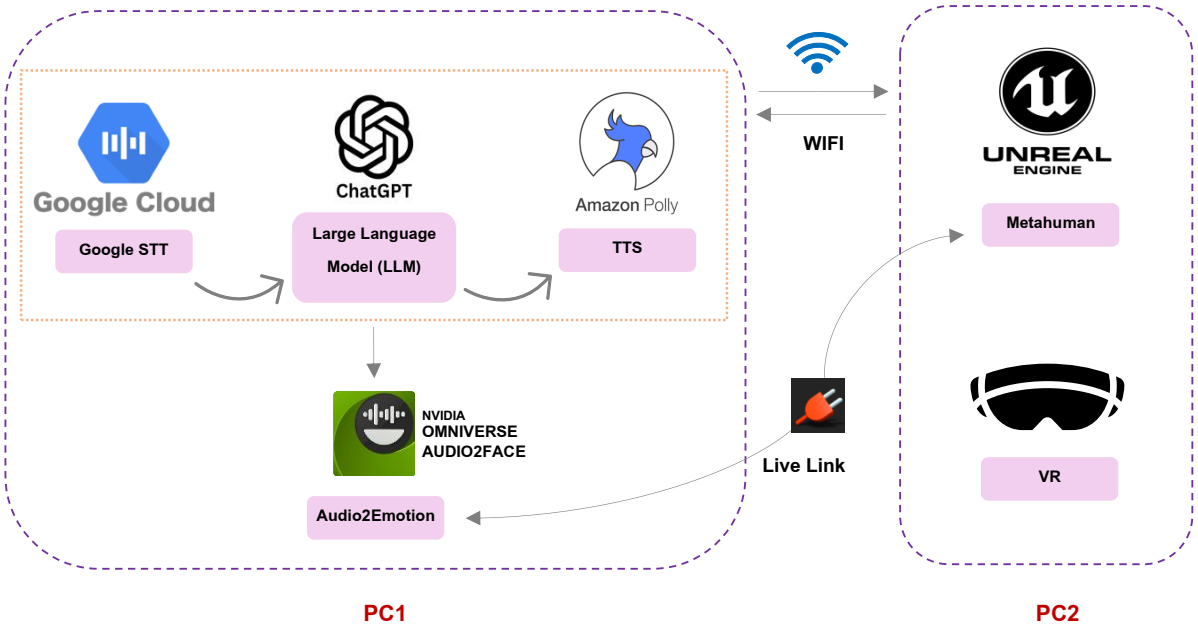


Figure 2.5: Architecture for real-time emotionally expressive child avatars. PC1 processes speech input (Google Cloud STT), generates responses (LLM), synthesizes speech (Amazon Polly TTS), and extracts emotional parameters (NVIDIA Audio2Face) to drive facial animation. PC2 renders a customizable Metahuman avatar (Unreal Engine) driven by this real-time emotional data via Wi-Fi. The Live Link plugin ensures synchronization, and Metahuman's anatomical rigging enables nuanced expressions. Supports desktop and VR environments.

Therefore, to address the limitations of prior emotional expression systems in AI avatars, this research introduces and evaluates a novel real-time architecture, as presented in Article IX. This system fuses Unreal Engine 5 (UE5) MetaHuman software [70] for rendering with NVIDIA Omniverse Audio2Face [71] to translate vocal prosody into high-fidelity facial expressions on photorealistic child avatars. UE5's advanced rendering pipeline, including features like Nanite geometry, the Lumen lighting system, and subsurface scattering for realistic skin, excels in real-time performance and film-quality visuals, essential for interactive and immersive systems. MetaHuman Creator allows for the generation of fully rigged, customizable characters, whose expressions can be driven by various inputs, including the audio-driven pipeline developed in this work.

This system leverages two core platforms: **Unreal Engine 5 (UE5) MetaHuman** [70] and **NVIDIA Omniverse Audio2Face** [71]. UE5 MetaHuman provides

the foundation for high-fidelity avatar rendering, utilizing advanced real-time graphics technologies like Nanite geometry, the Lumen lighting system, and subsurface scattering to achieve film-quality, customizable, and fully rigged characters. Complementing this, NVIDIA Omniverse Audio2Face serves as the generative AI engine, analyzing speech input to automatically produce synchronized lip movements and expressive facial animations in real-time. Audio2Face includes an intuitive character retargeting pipeline for seamless integration with MetaHumans and offers fine-grained control over emotional expression through facial action unit (FAU) adjustments and customizable emotion extraction settings. Together, these platforms enable the creation of visually realistic avatars capable of dynamic, audio-driven emotional expression suitable for interactive applications.

The implemented architecture, illustrated in Figure 2.5, utilizes a distributed two-PC setup to decouple language processing and speech synthesis (PC1) from GPU-intensive rendering (PC2), ensuring low-latency interaction. On PC1, user input is transcribed via Google Cloud STT (leveraging models like Google's foundation model Chirp for high accuracy across diverse accents and acoustic environments, with support for multilingual input), interpreted by an LLM (OpenAI's ChatGPT) for response generation, and converted to synthetic speech using Amazon Polly TTS, which offers lifelike multilingual voices and allows customization of intonation, pronunciation, and speaking style. Concurrently, NVIDIA Audio2Face analyzes the prosodic and spectral features of this synthesized speech to derive frame-level emotional parameters (e.g., blendshape weights or FAU activations for joy, fear, sadness). These parameters are transmitted via Wi-Fi to PC2, where Unreal Engine 5's MetaHuman framework renders a customizable child avatar. The emotional data drives the avatar's expressions in real-time via the Live Link plugin, mapping parameters to the MetaHuman Control Rig. This distributed design, while potentially subject to network or processing latencies, generally ensures low-latency synchronization and enables dynamic adjustments to emotional intensity (0–1 scale), critical for observing subtle affective cues. By unifying generative audio-driven animation with industry-standard character rendering, the system achieves significant emotional granularity and scalability.

As part of our investigation into **RO-5**, which examines the feasibility and impact of incorporating emotional expressiveness into AI-driven avatars through dynamic facial emotion expression, a user study was conducted. This study focused on assessing emo-



Figure 2.6: Example frames of the two avatars used in the user study: Amelia (female-presenting) and Emory (male-presenting). The top row displays their neutral facial expressions. The bottom rows show expressions of joy (left) and sadness (right) for each avatar, as used in the study.

tional clarity, facial realism, and empathic connection using two photorealistic MetaHuman avatars, Amelia and Emory (Figure 2.6), expressing joy, sadness, and anger. The study employed a between-subjects design with $N = 70$ participants recruited via Prolific, randomly assigned to either an audio+visual condition or a visual-only condition. Participants viewed six counterbalanced high-definition video clips and rated them on five-point Likert scales. In the audio+visual condition, speech was synthesized using two TTS systems: OpenAI's TTS* for the *Emory* avatar and NaturalReaders† for the *Amelia* avatar. Voice selections were made by two researchers based on how well they matched the avatars' visual and emotional characteristics. While child-like voices were available, young adult female voices were preferred, as the former were considered too juvenile for avatars designed to resemble adolescents. This decision is further supported by prior research [115], which found that female voices can effectively represent younger characters in training contexts.

*https://ttsopenai.com/
†https://www.naturalreaders.com/

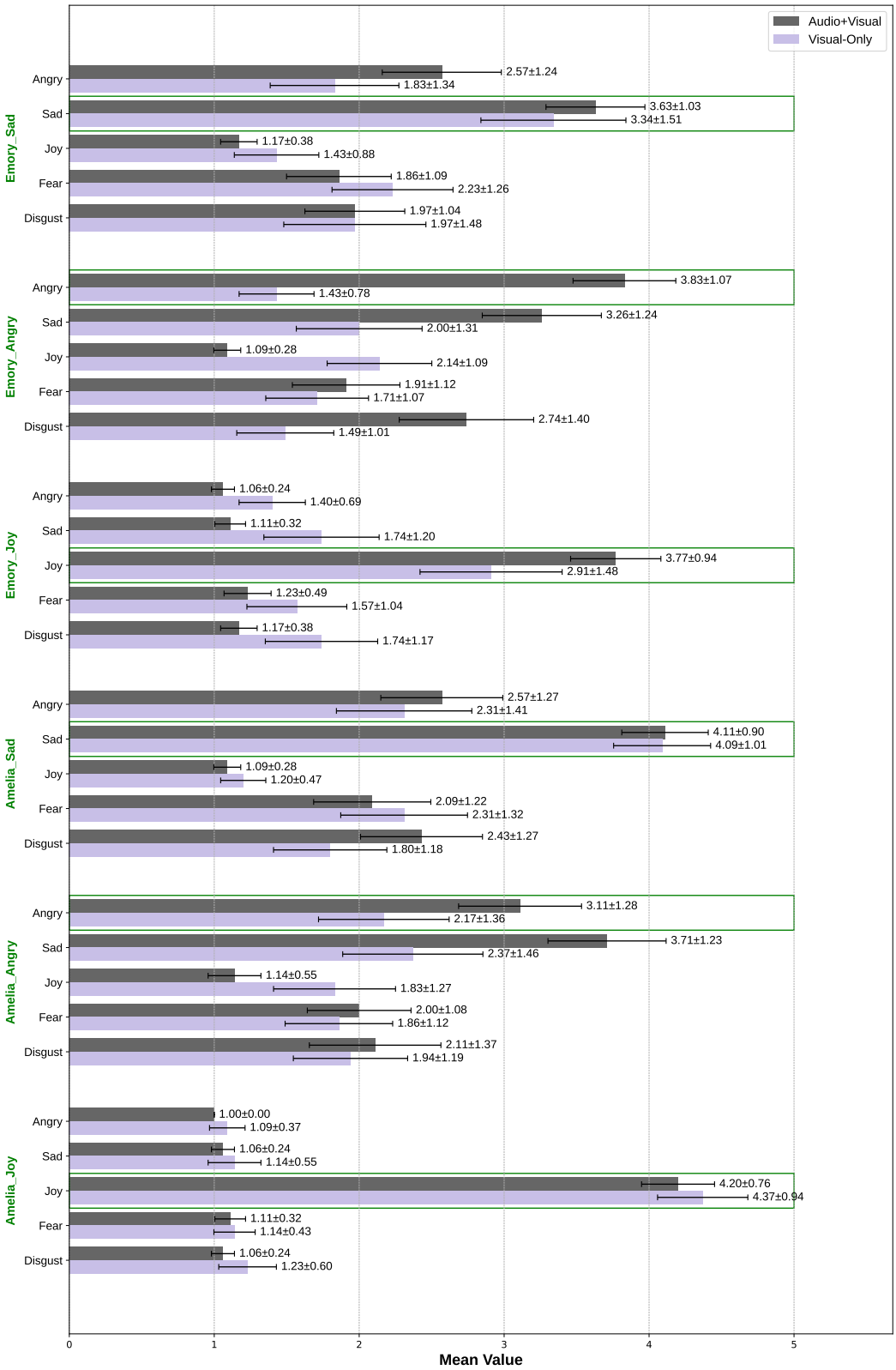


Figure 2.7: Mean ratings ($M \pm SD$) of emotional expressiveness across five emotions (Angry, Sad, Joy, Fear, Disgust) for the avatars *Emory* and *Amelia*, under both audio+visual and visual-only conditions. Bars represent group means; values next to each bar report the mean and standard deviation, while horizontal error bars indicate 95% confidence intervals (CI). Each video clip is labeled on the left, with color-coded bars distinguishing the experimental conditions. Green-highlighted boxes indicate target emotions critical for the study's recognition task.

Table 2.3: Ordinal mixed-effects model results predicting emotion recognition accuracy. Odds ratios (OR), 95% confidence intervals (CI), and p-values are reported. Significant effects ($p < .05$) are highlighted. *Reference levels:* Avatar = Amelia, Condition = visual-only, Emotion = intended. All model terms are interpreted as contrasts relative to these reference levels.

term	p.value	OR	OR_low	OR_high
Emotion ≠ Intended	0.000	0.039	0.024	0.063
Audio+Visual	0.437	1.310	0.663	2.586
Avatar: Emory	0.002	0.202	0.073	0.562
Audio+Visual × Emotion ≠ Intended	0.976	1.010	0.547	1.864
Emotion ≠ Intended × Avatar: Emory	0.000	6.191	3.299	11.618
Audio+Visual × Avatar: Emory	0.000	4.045	1.931	8.472
3-way Interaction	0.000	0.212	0.091	0.493

Table 2.4: Results from the ordinal mixed-effects model predicting perceived realism of avatars' facial expressions. Odds ratios (OR), 95% confidence intervals (CI), and p-values are presented. Significant effects ($p < .05$) are highlighted. *Reference levels:* Avatar = Amelia, Condition = Audio+Visual, Scenario = Angry. All contrasts are interpreted relative to these reference levels.

term	p.value	OR	OR.lower	OR.upper
Visual-only	0.526	1.568	0.391	6.294
Avatar: Emory	0.001	0.211	0.084	0.528
Scenario: Joy	0.722	0.850	0.348	2.077
Scenario: Sad	0.228	0.575	0.234	1.413
Visual-only × Avatar: Emory	0.055	3.584	0.972	13.208
Visual-only × Scenario: Joy	0.198	2.387	0.635	8.981
Visual-only × Scenario: Sad	0.023	4.554	1.232	16.829
Avatar: Emory × Scenario: Joy	0.216	2.254	0.622	8.165
Avatar: Emory × Scenario: Sad	0.875	1.108	0.310	3.963
Visual-only × Avatar: Emory × Scenario: Joy	0.081	0.188	0.029	1.225
Visual-only × Avatar: Emory × Scenario: Sad	0.115	0.229	0.037	1.433

Table 2.5: Results of mixed-effects cumulative-logit analyses comparing perceived realism and empathy between audio+visual and visual-only conditions. Significant results ($p < .05$) are highlighted.

	term	p.value	OR	OR_low	OR_high
Realism	Audio+Visual vs. Visual-Only	0.038	0.258	0.072	0.925
	Visual Appearance vs. Facial Expressions	0.765	0.865	0.333	2.246
	Interaction: Audio+Visual × Visual Appearance	0.587	1.434	0.391	5.264
Empathy	Audio+Visual vs. Visual-Only	0.397	0.693	0.295	1.614

This study highlights the essential role of audiovisual congruence in shaping participants' emotion recognition accuracy and perceived realism when evaluating child-like AI

avatars. Using an ordinal mixed-effects model (Table 2.3), the results showed that accurate recognition significantly improved when the expressed emotion aligned with the intended one (OR = 0.039, $p < .001$). Avatar morphology also had a substantial effect: participants were significantly more likely to misclassify Emory's expressions than Amelia's (OR = 0.202, $p = .002$). This suggests that sharper facial features may reduce emotional clarity.

Interestingly, while the addition of audio had no general effect on recognition, it significantly improved accuracy for Emory; the interaction between audio and avatar increased his recognition odds fourfold (OR = 4.045, $p < .001$). Descriptive means (Figure 2.7) support these findings: both avatars maintained near-ceiling accuracy for sadness and joy, but Emory's joy expression dropped by approximately 22% in the absence of sound, indicating that vocal prosody is crucial when facial cues are ambiguous.

Perceived realism followed a complementary pattern. According to the odds-ratio estimates from Table 2.4, Emory's expressions were judged to be significantly less lifelike than Amelia's (OR = 0.211, $p = .001$). Removing the audio track generally improved realism ratings for low-arousal expressions. Notably, in the visual-only condition, the perceived realism of Emory's sad expression increased significantly (OR = 4.554, $p = .023$), rising from 2.80 ± 0.90 to 3.43 ± 0.95 on the five-point scale. Similarly, Amelia's sad expression improved from 3.34 ± 1.00 to 4.00 ± 0.94. These results suggest that audiovisual desynchrony—particularly in lip-sync or prosodic alignment—can reduce perceived authenticity, and that silent presentations may mitigate such effects.

A broader post-exposure questionnaire (Table 2.5) further revealed that including audio decreased the likelihood of avatars being rated as highly realistic by approximately 74% (OR = 0.258, $p = .038$). However, this manipulation did not significantly affect empathic engagement (OR = 0.693, *n.s.*). In the audiovisual condition, realism judgments were primarily driven by voice tone, while in the visual-only condition, attention shifted toward facial detail. Participants appeared to shift their attention from auditory cues to facial detail when audio was absent, which may explain why realism ratings improved despite stable levels of empathy.

Taken together, these findings underscore that emotion recognition and perceived realism are highly sensitive to the alignment of voice, facial expression, and avatar design. In scenarios where perfect audiovisual synchronization cannot be ensured—particularly

for subtle, low-arousal emotions such as sadness—silent facial presentations can preserve credibility without diminishing empathic connection.

Conversely, high-arousal states like anger or emphatic joy require tightly synchronized audiovisual input to avoid misinterpretation. These outcomes support a flexible training protocol for **RO-5**, in which silent-face drills enhance sensitivity to static cues, while audio-enhanced scenarios reflect real-world complexity. Such an adaptive approach allows for tailoring emotional intensity and modality based on avatar morphology and expression type.

2.6 AI-Driven Visual Avatars for Effective Training

To address the core research question: ***"How can AI be used to develop a visually realistic child avatar-based system for investigative interview training with children?"***. This thesis systematically advances five ROs (**RO-1** to **RO-5**), each contributing distinct yet interconnected insights. Rather than reiterating findings detailed in preceding sections, this synthesis highlights how these objectives collectively address the overarching challenge.

RO-1 (Section 2.1) by establishing the technical foundation, resolving multimodal integration challenges (NLP, STT/TTS, visual synthesis) to enable scalable, real-time child avatars. **RO-2** (Section 2.2) revealed the importance of balancing anatomical accuracy and abstraction. This balance helps optimize user acceptance while avoiding uncanny valley effects. **RO-3** (Section 2.3) demonstrated complementary strengths of 2D and 3D environments, advocating for hybrid training strategies: immersive VR for empathy-driven scenarios and 2D interfaces for procedural mastery. **RO-4** (Section 2.4) achieved latency reductions of 80–90% through Whisper-based AFE, ensuring seamless audio-visual synchronization critical for realism. Finally, **RO-5** (Section 2.5) emphasized the central role of emotional expressiveness in interpreting psychosocial cues. Building on prior insights, Article IX introduced and evaluated a real-time multimodal system that significantly advanced how speech prosody and tone are translated into expressive facial animation, while also identifying key perceptual and technical challenges in achieving emotionally believable avatars.

Collectively, these contributions demonstrate that effective AI-driven training systems require not only technical robustness but also user-centered design suited to the real-world demands faced by CPS and law enforcement agencies. By harmonizing visual fidelity, environmental immersion, real-time responsiveness, and emotional granularity, the research establishes a paradigm for realistic yet adaptable avatars. In practical terms, this strengthens the skill sets of investigative teams: immersive simulations help practitioners refine age-appropriate questioning techniques, gain familiarity with the nuances of nonverbal cues, and apply evidence-based interview frameworks more consistently.

Beyond enhancing professional competencies, these improvements can have a direct, positive effect on children. When interviewers are better equipped to handle challenging conversations, the likelihood of gathering clear, reliable testimonies increases. This is particularly vital in cases with little or no physical evidence, as highlighted in Section 1.1. More accurate disclosures help reduce unnecessary re-interviews and lessen the trauma children experience throughout the investigative process. Importantly, scaled-up and standardized training also addresses organizational constraints. The systems described here can reduce reliance on full-scale role-plays with trained actors and mitigate disparities in training quality across different regions and agencies.

From a broader societal viewpoint, more efficient, empathetic, and evidence-based interviews facilitate earlier interventions, improve child welfare outcomes, and help lower the staggering economic and social costs linked to child abuse. Future work must continue to refine affective modeling and integrate hybrid learning environments. By offering a foundational framework, this thesis paves the way for more accessible, engaging, and ultimately transformative investigative interview training with children.

While initial design iterations benefited from feedback provided by CPS professionals, their direct involvement in later development phases was limited. This reduces the extent to which the current findings can be generalized to practical training contexts and highlights the need for stronger stakeholder engagement in future evaluations.

Chapter 3

Conclusion and Future Work

3.1 Conclusion

This thesis systematically addressed five interconnected ROs to provide comprehensive insights into the central research question regarding the use of AI-driven child avatars for effective investigative interview training with children.

Initially, **RO-1** focused on identifying significant challenges and essential considerations involved in the development of child avatar systems for investigative interview scenarios. Findings indicate that integrating advanced dialogue generation, visual rendering, and real-time interaction powered by AI is technically feasible and practically viable, under the evaluated conditions. This was achieved by addressing key design challenges: ensuring age-appropriate and emotionally expressive avatars without falling into the uncanny valley, synchronizing multimodal components (STT, TTS, NLP, and visual rendering) in real time, and balancing high-fidelity rendering with system performance. These trade-offs demonstrate both the feasibility of the system and the design constraints that shaped its development. This combined approach successfully addressed limitations encountered by previous avatar systems, such as dependency on prerecorded content or human intervention, demonstrating significant potential for fully autonomous, scalable training environments. Subsequently, **RO-2** investigated how varying levels of visual realism influence users' learning experience and acceptance of child avatars. Empirical evidence revealed that user preferences could not be reduced to a simple choice between photorealistic and stylized avatars. While photorealistic avatars were often perceived as more visually authentic and stylized avatars sometimes helped avoid the uncanny valley,

what ultimately mattered most was the consistency and stability of the avatars. Clear, smooth visuals without glitches or noise were rated as more effective, indicating that perceptual clarity outweighed visual style in shaping acceptance and learning outcomes. Building upon this, **RO-3** conducted a detailed comparative analysis between 2D and 3D virtual training environments in terms of effectiveness, usability, and user perceptions among CPS professionals. Results demonstrated that 3D virtual environments significantly enhanced immersion, situational realism, and participant engagement compared to their 2D counterparts. However, 2D platforms retained advantages in accessibility and ease of use. Thus, selecting between these environments should strategically balance immersive benefits against practical considerations, such as resource availability and training context. Further refinement occurred through **RO-4**, aimed at developing and refining talking-head avatar generation models through enhanced AFE, ensuring optimal technical performance and scalability. By evaluating several state-of-the-art AFE techniques, the research identified models capable of achieving high-quality real-time synchronization between audio inputs and avatar animations. This improvement in technical responsiveness directly contributed to more realistic and interactive avatar experiences, which improved the relevance and potential utility of the system for practical investigative training applications. Lastly, **RO-5** confirmed that emotional expressiveness plays a key role in creating believable and effective child avatars for training purposes. Building on this insight, a new real-time system was developed that integrates facial animation and audio-driven emotion synthesis. The system achieved a high degree of perceived realism, particularly in conveying emotions such as sadness and joy, even in the absence of audio. At the same time, the evaluations revealed that perceived realism and empathic impact were shaped not only by facial expressions but also by the coherence between voice and visual cues. These findings emphasize the importance of tight multimodal alignment in maximizing the emotional credibility of training simulations.

Building upon the summarized contributions, the following reflections highlight how this research specifically advances training for child investigative interviews and how its outcomes extend to broader professional domains.

This research advances the field of AI-driven training for child investigative interviews in several concrete ways. First, systematic exploration of visual realism demonstrated how varying levels of fidelity directly influence interviewer skill development. The find-

ings indicate that while photorealistic avatars enhance authenticity, stylized designs can reduce the risk of the uncanny valley and foster greater user comfort, which is particularly important in sensitive training contexts. Second, the integration of emotionally expressive avatars provides trainees with opportunities to recognize and respond to nuanced psychosocial cues, thereby strengthening empathy and improving their ability to manage emotionally charged situations. These contributions establish a clear pathway toward more effective, scalable, and ethically responsible training solutions for professionals who conduct child investigative interviews.

Beyond this domain, the framework developed in this dissertation also has broader implications for professional training in other high-stakes fields. The combined use of AI, visual realism, and immersive environments can be adapted to contexts such as medical consultations, psychotherapy, and law enforcement simulations, where professionals must balance technical expertise with interpersonal sensitivity. By demonstrating how AI-driven avatars can replicate complex emotional and communicative dynamics, this work provides a transferable foundation for advancing training systems across multiple professions that require both technical skill and empathic engagement.

At the same time, a limitation of this work lies in the degree of stakeholder involvement. The input of CPS professionals and trainees primarily informed the design and evaluation of the initial Unity-based visual system, but their involvement was less extensive in the later development stages. As a result, the broader practical impact for end-users remains to be validated through more systematic and longitudinal studies. Strengthening stakeholder co-design and evaluation will therefore be an important direction for future research to ensure that the developed systems fully address the needs and practices of those who conduct child investigative interviews. By acknowledging this limitation, the thesis not only reflects critical self-assessment but also outlines a concrete pathway for translating research contributions into sustainable real-world impact.

3.2 Future Work

This thesis has systematically explored the development and evaluation of AI-driven child avatars for investigative interview training with children, yielding significant insights and a robust foundation for future research. The directions proposed aim to build upon these

findings and address remaining challenges, primarily focusing on enhancing emotional expressiveness and realism in AI avatars. An essential next step involves comprehensive user evaluations of the latest avatar system, which incorporates NVIDIA Omniverse Audio2Face and Unreal Engine MetaHuman, to assess its impact on training effectiveness, user engagement, and perceived realism. Future work will also explore more detailed affective taxonomies and advanced context-aware feedback mechanisms, developing systems that can interpret and respond to subtle emotional cues from trainees and adapt the avatar's emotional state accordingly. Furthermore, the integration of multimodal datasets combining text, audio, and visual cues is essential to refine affective modeling, helping to overcome data scarcity and improve the replication of nuanced human-like emotions. Persistent challenges, such as audiovisual desynchronization and the subtlety of emotional expressions, will be addressed by exploring different animation tool settings to achieve more natural and less artifact-prone animations.

Advancing visual realism and user acceptance remains a key area for future investigation. This includes continuing to investigate the optimal balance between photorealistic and stylized avatars, with a focus on methodologies for tailoring avatar appearance to specific training objectives, user characteristics, and interaction contexts to maximize engagement while mitigating the Uncanny Valley effect. Additionally, further refinement of avatar morphology is needed to address issues like disproportionate features or unnatural movements, which can disrupt immersion and negatively impact user experience. The technical backbone of the avatar system, particularly talking-head models, will also see improvement. While the integration of Whisper significantly reduced latency, ongoing advancements in AFE models will be continuously evaluated and integrated to further enhance the responsiveness and realism of audio-visual synchronization. Efforts will also be made to enhance the avatar's language module to achieve better contextual awareness of interviewer prompts, which is important for generating more relevant emotional and verbal responses and reducing misclassifications.

The development and evaluation of hybrid training environments are also paramount, as findings suggest that a combination of 2D and 3D environments may offer the most comprehensive training solution. Future work will involve designing and implementing adaptive training systems that strategically integrate both 2D desktop and 3D VR modalities, leveraging VR for its immersive and emotionally engaging capabilities in specific

scenarios, while utilizing 2D interfaces for foundational skill development, accessibility, and prolonged use. Rigorous comparative studies will be conducted to evaluate the effectiveness of these hybrid models in achieving diverse training outcomes for professionals with varying levels of experience. To maximize the societal benefit of this research, future efforts will also consider broadening the scope of child avatar representations (e.g., age, gender, emotional predispositions) and the variety of interview scenarios to ensure the training is applicable to a wider range of real-world situations and promotes inclusivity.

Investigating the integration of the AI avatar system into existing professional training curricula for child protective services and law enforcement agencies is another important direction. Longitudinal studies are needed to assess the long-term impact of this training on interviewer skills, the quality of investigative interviews, and ultimately, child welfare outcomes.

Furthermore, the findings from the latest system iteration, integrating NVIDIA Omniverse Audio2Face with Unreal Engine MetaHuman and presented in a submitted journal article, are intended to advance the academic understanding and practical development of emotionally expressive AI avatars for training applications.

Finally, as AI and VR technologies continue to evolve, so too must the ethical frameworks guiding their application. This necessitates a continuous review and refinement of ethical protocols related to data privacy, the generation and use of synthetic media (especially concerning deepfake advancements), and the psychological safety of trainees engaging with highly immersive and potentially distressing VR scenarios. By pursuing these avenues of future work, the field can continue to advance the development of sophisticated, ethical, and highly effective AI-driven training tools. These tools will better prepare professionals for the critical task of conducting investigative interviews with children, ultimately contributing to improved child welfare and more effective justice systems.

3.3 Final Remarks

This doctoral journey has been dedicated to exploring the multifaceted intersection of visual realism, AI, and immersive technologies for a profoundly important application: enhancing the training of professionals who conduct investigative interviews with children. The research presented in this dissertation has sought to unravel how the visual

and interactive fidelity of AI-driven avatars can create more authentic, engaging, and ultimately more effective training environments. It is my sincere hope that the insights gained and the systems developed herein will contribute to empowering child welfare and law enforcement professionals, equipping them with refined skills to approach these sensitive interviews with greater confidence and efficacy. As AI and immersive technologies continue their rapid advancement, the aspiration is that this work will serve as a stepping stone towards a future where training tools become ever more sophisticated and accessible, consistently leading to improved evidence gathering and, most importantly, better support and outcomes for children who have experienced maltreatment. The wellbeing of these children has been, and will continue to be, the paramount inspiration behind this endeavor.

Bibliography

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[3] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. IEEE, 2021.

[4] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[5] Susan Hillis, James Mercy, Adaugo Amobi, and Howard Kress. Global prevalence of past-year violence against children: a systematic review and minimum estimates. *Pediatrics*, 137(3), 2016.

[6] Marije Stoltenborgh, Marian J Bakermans-Kranenburg, Marinus H Van Ijzendoorn, and Lenneke RA Alink. Cultural–geographical differences in the occurrence of child physical abuse? a meta-analysis of global prevalence. *International Journal of Psychology*, 48(2):81–94, 2013.

[7] Diane DePanfilis. *Child neglect: A guide for prevention, assessment, and intervention*. US Department of Health and Human Services, Adminsitration for Children and ..., 2006.

[8] Cathy Spatz Widom. Longterm consequences of child maltreatment. In *Handbook of child maltreatment*, pages 225–247. Springer, 2013.

[9] U.S. Department of Health & Human Services, Administration for Children, Families, Youth Administration on Children, Families, and Children's Bureau. Child maltreatment 2020, 2022. Accessed: 16 Feb 2025.

[10] Xiangming Fang, Derek S Brown, Curtis S Florence, and James A Mercy. The economic burden of child maltreatment in the united states and implications for prevention. *Child abuse & neglect*, 36(2):156–165, 2012.

[11] Wendy A Walsh, Lisa M Jones, Theodore P Cross, and Tonya Lippert. Prosecuting child sexual abuse: The importance of evidence type. *Crime & Delinquency*, 56(3):436–454, 2010.

[12] Kamala London, Maggie Bruck, Daniel B Wright, and Stephen J Ceci. Review of the contemporary literature on how children report sexual abuse to others: Findings, methodological issues, and implications for forensic interviewers. *Memory*, 16(1):29–47, 2008.

[13] Diana M Elliott and John Briere. Forensic sexual abuse evaluations of older children: Disclosures and symptomatology. *Behavioral Sciences & the Law*, 12(3):261–277, 1994.

[14] Mark D Everson and Jose Miguel Sandoval. Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements. *Child abuse & neglect*, 35(4):287–298, 2011.

[15] Michael E Lamb, Irit Hershkowitz, Yael Orbach, and Phillip W Esplin. *Tell me what happened: Structured investigative interviews of child victims and witnesses*, volume 56. John Wiley & Sons, 2011.

[16] Michael E Lamb, Yael Orbach, Irit Hershkowitz, Phillip W Esplin, and Dvora Horowitz. A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the nichd investigative interview protocol. *Child abuse & neglect*, 31(11-12):1201–1231, 2007.

[17] Chris Newlin, Linda Cordisco Steele, Andra Chamberlin, Jennifer Anderson, Julie Kenniston, Amy Russell, Heather Stewart, and Viola Vaughan-Eden. *Child forensic interviewing: Best practices*. US Department of Justice, Office of Justice Programs, Office of Juvenile . . . , 2015.

[18] Gunn-Astrid Baugerud, Miriam S Johnson, Helle BG Hansen, Svein Magnussen, and Michael E Lamb. Forensic interviews with preschool children: An analysis of extended interviews in norway (2015–2017). *Applied cognitive psychology*, 34(3):654–663, 2020.

[19] Gunn Astrid Baugerud, Ragnhild Klingenberg Røed, Helle BG Hansen, Julie Schøning Poulsen, and Miriam S Johnson. Evaluating child interviews conducted by child protective services workers and police investigators. *The British Journal of Social Work*, 53(5):2784–2803, 2023.

[20] Ann-Christin Cederborg, Yael Orbach, Kathleen J Sternberg, and Michael E Lamb. Investigative interviews of child witnesses in sweden. *Child abuse & neglect*, 24(10):1355–1361, 2000.

[21] Kathleen Coulborn Faller. Forty years of forensic interviewing of children suspected of sexual abuse, 1974–2014: Historical benchmarks. *Social Sciences*, 4(1):34–65, 2014.

[22] Mairi S Benson and Martine B Powell. Evaluation of a comprehensive interactive training system for investigative interviewers of children. *Psychology, public policy, and law*, 21(3):309, 2015.

[23] Mireille Cyr and Michael E Lamb. Assessing the effectiveness of the nichd investigative interview protocol when interviewing french-speaking alleged victims of child sexual abuse in quebec. *Child Abuse & Neglect*, 33(5):257–268, 2009.

[24] Martine B Powell and Mark Barnett. Elements underpinning successful implementation of a national best-practice child investigative interviewing framework. *Psychiatry, Psychology and Law*, 22(3):368–377, 2015.

[25] Francesco Pompedda, Jan Antfolk, Angelo Zappalà, and Pekka Santtila. A combination of outcome and process feedback enhances performance in simulations of child sexual abuse interviews using avatars. *Frontiers in psychology*, 8:1474, 2017.

[26] Martine B Powell, Ronald P Fisher, and Carolyn H Hughes-Scholes. The effect of intra-versus post-interview feedback during simulated practice interviews about child abuse. *Child abuse & neglect*, 32(2):213–227, 2008.

[27] Martine B Powell, Ronald P Fisher, and Carolyn H Hughes-Scholes. The effect of using trained versus untrained adult respondents in simulated practice interviews about child abuse. *Child Abuse & Neglect*, 32(11):1007–1016, 2008.

[28] Melanie Lawrie, Sonja P Brubacher, Martine B Powell, and David Boud. Forensic interviewers' perceptions of the utility of mock interviews with trained actors as a training tool for child interviewing. *Child abuse & neglect*, 106:104553, 2020.

[29] Melanie Lawrie, Sonja P Brubacher, Becky Earhart, Martine B Powell, Linda C Steele, and David Boud. Testing the effectiveness of a blended vulnerable witness training for forensic interviewers. *Journal of Family Trauma, Child Custody & Child Development*, 18(3):279–297, 2021.

[30] Martine B Powell, Sonja P Brubacher, and Gunn Astrid Baugerud. An overview of mock interviews as a training tool for interviewers of children. *Child Abuse & Neglect*, 129:105685, 2022.

[31] Annabelle Nicol, Zsofia A Szojka, Christopher D Watkins, Fiona Gabbert, and David La Rooy. A systematic examination of actor and trainee interviewer behaviour during joint investigative interviewing training. *Journal of Police and Criminal Psychology*, 38(3):593–606, 2023.

[32] Michael E Lamb, Deirdre A Brown, Irit Hershkowitz, Yael Orbach, and Phillip W Esplin. *Tell me what happened: Questioning children about abuse*. John Wiley & Sons, 2018.

[33] Olga Chernikova, Nicole Heitzmann, Matthias Stadler, Doris Holzberger, Tina Seidel, and Frank Fischer. Simulation-based learning in higher education: A meta-analysis. *Review of educational research*, 90(4):499–541, 2020.

[34] Helen Crompton, Matthew Bernacki, and Jeffrey A Greene. Psychological foundations of emerging technologies for teaching and learning in higher education. *Current Opinion in Psychology*, 36:101–105, 2020.

[35] Lijia Chen, Pingping Chen, and Zhijian Lin. Artificial intelligence in education: A review. *Ieee Access*, 8:75264–75278, 2020.

[36] Martine B Powell, Belinda Guadagno, and Mairi Benson. Improving child investigative interviewer performance through computer-based learning activities. *Policing and Society*, 26(4):365–374, 2016.

[37] Mairi Benson and Martine Powell. Organisational challenges to delivering child investigative interviewer training via e-learning. *International journal of police science & management*, 17(2):63–73, 2015.

[38] Sharon Casey and Martine B Powell. Usefulness of an e-simulation in improving social work student knowledge of best-practice questions. *Social work education*, 41(6):1253–1271, 2022.

[39] Kevin Charles Dalli. Technological acceptance of an avatar based interview training application: The development and technological acceptance study of the avbit application. Digitala Vetenskapliga Arkivet, Master Dissertation, Linnaeus University, Sweden, 2021.

[40] Belinda Guadagno and Martine Powell. E-simulations for the purpose of training forensic (investigative) interviewers. In *Professional education using e-simulations: Benefits of blended learning design*, pages 71–86. IGI Global, 2012.

[41] Ragnhild Klingenberg Røed, Martine B Powell, Michael A Riegler, and Gunn Astrid Baugerud. A field assessment of child abuse investigators' engagement with a child-avatar to develop interviewing skills. *Child Abuse & Neglect*, 143:106324, 2023.

[42] Francesco Pompedda. *Training in investigative interviews of children: Serious gaming paired with feedback improves interview quality.* Phd thesis, Åbo Akademi University, 2018.

[43] Niels Krause, Francesco Pompedda, Jan Antfolk, Angelo Zappalá, and Pekka Santtila. The effects of feedback and reflection on the questioning style of untrained interviewers in simulated child sexual abuse interviews. *Applied Cognitive Psychology*, 31(2):187–198, 2017.

[44] Francesco Pompedda, Angelo Zappalà, and Pekka Santtila. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law*, 21(1):28–52, 2015.

[45] Shumpei Haginoya, Shota Yamamoto, Francesco Pompedda, Makiko Naka, Jan Antfolk, and Pekka Santtila. Online simulation training of child sexual abuse interviews with feedback improves interview quality in japanese university students. *Frontiers in psychology*, 11:998, 2020.

[46] Ming Meng, Yufei Zhao, Bo Zhang, Yonggui Zhu, Weimin Shi, Maxwell Wen, and Zhaoxin Fan. A comprehensive taxonomy and analysis of talking head synthesis: Techniques for portrait generation, driving mechanisms, and editing. *arXiv preprint arXiv:2406.10553*, 2024.

[47] Katja Zibrek, Elena Kokkinara, and Rachel McDonnell. The effect of realistic appearance of virtual characters in immersive environments-does the character's personality play a role? *IEEE transactions on visualization and computer graphics*, 24(4):1681–1690, 2018.

[48] Dongsik Jo, Ki-Hong Kim, and Gerard Jounghyun Kim. Effects of avatar and background representation forms to co-presence in mixed reality (mr) tele-conference systems. In *SIGGRAPH ASIA 2016 virtual reality meets physical reality: modelling and simulating virtual humans and environments*, pages 1–4. 2016.

[49] Marc Erich Latoschik, Daniel Roth, Dominik Gall, Jascha Achenbach, Thomas Waltemate, and Mario Botsch. The effect of avatar realism in immersive social virtual realities. In *Proceedings of the 23rd ACM symposium on virtual reality software and technology*, pages 1–10, 2017.

[50] Maia Garau. *The impact of avatar fidelity on social interaction in virtual environments*. University of London, University College London (United Kingdom), 2003.

[51] Daniel Roth, Jean-Luc Lugrin, Dmitri Galakhov, Arvid Hofmann, Gary Bente, Marc Erich Latoschik, and Arnulph Fuhrmann. Avatar realism and social interaction quality in virtual reality. In *2016 IEEE virtual reality (VR)*, pages 277–278. IEEE, 2016.

[52] Samah Mansour, Mostafa El-Said, Carolyn Rude-Parkins, and Jagadeesh Nandigam. The interactive effect of avatar visual fidelity and behavioral fidelity in the collaborative virtual reality environment on the perception of social interaction. *WSEAS Transactions on Communications*, 5(8):1501–1509, 2006.

[53] Frank Biocca. The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of computer-mediated communication*, 3(2):JCMC324, 1997.

[54] Nissoite Abdou, Milana Dzhamalova, Samuel Désirée, Théo Le Goff, Bruno Mesz, and Jean-Christophe Sakdavong. The mediating effect of presence on musical emotions in virtual environments. In *International Conference on Extended Reality*, pages 3–20. Springer, 2024.

[55] Andrea Stevenson Won, Jeremy Bailenson, Jimmy Lee, and Jaron Lanier. Homuncular flexibility in virtual reality. *Journal of Computer-Mediated Communication*, 20(3):241–259, 2015.

[56] A McMahan. Immersion, engagement, and presence in the video game theory reader, by mark jp wolf and bernard perron, 67-76, 2003.

[57] Eduard Zell, Carlos Aliaga, Adrian Jarabo, Katja Zibrek, Diego Gutierrez, Rachel McDonnell, and Mario Botsch. To stylize or not to stylize? the effect of shape and material stylization on the perception of computer-generated faces. *ACM Transactions on Graphics (TOG)*, 34(6):1–12, 2015.

[58] Ferran Argelaguet, Ludovic Hoyet, Michaël Trico, and Anatole Lécuyer. The role of interaction in virtual embodiment: Effects of the virtual hand representation. In *2016 IEEE virtual reality (VR)*, pages 3–10. IEEE, 2016.

[59] Steve Benford, John Bowers, Lennart E Fahlén, Chris Greenhalgh, and Dave Snowdon. User embodiment in collaborative virtual environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 242–249, 1995.

[60] Steve Benford, John Bowers, Lennart E Fahlén, Chris Greenhalgh, and Dave Snowdon. Embodiments, avatars, clones and agents for multi-user, multi-sensory virtual worlds. *Multimedia Systems*, 5:93–104, 1997.

[61] Thomas Waltemate, Dominik Gall, Daniel Roth, Mario Botsch, and Marc Erich Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE transactions on visualization and computer graphics*, 24(4):1643–1652, 2018.

[62] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100, 2012.

[63] Mincheol Shin, Se Jung Kim, and Frank Biocca. The uncanny valley: No need for any further judgments when an avatar looks eerie. *Computers in Human Behavior*, 94:100–109, 2019.

[64] Harry Brenton, Marco Gillies, Daniel Ballin, and David Chatting. The uncanny valley: does it exist. In *Proceedings of conference of human computer interaction, workshop on human animated character interaction*. Citeseer, 2005.

[65] Matias Volante, Sabarish V Babu, Himanshu Chaturvedi, Nathan Newsome, Elham Ebrahimi, Tania Roy, Shaundra B Daily, and Tracy Fasolino. Effects of virtual human appearance fidelity on emotion contagion in affective inter-personal simulations. *IEEE transactions on visualization and computer graphics*, 22(4):1326–1335, 2016.

[66] Brian Ries, Victoria Interrante, Michael Kaeding, and Lane Phillips. Analyzing the effect of a virtual avatar's geometric and motion fidelity on ego-centric spatial perception in immersive virtual environments. In *Proceedings of the 16th ACM symposium on virtual reality software and technology*, pages 59–66, 2009.

[67] Pegah Salehi, Abdolah Chalechale, and Maryam Taghizadeh. Generative adversarial networks (gans): An overview of theoretical model, evaluation metrics, and recent developments. *arXiv preprint arXiv:2005.13178*, 2020.

[68] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[69] Unity Technologies. Unity real-time development platform, 2025. Accessed: 2025-05-20.

[70] Yang Pan, KiHong Kim, JuneSok Lee, YuanZi Sang, and JiIn Cheon. Research on the application of digital human production based on photoscan realistic head 3d scanning and unreal engine metahuman technology in the metaverse. *International journal of advanced smart convergence*, 11(3):102–118, 2022.

[71] NVIDIA. NVIDIA Audio2Face 3D. https://build.nvidia.com/nvidia/audio2face-3d, 2025. Accessed: 2025-04-04.

[72] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

[73] Xiachong Feng, Longxu Dou, Ella Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma, and Lingpeng Kong. A survey on large language model-based social agents in game-theoretic scenarios. *arXiv preprint arXiv:2412.03920*, 2024.

[74] Carmit Katz, Irit Hershkowitz, Lindsay C Malloy, Michael E Lamb, Armita Atabaki, and Sabine Spindler. Non-verbal behavior of children who disclose or do not disclose child abuse in investigative interviews. *Child abuse & neglect*, 36(1):12–20, 2012.

[75] Tony W Buchanan, Jacqueline S Laures-Gore, and Melissa C Duff. Acute stress reduces speech fluency. *Biological psychology*, 97:60–66, 2014.

[76] Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, 467:73–82, 2022.

[77] Wenxiang Jiao, Michael Lyu, and Irwin King. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8002–8009, 2020.

[78] Yeun Joo Lee and Yong Gu Ji. Effects of visual realism on avatar perception in immersive and non-immersive virtual environments. *International Journal of Human–Computer Interaction*, 41(7):4362–4375, 2025.

[79] André Queirós, Daniel Faria, and Fernando Almeida. Strengths and limitations of qualitative and quantitative research methods. *European journal of education studies*, 2017.

[80] Gordana Dodig-Crnkovic. Scientific methods in computer science. In *Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Suecia*, pages 126–130. sn, 2002.

[81] Peter J. Denning, Douglas E Comer, David Gries, Michael C. Mulder, Allen Tucker, A. Joe Turner, and Paul R Young. Computing as a discipline. *Computer*, 22(2):63–70, 1989.

[82] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[83] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS quarterly*, pages 75–105, 2004.

[84] Noemí Pereda, Georgina Guilera, Maria Forns, and Juana Gómez-Benito. The prevalence of child sexual abuse in community and student samples: A meta-analysis. *Clinical psychology review*, 29(4):328–338, 2009.

[85] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[86] Merle Spriggs. Children and bioethics: clarifying consent and assent in medical and research settings. *British medical bulletin*, 145(1):110–119, 2023.

[87] Vijaytha Muralidharan, Alyssa Burgart, Roxana Daneshjou, and Sherri Rose. Recommendations for the use of pediatric data in artificial intelligence and machine learning accept-ai. *NPJ Digital Medicine*, 6(1):166, 2023.

[88] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Tania Sari Bonaventura, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, et al. Deepfake media forensics: State of the art and challenges ahead. *arXiv preprint arXiv:2408.00388*, 2024.

[89] Sylvie Delacroix and Catrina Denvir. Virtually teaching ethics: Experiencing the discrepancy between abstract ethical stands and actual behaviour using immersive virtual reality. *S. Delacroix and C. Denvir, "Virtually teaching ethics*, 2019.

[90] Olivia Drewett, Gayle Hann, Marco Gillies, Carmel Sher, Sylvie Delacroix, Xueni Pan, Tara Collingwoode-Williams, and Caroline Fertleman. A discussion of the use of virtual reality for training healthcare practitioners to recognize child protection issues. *Frontiers in public health*, 7:255, 2019.

[91] Divine Maloney, Guo Freeman, and Andrew Robb. Social virtual reality: ethical considerations and future directions for an emerging research space. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 271–277. IEEE, 2021.

[92] Ben Kenwright. Virtual reality: ethical challenges and dangers [opinion]. *IEEE Technology and Society Magazine*, 37(4):20–25, 2018.

[93] Caroline Fertleman, Phoebe Aubugeau-Williams, Carmel Sher, Ai-Nee Lim, Sophie Lumley, Sylvie Delacroix, and Xueni Pan. A discussion of virtual reality as a new tool for training healthcare professionals. *Frontiers in public health*, 6:44, 2018.

[94] Michael Madary and Thomas K Metzinger. Real virtuality: A code of ethical conduct. recommendations for good scientific practice and the consumers of vr-technology. *Frontiers in Robotics and AI*, 3:180932, 2016.

[95] Jon Rueda and Francisco Lara. Virtual reality and empathy enhancement: Ethical aspects. *Frontiers in Robotics and AI*, page 160, 2020.

[96] Andy Nguyen, Ha Ngan Ngo, Yvonne Hong, Belle Dang, and Bich-Phuong Thi Nguyen. Ethical principles for artificial intelligence in education. *Education and information technologies*, 28(4):4221–4241, 2023.

[97] Vajira Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven A Hicks, Hugo L Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A Riegler. Singan-seg: Synthetic training data generation for medical image segmentation. *PloS one*, 17(5):e0267976, 2022.

[98] Saeed Shafiee Sabet, Cise Midoglu, Syed Zohaib Hassan, Pegah Salehi, Gunn Astrid Baugerud, Carsten Griwodz, Miriam Johnson, Michael Alexander Riegler, and Pål Halvorsen. Comparison of crowdsourced and remote subjective user studies: A case study of investigative child interviews. In *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2022.

[99] Syed Zohaib Hassan, Pegah Salehi, Michael Alexander Riegler, Miriam Sinkerud Johnson, Gunn Astrid Baugerud, PÅL Halvorsen, and Saeed Shafiee Sabet. A virtual reality talking avatar for investigative interviews of maltreat children. In *Proceedings of the 19th international conference on content-based multimedia indexing*, pages 201–204, 2022.

[100] Syed Zohaib Hassan, Saeed Shafiee Sabet, Pegah Salehi, Hayley Ko, Ingvild Riiser, Miriam Johnson, Gunn Astrid Baugerud, Michael Alexander Riegler, and Pål Halvorsen. A comparative study of interactive environments for investigative interview of a virtual child avatar. In *2022 IEEE International Symposium on Multimedia (ISM)*, pages 194–201. IEEE, 2022.

[101] Ragnhild Klingenberg Røed, Gunn Astrid Baugerud, Syed Zohaib Hassan, Saeed S Sabet, Pegah Salehi, Martine B Powell, Michael A Riegler, Pål Halvorsen, and Miriam S Johnson. Enhancing questioning skills through child avatar chatbot training with feedback. *Frontiers in Psychology*, 14:1198235, 2023.

[102] Gunn-Astrid Baugerud, Miriam S Johnson, Rachel Dianiska, Ragnhild K Røed, Martine B Powell, Michael E Lamb, Syed Zohaib Hassan, Saaed S Sabet, Steven Hicks, Pegah Salehi, et al. Using an ai-based avatar for interviewer training

at children's advocacy centers: Proof of concept. *Child maltreatment*, page 10775595241263017, 2024.

[103] Samaneh Taghizadeh. Benchmarking the user experience of different ai talking head generation. Master thesis, Oslo Metropolitan University, Oslo, Norway, 2023. Department of Computer Science, Faculty of Technology, Art and Design.

[104] Alexander William Ingvarsson Hals. Impact of different type of child avatar interactions on user quality of experience. Master thesis, Oslo Metropolitan University, Oslo, Norway, 2022. Department of Computer Science, Faculty of Technology, Art and Design.

[105] Gunn Astrid Baugerud, Miriam S Johnson, Ragnhild Klingenberg Røed, Michael E Lamb, Martine Powell, Vajira Thambawita, Steven A Hicks, Pegah Salehi, Syed Zohaib Hassan, Pål Halvorsen, et al. Multimodal virtual avatars for investigative interviews with children. In *Proceedings of the 2021 ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 2–8, 2021.

[106] Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.

[107] Deepfakes. Deepfakes faceswap. https://github.com/deepfakes/faceswap, 2023. Accessed: 3 April 2025.

[108] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre De Brebisson, and Yoshua Bengio. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442*, 2017.

[109] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127:1767–1779, 2019.

[110] Syed Zohaib Hassan, Pegah Salehi, Ragnhild Klingenberg Røed, Pål Halvorsen, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pierre Lison, Michael Riegler, Michael E Lamb, Carsten Griwodz, et al. Towards an ai-driven talking avatar in virtual reality for investigative interviews of children. In *Proceedings of the 2nd Workshop on Games Systems*, pages 9–15, 2022.

[111] UMA Steering Group. Unity multipurpose avatar, 2025. Accessed: April 4, 2025.

[112] Crazy Minnow Studio. Salsa lipsync suite. https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442. Accessed: 2025-04-02.

[113] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.

[114] IBM. Ibm watson products. https://www.ibm.com/products. Accessed: 2025-04-02.

[115] Pegah Salehi, Syed Zohaib Hassan, Saeed Shafiee Sabet, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pål Halvorsen, and Michael A Riegler. Is more realistic better? a comparison of game engine and gan-based avatars for investigative interviews of children. In *Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 41–49, 2022.

[116] Pegah Salehi, Syed Zohaib Hassan, Myrthe Lammerse, Saeed Shafiee Sabet, Ingvild Riiser, Ragnhild Klingenberg Røed, Miriam S Johnson, Vajira Thambawita, Steven A Hicks, Martine Powell, et al. Synthesizing a talking child avatar to train interviewers working with maltreated children. *Big data and cognitive computing*, 6(2):62, 2022.

[117] Pegah Salehi, Syed Zohaib Hassan, Gunn Astrid Baugerud, Martine Powell, Miriam S Johnson, Dag Johansen, Saeed Shafiee Sabet, Michael A Riegler, and Pål Halvorsen. A theoretical and empirical analysis of 2d and 3d virtual environments in training for child interview skills. *IEEE Access*, 2024.

[118] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[119] Toonify. Toonify yourself. https://toonify.photos/, 2023. Accessed: 2025-04-04.

[120] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021.

[121] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020.

[122] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (ToG)*, 40(6):1–17, 2021.

[123] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.

[124] Barney Dalgarno and Mark JW Lee. What are the learning affordances of 3-d virtual environments? *British journal of educational technology*, 41(1):10–32, 2010.

[125] Caroline Kuhne, Eda D Kecelioglu, Steven Maltby, Rebecca J Hood, Brendon Knott, Elizabeth Ditton, Frederick Rohan Walker, and Murielle G Kluge. Direct comparison of virtual reality and 2d delivery on sense of presence, emotional and physiological outcome measures. *Frontiers in Virtual Reality*, 4:1211001, 2023.

[126] HMA Namrouti, C Sik-Lanyi, and T Guzsvinecz. Comparison between 2d and 3d icons as menus in virtual reality assessing the usability of the menus and user satisfaction.

[127] Jonghyeon Ka, Haryun Kim, Jihyung Kim, and Wooksung Kim. Analysis of virtual reality teaching methods in engineering education: assessing educational effectiveness and understanding of 3d structures. *Virtual Reality*, 29(1):17, 2025.

[128] Syed Zohaib Hassan, Saeed Shafiee Sabet, Michael Alexander Riegler, Gunn Astrid Baugerud, Hayley Ko, Pegah Salehi, Ragnhild Klingenberg Røed, Miriam Johnson, and Pål Halvorsen. Enhancing investigative interview training using a child

avatar system: a comparative study of interactive environments. *Scientific Reports*, 13(1):20403, 2023.

[129] Pegah Salehi, Syed Zohaib Hassan, Gunn Astrid Baugerud, Martine B Powell, M Cayetana López Cano, Miriam S Johnson, Ragnhild Klingenberg Røed, Dag Johansen, Saeed Shafiee Sabet, Michael A Riegler, et al. Immersive virtual reality in child interview skills training: A comparison of 2d and 3d environments. In *MMVE@ MMSys*, pages 1–7, 2024.

[130] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024.

[131] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024.

[132] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024.

[133] Haiyang Liu, Xingchao Yang, Tomoya Akiyama, Yuantian Huang, Qiaoge Li, Shigeru Kuriyama, and Takafumi Taketomi. Tango: Co-speech gesture video reenactment with hierarchical audio motion embedding and diffusion interpolation. *arXiv preprint arXiv:2410.04221*, 2024.

[134] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022.

[135] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7568–7578, 2023.

[136] Renan Vinicius Aranha, Cléber Gimenez Corrêa, and Fátima LS Nunes. Adapting software with affective computing: a systematic review. *IEEE Transactions on Affective Computing*, 12(4):883–899, 2019.

[137] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[138] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[139] Eanes Torres Pereira, Herman Martins Gomes, Luciana Ribeiro Veloso, and Moisés Roberto A Mota. Empirical evidence relating eeg signal duration to emotion classification performance. *IEEE Transactions on Affective Computing*, 12(1):154–164, 2018.

[140] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[141] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[142] Daniel McDuff and Mary Czerwinski. Designing emotionally sentient agents. *Communications of the ACM*, 61(12):74–83, 2018.

[143] Tarun Lalwani, Shashank Bhalotia, Ashish Pal, Vasundhara Rathod, and Shreya Bisen. Implementation of a chatbot system using ai and nlp. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Volume-6, Issue-3*, 2018.

[144] Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103, 2000.

[145] Rachael E Jack and Philippe G Schyns. The human face as a dynamic tool for social communication. *Current Biology*, 25(14):R621–R634, 2015.

[146] Michel Cabanac. What is emotion? *Behavioural processes*, 60(2):69–83, 2002.

[147] Stanley Schachter and Jerome Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5):379, 1962.

[148] Jonathan Gratch and Stacy Marsella. Evaluating the modeling and use of emotion in virtual humans. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, volume 1, pages 320–327. IEEE Computer Society, 2004.

[149] Jonathan Gratch and Stacy Marsella. Lessons from emotion psychology for the design of lifelike characters. *Applied Artificial Intelligence*, 19(3-4):215–233, 2005.

[150] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021.

[151] Foivos Paraperas Papantoniou, Panagiotis P Filntisis, Petros Maragos, and Anastasios Roussos. Neural emotion director: Speech-preserving semantic control of facial expressions in" in-the-wild" videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18781–18790, 2022.

[152] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 716–731. Springer, 2020.

[153] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[154] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–13, 2023.

[155] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20687–20697, 2023.

Part II

Research Papers

Article I: Multimodal Virtual Avatars for Investigative Interviews with Children

Authors: Gunn Astrid Baugerud, Miriam S. Johnson, Ragnhild Klingenberg Røed, Michael E. Lamb, Martine Powell, Vajira Thambawita, Steven A. Hicks, *Pegah Salehi*, Syed Zohaib Hassan, Pål Halvorsen, Michael A. Riegler

Candidate's contributions: The candidate contributed to the paper by reviewing literature on visual avatar technologies and conducting a technical evaluation of face-swapping and lip-syncing tools, focusing on their limitations in generating age-appropriate child avatars. They also participated in drafting sections on technical implementation (e.g., lip-syncing challenges, anatomical mismatches in adult-to-child face-swapping) and contributed to the conceptual framework for integrating multimodal AI components, including visual generation, dialogue systems, and speech processing.

Published in: Proceedings of the ACM Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR) - Taipei Taiwan - November 16-19, 2021

DOI: 10.1145/3463944.3469269

Contributed to: This work contributed to the **RO-1**.

Multimodal Virtual Avatars for Investigative Interviews with Children

Gunn Astrid Baugerud
gunnba@oslomet.no
Oslo Metropolitan University
Oslo, Norway

Miriam S. Johnson
mirsin@oslomet.no
Oslo Metropolitan University
Oslo, Norway

Ragnhild Klingenberg Røed
rar@oslomet.no
Oslo Metropolitan University
Oslo, Norway

Michael E. Lamb
mel37@cam.ac.uk
University of Cambridge
Cambridge, UK

Martine Powell
martine.powell@griffith.edu.au
Griffith University
Brisbane, Australia

Vajira Thambawita*
vajira@simula.no
SimulaMet
Oslo, Norway

Steven A. Hicks*
steven@simula.no
SimulaMet
Oslo, Norway

Pegah Salehi*
pghsalehi@gmail.com
SimulaMet
Oslo, Norway

Syed Zohaib Hassan*
syed@simula.no
SimulaMet
Oslo, Norway

Pål Halvorsen*
paalh@simula.no
SimulaMet
Oslo, Norway

Michael A. Riegler†
michael@simula.no
SimulaMet
Oslo, Norway

ABSTRACT

In this article, we present our ongoing work in the field of training police officers who conduct interviews with abused children. The objectives in this context are to protect vulnerable children from abuse, facilitate prosecution of offenders, and ensure that innocent adults are not accused of criminal acts. There is therefore a need for more data that can be used for improved interviewer training to equip police with the skills to conduct high-quality interviews. To support this important task, we propose to research a training program that utilizes different system components and multimodal data from the field of artificial intelligence such as chatbots, generation of visual content, text-to-speech, and speech-to-text. This program will be able to generate an almost unlimited amount of interview and also training data. The goal of combining all these different technologies and datatypes is to create an immersive and interactive child avatar that responds in a realistic way, to help to support the training of police interviewers, but can also produce synthetic data of interview situations that can be used to solve different problems in the same domain.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Interactive learning environments**; **E-learning**.

KEYWORDS

Virtual avatar ; artificial intelligence ; chatbots ; generated video ; speech-to-text-to-speech ; interview training

ACM Reference Format:
Gunn Astrid Baugerud, Miriam S. Johnson, Ragnhild Klingenberg Røed, Michael E. Lamb, Martine Powell, Vajira Thambawita, Steven A. Hicks, Pegah Salehi, Syed Zohaib Hassan, Pål Halvorsen, and Michael A. Riegler. 2021. Multimodal Virtual Avatars for Investigative Interviews with Children. In *Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3463944.3469269

1 INTRODUCTION

Child abuse and maltreatment is a severe problem worldwide, and has enormous immediate and long-term repercussions. This has led to a cascade of consequences such as death, cognitive impairments, behavioral and social problems, substance abuse, delinquency, disruptions in neurobiological development, and serious mental health problems [49]. Child sexual abuse alone is identified by the WHO as substantively contributing to the global burden of disease [37]. To address this problem, law enforcement and Child Protective Services (CPS) play important roles in safeguarding children to promote their welfare and protect them from harm [16]. In forensic interviews with children, they are often both victims and principal witnesses to the criminal activity. The lack of corroborative evidence and information gained in investigative interviews thus plays a crucial role in the investigation of child abuse [34].

*Also affiliated with Oslo Metropolitan University, Norway
†Also affiliated with UIT The Arctic University of Norway, Tromsø, Norway

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICDAR '21, August 21–24, 2021, Taipei, Taiwan
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8529-9/21/08…$15.00
https://doi.org/10.1145/3463944.3469269

Particularly in cases of alleged domestic violence and child sexual abuse (CSA), the child is usually the only available source of information about the abusive incidents [8]. Most sexually abused children will not have any physical signs of abuse [3]. This means that the progress of the police investigation depends upon the child's account of the incident, and the interviewer's ability to maximize the quality and quantity of the information obtained from the child. The investigative interview is a vital component of a child abuse investigation and a prime opportunity for investigators to elicit accurate and detailed information from alleged victims of child abuse. This is a challenging and complex task and it requires the interviewer to understand children's vulnerability, memory, communicative, and social development, in addition to the importance of getting the interviewer to use supportive and appropriate questioning styles such as free recall questions and avoiding suggestive and closed questions [32, 33]. More than three decades of research have given us invaluable knowledge about how to conduct these interviews and what constitutes "best practice guidelines" [11, 12]. These guidelines imply that police officers interviewing children should use open-ended questions and avoid option-posing and suggestive questions [10, 34].

However, several international studies from different countries have revealed that poor quality investigative interviews are a widespread problem [11, 24], even leading to invalidation of evidence and dismissal of cases due to formal errors. Specifically, research and examples from practice suggest that practitioners often fail to follow best-practices interview guidelines despite huge investments in training programs [31]. The lack of both resources and training opportunities in realistic interview settings brings forth the need to develop cost-effective training programs that are more interactive than the traditional classroom models. For this to be effective, training must also be associated with spaced practice and systematic, detailed, and continuous feedback on an individual level [7]. Recent research has demonstrated that interactive computer-based learning activities can improve investigative interviewers' performance in questioning children [39]. When high-quality interviews are conducted, children as young as three years old can provide detailed and accurate statements of stressful events they have experienced [6, 14, 20]. However, children's metacognitive awareness is still developing throughout the pre-school and early school years, and young children are more vulnerable to suggestive interview strategies, confusion arising from complex questioning, and lack of support during the interview [19].

In fact, the quality of the interviewer's questioning may be a more important factor in eliciting informative narratives from children than the individual cognitive and social competence of the child [17, 30]. In particular, interviewers may find it difficult to adhere to 'best-practice' method when interviewing pre-school children [5].

Recent research has found that the use of gaming and avatars in interview training may be beneficial, and in particular, when interviewer training is combined with feedback and delivered over longer time periods [32, 38]. Currently, however, there are no systems that allow realistic avatar interviewees to respond verbally to any question asked by the interviewer. Interview practice currently requires live, trained actors, making training very costly to deliver - it can take up to three months to train an actor, and hourly rates are high. With respect to the future training of personnel, recent developments in technology have to a large degree changed the field of image and video manipulation (e.g., [22, 52]). Modern open software frameworks such as TensorFlow [2] and Keras [13] have enabled a large growth in artificial intelligence (AI) research where generative adversarial networks (GANs) [18], and autoencoders [41] can be used to manipulate images and videos in a way that is almost undetectable.

Our idea is to use this type of technology to develop realistic avatars. Such an avatar will be a virtual minor, providing a safe medium for police officers to train for the interview situation. This may have significant implications for the quality of child forensic interviews, i.e., improving the quality of simulated investigative interviews will improve the quality of real ones. AI-based training programs could simulate a range of realistic scenarios that police officers face when conducting child forensic interviews, and the few studies conducted using computer-assisted training and learning show promising results [9, 28]. In this paper, we describe our work-in-progress idea to develop a child forensic interview avatar. Figure 1 gives an overview of the complete architecture and interaction flow between the interviewer and the avatar. We present some related work in section 2 before we detail our idea in section 3. Further, we present our initial work in section 4. Finally, section 5 concludes the paper.

2 BACKGROUND AND RELATED WORK

Work related to the proposed research involves different fields and methods, which are discussed in this section.

2.1 Interview Teaching

Although many of the training programs worldwide are based on the same body of research and professionally endorsed recommendations (e.g., American Professional Society on the Abuse of Children [4], Home Office [21]; Norway, Justis-og beredskapsdepartementet [25]), there are wide variations in training, and most countries have more informal strategies and training schedules [40]. Even if the knowledge exists on how to conduct these interviews, there seems to be a huge problem for professionals to follow 'best practice' guidelines and avoid contaminating childrens' reports about their experiences [12, 27, 35, 43]. This may have a serious impact as the interview of the child might be seen as unreliable and could even be discounted in court proceedings [10].

Previously, Finland, Norway, Sweden, the UK, and the USA have developed classroom-based training programs to reduce the gap between research-based *best practices* and actual interviewer practice to improve faithful implementation of the interview methods. However, this type of training gives little room for individualized feedback, and ongoing training [31].

To be effective, training must include several features: it must occur over time, learned skills must be practices in context, and trainees must receive prompt, individual feedback over time [7, 17]. Classroom training cannot easily provide all these features in ways that are both affordable and accessible to those requiring training, who are often working full-time and have complicated schedules. Recent studies show that interactive computer-based learning activities can improve the investigators interviewing performance [31]. Researchers in Australia and Finland have explored the uses of

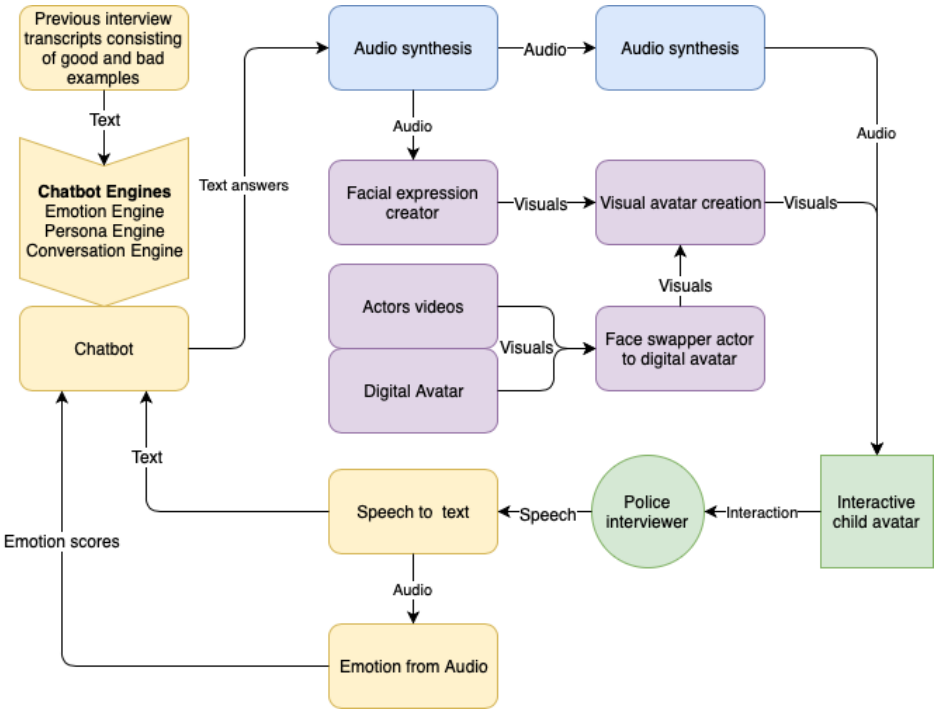


Figure 1: System architecture. Green blocks mark the interactive parts, yellow is text related, blue is audio, and purple marks the part of the system related to the visualization.

e-learning and avatars in interview- training and found them to be beneficial, in particular when interviewer training is delivered over longer time periods [9, 38, 39]. The Australian group developed an interactive training system, using mock interviews with trained actors, and simulated interviews with a child avatar. The participants were given online training in choosing effective questions, and engaged in simulated interviews with a virtual avatar several times a week. Results revealed enhanced interview quality post-training [7, 9, 39]. Interview training with avatars in combination with providing trainees with regular feedback also improved interview quality [28, 38]. Some studies using e-learning and avatars in training interview techniques currently exist in Australia and Finland. These studies have shown the benefits of replacing traditional classroom teaching with computer-assisted technology training programs, in particular when interviewer training is delivered over longer time periods [32]. The Australian group developed an interactive training system using mock interviews with real actors, and simulated interviews with a virtual child avatar as part of the program's learning activities. Results revealed enhanced interview quality post-training [7, 9, 39]. In two recent studies, the researchers used different computer-based child avatars with predefined memories about child sexual abuse. The combination of simulated interviews with avatars in addition to giving participants regular feedback, improved the interview quality. The participants used more recommended questions and fewer potentially damaging strategies [28, 38]. To improve interviewer skills, practice is the most fundamental activity. However, prior avatars have not provided the necessary level of interviewing practice because the technology did not allow avatar interviewees to answer any question. That is, the interviewers could only choose questions from a multiple-choice list. This means that interviewers gained experience identifying and selecting the best questions to ask, but did not gain experience recalling and asking them. For practice, trainees need to spontaneously generate the questions themselves. Currently, the only way to provide this practice is with live actors, which is very expensive and difficult to standardize.

2.2 Chatbots

Chatbots are typically used in applications such as e-commerce customer service, call centers, and internet gaming, where the chatbot typically uses machine learning to conduct a conversation via auditory or textual methods [50]. For the case of mimicking a conversation with a potentially abused child, we need to go beyond a simple question-answering version of chatbots [48]. To a certain extent, we will need emotions and feelings included in the conversations. Chatbots are able to have emotional conversations to a certain level, as demonstrated in the example presented in [51]. In addition to the conversation and emotions, we also need a person context. This will be important to increase the immersion of the trainee into the conversation. Persona-based neural conversation models are quite advanced and possible to create [36]. As one can see from the related work that has been discussed, the various available building blocks should allow a realistic child chatbot to

be developed. The challenge will be to combine the different parts into one entity.

2.3 Text-to-Speech and Speech-to-Text

Building a textual chatbot is one part of the challenge, but after a viable bot is trained, the textual conversation should be transformed at some point to audio. Text-to-speech research has developed considerably in recent years. An example is Google's Wavenet [47], which allows very realistic text to speech transformation using generative models. These methods are state of the art and also working efficiently, which makes them perfect for the presented system [26].

2.4 Avatars

Avatars are frequently used in various applications such as computer games. One way to generate an avatar is to use 3D models, which have recently become very realistic. However, in our scenario, we need a frontal view only. A large, complex 3D model of various avatars might be too costly and complex, and possibly not sufficiently realistic. Therefore, due to the potential of recent AI advances in the area of image and video manipulation, we investigate how an avatar can be built by live generation of fake videos.

2.5 Image Manipulation

One of the most interesting techniques to create such avatars is what currently goes under the term DeepFake, where one can combine and superimpose existing images and videos onto source images or videos using GANs. In this respect, the combination of source videos results in a fake video that shows an action or an event that never occurred in reality, e.g., creating fake news. For example, in January 2018, the desktop application FakeApp was launched, allowing users to easily create and share videos with faces swapped using an artificial neural network. Several frameworks are proposed and available. Face2Face [45] enables real-time facial reenactment of a target video sequence, i.e., animating the facial expressions of the target video by a source actor and re-rendering the manipulated output video in a photo-realistic fashion. Another software is pix2pix[1] [22] which uses conditional adversarial networks as a general-purpose solution to image-to-image translation. The DeepFaceLab[2] is another software tool using deep learning to identify and swap faces in videos, i.e., making fake videos. For example, the method was used to combine an image of Donald Trump and Nicolas Cage, putting Cage's face onto Trump's head[3]. GANimation[4] is a GAN conditioning tool based on Action Unit annotations to express anatomical facial movements[5] defining a human expression [42]. Finally, ICface is a versatile model for face animators that is driven by human interpretable control signals. ICface compared to GANimation modifies not only the facial expressions but also the head pose [46].

[1]https://phillipi.github.io/pix2pix/
[2]https://github.com/iperov/DeepFaceLab
[3]https://www.youtube.com/watch?time_continue=30&v=2R_aqHBClUQ
[4]https://github.com/albertpumarola/GANimation
[5]https://www.youtube.com/watch?v=AmUC4m6w1wo

3 AN AI-BASED INTERVIEW AVATAR

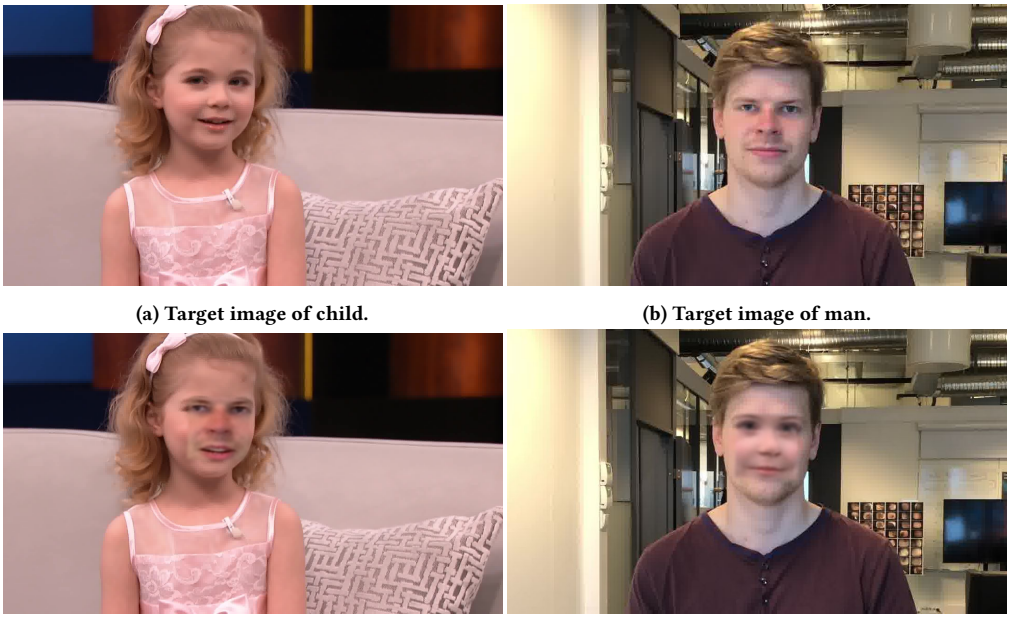
The planned study will use a randomized pre-post-training design to assess the effectiveness of an e-learning training program designed to improve the quality of investigative interview practice within two organizations: the police and the child protective services. Pre- and post-assessment interviews will be transcribed and subjected to an internationally recognized criteria-based analysis based on the NICHD-protocol and used in several studies that evaluate the quality in investigative interviews [11, 24].The coding scheme will differentiate of three main interviewer question categories: open-ended, directive, and suggestive questions. An examination of pre-and post-test comparisons will indicate whether the training significantly improved interviewer skills. The training program consists of two sessions. First, participants will complete a range of interactive exercises organized into 15 modules of approximately 6 hours each, focusing on interviewer skills such as recognizing the various question types, understanding child development in relation to memory and language, and learning best-practice interview technique, etc. Second, interactive computer-based learning activities that involve simulated interviews with a computer-generated child avatar, paired with feedback will be used. The avatar will consolidate and measure new knowledge gained during interviewer training because it provides a standardized measure of performance which can be applied easily in a range of contexts. As described above, the technological advances in multiple areas enable better and more realistic avatars. Our idea of the future police interviewer training system, see again figure 1, is based on an integration and combination of several areas, and the proposed avatar will consist of two main components. The first will be an interactive chatbot that will be trained using hours of 600 real investigated transcribed interviews from an ongoing National study. Then, our interview avatar concept is based on the DeepFake ideas where we generate an avatar model using child actors, where the system spontaneously generates videos using GANs to create realistic content like the "fake Obama video"[6]. The avatar's facial movements will be based on the chatbot's responses during the simulated interview, also taking into account emotion, giving a real-time re-enactment of a child's facial expressions.

4 ONGOING INITIAL WORK

Our preliminary efforts have been focused on research and the design of a proof of concept. Being a prototype, we simplify the scope to a system where we can manipulate a video or image of a person into lip-syncing a provided audio clip. We have experimented using the tools Faceswap [15] and ObamaNet [29], and the system described in Jamaludin et al. [23]. In this section, we will discuss our initial experiences and lessons learned using these tools.

We first tried the tool Faceswap [15]. This tool advertises being "the leading free and Open Source multi-platform Deepakes software" [1]. Faceswap can take two images and swap their faces. While this is not exactly our use case of lip-syncing the avatar itself, it could provide an alternative approach by projecting lips onto an avatar. Unfortunately, the tool does not work for our case. The tool requires a full video and many images of the face to be

[6]https://www.youtube.com/watch?v=gLoI9hAX9dw



(a) Target image of child.

(b) Target image of man.

(c) Face of fig. 2b on fig. 2a.

(d) Face of fig. 2a on fig. 2b.

Figure 2: Swapping faces using Faceswap [15].

transferred. Additionally, the tool cannot do processing in real-time. Initial results as seen in figure 2 show that Faceswap does not give us a realistic result when using the face of a man on a child. Using a man and a child who closely resemble could give a better result. For our use case, we can not rely on close resemblance. Therefore, Faceswap seems unsuited for achieving a realistic-looking avatar for our use case.

For realistic avatars, it seems better to manipulate the facial features of the subject in an image or video instead of pasting the features of a different subject such as in Faceswap. This approach is used for manipulating videos of President Barrack Obama[7] [29, 44]. We explored two methods: ObamaNet by Kumar et al. [29] and the code published by Suwajanakorn et al. [44]. Both methods move the mouth region of a video of Obama to sync with a given audio input. We could not get the code by Suwajanakorn to work. We moved on to ObamaNet, but the resulting video had an out-of-sync mouth movement. Further fine-tuning of the method might improve this. ObamaNet uses a hand-crafted 3D model to make the result more realistic. However, we need a method that works for child subjects and not only Obama. Modifying the method and hand-crafting child 3D models could be a way to achieve this.

Creating hand-crafted models could be a viable approach to creating a realistic avatar. However, the work by Jumaludin et al. [23] introduces a different approach[8]. Instead of relying on modeling, the method generates a video of a talking person based on a still image of the target. The method takes an audio file of a speech segment as input together with a still image of the target face and generates a video of the target face lip-synced with the audio. The authors state that the method runs in real-time and applies to both audios and faces not seen at training time. However, the code is not freely available, so we must re-implement the code based on the paper. We are not done with this process, meaning that we still need to work on implementing models and collecting data for both the visual avatar and the chatbots in order to be able to develop a realistic avatar.

[7] https://www.youtube.com/watch?v=cQ54GDm1eL0
[8] https://www.youtube.com/watch?v=-tOepR2IJdM

5 CONCLUSION

We have presented our ongoing work at the intersection of social science, psychology, and computer science, with the goal of creating a virtual realistic avatar using various multimedia content to support the training of investigative interviewers and the creation of synthetic data. This project is highly innovative, and offers the potential of methodological developments and applications, and has a huge potential to influence police and CPS training. The proposed system may involve small individual contributions of different components, but the large technical contribution will be the integration of systems and the interaction of all components with each other and with the police interviewers. Children can provide valuable information in a good quality investigative interview, and there is an urgent need for improvements, as well as the implementation of new and more effective analysis systems to be able to help and intervene on behalf of abused and maltreated children. Thus, the current project has a huge potential to focus future interviewer training, create data, and enhance the quality of investigative interviews, and may thus have important implications for children in need.

ACKNOWLEDGMENTS

This research is sponsored by the Research Council of Norway (RCN), project number #314690 ("Interview training of child-welfare and law-enforcement professionals interviewing maltreated children supported via artificial avatars").

REFERENCES

[1] 2019. Home of the leading free deepfakes faceswapping software. https://www.faceswap.dev/

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[3] Joyce A. Adams, Karen Farst, and Nancy D. Kellogg. 2018. Interpretation of Medical Findings in Suspected Child Sexual Abuse: An Update for 2018. *Journal of Pediatric and Adolescent Gynecology* 31 (2018), 225–231.

[4] American Professional Society on the Abuse of Children. 2002. *Guidelines for psychosocial evaluation of suspected sexual abuse in young children (Rev. ed.).* American Professional Society on the Abuse of Children, Chicago, IL: APSAC.

[5] Gunn-Astrid Baugerud, Miriam S Johnson, Helle BG Hansen, Svein Magnussen, and Michael E Lamb. 2020. Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015–2017). *Applied Cognitive Psychology* 34, 3 (2020), 654–663.

[6] Gunn Astrid Baugerud, Svein Magnussen, and Annika Melinder. 2014. High accuracy but low consistency in children's long-term recall of a real-life stressful event. *Journal of experimental child psychology* 126 (2014), 357–368.

[7] Mairi Benson and Martine Powell. 2015. Organizational challenges to delivering child investigative interviewer training via e-learning. *International Journal of Police Science and Management* 17 (2015), 63–73.

[8] Deirdre Brown and Michael Lamb. 2019. Forks in the road, routes chosen, and journeys that beckon: A selective review of scholarship on childrenss testimony. *Applied cognitive psychology* (2019), Special issue article 2018.

[9] Sonja Brubacher, Martine Powell, Helene Skouteris, and Belinda Guadagno. 2015. The effects of e-simulation interview training on teachers' use of open-ended questions. *CAN* 43 (2015), 95–103.

[10] Ann-Christin Cederborg, Charlotte Alm, Djaildes Lima da Silva Nises, and Michael E Lamb. 2013. Investigative interviewing of alleged child abuse victims: An evaluation of a new training programme for investigative interviewers. *Police Practice and Research* 14, 3 (2013), 242–254.

[11] Ann-Christin Cederborg, Yael Orbach, Kathleen Sternberg, and Michael Lamb. 2000. Investigative interviews of child witnesses in Sweden. *CAN* 24 (2000), 1355–1361.

[12] Ann-Christin Cederborg, Yael Orbach, Kathleen J Sternberg, and Michael E Lamb. 2000. Investigative interviews of child witnesses in Sweden. *Child Abuse & Neglect* 24, 10 (2000), 1355–1361.

[13] François Chollet et al. 2015. Keras: Deep learning library for theano and tensorflow. *URL: https://keras. io/k* (2015).

[14] Mireille Cyr and Michael Lamb. 2009. Assessing the Effectiveness of the NICHD Investigative Interview Protocol when Interviewing French-speaking Alleged Victims of Child Sexual Abuse in Quebec. *CAN* 33 (2009), 257–268.

[15] deepfakes. 2019. Faceswap. https://github.com/deepfakes/faceswap

[16] Louise Dixon, Daniel F Perkins, Catherine Hamilton-Giachritsis, and Leam A Craig. 2017. *The Wiley Handbook of what Works in Child Maltreatment: An Evidence-based Approach to Assessment and Intervention in Child Protection.* John Wiley & Sons.

[17] Katarina Finnil a, Nina Mahlberg, Pekka Santtila, Kenneth Sandnabba, and Pekka Niemi. 2003. Validity of a Test of Children's Suggestibility for Predicting Responses to Two Interview Situations Differing in Their Degree of Suggestiveness. *Journal of Experimental Child Psychology* 85 (2003), 23–49.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[19] Gail Goodman, Deborah Goldfarb, Jia Chong, and Lauren Goodman-Shaver. 2014. Children's Eyewitness Memory: The Influence of Cognitive and Socio-Emotional Factors. *Roger Williams University Law Review* 19, 2 (2014), Article 7.

[20] Gail S. Goodman, Jodi A. Quas, and Christin M. Ogle. 2010. Child Maltreatment and Memory. *Annual Review of Psychology* 61, 1 (2010), 325–351. https://doi.org/10.1146/annurev.psych.093008.100403

[21] Home Office. 2007. Achieving best evidence in criminal proceedings: Guidance on interviewing victims and witnesses, and using special measures.

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Proc. of CVPR*.

[23] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. 2019. You Said That?: Synthesising Talking Faces from Audio. *International Journal of Computer Vision* (13 Feb 2019). https://doi.org/10.1007/s11263-019-01150-y

[24] Miriam Johnson, Svein Magnussen, Christian Thorenson, Kyrre Lonnum, Lisa Victoria Burrell, and Annika Melinder. 2015. Best practice recommendations still fail to result in action: A national 10-year follow-up study of investigative inter-views in CSA cases. *Applied Cognitive Psychology* 29 (2015), 661–668.

[25] Justis- og beredskapsdepartementet. 2015. *Forskrift om avhør av barn og andre særlig sårbare fornærmede og vitner (tilrettelagte avhør) - Regulations of investigative interviews of children and other particularly vulnerable victims and witnesess.* Justis- og beredskapsdepartementet, lovdata.no.

[26] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient Neural Audio Synthesis. *arXiv preprint arXiv:1802.08435* (2018).

[27] Julia Korkman, Pekka Santtila, and N Kenneth Sandnabba. 2006. Dynamics of verbal interaction between interviewer and child in interviews with alleged victims of child sexual abuse. *Scandinavian journal of psychology* 47, 2 (2006), 109–119.

[28] Niels Krause, Francesco Pompedda, Jan Antfolk, Angelo Zappala, and Pekka Santtila. 2017. The Effects of Feedback and Reflection on the Questioning Style of Untrained Interviewers in Simulated Child Sexual Abuse Interviews. *Applied Cognitive Psychology* 31, 2 (2017), 187–198.

[29] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. 2018. ObamaNet: Photo-realistic lip-sync from text. *CoRR* abs/1801.01442 (2018). arXiv:1801.01442 http://arxiv.org/abs/1801.01442

[30] David La Rooy, Michael Lamb, and Amina Memon. 2011. Forensic interviews with children in Scotland: A survey of interview practices among police. *Journal of Police and Criminal Psychology* 26 (2011), 26–34.

[31] Michael Lamb. 2016. Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink? *American Psychologist* 71 (2016), 710–718.

[32] Michael Lamb, Deirdre Brown, Irit Hershkowitz, Yale Orbach, and Phillip Esplin. 2018. *Tell Me What Happened : Questioning Children about Abuse.* J.&W.&Sons.

[33] Michael E Lamb and Deirdre A Brown. 2006. Conversational apprentices: Helping children become competent informants about their own experiences. *British Journal of Developmental Psychology* 24, 1 (2006), 215–234.

[34] Michael E Lamb, David J La Rooy, Lindsay C Malloy, and Carmit Katz. 2011. *Children's testimony: A handbook of psychological research and forensic practice.* Vol. 53. John Wiley & Sons.

[35] Michael E Lamb, Yael Orbach, Kathleen J Sternberg, JAN Aldridge, Sally Pearson, Heather L Stewart, Phillip W Esplin, and Lynn Bowler. 2009. Use of a structured investigative protocol enhances the quality of investigative interviews with alleged victims of child sexual abuse in Britain. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 23, 4 (2009), 449–467.

[36] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155* (2016).

[37] World Health Organization et al. 2009. *Global health risks: mortality and burden of disease attributable to selected major risks.* Geneva: World Health Organization.

[38] Francesco Pompedda, Angelo Zappala, and Pekka Santtila. 2015. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime and Law* 21 (2015), 28–52.

[39] Martine Powell, Belinda Guadagno, and Mairi Benson. 2016. Improving child investigative interviewer performance through computer-based learning activities. *Policing and Society* 26 (2016), 365–374.

[40] Heather L Price and Kim P Roberts. 2011. The effects of an intensive training and feedback program on police and social workers' investigative interviews of children. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 43, 3 (2011), 235.

[41] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. Variational autoencoder for deep learning of images, labels and captions. In *Proc. of NIPS*. 2352–2360.

[42] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. 2018. GANimation: Anatomically-aware Facial Animation from a Single Image. In *Proc. of ECCV*.

[43] Kathleen J Sternberg, Michael E Lamb, and Irit Hershkowitz. 1996. Child sexual abuse investigations in Israel. *Criminal Justice and Behavior* 23, 2 (1996), 322–337.

[44] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Trans. Graph.* 36, 4, Article 95 (July 2017), 13 pages. https://doi.org/10.1145/3072959.3073640

[45] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, , and Matthias Niessner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. of CVPR*.

[46] Soumya Tripathy, Juho Kannala, and Esa Rahtu. 2020. Icface: Interpretable and controllable face reenactment using gans. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3385–3394.

[47] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499* (2016).

[48] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).
[49] Cathy Spatz Widom. 2014. Longterm consequences of child maltreatment. In *Handbook of child maltreatment*. Springer, 225–247.
[50] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *CIM* 13, 3 (2018), 55–75.
[51] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proc. of IEEE ICCV*.

Article II: Towards an AI-driven Talking Avatar in Virtual Reality for Investigative Interviews of Children

Authors: Syed Zohaib Hassan, *Pegah Salehi*, Ragnhild Klingenberg Røed, Pål Halvorsen, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pierre Lison, Michael Riegler, Michael E. Lamb, Carsten Griwodz, Saeed Shafiee Sabet

Candidate's contributions: The candidate played a central role in conceptualizing and designing the user study methodology to evaluate the VR-based AI child avatar system. She led the evaluation process, conducting comprehensive assessments of the prototype in terms of system performance, user experience, and training effectiveness for child protection professionals. As a primary contributor, the candidate authored key sections of the manuscript, including methodology, system architecture, and interpretation of user study results. She also ensured alignment between the technical implementation (e.g., VR integration, Unity game engine framework) and the practical training objectives for investigative interviews.

Published in: Proceedings of the Workshop on Games Systems (GameSys) - Athlone, Ireland - June 14-17, 2022

DOI: 10.1145/3534085.3534340

Contributed to: This work contributed to the **RO-1**.

Towards an AI-Driven Talking Avatar in Virtual Reality for Investigative Interviews of Children

Syed Zohaib Hassan
syed@simula.no
SimulaMet

Pegah Salehi
pegah@simula.no
SimulaMet

Ragnhild Klingenberg Røed
rar@oslomet.no
OsloMet

Pål Halvorsen
paalh@simula.no
SimulaMet

Gunn Astrid Baugerud
gunnba@oslomet.no
OsloMet

Miriam Sinkerud Johnson
mirsin@oslomet.no
OsloMet

Pierre Lison
plison@nr.no
Norwegian Computing Center

Michael Riegler
michael@simula.no
SimulaMet

Michael E. Lamb
mel37@cam.ac.uk
University of Cambridge

Carsten Griwodz
griff@ifi.uio.no
University of Oslo

Saeed Shafiee Sabet
saeed@simula.no
SimulaMet

ABSTRACT

Artificial intelligence (AI) and gaming systems have advanced to the stage where the current models and technologies can be used to address real-world problems. The development of such systems comes with different challenges, e.g., most of them related to system performance, complexity and user testing. Using a virtual reality (VR) environment, we have designed and developed a game-like system aiming to mimic an abused child that can help to assist police and child protection service (CPS) personnel in interview training of maltreated children. Current research in this area points to the poor quality of conducted interviews, and emphasises the need for better training methods. Information obtained in these interviews is the core piece of evidence in the prosecution process. We utilised advanced dialogue models, talking visual avatars, and VR to build a virtual child avatar that can interact with users. We discuss our proposed architecture and the performance of the developed child avatar prototype, and we present the results from the user study conducted with CPS personnel. The user study investigates the users' perceived quality of experience (QoE) and their learning effects. Our study confirms that such a gaming system can increase the knowledge and skills of the users. We also benchmark and discuss the system performance aspects of the child avatar. Our results show that the proposed prototype works well in practice and is well received by the interview experts.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
GameSys '22, June 14, 2022, Athlone, Ireland
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3534085.3534340

CCS CONCEPTS

• **Human-centered computing** → *User studies*; **Virtual reality**; • **Computing methodologies** → *Natural language processing*; **Animation**.

KEYWORDS

Avatar, Virtual Reality (VR), AI, Dialogue Model, Child Protection Services (CPS), Generative Adversarial Networks (GANs), Quality of Experience (QoE)

ACM Reference Format:
Syed Zohaib Hassan, Pegah Salehi, Ragnhild Klingenberg Røed, Pål Halvorsen, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pierre Lison, Michael Riegler, Michael E. Lamb, Carsten Griwodz, and Saeed Shafiee Sabet. 2022. Towards an AI-Driven Talking Avatar in Virtual Reality for Investigative Interviews of Children . In *2nd edition of the Game Systems Workshop (GameSys '22), June 14, 2022, Athlone, Ireland*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3534085.3534340

1 INTRODUCTION

Sexual and physical abuse of children is unfortunately a widespread problem in our society, with estimated prevalence rates suggesting that such abuse affects the lives of 300 million children around the world [41]. Although abuse seldom leads to the death of the child, it often causes long-term psychological and physical health problems [3, 5, 35, 38]. Recognizing and preventing child abuse is thus of vital importance for the child protection services (CPS) and police personnel. In cases of child abuse, children are not only the victims of the alleged crime, but often also the sole witnesses. Thus, in order to obtain reliable and detailed accounts of the alleged crimes and to help these children, investigative agencies must correctly perform high-quality interviews [18].

Research over the past three decades on children's cognitive development, combined with knowledge derived from many thousands of interviews conducted by police officers, social workers, medical professionals, and clinicians, has contributed to the development of 'best practice investigative interview guidelines. These

recommendations and guidance advise police and social service agencies around the world on how to conduct better investigative interviews with young victims [17, 22, 33]. Today, there is an international consensus among researchers and professionals that best practice is to ask open-ended question (e.g., "what happened after that?" where the child gets no clues that could influence the answer it gives). On the other hand, avoid asking suggestive or leading questions (e.g., "the man touched your private part, didn't he?" where clues are given and can lead the child to tell a certain story). Following the best practice guidelines is the only way to elicit reliable and detailed accounts of the incidents which are admissible in court [16]. It also decreases the risk of false reports, which can result in an innocent persons being charged and convicted [17]. Hence, the interview skills of CPS and police personnel greatly affect the outcomes of courts trials regarding abused children.

However, studies conducted in countries all over the world show that the quality of the interviews conducted with maltreated children is consistently poor, and interviewers often fail to follow the recommended best practices [1, 4, 13, 37]. Several studies have shown that only 2-8% of the questions asked in the average investigative interviews are open-ended, whereas suggestive or leading questions comprise from 16-60% of the questions asked [17].

Studies have shown that computer-based interactive learning can further improve investigative interviewing skills [26], this is not enough. However, the development of dynamic avatars has opened a new dimension for the transfer of knowledge. Currently, systems that employ such avatars lack generality and realism, using a multiple-choice questions format in which interviewers choose the best option offered. Although this does help interviewers select the right questions to ask, it does not help interviewers to verbalize and remember questions or question types [40]. Another issue with current avatars is that they do not seem visually realistic and do not show any emotions [23, 43].

This paper proposes an interview-training system using realistic avatars created using a game engine and virtual reality (VR). This system offers effective training of professionals in CPS and law enforcement, enabling them to consistently conduct high-quality investigative interviews of children who are alleged victims of sexual or physical violence. Users of this system will talk to an artificial abused child, having a story to tell, with the goal to elicit information which can be used in the court as evidence. Success of the interaction is judged based on the type of questions asked and closure of the story. In particular, we focus on the system challenges and quality aspects of such an avatar. Moreover, we describe the architecture of the proposed child avatar system, experiment with the system performance for both functionality and resource usage, and investigate the system's potential to enhance the competence of CPS and law enforcement professionals effectively. Our system is evaluated objectively by assessing its performance aspects and subjectively by conducting a user study targeting the various learning and quality of experience (QoE) aspects. A test based on the ITU-T Recommendation P.809 [11] was designed and conducted: an interactive test using experts in the field, to investigate the user experience and potential of the system for training CPS personnel. It shows that the proposed prototype is working well in practice and is well received by the experts.

In summary, the main contributions of our work are:(1) Development of the dialogue model to mimic an abused child; (2) Development of the prototype of an artificial intelligence (AI)-based child avatar in VR using the Unity game engine; (3)Evaluating the prototype in terms of system performance, user experience and learning effects.

2 RELATED WORKS

A large amount of research has been conducted to enhance the training of the CPS professionals. Powell et al [26] introduce a program to improve child investigative interviewer performance made up of computer-based training activities like watching videos, reading material and quizzes. These activities are followed by 10 minute mock assessment interviews where a trained actor play the role of a child. Empowering Interviewer Training (EIT) [25] is another program that simulates investigative interview. In EIT, child memories and responses are predefined, and rule-based algorithms are employed to select responses. An operator chooses a prerecorded video clip with different emotions which is shown to the users, based on the selected response. Moreover, Linnæus university and AvBIT Labs have developed an online interview setting, using prerecorded videos of a child avatars and audio responses. An avatar controller using the Wirecast software controls emotion and responses that should be shown to the user via the Skype interface [6, 12]. Although these systems have shown improvement in transferring investigative interviewing skills, these system are too rigid in the response generation, lack generality and have a human input during the interactions which make them expensive and harder to operate. In this paper, we propose an AI-driven system in which the talking avatar can dynamically handle the questions and can provide a higher realism for the interviewers. In addition, by removing human input in the system, it would be cost effective by saving the cost of training or hiring operator/actors. Lack of the need for human operators will lead to 24/7 availability for frequent training sessions. Such a system is realizable to conceive by employing different system components from the field of AI such as dialogue models, generation of digital visual avatars, and presenting that in a VR environment. In the rest of this section, we discuss the state of the art in each of these system components.

2.1 Conversations

Dialogue models have attracted significant attention in recent years for replacing systems that communicate with humans. They have been around for years now and have mainly been employed in the customer service, commercial, and health sectors. We are exploring the use of dialogue models in educating and enhancing learning. Most research has employed them as assistants that answer students' questions and help them accordingly [9, 20, 36].

The introduction of transformers [39], a deep learning model which employs an encoder-decoder framework with a self-attention mechanism, accelerated the development of language generation models like the GPT series [2, 27, 28] and the Bidirectional Encoder Representations from Transformers (BERT) model [7]. Before the introduction of transformers, natural language processing (NLP) models mostly relied on recurrent neural networks like gated recurrent units (GRUs) and Long short-term memory (LSTMs) with

an additional attention mechanism. Dialogue models developed using these models and large open-domain conversational corpora have improved the quality of responses generated. However, non-coherent, off-topic, and uninformative response generation, especially for domain-specific cases, remains a challenge for these models.

2.2 Vitual Reality

Virtual Reality is an immersive media technology that users experience the virtual environment created entirely in a computer simulation. Qualinet Whitepaper [24] classifies the immersive applications based on the level of interaction and number of senses involved in the application. Although more senses do not necessarily lead to higher immersion [8], information can be conveyed faster using more senses, which accelerates the creation of immersion. The levels of interaction while consuming media can be anywhere between being passive and interactive. A proper interactive system can make media consumption more immersive [8]. However, it does not mean that interaction always increases the level of immersion [8], i.e., giving freedom to the users sometimes can make the story even chaotic [29], and as a result, less immersive. VR has been used in many applications such as gaming, omnidirectional video [42], remote control and industrial applications, and health. In this paper, we will show how it can be merged with a dialogue model to create an interactive storytelling system that is beneficial for training CPS professionals.

3 THE AI-BASED CHILD AVATAR

This section discusses details about the developed child avatar, with different components integrated. Figure 1 shows the high-level overview of the architecture, where language module is a dialogue model developed on Rasa[1] is placed in the back-end, the font-end visual component is running in an Oculus Quest device developed in the Unity game engine, and the auditory component is based on the IBM Watson cloud service.

3.1 Visual

The visual part of the child avatar system was developed using the Unity game engine, with results shown to users using Oculus Quest 2. The child avatar was developed using the open-source project Unity Multipurpose Avatar (UMA)[2] in which character customization to combine meshes and textures are possible. Afterward, the voices were matched to the avatar using the Unity game engine asset Salsa Suit[3] to generate eye, head, and lip movements synced to the voices. In addition to facial movements, animations were added to the avatar's hand and neck to create gestures similar to a child talking by employing some simple animations which were repeating.

3.2 Language

A dialogue model was developed and deployed at the back-end using Rasa, an open-source framework to develop solutions to automate text or voice-based conversations. The decision to use

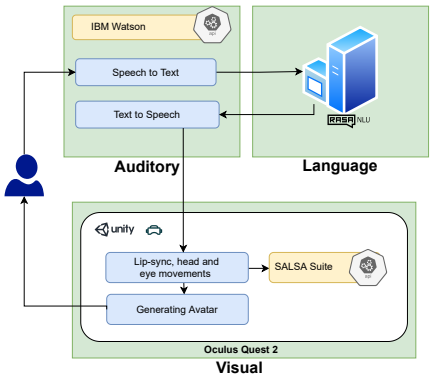


Figure 1: Architecture of the proposed child avatar.

Rasa to develop our dialogue model was motivated by the lack of available data and the desire to validate it as proof of concept for our proposed solution. We had two hundred transcripts of well-conducted training interviews from the Centre for Investigative Interviewing at Griffith University, Australia [26]. This dataset contained the conversations between a 5-7 years old child and interview trainees in a mock interview setting (CPS and police). A professional actor mimicked the child in this case.

3.3 Auditory

We tested different speech synthesis services with the main requirement to sound as child-like as possible. Based on this, we chose to use the IBM Watson services for text-to-speech[4] (TTS) and speech-to-text[5] (STT) synthesis. Watson TTS and STT are cloud service APIs that serve as communicative bridges between the language (back-end) and Visual (front-end) components. The user communicates with the front-end verbally, with the question uttered by the user sent to the IBM STT API to be transcribed with the response, then forwarded to the back-end. At the back end, the dialogue model processes this user utterance and generates an appropriate response. This response is then sent to the IBM TTS API, which then sends the generated audio response to the user at the front end.

3.4 Integration

Ngrok[6] services were used to make the back-end accessible over the internet, with tunneling used to make local servers available over the Internet. The VR front-end accessed the Rasa REST API using an HTTP Post method. The HTTP Rest APIs for the IBM Watson STT and TTS services were accessed using generated API keys and endpoint URLs for each service.

4 SYSTEM PERFORMANCE

In this section, we discuss the computational performance of our implementation. Computers with similar specifications were used at both the front-end and back-end. The system used was an HP

[1]https://rasa.com/
[2]https://github.com/umasteeringgroup/UMA
[3]https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442
[4]https://www.ibm.com/cloud/watson-text-to-speech
[5]https://www.ibm.com/no-en/cloud/watson-speech-to-text
[6]https://ngrok.com/

Pavilion Gaming Desktop TG01-0 model with an AMD Ryzen 7 3700X processor, 16 GB physical memory, an 8 GB NVIDIA GeForce RTX 2060 Super graphics processing unit (GPU), and a 512 GB SSD with Windows 10 operating system running at each end.

Front-end: The Unity3D and Oculus windows applications[7] running with Oculus Link were the primary applications run at the front-end. Utilization of resources was observed over each batch of 10 seconds. The average total CPU usage was 9.7%, and the physical memory load was 48.5% (7912 MB). The observed average GPU memory usage was 74%, GPU core load was 39%, and GPU video engine load was 29.5%. Both of these applications are mostly GPU dependent, and our off-the-shelf system is able to handle the load sufficiently. Both of these applications are mostly GPU dependent, and our system is able to handle the load of these off-the-shelf sufficiently.

Back-end: The Rasa REST API, Rasa action server and ngrok tunneling services were the primary applications running in the back-end. The utilization of resources was observed over each batch of 2-minute conversations with a dialogue model. The average total CPU usage was 1%, and the physical memory load was 35% (5723 MB). The average GPU memory usage was 4.8%, GPU core load was 4%, and GPU video engine load was 0%. The Rasa back-end was not very resource-intensive w.r.t to both CPU and GPU. The only use of the GPU occurred when the intent classification models were called using tensorflow backend.

Text to Speech: The average, max and min times from sending the text to IBM Watson TTS API to receiving the audio response were 1179 ms, 3151 ms and 429 ms, respectively.

Speech to Text: These times were calculated from sending the 10-second audio to IBM Watson STT API till receiving of the text. The SST API processes the audio snip-it as a stream while transcribing and seeks to optimize the textual response over the duration of the audio. Average, max and min times were 12 sec 544 ms, 13 sec 439 ms, and 12 sec 9 ms, respectively. These time measurements includes the length of the 10-second audio as well.

Overall, the system performance results show that an average PC can run the components for the current iteration. Although the prototype is designed with the game engine, it is not as sensitive to time delay as games where a delay of around 200 ms can destroy the user experience [30]. In the case of interviewing a child, delays up to even a few seconds can be considered normal due to the child's hesitation. In the next section, we discuss the system's responsiveness to evaluate this aspect in more detail.

5 INTERACTIVE USER STUDY

This study was an interactive test paradigm based on ITU-T Rec. P.809 which suggests that participants should interact with the system in a designed scenario using a specific assessment method. In this study, users were asked to interact with the system by interviewing the child avatar. This study was designed to evaluate the user's QoE as well as to investigate how effectively the system enhanced the learning experience and the acquisition of knowledge and skills about communicating with abused children. In summary, the main research questions of this study are the following:

- **RQ1:** How is the experts' QoE from using the system?
- **RQ2:** Can such a system enhance the knowledge and skills in interviewers maltreated children?
- **RQ3:** What are the most important quality aspects in the success of such a system?

5.1 Experiment Design

The invited participants were all experienced in the field of child protection because users with no-prior experience do not have the competency either to ask the right questions or to judge the quality of the interview. In this study, users (CPS workers and child welfare students) interacted with the child avatar prototype by wearing a head-mounted display (Oculus Quest 2) and engaging with the avatar in a VR environment in which the child avatar performed spoken conversations using a STT and TTS system as discussed in Section 3. A screenshot of the subject wearing the VR headset and interacting with the child avatar[8] is shown in Figure 2.



Figure 2: A screenshot of the VR user-view of the child avatar.

5.1.1 Demographic of the participants. In this study, 11 experienced subjects, including seven females, three males, and one transgender participated. They were first asked to rate their experience with CPS and VR using a 10-point scale ranging from not experienced to expert. The mean (M) participant's experience in CPS was 6.18 (standard deviation (SD)=2.67). However, these experts in the field of CPS were inexperienced users of VR (M=1.18, SD=.45), i.e., most of the participants were experiencing VR for the first time. For that reason and for the sake of training, the participants interviewed the child avatar twice in the VR environment. Fortunately, none of the participants reported experiencing any motion sickness after using the prototype. Two users expressed confusion in the first interview, but claimed that they could adapt to the situation.

5.1.2 Questionnaire. After each interview, the experts were asked to fill out a post-experience questionnaire as well as to answer a few open-ended questions targeting their opinion about the use of the virtual avatar for training. The questionnaire was designed to measure overall user experience, the interactivity of the system (Responsiveness from GIPS [34] as discussed in [31]), flow items in iGEQ [10] as well as four questions proposed by social science experts regarding learning effects. The first items in the questionnaire used a 7-point continuous scale, known as Extended Continuous-ACR scale as was recommended in ITU-T Rec. P.809, and the rest of the questions used 5-point Likert scale. However, the EC-ACR is also based on a Likert scale, which uses five labels ranging from "strongly disagree" to "strongly agree". In order to transfer it back

[7] https://www.oculus.com/setup/

[8] https://www.dropbox.com/s/k4yokutx27btt2b/AvatarDemo.mp4?dl=0

to a 5-point ACR, the scale transformation [15] was used. Table 1 provides an overview of the items in the questionnaire.

5.1.3 Story. To setup the story-line for this experiment, transcripts of interviews were clustered based on different personas. Availability of multiple numbers of different transcripts and enough details to generate conversations of reasonable length was used to select the persona for this study. The child avatar in this story is named Theresa, and she is five years old. She visits an adult friend, named Jerry, every Tuesday, and there is suspicion of sexual harassment or abuse on her last visit. Her mother has brought her to the child protection agency to be interviewed about the suspected abuse. Participants of the interactive study interviewed the child avatar verbally with the goal to elicit her account of the incident.

Table 1: List of questionnaire items [34], [10]

Overall QoE
QoE: How would you rate your overall quality of experience?
Responsiveness
RE1: I noticed delay between my actions and the outcomes.
RE2: The responsiveness of my inputs was as I expected.
RE3: My inputs were applied smoothly.
Flow
FL1: I forgot everything around me.
FL2: I felt completely absorbed.
Learning Effects
Communication: It improves knowledge and skills in communication.
Self-efficacy: It can enhance my self-efficacy
Learning Experience: It would enhance my learning experience.
Pedagogical Benefits: It has pedagogical benefits.

5.2 Evaluation and Results

This section discusses the results of the conducted experiment. The results are discussed based on various aspects, namely the user experience, learning effect, expert comments, and importance of the quality aspects.

5.2.1 RQ1: User Experience Aspects. Figure 3 shows the user ratings of QoE, Flow, and Responsiveness in their first and second interviews. Overall, we can see a trend towards improvement in the user experience after training with the system. For the overall QoE, the first interview had a $M = 3.62$, $SD = 1.25$, and the second interview created a Mean Opinion Score (MOS) of $M = 3.92$, $SD = .71$, which implies a good QoE on the ACR scale. However, a paired-sample t-test showed that there was no significant difference between the MOSs of the QoE in the first and second interviews $t(10) = -1.47$, $p = .17$. The results were similar for Flow with the experts feeling more immersed in the second interview ($M = 3.45$, $SD = .57$) than in the first ($M = 3.0$, $SD = .89$), but the difference was not significant ($t(10) = -1.66$, $p = .12$). However, the user responsiveness significantly improved from the first interview ($M = 2.78$, $SD = .42$) to the second interview ($M = 3.18$, $SD = .45$, $t(10) = -3.13, p = .01*$) because the participants had to press a button each time they finished a question. This contrasts with natural speaking, and participants needed time to adapt to the system. In addition, because of the button pressing, the participants felt there was a delay between their actions and the outcome of the system, underlining the need to improve the responsiveness of the system. A button-less experience should be implemented as a future improvement.

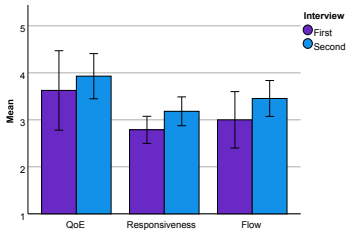


Figure 3: Bar-plot (95% confidence interval) of means showing quality assessments in the first and second rounds of testing.

5.2.2 RQ2: Learning Effects. Figure 4 shows the bar-plot of user ratings for all items regarding learning effects in the questionnaire. One can observe that, in all the aspects, experts agreed that the virtual avatar could help them to improve themselves. 72% of the participants stated that such an AI-based child avatar could help them to acquire knowledge and skills in communication, 81% opined that the avatar could enhance their self-efficacy, and all the participants agreed that the avatar could improve their learning experience.

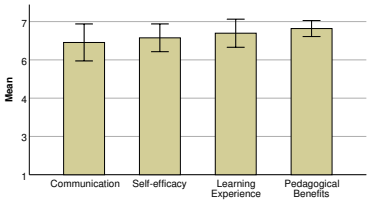


Figure 4: Barplot (95% confidence interval) of means for learning effects.

5.2.3 RQ3: Importance of the Quality Aspects. The experts were asked about their opinions on the importance of each quality aspect. This could help us to further improve the system with respect to the requirements and expectations of actual users of the child avatar system. Experts were asked to sort the importance of the aspects *interactivity, impassiveness, realism, presence, escapism, and investigative* from 1 to 6 (least important to most important). Figure 5 shows the importance of each aspect based on the experts' judgment. Interactivity, realism, and presence were rated as more important, identifying our next steps for developing the child avatar system.

5.2.4 Expert Comments and Suggestions. In addition to the questionnaire, experts were asked to write a paragraph about the child avatar system. The written texts were gathered and coded to extract quantifiable results. Six of 11 participants mentioned that the virtual avatar was helpful/useful and could assist them, six experts stated that the virtual avatar is very important/valuable, 3 experts stated that the child avatar provided realistic training, and all the participants wrote that training/practicing with the system would enhance their knowledge/insights. Interestingly, one of the experts said, "I already learned in this first conversation how child conversations can be improved." However, this study aimed only to evaluate the first version of the system and gather comments from the experts rather than to teach them. In addition to the written texts, users' feedback during the experiment was collected when they were thinking out loud. Some of the participants expressed confusion due to the need to press the button each time after finishing their questions. A couple of participants expressed the urge to take notes, which was impossible due to the VR head-mounted display constraints.

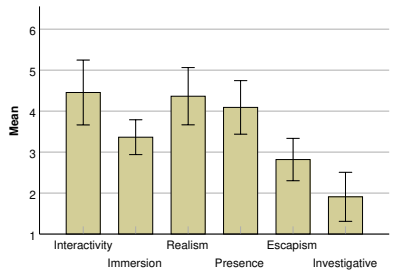


Figure 5: Bar-plot (95% confidence interval) of mean of the importance of aspects.

6 DISCUSSION AND FUTURE WORK

This paper proposed an AI-driven talking avatar in VR that could understand and respond to human language to emulate an investigative interview with maltreated children. As a prof of concept, we presented the system architecture of the prototype, and using a subjective study with expert participants. In addition, the aim was to have a qualitative study with experts where they are able to express their opinion with open questions rather than a quantitative study. The study showed that the system can improve the knowledge and skills of the CPS and law enforcement personnel. In addition, it was shown that having the system in the VR can create a higher level of QoE and helps to increase the realism and presence that were rated important by experts. We targeted a specific group of participants who either have active or passive experience in interacting with abused children and are familiar with current teaching/training programs. The low number of participant is attributed to the fact that it is hard to manage such professionals to participate in our research as they had their professional commitments as well. However, high agreement between the experts indicates that even higher number of such professionals would have produced similar results.

The current implementation of our system is meant to serve as a baseline for our future work in this research area. We are currently working to develop a more generalizable and flexible dialogue model using GPT-3, coupled with deterministic models built on annotated conversational data for feedback mechanism and prompt generation. Based on the experts' feedback, the system's responsiveness has to be improved, and a button-less experience is planned. Moreover, note-taking during the interviews was pointed out as a critical activity an interviewer often perform during the interview, which we plan to address in future work. Emotional recognition of child responses can be exploited to alter the features of child audio like tone or pitch to give a more realistic experience and visual expression of these sentiments. Recent research has shown that there is a high correlation between non-verbal and verbal expressions of emotions [14]. Contrast to that, experienced experts are of the opinion that maltreated children show and deal with emotions differently, and some traumatized children do not show any emotions at all. Therefore, displaying emotion during the conversation should be within the context of children with different personas. To create realistic personas, there is a need for a video dataset of such interviews, which is not available to us at the time, as the data is very sensitive so it's hard to get hold of it. We are collaborating with our partners in CPS and police from two different countries to get access to this data, and to be able to add emotions to audio/video into the system in future works. Furthermore, we will also investigate the use of realistic talking avatars in more detail and whether the uncanny valley effect [19, 21] exists in the context of such child interviews, meaning that users feel a sense of anxiety or discomfort in response to a highly realistic avatars resulting in low QoE.

7 CONCLUSION

In this paper, an implementation of our proposed VR-based interview training system was described. It is developed using the Unity game engine and a dialogue model developed using RASA. Audio communication with system was synthesised using IBM text-to-speech and speech-to-text services. The game system was discussed and evaluated objectively in terms of system performance and subjectively using an interactive user study. The participants of this study were CPS experts, and the results show that the child avatar system is well received by the experts and can help them obtain better communication skills for interviewing abused children. In addition, our findings identified *interactivity*, *realism* and *presence* as the important quality aspects to target a successful training program. Using the results of this study, a second prototype will be designed to create more presence and realism for users, which validates our initial hypothesis of using VR over other interfaces [32]. Finally, we conclude, from the observed system performance of our first avatar version, that a standard PC can handle the system's processing requirements.

ACKNOWLEDGMENTS

This research is sponsored by the Research Council of Norway, project number #314690.

REFERENCES

[1] Gunn-Astrid Baugerud, Miriam S Johnson, Helle BG Hansen, Svein Magnussen, and Michael E Lamb. 2020. Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015–2017). *Applied Cognitive Psychology* 34, 3 (2020), 654–663.

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[3] Nathalie Carrick, Jodi A Quas, and Thomas Lyon. 2010. Maltreated and nonmaltreated children's evaluations of emotional fantasy. *Child Abuse & Neglect* 34, 2 (2010), 129.

[4] Ann-Christin Cederborg, Yael Orbach, Kathleen J Sternberg, and Michael E Lamb. 2000. Investigative interviews of child witnesses in Sweden. *Child abuse & neglect* 24, 10 (2000), 1355–1361.

[5] Dante Cicchetti and Sheree L Toth. 2005. Child maltreatment. *Annu. Rev. Clin. Psychol.* 1 (2005), 409–438.

[6] Kevin Charles Dalli. 2021. Technological Acceptance of an Avatar Based Interview Training Application: The development and technological acceptance study of the AvBIT application.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Pierre Gander. 1999. *Two myths about immersion in new storytelling media.* Lund University.

[9] Sebastian Hobert. 2019. Say hello to 'coding tutor'! design and evaluation of a chatbot-based learning system supporting students to learn to program. (2019).

[10] Wijnand A IJsselsteijn, Yvonne AW de Kort, and Karolien Poels. 2013. The game experience questionnaire. *Eindhoven: Technische Universiteit Eindhoven* 46, 1 (2013).

[11] ITU-T Recommendation P.809. 2018. *Subjective Evaluation Methods for Gaming Quality.* International Telecommunication Union, Geneva.

[12] David Johansson. 2015. Design and evaluation of an avatar-mediated system for child interview training.

[13] Miriam Johnson, Svein Magnussen, Christian Thoresen, Kyrre Lønnum, Lisa Victoria Burrell, and Annika Melinder. 2015. Best practice recommendations still fail to result in action: A national 10-year follow-up study of investigative interviews in CSA cases. *Applied Cognitive Psychology* 29, 5 (2015), 661–668.

[14] Yael Karni-Visel, Irit Hershkowitz, Michael E Lamb, and Uri Blasbalg. 2021. Nonverbal emotions while disclosing child abuse: the role of interviewer support. *Child maltreatment* (2021), 10775595211063497.

[15] Friedemann Köster, Dennis Guse, Marcel Wältermann, and Sebastian Möller. 2015. Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech. *Fortschritte der Akustik, DAGA* (2015).

[16] Michael E Lamb. 2016. Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink? *American psychologist* 71, 8 (2016), 710.

[17] Michael E Lamb, Deirdre A Brown, Irit Hershkowitz, Yael Orbach, and Phillip W Esplin. 2018. *Tell me what happened: Questioning children about abuse.* John Wiley & Sons.

[18] Michael E Lamb, Yael Orbach, Irit Hershkowitz, Phillip W Esplin, and Dvora Horowitz. 2007. A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child abuse & neglect* 31, 11-12 (2007), 1201–1231.

[19] Karl F MacDorman, Robert D Green, Chin-Chang Ho, and Clinton T Koch. 2009. Too real for comfort? Uncanny responses to computer generated faces. *Computers in human behavior* 25, 3 (2009), 695–710.

[20] Fernando A Mikic-Fonte, Martín Llamas-Nistal, and Manuel Caeiro-Rodríguez. 2018. Using a Chatterbot as a FAQ Assistant in a Course about Computers Architecture. In *2018 IEEE Frontiers in Education Conference (FIE).* IEEE, 1–4.

[21] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

[22] Chris Newlin, Linda Cordisco Steele, Andra Chamberlin, Jennifer Anderson, Julie Kenniston, Amy Russell, Heather Stewart, and Viola Vaughan-Eden. 2015. *Child forensic interviewing: Best practices.* US Department of Justice, Office of Justice Programs, Office of Juvenile ....

[23] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[24] Andrew Perkis, Christian Timmerer, Sabina Baraković, Jasmina Baraković Husić, Søren Bech, Sebastian Bosse, Jean Botev, Kjell Brunnström, Luis Cruz, Katrien De Moor, et al. 2020. QUALINET white paper on definitions of immersive media experience (IMEx). *arXiv preprint arXiv:2007.07032* (2020).

[25] Francesco Pompedda, Angelo Zappalà, and Pekka Santtila. 2015. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law* 21, 1 (2015), 28–52.

[26] Martine B Powell, Belinda Guadagno, and Mairi Benson. 2016. Improving child investigative interviewer performance through computer-based learning activities. *Policing and Society* 26, 4 (2016), 365–374.

[27] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[29] Marie-Laure Ryan. 1999. Immersion vs. interactivity: Virtual reality and literary theory. *SubStance* 28, 2 (1999), 110–137.

[30] Saeed Shafiee Sabet, Steven Schmidt, Saman Zadtootaghaj, Carsten Griwodz, and Sebastian Möller. 2020. Delay sensitivity classification of cloud gaming content. In *Proceedings of the 12th ACM International Workshop on Immersive Mixed and Virtual Environment Systems.* 25–30.

[31] Saeed Shafiee Sabet, Steven Schmidt, Saman Zadtootaghaj, Babak Naderi, Carsten Griwodz, and Sebastian Möller. 2020. A latency compensation technique based on game characteristics to mitigate the influence of delay on cloud gaming quality of experience. In *Proceedings of the 11th ACM Multimedia Systems Conference.* 15–25.

[32] Pegah Salehi, Syed Zohaib Hassan, Saeed Shafiee Sabet, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Michael A. Riegler, and Pål Halvorsen. 2022. Is More Realistic Better? A Comparison of Game Engine and GAN-based Avatars for Investigative Interviews of Children. *Workshop on Intelligent Cross-Data Analysis and Retrieval* (2022).

[33] Karen J Saywitz, Thomas D Lyon, and Gail S Goodman. 2017. 19 When Interviewing Children: A Review and Update. *The APSAC handbook on child maltreatment* (2017), 310.

[34] Steven Schmidt. 2021. *Assessing the Quality of Experience of Cloud Gaming Services.* P.h.D. thesis, Technische Universität Berlin.

[35] Jack P Shonkoff. 2016. Capitalizing on advances in science to reduce the health consequences of early childhood adversity. *JAMA pediatrics* 170, 10 (2016), 1003–1007.

[36] Sharob Sinha, Shyanka Basak, Yajushi Dey, and Anupam Mondal. 2020. An educational Chatbot for answering queries. In *Emerging Technology in Modelling and Graphics.* Springer, 55–60.

[37] Kathleen J Sternberg, Michael E Lamb, Yael Orbach, Phillip W Esplin, and Susanne Mitchell. 2001. Use of a structured investigative protocol enhances young children's responses to free-recall prompts in the course of forensic interviews. *Journal of applied psychology* 86, 5 (2001), 997.

[38] Sheree L Toth and Dante Cicchetti. 2013. A developmental psychopathology perspective on child maltreatment. *Child maltreatment* 18, 3 (2013), 135–139.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems.* 5998–6008.

[40] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).

[41] WHO. 2020. *Child maltreatment.* https://www.who.int/news-room/fact-sheets/detail/child-maltreatment

[42] Mai Xu, Chen Li, Zhenzhong Chen, Zulin Wang, and Zhenyu Guan. 2018. Assessing visual quality of omnidirectional videos. *IEEE transactions on circuits and systems for video technology* 29, 12 (2018), 3516–3530.

[43] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

Article III: Is More Realistic Better? A Comparison of Game Engine and GAN-based Avatars for Investigative Interviews of Children

Authors: *Pegah Salehi*, Syed Zohaib Hassan, Saeed Shafiee Sabet, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pål Halvorsen, Michael A Riegler

Candidate's contributions: The candidate led the design and implementation of the user study comparing avatar realism using game engine-generated and GAN-based avatars. She was responsible for constructing the experimental setup, including the selection, development, and customization of both Unity-engine avatars and GAN-based talking-head avatars. The candidate also directed the analysis of participant feedback, focusing on perceived realism, user experience, and lip-sync accuracy. As the primary author, the candidate played a central role in writing the manuscript, with particular emphasis on the methodology, results interpretation, and discussion of how visual realism influences engagement and effectiveness in investigative interview training with children.

Published in: Proceedings of the 3rd ACM Workshop on Intelligent CrossData Analysis and Retrieval (ICDAR) - Newark, NJ, USA - June 27-30, 2022

DOI: 10.1145/3512731.3534209

Contributed to: This work contributed to the **RO-2**.

Is More Realistic Better? A Comparison of Game Engine and GAN-based Avatars for Investigative Interviews of Children

Pegah Salehi
SimulaMet
Norway

Syed Zohaib Hassan*
SimulaMet
Norway

Saeed Shafiee Sabet
SimulaMet
Norway

Gunn Astrid Baugerud
OsloMet
Norway

Miriam Sinkerud Johnson
OsloMet
Norway

Pål Halvorsen*
SimulaMet
Norway

Michael A. Riegler†
SimulaMet
Norway

ABSTRACT

The success of investigative interviews with maltreated children is often defined by the interviewer's ability to elicit a reliable and coherent account of the alleged incident from the child. Research shows that a child avatar mimicking a maltreated child can improve interviewers' performance in conducting these interviews. The realism of such a child avatar is considered one of the most critical factors. Based on this, the current study aims to generate realistic child avatars in real-time that utilize multimodal data and different components from artificial intelligence. This paper discusses the subjective findings of a study of two types of child avatar videos; animated avatars created using the Unity game engine and photorealism talking-head avatars using Generative adversarial networks (GANs). The results show that although the state-of-the-art GAN-generated avatars are significantly more realistic, they do not necessarily create a better experience, as most of the participants prefer talking to animated avatars.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; • **Computing methodologies** → **Machine learning**; **Image and video acquisition**; **Adversarial learning**; *Natural language processing*; • **Applied computing** → **Interactive learning environments**.

KEYWORDS

Child Protection Services (CPS), Uncanny Valley, Digital Human, Generative Adversarial Networks (GANs).

*Also affiliated with OsloMet, Norway
†Also affiliated with UIT The Arctic University of Norway, Norway

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICDAR '22, June 27–30, 2022, Newark, NJ, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9241-9/22/06…$15.00
https://doi.org/10.1145/3512731.3534209

ACM Reference Format:
Pegah Salehi, Syed Zohaib Hassan, Saeed Shafiee Sabet, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pål Halvorsen, and Michael A. Riegler. 2022. Is More Realistic Better? A Comparison of Game Engine and GAN-based Avatars for Investigative Interviews of Children. In *Proceedings of the 3rd Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512731.3534209

1 INTRODUCTION

Sexual and physical abuse of children is a significant problem and affects children around the world. According to an estimate by the World Health Organization (WHO), 300 million children's lives are affected globally by abuse and maltreatment [24]. The problems stemming from sexual and physical abuse can have long-lasting, damaging effects on children's behavior, their psychological development and adjustment, as well as increase their risk of early death [4]. Child protection services (CPS) and police play a key role in ensuring that these abused and mistreated children are better protected and treated. As a part of the child abuse investigation, these agencies must interview the children about their experiences of abuse and maltreatment. Most often in these cases, the children are the only witnesses of criminal incidents, and it becomes critical that professionals conduct investigative interviews in accordance with best-practice guidelines. A high-quality interview is important in court proceedings and is associated with a higher chance of convicting alleged perpetrators of these crimes. However, field studies of real-life investigative interviews show that the quality of these interviews is consistently poor and there is a huge gap between best practice recommendations and successful implementation of skills [3, 10].

Hence, several training methodologies have been designed to provide the necessary knowledge to conduct these interviews in accordance with the best practice recommendations. Classroom training is one of the most common approaches to transfer knowledge to police and CPS personnel. Training consisting of computer-based activities, such as reading materials, watching videos and taking quizzes is also employed [18]. After these training programs, participants go through mock-interview exercises where an instructor/professional actor role-plays an abused child and they have to

interview them. Recent research has demonstrated that interactive computer-based learning activities can improve investigative interviewers' performance in questioning children [18]. We are proposing an investigative training platform developed using a digital child avatar, the first prototype of the system is presented in [9] where we utilized different advanced technologies such as dialogue models, virtual reality, text-to-speech (TTS) and speech-to-text (STT) components to create an interactive child avatar. The findings indicated that such a system might have the potential to increase the knowledge and skills of users. Furthermore, experts identified the realism of the child avatar as one of the most important factors in their experience. Therefore, this paper tries to increase the realism of avatars by introducing more GAN-generated avatars and investigates whether more realistic avatars can necessarily create a better user experience.

Interaction with the virtual avatar requires techniques that are able to generate high-fidelity talking-head frames where the human observers do not lose trust and empathy with the avatar. According to the Uncanny Valley [15, 17] theory, if an avatar is humanoid but imperfect, its non-human features create strange feelings in the audience that cause feelings of uneasiness and even disgust. Thus, we investigate whether the uncanny valley exists in the context of the child interview. We took two different approaches to create several child avatars; i) Animated avatars created with the Unity game engine and ii) human-like avatars generated using three Generative Adversarial Networks (GANs) techniques, namely PC-AVS [29], MakeItTalk [30] and LSP [14]. In addition to the visual of the avatars, different human and computer-generated voices were combined with each character to generate multiple versions of each avatar. Following that, a subjective study was designed and conducted to evaluate various aspects of each avatar. The user study evaluates the realism of avatars with a passive viewing-and-listening test, investigating based on their appearance and talking. Findings from this study can assist us in choosing the right combination of voices and the method of creating the visual avatar. In summary, our main research questions are:

- **RQ1:** Which GAN-based method creates a more realistic avatar?
- **RQ2:** Which animated avatar is the more realistic one?
- **RQ3:** Can computer-generated voices be as realistic as human voices?
- **RQ4:** Which avatars are the best, avatars created with a GANs or a game engine?

The rest of the paper is organized as follows: in Section 2, related work on the talking-head generation is reviewed, and in Section 3, we explain the different parts of user study execution. Section 4 shows our subjective results. Section 5 discusses the results, limitations and suggestions for future research. Finally, Section 6 concludes the paper and highlights the main findings.

2 RELATED WORKS

Recent advances in deep neural networks have helped make a lot of progress in a high-quality visual content generation, one of the most important of these is the generation of digital humans for a wide range of applications such as news broadcasting, virtual video conferencing, online shopping, visual dubbing in movies, and filmmaking. According to the survey paper by Sha et al. [20], person generation can be divided into three main tasks talking-head generation (face), pose-guided person generation (pose), and garment-oriented person generation (cloth). Since only the neck and the shoulder parts of a person are required to conduct an interview, our focus in this project is on talking-head generation. This approach can synthesize natural lip synchronization, head movements, and micro-expression from a specific person by capturing one or more input signals. Depending on the input data type, talking-head generation can be based on both audio and video. Many audio-driven works avoid directly mapping from audio to image, but instead first, map audio to an intermediate step like 2D facial landmarks [6, 14, 16, 22, 30] or 3D face shapes [5, 19, 21, 23, 27], and then synthesize photorealistic renderings. Audio-driven talking-head generation is inherently difficult because it synthesizes mouth movements based only on an audio stream. Therefore, some of the proposed methods control the motions of a subject with a driving video that are categorized as video-driven methods (also known as face reenactment) [8, 25, 28, 29]. Some models also focus on training their model in a specific person [2, 13]. Audio-driven methods have received more attention in computer graphics because they aim to generate a single universal talking-head model for a variety of subjects. To investigate GAN-based methods in this study, we adopted three recent state-of-the-art methods among the prior research, including: PC-AVS [29], MakeItTalk [30] and LSP [14]. MakeItTalk proposes a generated model using self-attention based LSTM that is able to the disentanglement of content and style in audio and leads to creating a speaker-aware talking-head, on the other due to relying on precise facial landmarks, it causes create eye blinks but it suffers from inaccurate shapes of the mouth. PC-AVS, without relying on any intermediate information such as landmarks and 3D face shapes, implicitly devises a pose code that is free of mouth shape or identity and then modularizes audio-visual representations into spaces of audio content, head movement, and identity. PC-AVS generates frontalizing talking faces and achieves more accurate lip-sync results than Makeittalk. However, PC-AVS can not be as successful as MakeItTalk in creating blink eye and preserving identity. Another disadvantage of the PC-AVS generated videos have a disordered texture and resolution changes a lot. Beyond these methods, LSP [14] introduces a live system that steps forward to practical applications, it generates personalized talking-head with natural head poses, lip-sync and eye blinks only driven by audio signals at over 30 fps.

Despite recent advances, synthesizing an interactive talking avatar with little or no manual labor remains an open challenge. In the present methods, emotions in their words are not in line with the face's movement, and face attributes have weak correlations with the phonetic signal, e.g., the head motion and the eye blinking are not natural and realistic. In addition, the performance of these methods in terms of realism and user experience is unknown. These methods generate human-like avatars that look lifelike, but they are imperfect in comparison to real humans. Based on the uncanny valley theory [15, 17], an imperfect human-like avatar might create a bad experience for users. This paper investigates these methods in more detail, discussing the hypothesis regarding the uncanny valley and realism of the avatars.

3 EXPERIMENT DETAILS

A crowdsourcing user study was conducted to answer the research questions. In particular, this study investigates the user experience and realism of 28 videos generated in the combination of synthetic and human voices with visual avatars generated by GAN-based techniques and the Unity game engine approaches. First, the test conditions containing these 28 videos are elaborated, and we then discuss the questionnaire used in the study. Finally, the details of the subjective study and its conditions are explained.

3.1 Video Dataset

The data used in this experiment is generated by two different GAN-based and Unity game engine approaches. These characters were generated using methods based on either GANs or animations using the Unity game engine. Some examples are shown in Figure 1. After combining these avatars with computer-generated voices from IBM Watson and human voices, we created 28 talking avatars, including 10 animated avatars and 18 GAN-generated avatars.

In the first approach, the child avatars were developed using the open-source project **Unity Multipurpose Avatar (UMA)**[1] system. We performed character customization by combining meshes and textures, and the audios were matched to the avatar. i.e., to generate eye, head, and lip movements in sync with the voices, using the Unity game engine asset SALSA LipSync Suite[2]. In the second approach, we used three methods: MakeItTalk, PC-AVS and LSP. GAN-based techniques proposed for audio-to-video translation are scientifically known as talking-head generation. In this way, an arbitrary audio stream is given to the network as input aims to synthesize lip-synced videos.

We picked three state-of-the-art open-source models that support driving a static portrait with arbitrary audio among the available methods. In particular, we include the following methods: **PC-AVS** [29], which takes one facial image as an input and generates a talking-head whose poses are controlled by another pose source video; **MakeItTalk** [30], trained on VoxCeleb2 [7] dataset, generates expressive talking-head videos from an audio stream and a portrait face image as the only input, that can generalize well on unseen facial images; and **LSP** [14], that initially requires to be trained on a several-minute length video of the desired character, then by capturing arbitrary audio, generates photo-realistic talking-head video in real-time. For the voice of the avatars, two models from IBM Watson TTS and two natural human voices were recorded, one male and one female. Portrait images were used for input of both PC-AVS and MakeItTalk provided using StyleGAN [11, 12]. StyleGAN generates ultra-realistic photos of humans who do not exist. We generated a number of portrait photos using the StyleGAN website and chose six photos of children, three girls and three boys, as shown in the human face category in Figure 1. Finally, 28 ten-second videos clips of the avatars were obtained to conduct the user study. The video dataset is publicly available at Google Drive[3].

The videos were presented to the participants in a random order created by Latin Square design [1]. The subjective study followed a within-subject design where all the participants watched all the 28 video clips. In addition to the 28 videos, to anchor the user experience, we created an unrealistic avatar with a low level of details as an anchoring condition. We showed that always as the first condition to the participant.

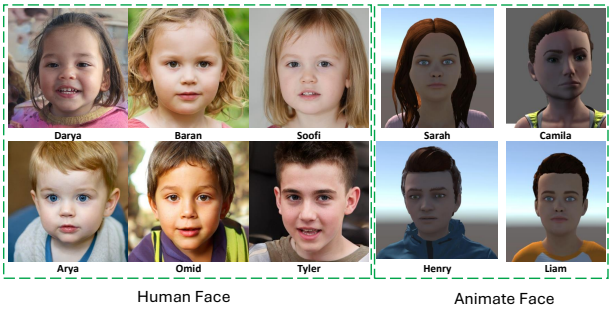


Figure 1: Human and animated faces used to create GAN-based and Unity game engine-based avatars, respectively.

3.2 Demographic information of Participants

The study was conducted with crowdsourcing, and Microworker[4] was used as the recruiting platform environment with users directed to a questionnaire tool hosted on a separate server containing the videos. In order to ensure the validity and reliability of the collected data, only the crowdworkers with high approval rate who were the best performers were invited to the test.

Overall, 39 crowdworkers provided valid results in this study, including 10 females, 27 males, and two with other genders. The crowdworkers aged between 19 and 54 years ($Median = 28$, $Mean = 29.58$, $SD = 8.36$) were geographically distributed equally between Europe, Asia, and America (North and South). 15 of the participants used a desktop monitor, 16 a laptop, seven a mobile or tablet, and one a monitor larger than 30 inches.

3.3 Questionnaire

In the experiment, the participants were asked to watch videos presented above and rate a series of questions according to their experience. The items in the questionnaire were inspired by the work conducted by Wilson et al. [26] which proposes a questionnaire for assessment of realism in stimuli. Table 1 shows the list of these questions on the questionnaire. Each participant was asked to rate from 1-5 on the Likert-scale for each query: realism in talking, appearance of the avatar, overall user experience with the avatar, and the lip synchronization quality. In addition to the post-experience questionnaire which was asked after watching each video, participants were asked to answer a post-test questionnaire after finishing the test and watching all the videos. The list of the questions in the post-test questionnaire is presented in Table 2.

4 EVALUATION AND RESULTS

Eventually, the talking child avatars will be used directly by users for investigative interviews. Therefore, subjective evaluation of humans is an essential step. Using the subjective scores collected in the subjective experiment detailed in the previous section, this section answers the research questions presented towards the end

[1] https://github.com/umasteeringgroup/UMA
[2] https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442
[3] https://drive.google.com/drive/folders/1jRX_lmvrRcxCbNEKoKGW4qlSvaxoOSea
[4] microworkers.com/

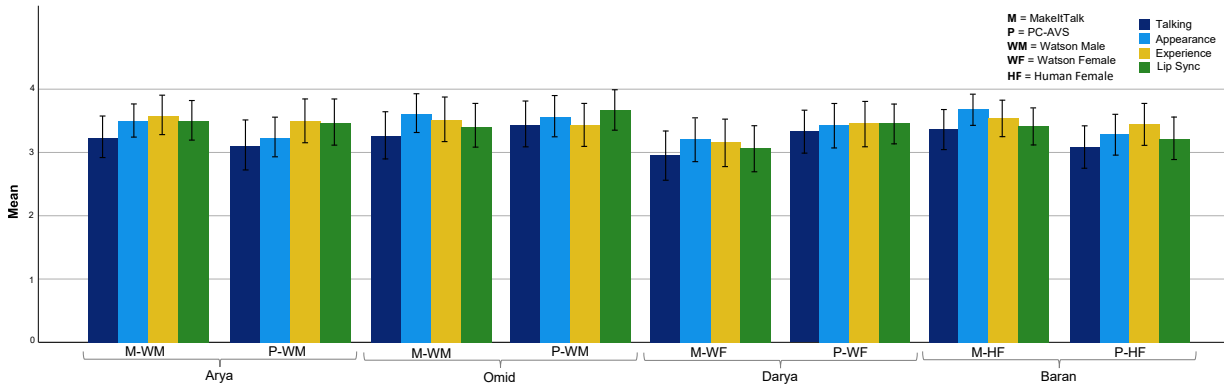


Figure 2: Comparison of MakeItTalk and PC-AVS methods.

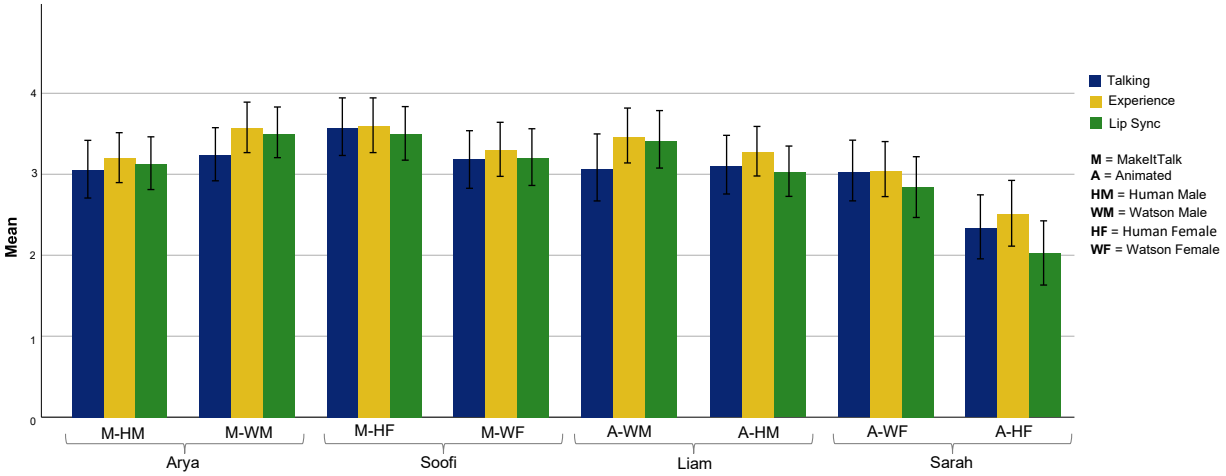


Figure 3: Comparison the voices generated by a computer and human.

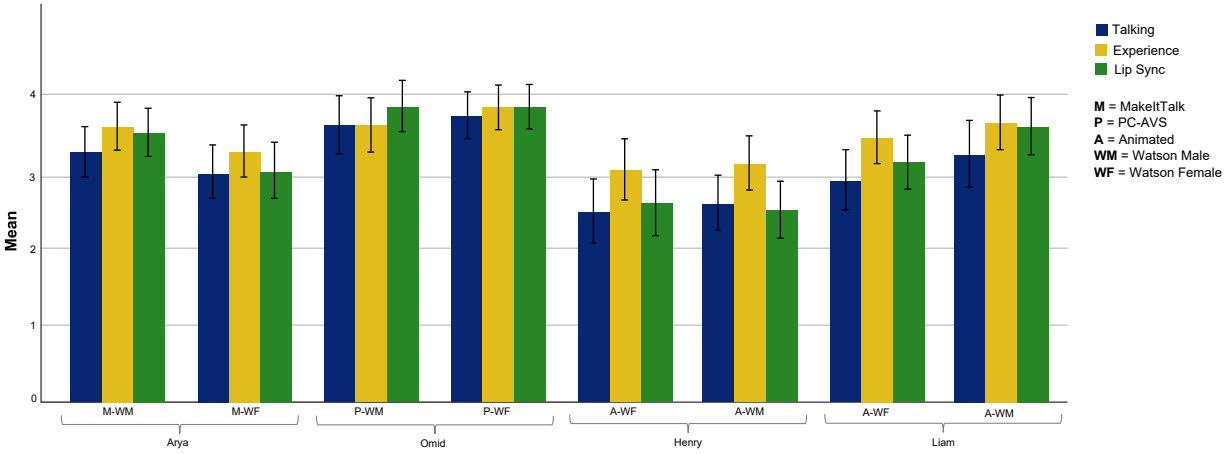


Figure 4: Comparison of using the voice of a girl for a boy.

Table 1: List of items in post-experience questionnaire

Code	Label
Talking	How realistic the avatar was talking?
Appearance	How realistic was the appearance of the avatar?
Experience	How was your overall experience with the avatar?
Lips Sync	How were the audio and mouth/lips synced?

Table 2: Post-test questionnaire to compare the avatars.

Questions	Label
Q1	Which avatars had the most realistic appearance?
Q2	Which avatars do you like the most?
Q3	Considering the avatars' talking, overall which was the most realistic?
Q4	For a conversation (taking to a computer), which avatar would you prefer to use?
Open	Please describe your reasons and feelings towards different types of avatars.

of the Section 1, each of them has been discussed in a distinct subsection.

4.1 RQ1: GAN-generated Avatars

As explained in Section 3.1, we selected three state-of-the-art open-source GAN-based methods for this study. This research question investigates the difference between all three models from different aspects. At first, the aim is to compare MakeItTalk [30] and PC-AVS [29] models to find out which model performs the best and if the performance of the model depends on the selected facial image of a character.

The bar-plots in Figure 2 shows the user ratings, where there is no noticeable difference between the two models MakeItTalk and PC-AVS for all four characters Arya, Omid, Darya and Baran. To examine this comparison, four repeated measure factorial Analysis of variance (ANOVA) tests were conducted to compare the main effect of models and characters and interactions between them on all four quality aspects.

Based on all four quality aspects presented in Table 1, there was no difference between the two models (MakeItTalk and PC-AVS), and the characters (the selected facial image). For the quality aspect for *Talking*, there was no significant main effect of model p-value (p) more than <.05 ($p < .6$), no significant main effect of character ($p < .5$) and no significant interaction between model and character ($p < .06$). For *Appearance*, there was no main effect of model ($p < .09$), no main effect of characters ($p < .21$) and no interaction between model and characters ($p < .09$). For *Overall experience*, there was no main effect of model ($p < .86$), no main effect of characters ($p < .29$) and no interaction between model and characters ($p < .38$). And for lip-sync, there was no main effect of model ($p < .52$), no main effect of characters ($p < .17$) and no interaction between model and characters ($p < .07$). PC-AVS can generate accurate lip-sync but can not preserve the identity and resolution of the input facial image, it also generates talking-head no blink actions. In contrast, MakeItTalk can blink actions and also preserves resolution of the input facial image, but it does not succeed in synchronize mouth shape. Accordingly, we expected MakeItTalk to be more realistic from the user's point of view. We also investigate the importance of the head movement in perceived realism of the avatar. Natural head movement in real world is one of the key factors in high-quality communication. In fact, human can easily express their feelings by talking with fixed head pose or nonverbal communication with head movement. However, generating lip-synced videos with audio to control head motions is challenging. MakeitTalk compared to most existing methods that generate facial animation with a fixed head pose, can generate subtle head motions. These motions involve many repetitive behaviors and movement patterns only slightly swing around the initial state. In contrast, PC-AVS uses some short target video and an audio stream as input to transfer head pose, which creates natural and diverse rhythmic head motions according to an input video clip.

In this inference, two avatars were synthesized using Tyler characters and different videos to control head motion, one of which had a strong head movement (named P-WM1) and the other a slight head movement (named P-WM2). However, contrary to what we expected, there is no significant effect of head movement found between the P-WM1 and P-WM2 avatars based on all aspects. Table 3 shows the results of the t-test investigating the effect of head movement for all the quality aspects.

Next, LSP and MakeItTalk are compared. LSP and MakeItTalk leverage facial landmarks as an intermediate representation, first predict facial landmarks from the speech content and then generate images conditioned on synthesized face landmarks. This technique leads to generation of blinking movements, personalized head movements based on audio stream (identify-preserving), and high visual quality results. MakeItTalk is also able to create non-photorealistic cartoon animations because their labeled 68 facial landmarks correspond to human faces. The user study results with facial landmarks for LSP and MakeItTalk are shown in Figure 7. Unlike MakeItTalk, LSP is not generalized for unseen facial images and requires model training for each specific image. To compare the results, a repeated measure t-test was conducted to compare the mean of quality aspects *Talking*, *Appearance*, *Overall experience* and *Lips sync* in LSP and MakeitTalk model. For realism in *Talking* there was a significant difference found between the mean of landmark created with LSP ($Mean(M) = 2.84, StandardDeviation(SD) = 1.15$) and MakeitTalk ($M = 2.39, SD = 1.00$), $t(37) = 2.38, p = .02$. For appearance, a significant difference was found between mean of LSP ($M = 2.45, SD = 1.2$) and MakeitTalk ($M = 2.05, SD = 1.13$), $t(37) = 2.49, p = .017$. For the overall experience, there was a significant difference found between the mean of landmark created with LSP ($M = 3.00, SD = 1.13$) and MakeitTalk ($M = 2.58, SD = 1.28$), $t(37) = 2.24, p = .031$. Finally, for lip sync, there was a significant difference found between the mean of landmark created with LSP ($M = 3.26, SD = .97$) and MakeitTalk ($M = 2.58, SD = 1.28$), $t(37) = 3.25, p = .002$.

To summarise, RQ1 investigated; i) comparing model PC-AVS and MakeItTalk, ii) comparing the landmark avatars created by MakeItTalk and LSP model, and iii) investigating the impact of head movement in GAN-generated avatars. The results show that i) there is no difference between PC-AVS and MakeItTalk, and they are equally rated in all four dimensions, ii) the landmark avatars created by the LSP model are rated significantly higher than the MakeItTalk model, and finally, iii) showed that contrary to our

Table 3: Results of the t-test investigating the effect of head movement in all four quality aspects.

Quality	P-WM2	P-WM1	t(37)	Sig.
Overall QoE	M=3.61, SD=1.10	M=3.63, SD=0.97	.17	.86
Appearance	M=3.53, SD=1.13	M=3.45, SD=1.15	.59	.55
Talking	M=3.39, SD=1.15	M=3.37, SD=1.07	.16	.86
Lip Sync	M=3.79, SD=1.04	M=3.63, SD=0.97	1.43	.16

expectation, the head movement does not improve the realism or the user experience.

4.2 RQ2: Animated Avatars

Figure 5 shows the quality ratings for the animated avatars. To identify the best character created by the game engine, a one-way repeated measure ANOVA was conducted for the mean of quality aspects overall experience, and appearance. Since all the animated characters were using the same technique for lip-sync generation using Salsa component in the Unity game engine, the lip-sync quality is not investigated in the data analyses of this section.

The result of repeated measure ANOVA indicates a significant difference between mean of realism in *Talking* of different animated characters, $F(2.56, 89.59) = 14.91, p < .001$. A post-hoc test using a Bonferroni correction was conducted to compare the characters and find the best character for each gender and overall. The results are presented in Table 4. For male characters, results show a trend in higher realism in *Talking* of Liam than Henry, and for females that Sarah talks significantly more realistic than Camilla. However, there was no significant difference found between the Liam and Sarah character.

The *Overall experience* also was effected by different characters, where repeated measure ANOVA indicated that there exists a significant difference between the means of *Overall experience* in four characters, $F(3, 105) = 10.71, p < .001$. The post-hoc pair-wise comparison is reported in Table 4. Results indicate a similar pattern as for *Talking*, where Liam created a significantly higher user experience than Henry for male characters and Sarah had a significantly higher experience than Camilla for female characters. However, there was no difference between Liam and Sarah.

For realism in *Appearance*, ANOVA showed a significant difference between the characters, $F(3, 105) = 6.22, p = .001$. The post-hoc test using a Bonferroni correction are reported in Table 4. The paired comparison revealed that unlike realism in talking and overall experience, there is no significant difference between the characters in terms of the realism of the appearance. The only significant difference was between Liam and Camila with $p = .002$, comparing the worst-rated female and the best-rated male characters. User study results on animated avatars created with the Unity game engine are shown in Figure 5.

To summarise, RQ2 compared four different avatars created with the game engine and found the best-rated avatars. Characters Sarah and Liam illustrated in Figure 1 were rated highest, and they were significantly better in terms of realism in talking and user experience than Henry and Camilla. However, there were no differences between Liam and Sarah or Henry and Camila.

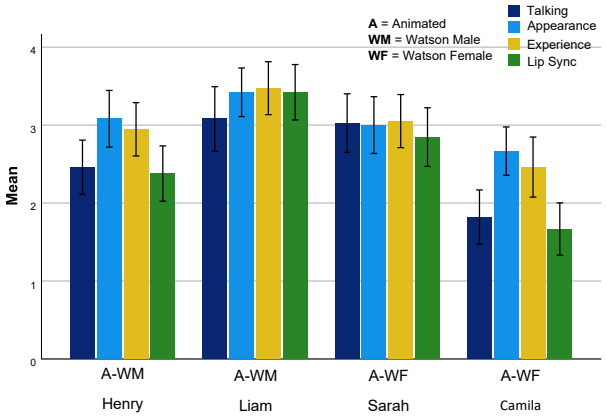


Figure 5: Comparison of Unity-animated avatars

Table 4: Results of PostHoc test with Bonferroni correction comparing the animated characters.

Animated male characters			
Quality	**Henry**	**Liam**	**Sig.**
Overall QoE	M=2.97 , SD=1.02	M=3.44 , SD=1.02	.02*
Talking	M= 2.46, SD=1.04	M=3.08 , SD=1.26	.059
Appearance	M=3.08 , SD=1.10	M=3.44 , SD=.93	.24
Animated female characters.			
Quality	**Sarah**	**Camila**	**Sig.**
Overall QoE	M=3.11 , SD=1.0	M=2.5 , SD=1.2	.012*
Talking	M=3.03 , SD=1.15	M=1.7 , SD=.96	<.001*
Appearance	M=3.08 , SD=1.10	M=2.64 , SD=.96	.24
Best animated female and male characters.			
Quality	**Sarah**	**Liam**	**Sig.**
Overall QoE	M=3.11 , SD=1.00	M=3.44 , SD=1.02	.26
Talking	M=3.03 , SD=1.15	M=3.03 , SD=1.15	1.0
Appearance	M=3.08 , SD=1.10	M=3.44 , SD=.93	1.0

4.3 RQ3: Synthetic and Natural Voices

While RQ1 and RQ2 were focused on the visual side of the avatar, this research question investigates the realism in the voice of the avatars. The test condition included synthetic voices generated by IBM Watson TTS and recorded voices of natural speech, each of them for two different genders. This section investigates two sub-questions of whether the computer-generated voices generated can be as realistic as human voices and the appropriateness of the voices associated with each character. Figure 3 shows the results of the user study to compare the voices generated by a computer and human.

To answer the research question comparing the synthetic and human voices, a factorial ANOVA was conducted over the eight avatars to compare the impact of model (GAN-generated and animated), gender (female and male) and voice (synthetic and natural) and interaction between them on the realism of *Talking* of the characters.

The factorial ANOVA reveals that there is a significant main effect of Model $F(1, 35) = 5.98, p = .02$, but there is no significant main effect for gender $F(1, 35) = .5, p = .48$ and type of voices

$F(1, 35) = 1.39, p = .24$, meaning that the computer generated voices are rated equally as human voices and there was no significant interaction effect between gender and type of voice.

In this test, type of voices always matched the gender, i.e., for a male character the voices were also male and for female characters the voices were also female. Our pilot study suggested that female voice seems to work better on male child characters. A factorial ANOVA was conducted over the mean of realism in Talking to test the hypothesis. The results indicate selecting a female voice for a boy is as good as boy voice and there is no significant main effect for voices (p=.67). This means that the generation of the child avatars for both genders could be possible by developing only one voice.

To summarise, RQ3 showed that using natural voices does not create any extra realism for virtual avatars than the computer-generated voices. Furthermore, it was revealed that a female's voice could be used for both boy and girl child avatars, and it can create the same level of realism and experience for users.

4.4 RQ4: Generated avatars using a GAN vs. the Unity game engine

This research question compares the best GAN-generated avatars created by GAN techniques derived from RQ1 and the best animated avatars from RQ2. The best avatars were selected based on the realism in *appearance* and *talking* of the avatars. Since the system will eventually use computer-generated voices and were rated as good as natural voices, we only considered avatars with IBM Watson voice.

In the GAN-generated avatars, the two avatars of Omid and Darya from model PC-AVS were rated as most realistic, and as discussed in RQ2, Liam and Sarah were the most realistic once in the animated avatars. A factorial ANOVA was conducted to compare these avatars to compare human face and animated face videos in both genders. Although there was a general trend on more realism of videos created by GAN based on the subjective experience, there was no significant main effect for realism in *Talking* $F(1, 37) = 2.54, p = .12$, *Appearance* $F(1, 37) = 2.66, p = .11$, and *Overall experience* $F(1, 37) = 1.09, p = .30$.

To summarise, RQ4 showed that although avatars generated by GANs more realistic, they did not necessarily create a higher quality of experience and realism for the users.

4.5 Qualitative Results

In addition to the post-experience questionnaire, which consisted of four quality aspects, we used a 4-item post-test questionnaire followed by an open-question. We directly asked participants to choose between animated and GAN-generated avatars and provide the reasons behind their selections. These items are coded with Q1 to Q4 as shown in Table 2. The results are presented in Table 5, where the total number of votes per avatar is shown for each item. The results indicate that 64% of participants find the appearance (Q1) of GAN-generated avatars more realistic, 58% found the overall talking (Q3) of GAN-generated avatars more realistic and 58% liked (Q2) the GAN-generated avatars more. However, when participants

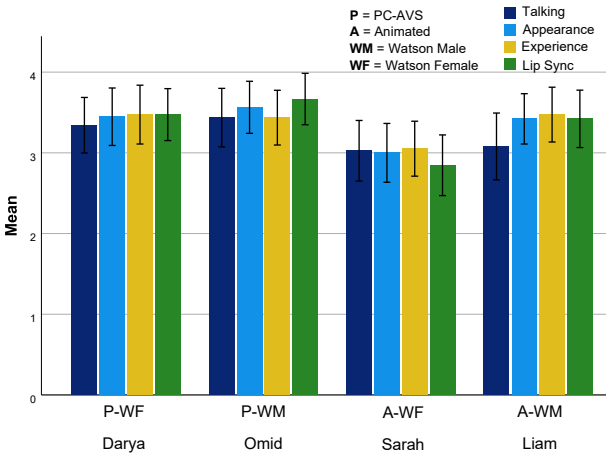


Figure 6: Results of the user study for the evaluation of two of the best girl and boy characters created for both the GAN-Based and game engine-based approaches.

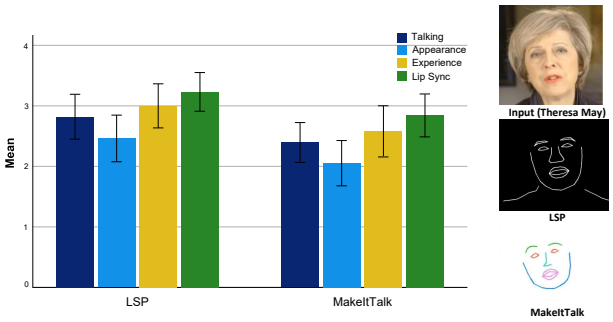


Figure 7: Comparison of faical landmarks from LSP and MakeItTalk.

were asked about their preference for using the avatar in a conversation with computer (Q4), 51% of the participants prefer to use the animated avatars.

In the open-ended question, we asked them to express why they preferred one over the other. Almost all the participants expressed that the avatars generated with GANs are more realistic. Four participants mentioned that the lip sync was better in the GAN-generated avatars than in the animated avatars, i.e., one participant even mentioned that "The GAN-generated avatars could trick me into that an actual person is talking." Interestingly, three participants explicitly mentioned that the talking faces looked scary and one participant mentioned that there was something off-putting and it felt wrong. An uncanny valley can explain why GAN-generated avatars are innately unlikable unless they are perfect and natural.

Table 5: The results of the user study to evaluate the number of votes in response the questions.

Question	GAN-generated	Animated
Q1	25	14
Q2	23	16
Q3	23	16
Q4	19	20

5 DISCUSSION AND FUTURE WORK

This paper presented the results of a subjective study, where various animated avatars created using the Unity game engine and GAN-generated avatars created by the state-of-the-art GANs-based techniques were compared. The work on RQ1, comparing GAN-based approaches, showed no significant difference between avatars created by PC-AVS and MakeitTalk. However, the comparison between the facial landmark of LSP and MakeitTalk methods indicated that the facial landmarks created by LSP methods are more realistic. In future work, it is planned to train the LSP method to be able to generate GAN-generated avatars. Furthermore, the MakeitTalk and PC-AVS methods are not real-time, but we are planning to work on real-time processing support in our future work.

RQ2 investigates the best animated avatar, which is going to be used in the next iteration of the prototype and improve the system illustrated in [9]. It was shown that, for male characters Liam and the female characters Sarah can create the best realism and user experience. However, there was no significant difference among the best characters of each gender.

RQ3 investigated the voice of the avatars. It was shown that the synthetic voices created by IBM Watson TTS are as realistic as the natural voices and can make a similar user experience for users of the system. Moreover, it was shown that female synthetic voices could be used for both boy's and girl's avatars and create a realistic experience. In the future iteration of the prototype, we will use the computer-generated female voices for both characters as it was shown to create a better user experience.

Furthermore, RQ4 investigated the difference between GAN-generated and animated avatars. Contrary to our expectations, animated and GAN-generated avatars were rated similarly for user experience and realism. There were no significant differences between the best animated and GAN-generated avatars. The results are more surprising when more than half of the participants stated that they prefer to have a conversation with an animated avatar, even though they are less realistic in terms of appearance and the overall talking of the avatar. The reason behind that is reflected in the open questions, where a few participants stated that the GAN-generated avatars look scary. This can be evidence of the existence of the uncanny valley theory, which states that if an avatar is humanoid but imperfect, its non-human features create strange feelings in the audience that cause feelings of uneasiness and even disgust.

Overall, generating video from audio is a challenging problem. This is due to facts that mapping goes from a lower-dimensional (audio) to a higher-dimensional (video) signal, and human eyes are susceptible to visual artifacts in images and videos (the so-called Uncanny valley effect). These challenges pose stringent requirements on the talking-head models. In our ongoing and future works, we target to improve the realism of the GAN-generated avatars in a way that creates fewer artifacts. In addition, we will experiment with mixing both approaches, whereby by feeding an animated facial image as the input, the talking head can be generated. In this way, the avatar does not have a human face while the lip sync and the head movements are good.

6 CONCLUSION

The ability of interviewers to retrieve essential information from a maltreated and abused child is highly dependent on their interviewing skills. Interview-training systems for CPS and law enforcement professionals that use realistic interactive avatars can effectively improve conversation skills in this context. To prototype such a system, we adopted two different approaches to create several child avatars, i.e., animated avatars created with the Unity game engine and human-like avatars generated using the PC-AVS, MakeItTalk and LSP GAN techniques. We also compared real and computer-generated voices of the avatars. To compare and evaluate the approaches, we designed a user study on 28 generated videos to perform the assessment based on subjective score evaluation. The study evaluated several GAN and game engine methods for creating virtual avatars. Moreover, the results of the study indicate the existence of an uncanny valley, where participants preferred talking to the animated avatars rather talking to the more realistic state-of-the-art avatars generated by GAN. Furthermore, it was shown that using natural voices for virtual avatars does not create more realism and better user experiences compared to computer-generated voices.

ACKNOWLEDGMENTS

This research is sponsored by the Research Council of Norway, project number #314690 ("Interview training of child-welfare and law-enforcement professionals interviewing maltreated children supported via artificial avatars").

REFERENCES

[1] Mike Allen. 2017. *The SAGE encyclopedia of communication research methods.* SAGE publications.
[2] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European conference on computer vision (ECCV).* 119–135.
[3] Gunn-Astrid Baugerud, Miriam S Johnson, Helle BG Hansen, Svein Magnussen, and Michael E Lamb. 2020. Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015–2017). *Applied Cognitive Psychology* 34, 3 (2020), 654–663.
[4] Nathalie Carrick, Jodi A Quas, and Thomas Lyon. 2010. Maltreated and nonmaltreated children's evaluations of emotional fantasy. *Child Abuse & Neglect* 34, 2 (2010), 129.
[5] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. 2020. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision.* Springer, 35–51.
[6] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7832–7841.
[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622* (2018).
[8] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. 2020. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10893–10900.

[9] Syed Zohaib Hassan, Pegah Salehi, Ragnhild Klingenberg Røed, Pål Halvorsen, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pierre Lison, Michael Riegler, Michael E. Lamb, Carsten Griwodz, and Saeed Shafiee Sabet. 2022. Towards an AI-Driven Talking Avatar in Virtual Reality for Investigative Interviews of Children. *2nd edition of the Game Systems Workshop (GameSys '22)* (2022).

[10] David Johansson. 2015. Design and evaluation of an avatar-mediated system for child interview training.

[11] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[13] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

[14] Yuanxun Lu, Jinxiang Chai, and Xun Cao. 2021. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–17.

[15] Karl F MacDorman, Robert D Green, Chin-Chang Ho, and Clinton T Koch. 2009. Too real for comfort? Uncanny responses to computer generated faces. *Computers in human behavior* 25, 3 (2009), 695–710.

[16] Moustafa Meshry, Saksham Suri, Larry S Davis, and Abhinav Shrivastava. 2021. Learned Spatial Representations for Few-shot Talking-Head Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13829–13838.

[17] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

[18] Martine B Powell, Belinda Guadagno, and Mairi Benson. 2016. Improving child investigative interviewer performance through computer-based learning activities. *Policing and Society* 26, 4 (2016), 365–374.

[19] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando De la Torre, and Yaser Sheikh. 2021. Audio-and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 41–50.

[20] Tong Sha, Wei Zhang, Tong Shen, Zhoujun Li, and Tao Mei. 2021. Deep Person Generation: A Survey from the Perspective of Face, Pose and Cloth Synthesis. *arXiv preprint arXiv:2109.02081* (2021).

[21] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. 2022. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security* (2022).

[22] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.

[23] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*. Springer, 716–731.

[24] WHO. 2020. *Child maltreatment*. https://www.who.int/news-room/fact-sheets/detail/child-maltreatment

[25] Olivia Wiles, A Koepke, and Andrew Zisserman. 2018. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*. 670–686.

[26] Erin Wilson, David G Hewett, Brian C Jolly, Sarah Janssens, and Michael M Beckmann. 2018. Is that realistic? The development of a realism assessment questionnaire and its application in appraising three simulators for a gynaecology procedure. *Advances in Simulation* 3, 1 (2018), 1–7.

[27] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. 2020. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137* (2020).

[28] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9299–9306.

[29] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4176–4186.

[30] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–15.

Article IV: Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children

Authors: *Pegah Salehi*, Syed Zohaib Hassan, Myrthe Lammerse, Saeed Shafiee Sabet, Ingvild Riiser, Ragnhild Klingenberg Røed, Miriam S Johnson, Vajira Thambawita, Steven A Hicks, Martine Powell, Michael E Lamb, Gunn Astrid Baugerud, Pål Halvorsen, Michael A Riegler

Candidate's contributions: The candidate coordinated the overall development and structure of the paper, ensuring consistency across contributions from multiple co-authors. She played a central role in the design and execution of the user study comparing GAN-based and Unity-generated child avatars. Her responsibilities included the experimental setup, participant coordination, and data collection. She also led the analysis of participant feedback, with a focus on perceived visual realism, talking realism, and user experience. As the primary contributor to the manuscript, the candidate authored key sections detailing the study methodology and interpretation of results, providing insights that informed the broader development of avatar-based training for CPS and law enforcement professionals.

Published in: Big data and cognitive computing (BDCC), vol. 6, pp. 62 (2022)

DOI: 10.3390/bdcc6020062

Contributed to: This work contributed to the **RO-2**.

big data and cognitive computing

MDPI

Article

Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children

Pegah Salehi [1,*], Syed Zohaib Hassan [1], Myrthe Lammerse [2], Saeed Shafiee Sabet [1], Ingvild Riiser [2], Ragnhild Klingenberg Røed [2], Miriam S. Johnson [2], Vajira Thambawita [1], Steven A. Hicks [1], Martine Powell [3], Michael E. Lamb [4], Gunn Astrid Baugerud [2], Pål Halvorsen [1,2] and Michael A. Riegler [1,5]

1 SimulaMet, 0167 Oslo, Norway; syed@simula.no (S.Z.H.); saeed@simula.no (S.S.S.); vajira@simula.no (V.T.); steven@simula.no (S.A.H.); paalh@simula.no (P.H.); michael@simula.no (M.A.R.)
2 Department of Computer Science, Oslo Metropolitan University, 0130 Oslo, Norway; myrthela@oslomet.no (M.L.); ingvildr@oslomet.no (I.R.); rar@oslomet.no (R.K.R.); mirsin@oslomet.no (M.S.J.); gunnba@oslomet.no (G.A.B.)
3 Centre for Investigative Interviewing, Griffith Criminology Institute, Griffith University, Brisbane 4122, Australia; martine.powell@griffith.edu.au
4 Deaprtment of Psychology, University of Cambridge, Cambridge CB2 3RQ, UK; mel37@cam.ac.uk
5 Department of Computer Science, University of Tromsø, 9037 Tromsø, Norway
* Correspondence: pegah@simula.no; Tel.: +47-92097694

Abstract: When responding to allegations of child sexual, physical, and psychological abuse, Child Protection Service (CPS) workers and police personnel need to elicit detailed and accurate accounts of the abuse to assist in decision-making and prosecution. Current research emphasizes the importance of the interviewer's ability to follow empirically based guidelines. In doing so, it is essential to implement economical and scientific training courses for interviewers. Due to recent advances in artificial intelligence, we propose to generate a realistic and interactive child avatar, aiming to mimic a child. Our ongoing research involves the integration and interaction of different components with each other, including how to handle the language, auditory, emotional, and visual components of the avatar. This paper presents three subjective studies that investigate and compare various state-of-the-art methods for implementing multiple aspects of the child avatar. The first user study evaluates the whole system and shows that the system is well received by the expert and highlights the importance of its realism. The second user study investigates the emotional component and how it can be integrated with video and audio, and the third user study investigates realism in the auditory and visual components of the avatar created by different methods. The insights and feedback from these studies have contributed to the refined and improved architecture of the child avatar system which we present here.

Keywords: Child Protection Services (CPS); interview training; virtual child avatar; generative pre-trained transformer 3 (GPT-3); generative adversarial networks (GANs)

check for updates

Citation: Salehi, P.; Hassan, S.Z.; Lammerse, M.; Sabet, S.S.; Riiser, I.; Røed, R.K.; Johnson, M.S.; Thambawita, V.; Hicks, S.A.; Powell, M.; et al. Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children. *Big Data Cogn. Comput.* **2022**, *6*, 62. https://doi.org/10.3390/bdcc6020062

Academic Editor: Moulay A. Akhloufi

Received: 30 April 2022
Accepted: 21 May 2022
Published: 1 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Child sexual abuse (CSA), violence, and neglect are major global public health issues that have far-reaching immediate and long-term implications for the children involved, as well as for society in general [1]. Child abuse represents a risk to the child's existence and development. Research proves that it has major consequences for the child, leading to cognitive, behavioral, and social problems, as well as substance abuse, serious mental health problems, and death [2]. CSA alone is identified by the World Health Organization (WHO) as substantively contributing to the global burden of disease [3]. To investigate this problem, child protective services (CPS) and law enforcement personnel must interview the children concerned when planning responses to such abuse; the interviews thus play

important roles in safeguarding children to promote their welfare and protect them from harm [4]. Children are often interviewed as both victims and key witnesses to the abusive incidents [5]. As there is often a lack of corroborative evidence in these cases, informative interviews with child complainants play a crucial role in their investigation [6].

Most abused children do not have any physical signs of abuse [7]. This means that the progress of the investigation depends upon the informativeness of the child's account of the incident. Therefore, the interviewer's ability to ask good questions that maximize the quality and quantity of the information provided by the child is crucial. The investigative interview is thus a vital component of a comprehensive child abuse investigation and a prime opportunity for investigators to elicit accurate and detailed information from alleged victims of child abuse [8]. A large body of research on children's cognitive and social development, supplemented by field studies on children's ability to describe experienced stressful and traumatic events, have identified best practices for how investigative interviewing with maltreated children should be conducted [9]. These best-practice investigative interview guidelines provide interviewers with clear instructions regarding how child witnesses should be questioned and supported in a non-suggestive way during the interview to maximize the value of their testimony [10,11]. In particular, this includes offering the children open-ended prompts while avoiding forced-choice and suggestive questions because free recall questions encourage more accurate and longer responses from children [12,13] (e.g., "What happened next?" where the child is provided with no information that could influence their answers). Interviewers should also avoid suggestive or leading questions (e.g., "the person touched your private part. Is it true?" or any other question that might lead the child to tell a specific story). However, unfortunately most interviewers do not adhere to such best-practice guidelines [14].

Powell et al. [15] found that interactive computer-based learning activities can improve the effectiveness and productivity of investigative interviewers. Moreover, the area of Artificial Intelligence in Education (AIE) facilitates the learning process and transfers knowledge through humanoid robots or virtual avatars. In the same vein, there has been significant progress in recent years towards synthesizing realistic digital humans, avatars, characters, and agents [16]. The approach we intend to follow is to develop a realistic child avatar that involves the integration and interaction of different components with each other, including dialogue models, auditory, emotional and visual components of the avatar. Our initial prototypes [17,18] demonstrated the potential of such an avatar, and following the process of improving the interactive child avatar, we have employed cutting-edge technologies to synthesize a realistic child avatar; for example, we have used the RASA and GPT-3 for dialogues, the IBM Watson service for auditory, the GPT-3 and BART for emotions, and GANs and Unity game engine for the visual appearance of the avatar. In this paper, we systematically discuss the results obtained using these various tools and techniques.

In addition, this article compares and contrasts various state-of-the-art ways of incorporating multiple characteristics of the child avatar in three user studies. The first user study assessed the entire system and found that it was well received by experts who emphasized the importance of realism; the second user study looked at the emotional component and how it could be integrated with video and audio features; and the third user study looked at the realism in the auditory and visual components of the avatar created using various methods. Based on the findings and user comments from these three investigations, the paper describes the architecture of our child avatar system. In summary, the main contributions of our work are:

- An investigation of the potential learning effects and user experience with the system.
- An investigation of the realism of the synthetic voices compared to natural voices.
- An examination of emotion extraction with different models based on children's answers in investigative interviews.
- An investigation of the realism of several methods for generating the appearance of the talking avatar.

- An investigation of the system architecture regarding the integration and interaction of various system components.

The rest of the paper is organized as follows: Section 2 provides an overview of the state-of-the-art with respect to different system components and highlights the importance of the child avatar interview training system. Section 3 describes the system in overview and discusses the material and methods used to develop various components of the system. Section 4 evaluates the performance of various system components and discusses the results of our three user studies conducted using the system components, and Section 5 discusses the results and their limitations along with suggestions for future research. Finally, Section 6 concludes the paper and highlights the main findings.

2. Related Work

This section reviews related work on each of our research questions separately.

2.1. Investigative Interview Training

A number of training approaches have been developed to provide essential information about how to conduct these interviews in compliance with best-practice guidelines. One of the most prevalent methods of imparting knowledge to police officers and CPS workers involves traditional classroom-based teaching. However, a large body of research from all over the world of investigative interviewing of children [9] has shown that attempts to improve the quality of investigative interviewing using such traditional classroom-based instructions have not been productive [14]. Professionals tend not to follow recommended questioning strategies. Instead, they tend to ask many risky option-posing and suggestive questions rather than the open-ended questions that are preferred, resulting in unreliable children's reports about their experiences [19–22]. Following most training programs, participants are required to participate in mock-interview activities involving an instructor/professional actor portraying an abused child, with feedback provided by a trainer. Recently, the use of mechanical avatars in interviewer training have been shown to be advantageous, especially when integrated with feedback [9,15,23–26]. By training interviewers in the adoption of recommended best-practice interview strategies using a dynamic avatar, including question types that encourage children to make the fullest possible use of their cognitive and communicative abilities while avoiding strategies like suggestive questions, has the potential to be beneficial. Conducting simulated interviews may be particularly suitable for basic training, as well as practice and refresher training, which are known to be especially important for the maintenance of skills. Using dynamic avatars in training may be a very efficient way of acquiring and training practical skills. In addition, AI-based interview training is less expensive, can be used over an extended period of time and may be much more accessible to users than traditional in-person-training [15,27].

2.2. Emotions

Emotional intelligence is the ability to perceive, use, understand, and manage emotions [28]. The level of emotional intelligence that an investigative interviewer portrays influences the execution and performance of the emotional labor [29]. Which refers to jobs where employees are expected to recognize emotions and act accordingly. These jobs have guidelines and rules that need to be followed to ensure the quality of the work completed [30]. The results of an investigative interview with an adult depend on how officers handle the emotions of the interviewee. The outcome, the interviewees' well being, and therapeutic jurisprudence are all positively influenced if the interview is conducted in an emotionally intelligent way [31]. Research addressing emotional intelligence in child interviews is limited, but Albaek et al. [32] showed a need to address professionals' emotional distress in child abuse cases in one qualitative meta-synthesis. To train CPS workers and police officers to conduct investigative interviews effectively, it is important that the training module emphasizes the right kind of emotionally expressive responses to children's emotions. By having the avatar express different emotions, just like real children

would portray them, the interviewer can practice handling emotions in different scenarios by, for example, offering non-suggestive emotional support when interviewing children.

Research shows that by classifying emotions through the identification of facial cues, we can distinguish seven universal emotions [33–35]. These seven emotions are *joy, sadness, anger, disgust, contempt, fear, and surprise.* These emotions can be expressed both verbally as well as non-verbally. Katz and Hunter [36] and Karni-Visel et al. [37] classified non-verbal emotions based on real-life interviews and discovered that non-verbal expressions are 10 times more commonly expressed than verbal emotional expressions. However, traumatized children may not show any emotion, either positive or negative [38]. This phenomenon is called numbing. The paradoxical combination of expressing more non-verbal emotions and numbing creates a difficult landscape for the talking child avatar. Therefore, it is important to portray these emotions during the conversation with different degrees of expression based on different child avatar personas. We must pay attention to the extent to which the different personas express different emotions, as this can vary considerably between children. The expression of the emotions will be noticeable in both visual and auditory outputs with emotions, for example, changing facial expressions or the pitch of the voice.

2.3. Chatbot

Typical applications of chatbots are found in call centers, e-commerce customer services, and internet gaming. Chatbots in these sectors employ different machine learning algorithms to conduct auditory or textual conversations with end users [39]. Chatbots that mimic an allegedly abused child are much more complex to realize than simple question-answering or even open-ended social chatbots [40]. Social chatbots have become more realistic and advanced in the last few years. The goal of a social chatbot is to establish an emotional connection with the person who is using it. A social chatbot is not the same as an investigative training module, and it also must be able to recognize emotions and track emotional changes during a conversation. Verbal or non-verbal emotions are an important part of the conversation and can convey a lot of required information. However, there are some implementations of chatbots that can converse with a certain emotional component [41]. XiaoIce is a social chatbot developed by Microsoft [42] which detects the sentiment reflected in the input before responding. However, identifying, understanding and displaying emotions, and modeling the impact of emotions on the conversation quality are challenging tasks that we are currently addressing. In addition to the emotions expressed in the conversation, there is a need to design and generate different personas that have the ability not only to show emotions differently, but also to react to emotionally charged conversations. Li et al. [43] showed how a neural-based approach can model the different personas. Parametrization of emotions and personas into a chatbot could really improve realism and immersiveness for the end-users, i.e., in our case, CPS trainees. Therefore, this social chatbot aims to recognize emotions and present a consistent personality. This personality would involve acting according to a given age, gender, language, speaking style, attitude, level of knowledge, etc. We have started working on these individual aspects of the chatbots, and we discuss them in the sections below.

2.4. Auditory

A conversation is a two-way flow of information. For a human to communicate with a computer program using speech, one must find ways to translate between the textual and auditory domains. This requires so-called speech-to-text (STT) and text-to-speech (TTS) methods that work at low latencies to ensure a natural conversational flow. With the recent advances in function of CNNs [44] and recurrent neural networks (RNNs) [45], STT methods are approaching human parity [46]. Benchmarks such as LibriSpeech [47] and the 2000 NIST Speaker Recognition Evaluation [48] have shown that current state-of-the-art deep learning methods can recognize speech with very few mistakes [49,50]. As auditory speech is more information-rich than pure text, we may also extract paralinguistic information

such as emotions. Databases, such as IEMOCAP [51], can be used to train speech emotion classifiers. Examples of such classifiers include CNN long short-term memory (LSTM) networks [52] and CNNs [53], which take audio and/or (log-mel) spectrograms as input. TTS research has also developed considerably in recent years. Models such as Wavenet [54,55] and Waveglow [55] have shown immense promise in using generative models to produce very realistic speech from input transcripts. These methods are state-of-the-art and work efficiently, which makes them apposite for the system presented here [56].

2.5. Visual

Due to the free access to large-scale datasets and the rapid development of deep learning technologies, especially Generative Adversarial Network (GAN) [57] and its variants, unprecedented successes have been achieved in the generation of realistic multimedia content.

Among the recent innovations in humanoid avatars, the Uneeq (https://digitalhumans.com/, accessed on 24 May 2022) has designed a conversational AI system based on video and audio components with natural voice and highly realistic facial expressions, and the Soul Machines (https://www.soulmachines.com/, accessed on 24 May 2022) has developed a digital brain that emulates human cognitive processes.

In the following, we provide a basic classification for understanding the differences between various facial manipulation techniques regarding the amount of manipulation (as shown in Figure 1). A brief explanation of each of them is given below.

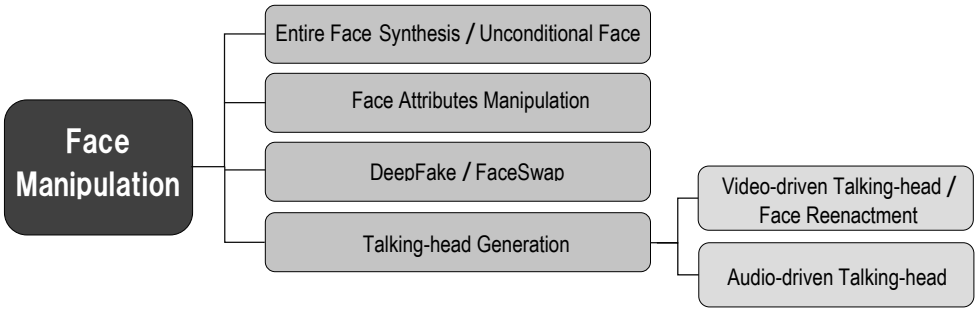


Figure 1. A comprehensive category of face manipulation techniques.

Entire Face Synthesis (or Unconditional Face Generation) means the production of non-existent portrait images in the real world, e.g., PGGAN [58] and StyleGAN (https://thispersondoesnotexist.com/, accessed on 24 May 2022) [59] fall into this manipulation, which generates ultra-realistic photos of humans who do not actually exist; even humans have difficulty assessing their realism.

Face Attributes Manipulation consists of editing some attributes of a face such as hair, skin color, eyeglasses, gender, and age as when using StarGAN [60]. Moreover, FaceApp (https://www.faceapp.com/, accessed on 24 May 2022) has popularized facial attribute editing as an application.

DeepFake (or FaceSwap) commonly replaces one person's face in an image or video with another person's. In this respect, the combination of source videos results in a fake video that shows an action or an event that never occurred in reality, for example, creating fake news [61]. In the same way, a digital avatar has been developed to swap faces in video chats (https://blog.siggraph.org/2021/01/ai-avatars-virtual-assistants-and-deepfakes-a-real-time-look.html/, accessed on 24 May 2022). The DeepFaceLab (https://github.com/iperov/DeepFaceLab, accessed on 24 May 2022) [62] is also an open sourced software package to build high fidelity face-swapping videos.

Talking-head Generation can synthesize precise lip synchronization, head pose motion, and natural facial expressions from a specific person by capturing one or more input signals. Depending on the input data type, the existing approaches to talking-head generation can be divided into two broad categories: audio-driven and video-driven. Many

audio-driven methods avoid straight mapping from audio to image, but instead first map audio to an intermediate step like 2D facial landmarks [63–67] or 3D face shapes [68–72], and then render photorealistic videos. Audio-driven talking-head generation is inherently difficult because it generates hand and mouth movements based only on an audio signal. Thus, some of the proposed works control the motions of a subject with one or more additional videos as input categorized as **video-driven** methods (a.k.a **face reenactment**) [73–78]. Similarly, Face2Face [73] enables real-time facial reenactment of a target video sequence, i.e., animating the facial expressions of the target video using a source actor and re-rendering the manipulated output video in a photorealistic fashion. Some models also focus on training their model using a specific person [79,80]. Audio-driven methods have received more attention in recent research because they aim to synthesize a universal talking-head model for various subjects and applications.

2.6. Child Interview Training Avatars

Empowering Interviewer Training (EIT) [27] is an investigative interview training program. Child responses are pre-defined in the system and responses are selected using a rule-based algorithm. Based on the selected response, prerecorded videos of children showing different emotions are chosen by a human operator and shown to the user. Pomedda et al. conducted multiple studies using this system to analyze the training effects after having multiple sessions of these mock interviews, the effects of feedback and reflection on improving the quality of the investigative interviews [24,81–83]. In Sweden, Linnæus university and AvBIT Labs have also introduced an online interview training system. They also use prerecorded audio responses and videos of a child avatar and human operator. the user is shown an appropriate video response with suitable emotions controlled using Wirecast software controls via the Skype interface [84,85]. Even though development and testing of these systems have successfully transferred the investigative skills needed to interview abused children, these systems are not dynamic in the response generation and have human input during the response selection phase. This makes them rigid and harder to operate.

We propose an AI-driven training system that can dynamically respond to the questions and provide a higher realism during the training interviews. In addition, we propose a system that would be completely independent of human input.

3. Materials and Methods

Figure 2 provides an overview of the entire architecture of the child investigative interview avatar. Our proposed idea for the interaction flow between different system components and the flow of data/information between them is well traceable in the architecture. Yellow parts mark the language model, where GPT-3 is trained on real and mock interviews, green parts are the emotional engine that extracts emotion from the generated texts and provides the input to be applied in the avatar's visual and auditory output, blue parts are the auditory system, which uses IBM Watson API to convert the text to audio and audio to text, and purple parts show the visual sections, where the generated audio and visual aspects of the avatar are combined to create a talking face.

Although all of these modules can be developed in isolation, integrating all the components would still be needed. This section discusses future research and the work in progress for each component.

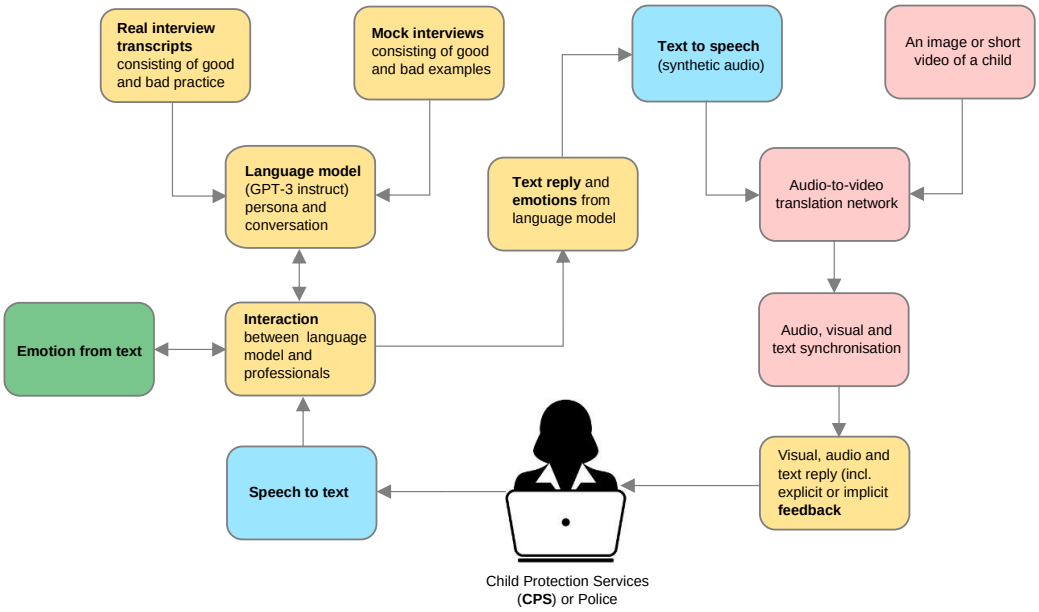


Figure 2. System architecture. Green blocks denote the interactive parts, yellow blocks are language-related, blue audio-related, and pink the parts of the system related to visualization.

3.1. Language

There are two sources of data, mock and real interviews, which will provide the training data for our chatbot. The Centre for Investigative Interviewing at Griffith University, Australia [15] provided 1000 transcripts of the mock interviews conducted as part of their investigative training for social workers, police officers and psychologists. In these mock interviews, a trained actor mimics an allegedly abused child. Real-life child investigative interviews will be added to the system at a later stage since we want to initially have a more rigid and controllable version of the avatar.

The current chatbot was developed using Rasa (https://rasa.com/, accessed on 24 May 2022). It provides an open source framework to develop the automated solutions for text-based conversations. The choice of RASA at this stage was motivated by the fact that we lack enough data to develop a solution from scratch and aim to develop a prototype for proof of concept. Additionally, with the small amount of data, RASA provides an environment to control the flow of conversation. At the time of developing the first chatbot, we had two hundred transcripts of well-conducted training interviews. The dataset we created comprises conversations between a child aged 5–7 years and an interviewer. RASA is powered by the TensorFlow [86] back-end framework. RASA has multiple modules which employ different deep learning models to develop a complete dialogue model. Each module takes in the training data in a module specific format. The NLU module employs the Dual Intent and Entity Transformer (DIET) [87] for training to predict intent and entities in the utterances jointly.

We extracted personas from the available transcripts by manually clustering them based on scenario and chose one of them to develop our RASA chatbot. We aim to build a chatbot that can show different personas and we are working with GPT-3 to achieve that. We plan to use fine-tuned GPT-3 on our transcripts to allow us to capture the behavior of children with different personas dynamically. We classify questions asked by interviewer into 15 different categories and child responses as productive and non-productive. We plan to use these data to develop deterministic models that can be employed for feedback mechanism and regulating GPT-3 during the conversation to alter the behavior of the child bot based on the type of question asked.

We conducted an interactive study that followed ITU-T Rec. P.809 [88] test paradigm which suggests that participants should be evaluated using a specified method after they interact with the system in a specific scenario. In this study, we invited participants from CPS agencies who had some experience interviewing abused children [17]. They were asked to interview the child avatar mimicking a six years old child allegedly sexually abused. The goal of this study was to evaluate the user's quality of experience and assess the system's capability to enhance the learning experience and the acquisition of knowledge and skills for communicating with abused children.

3.2. Auditory

We tested different speech synthesis services to see which one sounded the most childlike. We chose to use IBM Watson services for text-to-speech (https://www.ibm.com/cloud/watson-text-to-speech, accessed on 24 May 2022) (TTS) and speech-to-text (https://www.ibm.com/no-en/cloud/watson-speech-to-text, accessed on 24 May 2022) (STT) synthesis because of the range of synthetic voices and in-built options for pace and pitch adjustment. Watson TTS and STT are cloud service APIs that serve as communicative bridges between language (back-end) and visual (front-end) components. The user communicates with the front-end verbally, with the question uttered by the user sent to the IBM STT API to be transcribed with the response then forwarded to the back-end. The dialogue model processes the user utterance at the back end and generates the appropriate response. This response is then sent to the IBM TTS API, sending the audio response to the user at the front end.

3.3. Emotions

Unfortunately, it is difficult to obtain audiovisual field interviews due to privacy concerns. Therefore, the prediction of the emotions is purely based on written text rather than a combination of both textual, audio, and visual input from transcripts of mock interviews [15]. Due to the lack of annotated data, we started our experiments with a pre-trained transformer model for sequence classification. From the HuggingFace library (https://huggingface.co/, accessed on 24 May 2022) we used the zero-shot classification pipeline in combination with the BART large model. The BART large model [89] is an autoencoder for pretraining sequence-to-sequence models. This model is made specifically for GLUE tasks [90], with sentiment classification as a sub-task. We present the model with our set of labels. The model then returns the class labels with their corresponding probabilities. We also experimented with the use of GPT-3 [91] and with different approaches to predict and choose the correct sentiment from the seven options. The emotions were predicted using only one sentence, the whole story up until the current sentence, and using a sliding window with and without a threshold to restart the window. The window size experiments varied with window sizes of 3, 5, 7, 10, and 15. If a threshold is set and the single sentence prediction has a higher probability than a certain threshold, it restarts the window. The choice for including such a threshold was made to be able to recognize the potential for sudden significant changes in emotions. If a child, for example, starts telling about how they started crying, the sentiment turns very sad very quickly, and the context is thus of less influence.

We commenced using the before-mentioned seven universal emotions. However, we also performed the same experiments using only the four basic emotions of joy, sadness, anger, and fear [92] instead of the seven. The reason to change the subclass of emotions from four to seven was due to the low Intraclass Correlation Coefficient score on the seven dimension experiments, as it only scored 0.537 on average measures.

The importance of the emotion classification part becomes cognizable when it is integrated into the complete system. It plays an integral part in the visual and audio output by altering the output based on the specified emotion. We created an emotion pipeline that predicts both the emotional valence of the interviewer's input and the chatbot's output. Due to the closed-off RASA environment, it is not possible to directly classify the bot's

responses in the environment. However, it is possible to classify the input in the RASA environment. Using the requests package (https://github.com/psf/requests, accessed on 24 May 2022) in Python connected to the localhost address where the RASA bot is running, we are able to receive and classify both the human input as well as the RASA output. This is necessary because classifying the RASA response is of great importance for our system. The emotion pipeline receives these texts as input and then provides the input for the audio and visual part.

3.4. Visual

Our early efforts are dedicated to research and proof-of-concept development. To discover the right approach to generate a visual child avatar, we went through a trial-and-error procedure that involved three prototypes: **Faceswap**, **Unity Multipurpose Avatar (UMA)** and **Talking-head Generation**. The following is a description of each.

In the first prototype, we narrowed the scope to a system that can lip-synch a specified audio stream by manipulating a video or image of a person. Similarly, we chose Faceswap [93], which allows us to swap two people's faces, an open source deepfake software, that is based on two autoencoders with a shared encoder in which the encoder learns the common features of the source and the target faces, while the two decoders learn to generate the source and target faces. The method requires a full video and many images of the face to be transferred and each new face requires a separate neural network. In the second approach, the child avatars were developed using the open source project Unity Multipurpose Avatar (UMA) (https://github.com/umasteeringgroup/UMA, accessed on 24 May 2022) system. We customized the characters by merging meshes and textures, and the audios were synchronized with the avatar. Using the Unity game engine asset SALSA LipSync Suite (https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442, accessed on 24 May 2022), we generated eye, head, and mouth movements in sync with audio.

After reading a number of publications [94–96], we turned to audio-to-video translation methods using GANs for realistic avatar synthesis which were popularized when used to generate a fake video of Obama (https://www.youtube.com/watch?v=cQ54GDm1eL0, accessed on 24 May 2022) [63,97]. Following an examination of the two approaches of ObamaNet [97] and the method proposed by Suwajanakorn et al. [63] in our previous work [98], we experimented with the ICface [74], a lightweight model for face animators that is driven by human interpretable control signals. This method belongs to video-driven talking-head generation (a.k.a., face reenactment) because it uses another video as input to regulate expression, pose, mouth, eye, and eyebrows movements.

We next used two contemporary state-of-the-art procedures to obtain more results after collecting the necessary understanding regarding talking-head methods, including: PC-AVS [75] and MakeItTalk [65]. The same general approach was adopted for all methods: an audio stream generated by IBM Watson TTS was given to the network as input to synthesize lip-synced videos.

MakeItTalk was trained on the VoxCeleb2 [99] dataset, which contains video segments from a variety of speakers and can synthesis expressive talking-head videos with an audio stream and a portrait face image as the only inputs. It also generalizes well to unseen facial images. MakeItTalk presents a self-attention based LSTM able to disentangle content and style in audio and lead to the generation of a speaker-aware talking-head. PC-AVS takes a single facial image as an input and generates a talking-head whose poses are controlled by the pose on another source video. The method implicitly devises a posture code that is free of mouth shape or identification, then modularizes audio-visual representations into spaces for audio content, head movement, and identity without relying on any intermediate information such as landmarks and 3D face shapes.

We conducted a user study to assess the two promising Unity game engine and talking-head techniques. The findings are briefly addressed in the next section.

4. Results

Eventually, users learning to conduct expert investigative interviews will use the talking child avatars directly. Therefore, subjective evaluation of humans is an essential step. In this section, the extracted results are described for each language, auditory, emotional, and visual components.

4.1. Language

The first interactive study involved expert users interacting with a six-years-old sexually abused child to validate our first prototype as a proof of concept for the proposed training system for investigative interviewer training. The core of the system used RASA as the dialogue model. Results of this user study showed that it was well received by the CPS workers. 72% opined that it could help them obtain knowledge and skills for communicating with abused children and 81% stated that it could improve their self-efficacy. The current chatbot was developed specifically to practice an investigative interviewing methodology following best-practice guidelines and inspired by the research on interview methodology conducted by researchers at the US National Institute of Child Health and Human Development (NICHD) [11]. Although the dialogue model can answer questions about the child's life that might not be relevant to the alleged incident it is still not capable of having a generic conversation with the child. It is hard to model small talk with RASA, as it leads to many intent definitions that negatively affect the performance of the intent classification model. We believe we can deal with the shortcomings of RASA chatbot using GPT-3 to keep the story coherent while also being able to answer generic small talk. RASA provided more control over how we wanted the conversation to proceed, but it is not a scalable solution.

4.2. Auditory

The user study had synthetic voices generated by IBM Watson TTS and recorded voices of natural speech, each of them for two different genders, as test conditions. This section investigated whether computer-generated voices can be as realistic as human voices and the appropriateness of the voices associated with each character. Figure 3 shows the results of the user study that compared the voices generated by a computer and by a human. A factorial ANOVA compared animated and GAN-generated avatars, where each had both synthetic and natural voices. The factorial ANOVA shows no significant main effect of type of voices ($F(1, 35) = 1.39, p = 0.24$), meaning that the computer generated voices were rated equivalent to human voices and there was no significant difference between them.

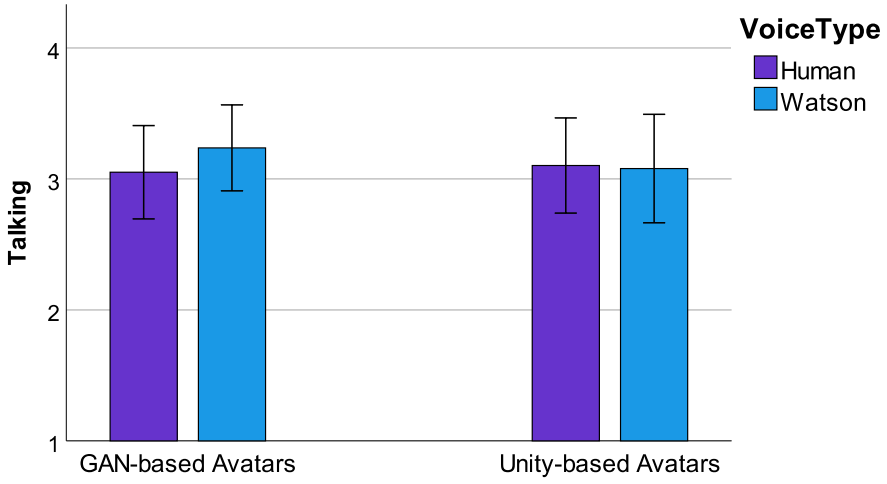


Figure 3. A comparison between natural and synthetic voices in animated unity-based and GAN-based avatars.

4.3. Emotions

As explained in Section 3, we started with both GPT-3 and BART predicting one sentence at a time. Comparing this with the results of a user study yielded promising results. However, these results do not always make sense in a larger context. Take the sentence "I was on the playground" as an example. This sentence is innocent and suggests enjoyment when viewed on its own, but it may no longer have that positive connotation in the context of abuse. The single-sentence prediction misses the importance of context. The opposite occurs when predicting the emotion based on the whole story. At some point, the model is not able to pick up subtle emotional changes since there is too much context. Thus, the single-sentence prediction is less complicated but not accurate since it misses important context. By contrast, the whole story has too much context, resulting in results that are not accurate either. The sliding window appeared to offer the perfect solution to this problem. Using a window consisting of between 3 and 7 sentences was expected to be ideal. However, the threshold did not work as expected. The hypothesis was that it would help spot sudden substantial emotional changes in the story, for example, if a child started crying. However, a restart of the window often occurred with sentences classified as enjoyment. Consequently, the model predictions got worse due to the implementation of the threshold.

It is challenging for humans to reach a consensus when classifying data excerpts with seven emotions as the options. There was unanimous agreement about only one of the twenty-one questions, that one being a single sentence excerpt. Although there was a clear winner for eleven questions, the Intraclass Correlation Coefficient (ICC) based on average measures was low on seven dimensions, scoring only 0.537 on average. The ICC describes how strongly elements in the same group resemble each other. An ICC between 0.40 and 0.59 is a moderate score. However, it is preferable to have an ICC score of at least 0.75 [100]. When we reduced the number of emotions to four, the participants were better at agreeing on the best fitting class. Reducing the number of categories resulted in more clear winners than before. The ICC also increased from 0.537 to 0.788: it was significantly higher with four dimensions than with seven.

A comparison between GPT-3 and BART was based on human annotations from a survey of 21 participants. 52.4% of the participants identified as female and 47.6% as male. Most of the participants were between the age of 26 and 35, but people between 18 and 25, 36 and 45, and 56 and 65 years also participated. The survey included eight single-sentence excerpts and 12 excerpts for each window size. The window sizes that we used were 3, 5, and 7.

The results showed that BART could predict more single sentences correctly, whereas GPT-3 was better at predicting sentences in context. One of the sentences in the user study was "We watched a movie and then we got some ice cream and then we went to bed.". The human consensus, the GPT-3 model, and BART, classified this as enjoyment. It becomes more difficult to classify sentences such as "It really hurt", which BART and the human participants classified as sadness, while GPT-3 considered it to be anger.

Sometimes both the models and the user study results were in agreement. This happened for the excerpt shown in Figure 4. It was unanimously considered to belong to the class fear.

However, there are also instances where both models made different classifications than human participants. This happened with the conversation shown in Figure 5. Both models only saw the responses that the child gave us and thus, classified this excerpt as enjoyment, whereas the participants saw the whole conversation and classified it as fear instead.

There is not always a reason why one or both models are wrong. Sometimes, it is a matter of interpretation. As mentioned above, GPT-3 is better when the context is involved, as was also the case in Figure 6, where both models predicted a different emotion based on the text given to them.

Interviewer	Child
Hm-hm.	And I really needed to go to the toilet.
Okay, and then what happened next?	I went to the toilet by myself because mum couldn't leave my baby brother by himself.
Ah-ha.	And then I went to the toilet by myself.
Okay, and tell me everything that happened from when you went to the toilet.	I went into the toilet and I didn't lock the door because I didn't want to get stuck in.
Hm-hm.	And then the bad man came inside.

Figure 4. Excerpt from the user study with window size 5 where both models are in agreement with the human opinion.

Interviewer	Child
Okay, so no one else was home. Where was everyone else?	Work.
At work, okay, and why was Mark there?	To play.
To play, okay. Okay, so you've told me quite a bit. Let me just doublecheck that I've got all of that right. So Mark was with you and you were playing Nintendo and you stood up and broke the controller. Then he got angry at you for that so you went to the garden and got a stick and then he hit you with it lots of times and then after that, he just left and he told you not to tell anyone.	Yes.

Figure 5. Excerpt from the user study with window size 3 for which both models were not in agreement with the human raters.

Interviewer	Child
Tell me more about the part where he put his hand in your undies.	He wrapped his hand around my doodle and then I said I didn't like it.
Hm-hm.	And then he laughed and took it out.
Okay, what happens next?	And then I went home.
Hm-hm, and was anyone else with you at uncle George?	No.
Hm-hm, and can you tell me more about the big boys game? Have you played it more than one time?	Yes.

Figure 6. Excerpt from the user study with window size 5 for which both GPT-3 and the human raters agreed that this should be classified as fear, while the BART model classified its as anger.

4.4. Visual

Here, we report on the findings and lessons learned using the tools introduced in Section 3.4. First, we investigated the deepfake models and why they are unsuitable for creating realistic avatars. Then, using a subjective study, we evaluated the realism of the talking-head techniques and avatars created by the game engine as the solution to be integrated into the child avatar system.

Faceswap [93] is the leading free and open source multi-platform Deepfakes software that uses computer graphics and visualization techniques. Faceswap is used to swap the faces of two people. According to the results of published research [98], when a man's visage is used on a child, Faceswap does not produce a realistic outcome. Although two people who are very similar in appearance could create a better result, we cannot rely on close likeness for our use case. As a result, Faceswap appears not to be suitable for creating a realistic-looking avatar.

ICface [74] can transfer the expressions from a driving video to a source face image (as shown in Figure 7) video-driven talking-head (refer to Section 2.5). Although ICface can be applied to arbitrary face images, the results are not as visually plausible. Moreover, manipulating expressions in images using human interpretable control signals such as head movement angles and Action Units (AUs) is tedious, time consuming, and causes repetitive facial expressions. Therefore, this method is not efficient enough to employ to generate a realistic avatar.

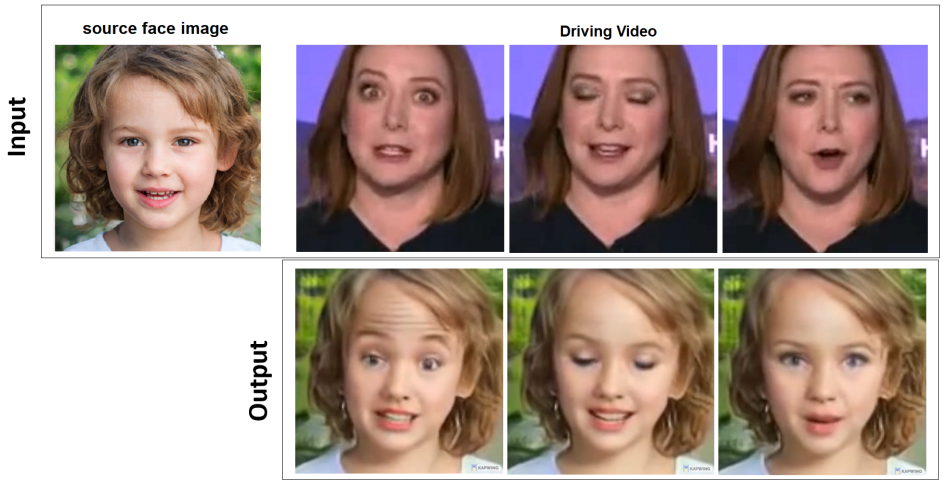


Figure 7. Given an arbitrary source face image generated by styleGAN [59,101] and a driving video, ICface [74] has generated the talking-head of a child.

We investigated the efficacy of the two approaches to talking-head generation using GANs and Unity game engine-based generation by conducting a crowdsourced user study. In both cases the audio stream used as input was obtained from IBM Watson (Introduced in Section 3.2) and recorded natural human voices. We generated 18 avatars using the two open source talking-head methods MakeItTalk [65] and PC-AVS [75], which can generalize well on any desired facial images. Here, StyleGAN [59,101] was utilized to create some child portrait images. Input facial images and several frames of video images thus generated are shown Figure 8. We also developed ten animated avatars using the Unity game engine. Finally, 28 ten-second video clips of the created avatars were provided using methods based on either GANs or the Unity game engine.

The study was conducted through crowdsourcing and Microworker (https://www.microworkers.com/, accessed on 24 May 2022) was used as a recruiting environment, with users referred to a questionnaire tool hosted on a separate server that contained the videos. To ensure the validity and reliability of the collected data, only high-performing crowdworkers who performed the best on the test were invited to participate.

Overall, 39 crowdworkers provided valid results in this study, including 10 women, 27 men, and two of other genders. The crowd workers aged 19 to 54 years (Median = 28, Mean = 29.58, SD = 8.36) were geographically evenly distributed between Europe, Asia and America (North and South).

Participants were asked to evaluate each video with regard to three statements: 'How realistic was the talking avatar?', 'How realistic was the avatar's appearance?', 'How was your overall experience with the avatar?' and 'How were the audio and mouth/lips synced?' on a scale between 1 to 5 (5-strongly agree, 4-agree, 3-neither agree nor disagree, 2-disagree, 1-strongly disagree). More details about the dataset and the study design are given in [18].

We first examine the difference between the MakeItTalk [65] and PC-AVS [75] as audio-driven models from different perspectives. From our point of view, PC-AVS can generate correct lip-synching, but it can not maintain the identity and resolution of the input facial image, and it also generates talking-heads with no blink motions. MakeItTalk, on the other

hand, can perform blink actions and maintain the resolution of the input facial image by depending on exact facial landmarks. Still, it does not accurately synchronize the shape of the mouth. Additionally, natural head movement is one of the most important aspects of effective communication in the real world. MakeitTalk can generate subtle head motions, unlike most existing approaches that generate face animation with a fixed head pose. These motions involve many repeating behaviors and movement patterns that only minimally swing around. In contrast, PC-AVS takes a brief target video and an audio stream as input to transfer the head position, resulting in more varied rhythmic head motions based on the input video clip. However, because this strategy depends on another short video clip, it is not particularly useful in our situation.

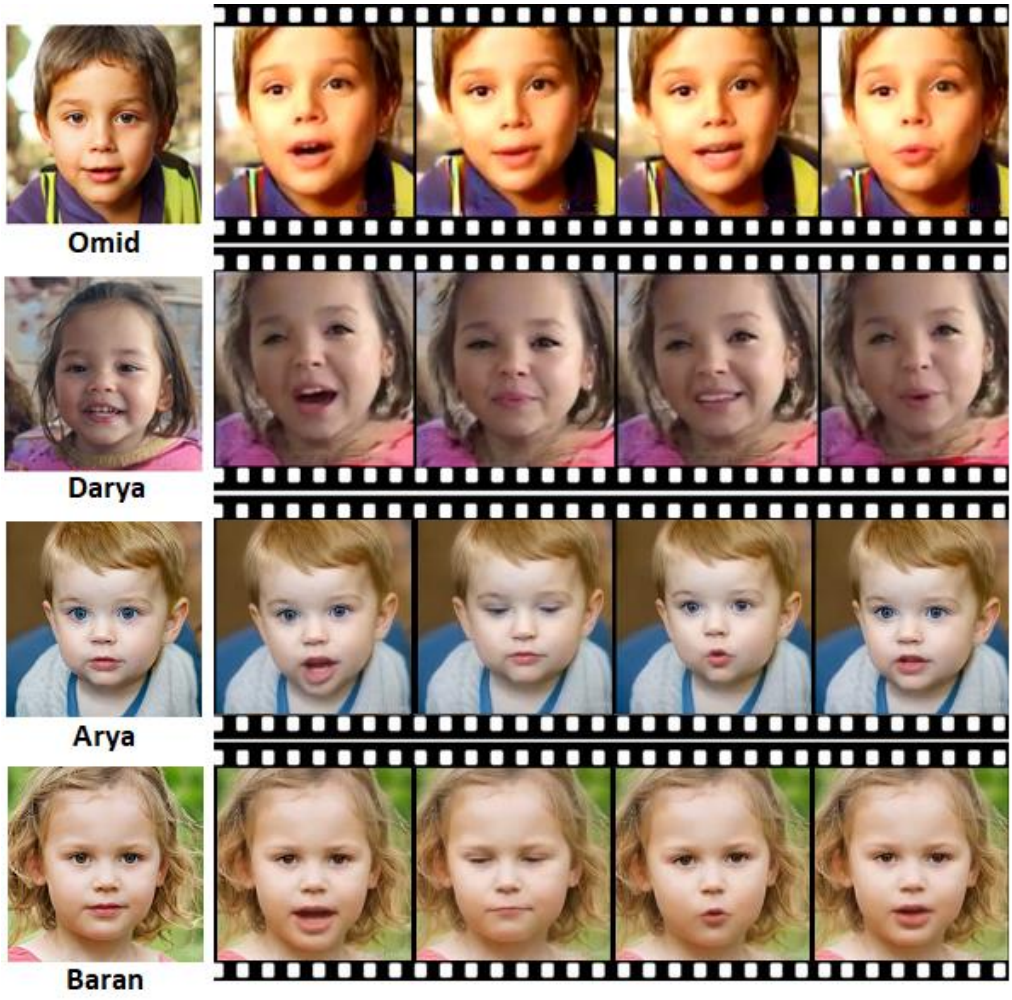


Figure 8. Illustration of a talking-head video generated using two methods, PCAVS [75] and MakeItTalk [65]. The input is an image generated using styleGAN and an audio generated using IBM Watson. The first two rows: PCAVS and the second two rows: MakeItTalk.

The bar-plots in Figure 9 depict the user ratings showing no significant difference between the MakeItTalk and PC-AVS models for four different characters. We expected MakeItTalk to appear more realistic from the user's perspective, but it did not. Four repeated measure factorial ANOVAs were used to compare the main effects of models and characters and their interactions on all four quality dimensions. According to all four quality criteria talking, appearance, experience and lips sync, there was no significant difference between the two models (MakeItTalk and PC-AVS), and the characters (the selected facial image). For the realism of *Talking*, there were no significant main effects of the model (p-value (p) more than <0.05) ($p < 0.6$), or characters ($p < 0.5$) and no significant interaction between model and characters ($p < 0.06$). For *Appearance*, no main effects were

observed for the model ($p < 0.09$), or characters ($p < 0.21$) and there was no significant interaction between model and characters ($p < 0.09$). For *Overall experience*, there was no main effect of model ($p < 0.86$), no main effect of characters ($p < 0.29$) and no interaction between model and characters ($p < 0.38$). And for lip-sync, there was no main effect of model ($p < 0.52$), no main effect of characters ($p < 0.17$) and no interaction between model and characters ($p < 0.07$).

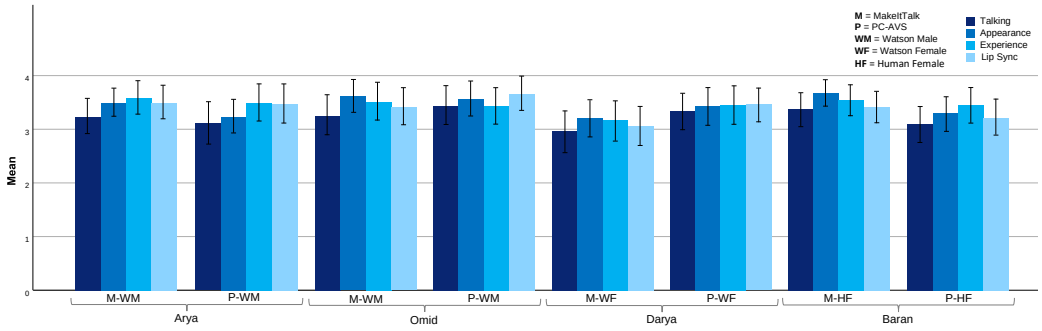


Figure 9. Bar-plot (95% confidence interval) for comparison of MakeItTalk [65] and PC-AVS [75].

Interaction with the virtual avatar necessitates techniques to generate high-fidelity talking-head frames while maintaining human observers' trust and empathy. According to Uncanny Valley theory [102,103], if an avatar is humanoid but imperfect, its non-human features instill confusion and even fear in the audience. Thus, we developed animated facial avatars using the Unity game engine to see whether the uncanny valley exists in the context of a child interview.

To investigate the avatars created by the game engine, a one-way repeated measure analysis of variance (ANOVA) for the mean of overall experience and appearance was performed. The lip-sync quality was not explored in the data analyses because all of the animated characters used the same technique for lip-sync generation utilizing the Salsa component in the Unity game engine. The animated avatars' quality ratings are shown in Figure 10.

Next, the best-rated avatars created with GAN were compared with the animated avatars created by the Unity game engine. The best avatars in GAN were Omid and Darya from model PC-AVS, which were selected based on how realistic their *Appearance* and *Talking* were rated. These two avatars are compared with Sarah and Liam's animated avatars created by the Unity game. Figure 10 shows the quality ratings for these four avatars. Sarah's character was rated slightly lower than the others, which might be due to weaker lip synchronization, but there was no statistically significant difference. A factorial ANOVA showed that although there was a general trend for more realism to be perceived in the videos created by GANs, there was no significant main effect of realism for *Talking* $F(1, 37) = 2.54, p = 0.12$, *Appearance* $F(1, 37) = 2.66, p = 0.11$ and *Overall experience* $F(1, 37) = 1.09, p = 0.30$. Overall, the results showed that although the avatars generated by GANs were slightly more realistic, they did not necessarily create a higher quality of experience and realism for the user.

In addition to the quality aspects, we asked participants an open question about their reasons for preferring one character over the other at the end of the questionnaire. Almost all of them agreed that the avatars generated by GANs were more realistic. Four participants stated that lip-synching in GAN-generated avatars was better than in animated avatars. One participant even stated, "The GAN-generated avatars could trick me into thinking that an actual person is talking." Three participants unequivocally stated that the talking faces appeared frightening, and one participant mentioned that something was off-putting and felt incorrect. This is because GAN-generated avatars are innately unlikable unless they are perfect and natural, as explainable by the uncanny valley theory.

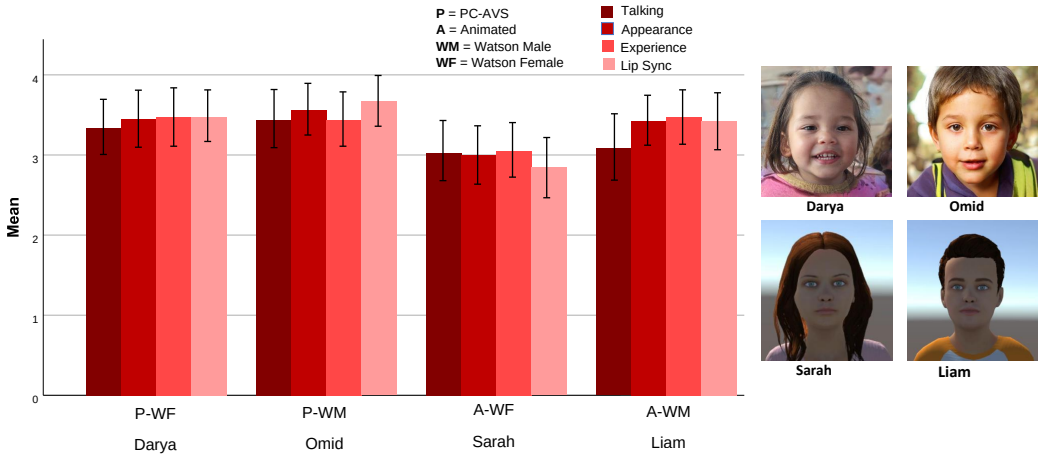


Figure 10. Bar-plot (95% confidence interval) to show results of the user study for the evaluation of two of the best female and male characters created for both the GAN-Based and game engine-based approaches.

5. Discussions and Future Work

Enhancing the disclosure of child abuse can facilitate earlier protection, prevention, and prosecution when children are victims of sexual, physical, and emotional abuse, thereby possibly having a huge impact on individuals' mental health and the future of society in general. The paper describes how an avatar-assisted mode of training might help CPS workers and police officers learn how to effectively interview sexually and physically abused children. Thanks to recent developments in AI, we proposed synthesizing a realistic digital child avatar, attempting to emulate an abused child. Our ongoing work focuses on integrating various components, including the language, auditory, emotional, and visual components of the avatar. The proposed system can dynamically respond to the questions and provide a higher level of realism during the training interviews and would be completely independent of human input. Therefore, unlike previous systems that were too rigid in their response generation, lacked generality, and required human input, our proposed system can generate dynamic responses without human operation. Such a system would be more dynamic and cost-effective because it lowers expenditures on human resources. This study discussed lessons and outcomes achieved by adopting various techniques. Moreover, three user studies were conducted to comprehensively evaluate a number of these methods.

The first user study showed that, although RASA provides an environment to manage the flow of dialogue within brief mock interviews, it is not easy to model small talk with RASA because it leads to many intent definitions, negatively affecting classification of the model's intent. We believe that by using GPT-3, we can address the shortcomings of the RASA chatbot, which will maintain the coherence of the story while also answering generic small talk. RASA gave us greater control over how the conversation flows, but it is not a robust solution as it has not been trained on open-domain data to sensibly answer the questions that are not in transcripts used during training.

There were also promising initial results concerning the emotional component. However, it is essential that more annotated data get analyzed and compared to the output of the existing models. We can incorporate the right model into our emotion pipeline. After that, we can start implementing the pipeline in both the auditory and visual components, following which users can evaluate what does and does not work. However, at this stage the biggest improvements would be made if there were annotated text data available in combination with the corresponding videos. It is also worth researching whether better results would be obtained if the interviewer's questions were also classified. Figure 5 shows this could be a positive influence. However, it may negatively affect other currently correctly classified examples, so it is worth investigating.

Another user study was conducted to explore the realism of the avatar's auditory and visual components. For the audio component, the avatars were tested using synthetic voices generated by computer and natural human voices. The results showed that synthetic voices generated by IBM Watson TTS can be as realistic as natural voices. Based on those results, we decided to use the synthetic voices for the avatar.

For the visual component, facial manipulations using the FaceSwap and video-driven talking-head methods were ineffective. We therefore turned to the Audio-driven talking-head using GANs and Unity game engine and then conducted a user study to compare them. Contrary to our predictions, GAN-generated and Unity game engine avatars were rated similarly. The best generated avatars using the two approaches were not rated differently. The results are even more startling because more than half of the participants claimed to prefer conversing with avatars developed using the Unity game engine, even though the avatar's appearance and overall talking were less realistic. A few participants noted that the GAN-generated avatars appeared to be frightening in response to an open question. This could be evidence of the presence of an uncanny valley [102,103]): if an avatar is humanoid but flawed, the viewer may experience weird feelings of uneasiness and even revulsion as a result of the avatar's non-human traits. We aim to improve the realism of GAN-generated avatars in the future so that fewer artifacts are produced. Furthermore, we will experiment with combining the two methodologies, with the talking-head formed using an animated facial image as input. Further, future works will explore the human influencing factors and their impact on the user experience. The human influencing factors include demographic characteristics such as age, gender and socio-economic background, physical and mental constitution, and the user's emotional state [104], all of which can have an impact on the user experience.

In addition, we intend to consider the versatile and novel LSP model [67] in future work. LSP introduces a real-time system that generates talking heads with natural head movement, lip-sync, and eye blinks using only audio signals at more than 30 frames per second. Furthermore, the system requires training on a several-minute video of the desired character. We can provide this video from a real human generated using GANs or an existing talking animation. Thus, LSP appears practical and ideal for developing a digital child avatar due to its high-fidelity video creation and real-time capabilities.

6. Conclusions

In forensic interviews, obtaining valuable information from a neglected or abused child requires the interviewer to have good interviewing skills. Our ongoing work aims to synthesize an interactive virtual child avatar in real-time, mimicking a child so that law enforcement and CPS workers can efficiently be trained to learn and improve these skills. This paper presented the interactive child avatar system, mimicking an abused child. The system is designed using different artificial intelligence-based technologies such as the avatar's language, auditory, emotional, and visual components. Furthermore, using three subjective studies, various system components were investigated. The results of the first user study showed that participants believed the child avatar system could effectively improve conversational skills and was well received by the CPS workers. Furthermore, the second study examined emotion extraction using different models and discussed how emotion could be integrated with auditory and visual components. The third user study first showed that the synthetic voices generated by computers could be as realistic as natural voices, and then investigated the realism of various techniques for generating child avatars. It was shown that the GAN-based and game engine-based avatars could create the most realistic avatars. Using the insights provided by these three user studies, the refined and improved architecture of the child avatar system was presented, and the integration and interaction of various components were discussed.

Author Contributions: Investigation, P.S., S.Z.H., M.L., S.S.S. and I.R.; Methodology, P.S., S.Z.H., M.L. and S.S.S.; Project administration, S.S.S., G.A.B., P.H. and M.A.R.; Writing—original draft, P.S., S.Z.H. and M.L.; Writing—review & editing, P.S., S.Z.H., M.L., S.S.S., I.R., R.K.R., M.S.J., V.T., S.A.H.,

M.P., M.E.L., G.A.B., P.H. and M.A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research is sponsored by the Research Council of Norway, project number #314690 ("Interview training of child-welfare and law-enforcement professionals interviewing maltreated children supported via artificial avatars").

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sethi, D.; Bellis, M.; Hughes, K.; Gilbert, R.; Mitis, F.; Galea, G. *European Report on Preventing Child Maltreatment*; World Health Organization, Regional Office for Europe: Geneva, Switzerland, 2013.
2. Widom, C.S. Longterm consequences of child maltreatment. In *Handbook of Child Maltreatment*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 225–247.
3. World Health Organization. *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks*; World Health Organization: Geneva, Switzerland, 2009.
4. Dixon, L.; Perkins, D.F.; Hamilton-Giachritsis, C.; Craig, L.A. *The Wiley Handbook of What Works in Child Maltreatment: An Evidence-Based Approach to Assessment and Intervention in Child Protection*; John Wiley & Sons: Hoboken, NJ, USA, 2017.
5. Brown, D.; Lamb, M. Forks in the road, routes chosen, and journeys that beckon: A selective review of scholarship on childrenss testimony. *Appl. Cogn. Psychol.* **2019**, *33*, 480–488. [CrossRef]
6. Lamb, M.E.; La Rooy, D.J.; Malloy, L.C.; Katz, C. *Children's Testimony: A Handbook of Psychological Research and Forensic Practice*; John Wiley & Sons: Hoboken, NJ, USA, 2011; Volume 53.
7. Adams, J.A.; Farst, K.; Kellogg, N.D. Interpretation of Medical Findings in Suspected Child Sexual Abuse: An Update for 2018. *J. Pediatr. Adolesc. Gynecol.* **2018**, *31*, 225–231. [CrossRef] [PubMed]
8. Newlin, C.; Steele, L.C.; Chamberlin, A.; Anderson, J.; Kenniston, J.; Russell, A.; Stewart, H.; Vaughan-Eden, V. *Child Forensic Interviewing: Best Practices*; US Department of Justice, Office of Justice Programs, Office of Juvenile: Washington, DC, USA, 2015.
9. Lamb, M.E.; Brown, D.A.; Hershkowitz, I.; Orbach, Y.; Esplin, P.W. *Tell Me What Happened: Questioning Children about Abuse*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
10. Lamb, M.E.; Orbach, Y.; Hershkowitz, I.; Esplin, P.W.; Horowitz, D. A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child Abus. Negl.* **2007**, *31*, 1201–1231. [CrossRef] [PubMed]
11. Powell, M.B.; Brubacher, S.P. The origin, experimental basis, and application of the standard interview method: An information-gathering framework. *Aust. Psychol.* **2020**, *55*, 645–659. [CrossRef]
12. Lyon, T.D. Interviewing children. *Annu. Rev. Law Soc. Sci.* **2014**, *10*, 73–89. [CrossRef]
13. Powell, M.B.; Hughes-Scholes, C.H.; Smith, R.; Sharman, S.J. The relationship between investigative interviewing experience and open-ended question usage. *Police Pract. Res.* **2014**, *15*, 283–292. [CrossRef]
14. Lamb, M. Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink? *Am. Psychol.* **2016**, *71*, 710–718. [CrossRef] [PubMed]
15. Powell, M.B.; Guadagno, B.; Benson, M. Improving child investigative interviewer performance through computer-based learning activities. *Polic. Soc.* **2016**, *26*, 365–374. [CrossRef]
16. Seymour, M.; Riemer, K.; Kay, J. Actors, avatars and agents: Potentials and implications of natural face technology for the creation of realistic visual presence. *J. Assoc. Inf. Syst.* **2018**, *19*, 4. [CrossRef]
17. Hassan, S.Z.; Salehi, P.; Røed, R.K.; Halvorsen, P.; Baugerud, G.A.; Johnson, M.S.; Lison, P.; Riegler, M.; Lamb, M.E.; Griwodz, C.; et al. Towards an AI-Driven Talking Avatar in Virtual Reality for Investigative Interviews of Children. In Proceedings of the 2nd Edition of the Game Systems Workshop (GameSys '22), Athlone, Ireland, 14–17 June 2022.
18. Salehi, P.; Hassan, S.Z.; Sabet, S.S.; Baugerud, G.A.; Johnson, M.S.; Riegler, M.; Halvorsen, P. Is More Realistic Better? A Comparison of Game Engine and GAN-based Avatars for Investigative Interviews of Children. In Proceedings of the ICDAR Workshop, ACM ICMR 2022, Newark, NJ, USA, 27–30 June 2022.
19. Cederborg, A.C.; Orbach, Y.; Sternberg, K.J.; Lamb, M.E. Investigative interviews of child witnesses in Sweden. *Child Abus. Negl.* **2000**, *24*, 1355–1361. [CrossRef]
20. Baugerud, G.A.; Johnson, M.S.; Hansen, H.B.; Magnussen, S.; Lamb, M.E. Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015–2017). *Appl. Cogn. Psychol.* **2020**, *34*, 654–663. [CrossRef]
21. Korkman, J.; Santtila, P.; Sandnabba, N.K. Dynamics of verbal interaction between interviewer and child in interviews with alleged victims of child sexual abuse. *Scand. J. Psychol.* **2006**, *47*, 109–119. [CrossRef] [PubMed]

22. Lamb, M.E.; Orbach, Y.; Sternberg, K.J.; Aldridge, J.; Pearson, S.; Stewart, H.L.; Esplin, P.W.; Bowler, L. Use of a structured investigative protocol enhances the quality of investigative interviews with alleged victims of child sexual abuse in Britain. *Appl. Cogn. Psychol. Off. J. Soc. Appl. Res. Mem. Cogn.* **2009**, *23*, 449–467. [CrossRef]
23. Brubacher, S.P.; Shulman, E.P.; Bearman, M.J.; Powell, M.B. Teaching child investigative interviewing skills: Long-term retention requires cumulative training. *Psychol. Public Policy Law* **2021**, *28*, 123–136. [CrossRef]
24. Krause, N.; Pompedda, F.; Antfolk, J.; Zappala, A.; Santtila, P. The Effects of Feedback and Reflection on the Questioning Style of Untrained Interviewers in Simulated Child Sexual Abuse Interviews. *Appl. Cogn. Psychol.* **2017**, *31*, 187–198. [CrossRef]
25. Haginoya, S.; Yamamoto, S.; Pompedda, F.; Naka, M.; Antfolk, J.; Santtila, P. Online simulation training of child sexual abuse interviews with feedback improves interview quality in Japanese university students. *Front. Psychol.* **2020**, *11*, 998. [CrossRef]
26. Haginoya, S.; Yamamoto, S.; Santtila, P. The combination of feedback and modeling in online simulation training of child sexual abuse interviews improves interview quality in clinical psychologists. *Child Abus. Negl.* **2021**, *115*, 105013. [CrossRef]
27. Pompedda, F.; Zappalà, A.; Santtila, P. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychol. Crime Law* **2015**, *21*, 28–52. [CrossRef]
28. Mayer, J. D.; Salovey, P. What is emotional intelligence? In *Emotional Development and Emotional Intelligence: Educational Implications*; Basic Books: New York, NY, USA, 1997; pp. 3–33.
29. Joseph, D.L.; Newman, D.A. Emotional intelligence: An integrative meta-analysis and cascading model. *J. Appl. Psychol.* **2010**, *95*, 54–78. [CrossRef]
30. Hochschild, A.R. *The Managed Heart: Commercialization of Human Feeling*; University of California Press: Berkeley, CA, USA; London, UK, 2012.
31. Risan, P.; Binder, P.E.; Milne, R.J. Emotional Intelligence in Police Interviews—Approach, Training and the Usefulness of the Concept. *J. Forensic Psychol. Pract.* **2016**, *16*, 410–424. [CrossRef]
32. Albaek, A.U.; Kinn, L.G.; Milde, A.M. Walking Children Through a Minefield: How Professionals Experience Exploring Adverse Childhood Experiences. *Qual. Health Res.* **2018**, *28*, 231–244. [CrossRef] [PubMed]
33. Ekman, P.; Friesen, W.V. A new pan-cultural facial expression of emotion. *Motiv. Emot.* **1986**, *10*, 159–168. [CrossRef]
34. Ekman, P.; Heider, K.G. The universality of a contempt expression: A replication. *Motiv. Emot.* **1988**, *12*, 303–308. [CrossRef]
35. Matsumoto, D. More evidence for the universality of a contempt expression. *Motiv. Emot.* **1992**, *16*, 363–368. [CrossRef]
36. Katz, L.F.; Hunter, E.C. Maternal meta-emotion philosophy and adolescent depressive symptomatology. *Soc. Dev.* **2007**, *16*, 343–360. [CrossRef]
37. Karni-Visel, Y.; Hershkowitz, I.; Lamb, M.E.; Blasbalg, U. Nonverbal Emotions While Disclosing Child Abuse: The Role of Interviewer Support. *Child Maltreatment* **2021**, *29*, 10775595211063497. [CrossRef]
38. Kerig, P.K.; Bennett, D.C.; Chaplo, S.D.; Modrowski, C.A.; McGee, A.B. Numbing of Positive, Negative, and General Emotions: Associations with Trauma Exposure, Posttraumatic Stress, and Depressive Symptoms Among Justice-Involved Youth: Numbing of Positive, Negative, or General Emotions. *J. Trauma. Stress* **2016**, *29*, 111–119. [CrossRef]
39. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *CIM* **2018**, *13*, 55–75.
40. Vinyals, O.; Le, Q. A neural conversational model. *arXiv* **2015**, arXiv:1506.05869.
41. Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
42. Zhou, L.; Gao, J.; Li, D.; Shum, H.Y. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *arXiv* **2019**, arXiv:1812.08989.
43. Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.P.; Gao, J.; Dolan, B. A persona-based neural conversation model. *arXiv* **2016**, arXiv:1603.06155
44. Tachibana, H.; Uenoyama, K.; Aihara, S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4784–4788.
45. Amberkar, A.; Awasarmol, P.; Deshmukh, G.; Dave, P. Speech recognition using recurrent neural networks. In Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 1–3 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
46. Xiong, W.; Droppo, J.; Huang, X.; Seide, F.; Seltzer, M.L.; Stolcke, A.; Yu, D.; Zweig, G. Toward Human Parity in Conversational Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2017**, *25*, 2410–2423. [CrossRef]
47. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779.
48. Sadjadi, O.; Greenberg, C.; Singer, E.; Mason, L.; Reynolds, D. *NIST 2021 Speaker Recognition Evaluation Plan*; NIST: Gaithersburg, MD, USA, 2021.
49. Zhang, Y.; Qin, J.; Park, D.S.; Han, W.; Chiu, C.C.; Pang, R.; Le, Q.V.; Wu, Y. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv* **2020**, arXiv:2010.10504.
50. Chung, Y.A.; Zhang, Y.; Han, W.; Chiu, C.C.; Qin, J.; Pang, R.; Wu, Y. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *arXiv* **2021**, arXiv:2108.06209.

51. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower Provost, E.; Kim, S.; Chang, J.; Lee, S.; Narayanan, S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
52. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323. [CrossRef]
53. Fayek, H.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Netw.* **2017**, *92*, 60–68. [CrossRef]
54. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4779–4783.
55. Prenger, R.; Valle, R.; Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–19 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3617–3621.
56. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Oord, A.V.D.; Dieleman, S.; Kavukcuoglu, K. Efficient Neural Audio Synthesis. *arXiv* **2018**, arXiv:1802.08435.
57. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
58. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
59. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8110–8119.
60. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8789–8797.
61. Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; Li, H. Protecting World Leaders Against Deep Fakes. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; Volume 1.
62. Perov, I.; Gao, D.; Chervoniy, N.; Liu, K.; Marangonda, S.; Umé, C.; Dpfks, M.; Facenheim, C.S.; RP, L.; Jiang, J.; et al. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv* **2020**, arXiv:2005.05535.
63. Suwajanakorn, S.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph. ToG* **2017**, *36*, 1–13. [CrossRef]
64. Chen, L.; Maddox, R.K.; Duan, Z.; Xu, C. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7832–7841.
65. Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; Li, D. Makelttalk: Speaker-aware talking-head animation. *ACM Trans. Graph. TOG* **2020**, *39*, 1–15. [CrossRef]
66. Meshry, M.; Suri, S.; Davis, L.S.; Shrivastava, A. Learned Spatial Representations for Few-shot Talking-Head Synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13829–13838.
67. Lu, Y.; Chai, J.; Cao, X. Live speech portraits: Real-time photorealistic talking-head animation. *ACM Trans. Graph. TOG* **2021**, *40*, 1–17. [CrossRef]
68. Yi, R.; Ye, Z.; Zhang, J.; Bao, H.; Liu, Y.J. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv* **2020**, arXiv:2002.10137.
69. Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; Nießner, M. Neural voice puppetry: Audio-driven facial reenactment. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 716–731.
70. Chen, L.; Cui, G.; Liu, C.; Li, Z.; Kou, Z.; Xu, Y.; Xu, C. Talking-head generation with rhythmic head motion. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 35–51.
71. Richard, A.; Lea, C.; Ma, S.; Gall, J.; De la Torre, F.; Sheikh, Y. Audio-and gaze-driven facial animation of codec avatars. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 11 August 2021; pp. 41–50.
72. Song, L.; Wu, W.; Qian, C.; He, R.; Loy, C.C. Everybody's talkin': Let me talk as you want. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 585–598. [CrossRef]
73. Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2387–2395.
74. Tripathy, S.; Kannala, J.; Rahtu, E. Icface: Interpretable and controllable face reenactment using gans. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2020; pp. 3385–3394.

75. Zhou, H.; Sun, Y.; Wu, W.; Loy, C.C.; Wang, X.; Liu, Z. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4176–4186.
76. Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; Wang, X. Talking face generation by adversarially disentangled audio-visual representation. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9299–9306.
77. Wiles, O.; Koepke, A.; Zisserman, A. X2face: A network for controlling face generation using images, audio, and pose codes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 670–686.
78. Ha, S.; Kersner, M.; Kim, B.; Seo, S.; Kim, D. Marionette: Few-shot face reenactment preserving identity of unseen targets. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10893–10900.
79. Bansal, A.; Ma, S.; Ramanan, D.; Sheikh, Y. Recycle-gan: Unsupervised video retargeting. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 119–135.
80. Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; Theobalt, C. Deep video portraits. *ACM Trans. Graph. TOG* **2018**, *37*, 1–14. [CrossRef]
81. Pompedda, F.; Antfolk, J.; Zappalà, A..; Santtila, P. A combination of outcome and process feedback enhances performance in simulations of child sexual abuse interviews using avatars. *Front. Psychol.* **2017**, *8*, 1474. [CrossRef]
82. Pompedda, F.; Palu, A.; Kask, K.; Schiff, K.; Soveri, A.; Antfolk, J.; Santtila, P. Transfer of simulated interview training effects into interviews with children exposed to a mock event. *Nordic Psychol.* **2020**, *73*, 43–67. [CrossRef]
83. Pompedda, F.; Zhang, Y.; Haginoya, S.; Santtila, P. A Mega-Analysis of the Effects of Feedback on the Quality of Simulated Child Sexual Abuse Interviews with Avatars. *J. Police Crim. Psychol.* **2022**, 1–14. [CrossRef]
84. Dalli, K.C. Technological Acceptance of an Avatar Based Interview Training Application: The Development and Technological Acceptance Study of the AvBIT Application, Master's Thesis. Linnaeus University, Växjö, Sweden, 2021.
85. Johansson, D. Design and Evaluation of an Avatar-Mediated System for Child Interview Training. Master's Thesis, Line University, Kanagawa, Japan, 2015.
86. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M. and Ghemawat, S. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.
87. Bunk, T.; Varshneya, D.; Vlasov, V.; Nichol, A. Diet: Lightweight language understanding for dialogue systems. *arXiv* **2020**, arXiv:2004.09936.
88. ITU-T Recommendation P.809. *Subjective Evaluation Methods for Gaming Quality*; International Telecommunication Union: Geneva, Switzerland, 2018.
89. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**, arXiv:1910.13461. https://doi.org/10.48550/ARXIV.1910.13461.
90. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, 1 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 353–355. [CrossRef]
91. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
92. Jack, R.E.; Garrod, O.G.B.; Schyns, P.G. Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time. *Curr. Biol.* **2014** *24*, 187–192. [CrossRef]
93. Deepfakes. github. 2019. Available online: https://github.com/deepfakes/faceswap (accessed on 20 May 2022).
94. Sha, T.; Zhang, W.; Shen, T.; Li, Z.; Mei, T. Deep Person Generation: A Survey from the Perspective of Face, Pose and Cloth Synthesis. *arXiv* **2021**, arxiv:2109.02081.
95. Zhu, H.; Luo, M.; Wang, R.; Zheng, A.; He, R. Deep audio-visual learning: A survey. *Int. J. Autom. Comput.* **2021**, *18*, 351–376. [CrossRef]
96. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
97. Kumar, R.; Sotelo, J.; Kumar, K.; de Brébisson, A.; Bengio, Y. ObamaNet: Photo-realistic lip-sync from text. *arXiv* **2018**, arXiv:1801.01442.
98. Baugerud, G.A.; Johnson, M.S.; Klingenberg Røed, R.; Lamb, M.E.; Powell, M.; Thambawita, V.; Hicks, S.A.; Salehi, P.; Hassan, S.Z.; Halvorsen, P.; et al. Multimodal virtual avatars for investigative interviews with children. In Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval, Taipei, Taiwan, 21 August 2021; pp. 2–8.
99. Chung, J.S.; Nagrani, A.; Zisserman, A. Voxceleb2: Deep speaker recognition. *arXiv* **2018**, arXiv:1806.05622.
100. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef] [PubMed]
101. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–19 June 2019; pp. 4401–4410.
102. Mori, M.; MacDorman, K.F.; Kageki, N. The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.* **2012**, *19*, 98–100. [CrossRef]

103. MacDorman, K.F.; Green, R.D.; Ho, C.C.; Koch, C.T. Too real for comfort? Uncanny responses to computer generated faces. *Comput. Hum. Behav.* **2009**, *25*, 695–710. [CrossRef]
104. Brunnström, K.; Beker, S.A.; De Moor, K.; Dooms, A.; Egger, S.; Garcia, M.N.; Hossfeld, T.; Jumisko-Pyykkö, S.; Keimel, C.; Larabi, M.C.; et al. *Qualinet white Paper on Definitions of Quality of Experience*; HAL: Bengaluru, India, 2013.

Article V: Enhancing Investigative Interview Training Using a Child Avatar System: A Comparative Study of Interactive Environments

Authors: Syed Zohaib Hassan, Saeed Shafiee Sabet, Michael Alexander Riegler, Gunn Astrid Baugerud, Hayley Ko, *Pegah Salehi*, Ragnhild Klingenberg Røed, Miriam Johnson, Pål Halvorsen

Candidate's contributions: The candidate contributed to the conceptualization and methodology design of the comparative study evaluating four interactive training environments. She assisted in developing the experimental protocol, including selection of evaluation metrics (QoE, presence, flow, and realism), and participated in conducting the user study. The candidate also contributed to developing the Unity-based 2D and VR environments for the child avatar system. Additionally, she helped review and edit the manuscript.

Published in: Scientific Reports, vol. 13, pp. 20403 (2023)

DOI: 10.1038/s41598-023-47368-2

Contributed to: This work contributed to the **RO-3**.

scientific reports

Check for updates

OPEN

Enhancing investigative interview training using a child avatar system: a comparative study of interactive environments

Syed Zohaib Hassan[1,6]✉, Saeed Shafiee Sabet[1,6], Michael Alexander Riegler[1,2,6], Gunn Astrid Baugerud[3,6], Hayley Ko[3,5,6], Pegah Salehi[1,6], Ragnhild Klingenberg Røed[3,6], Miriam Johnson[4,6] & Pål Halvorsen[1,2,6]

The impact of investigative interviews by police and Child Protective Services (CPS) on abused children can be profound, making effective training vital. Quality in these interviews often falls short and current training programs are insufficient in enabling adherence to best practice. We present a system for simulating an interactive environment with alleged abuse victims using a child avatar. The purpose of the system is to improve the quality of investigative interviewing by providing a realistic and engaging training experience for police and CPS personnel. We conducted a user study to assess the efficacy of four interactive platforms: VR, 2D desktop, audio, and text chat. CPS workers and child welfare students rated the quality of experience (QoE), realism, responsiveness, immersion, and flow. We also evaluated perceived learning impact, engagement in learning, self-efficacy, and alignment with best practice guidelines. Our findings indicate VR as superior in four out of five quality aspects, with 66% participants favoring it for immersive, realistic training. Quality of questions posed is crucial to these interviews. Distinguishing between appropriate and inappropriate questions, we achieved 87% balanced accuracy in providing effective feedback using our question classification model. Furthermore, CPS professionals demonstrated superior interview quality compared to non-professionals, independent of the platform.

Child abuse is a global concern that is known to have adverse effects on children's development and mental and physical health. Research has shown that children who are victims of abuse experience cognitive impairments, and both mental and physical health issues that can affect them throughout their entire lives[1]. In addition, meta-analytic studies have estimated that 22.6% of children experience physical abuse and 11.8% are subjected to sexual abuse before the age of eighteen[2,3], i.e., 8% of boys and 19% of girls face child sexual abuse (CSA)[4]. However, Andrews et al.[5] argue that CSA prevalence actually ranges from 2-62% for sexually abused children during childhood, with the reported prevalence varying due to different methods of data collection and assessing CSA, meaning that the incidence of CSA may be higher.

Less than 15% of reported CSA cases have physical evidence, with only 9% of cases having medical evidence[6]. Furthermore, studies have shown that in 70% of the CSA cases children are the only witnesses of the criminal incident[7,8]. Thus, investigative interviews of children are critical for the prosecution of these cases. Research has further shown that children are reliable witnesses when interviewed properly and in line with best-practice recommendations[9–12], i.e., developed by researchers and professionals to establish rapport and increase narrative details with the child[13]. These guidelines promote communication with the child through open-ended questions (e.g., "Tell me what happened."), active listening, and support[14,15]. To encourage children to express themselves in their own words, encourage elaborate and coherent information, and prevent any contamination of their original memory, interviewers should use open-ended questions[16]. However, multiple international studies have found

[1]Department of Holistic Systems, SimulaMet, 0167 Oslo, Norway. [2]Faculty of Technology, Art and Design, OsloMet, 0167 Oslo, Norway. [3]Faculty of Social Sciences, OsloMet, 0167 Oslo, Norway. [4]Faculty of Health Sciences, OsloMet, 0167 Oslo, Norway. [5]Oral Health Centre of Expertise in Eastern Norway, 0369 Oslo, Norway. [6]These authors contributed equally: Syed Zohaib Hassan, Saeed Shafiee Sabet, Michael Alexander Riegler, Gunn Astrid Baugerud, Hayley Ko, Pegah Salehi, Ragnhild Klingenberg Røed, Miriam Johnson, and Pål Halvorsen. ✉email: syed@simula.no

that these guidelines are rarely followed and that the quality of investigative interviews is quite poor[17–20]. The need to enhance the quality of investigative interviewing is evident and widely acknowledged in academic discourse. While some training programs have shown promise, they have not been effective in consistently improving interviewer behavior[21–23]. Recent studies, including one conducted in Norway over a 10-year period, suggest that despite advancements in training and innovation, there has been no significant improvement in interview quality[24]. To address this issue, it may be useful to design an interactive training program that can effectively train professionals in investigative interviewing.

This is precisely where the domain of computer science, specifically artificial intelligence (AI), can offer invaluable contributions. To the best of our knowledge, there are currently two child avatar training systems with the primary objective of enhancing the quality of interviews. Linnæus University and AvBIT Labs in Sweden developed an interview training system that employs prerecorded audio and video responses of an abused child. A human operator selects an appropriate video response for the user, which is displayed through Wirecast software controls on the Skype interface[25,26].

Pomedda et al.[27] presented Empowering Interviewer Training (EIT), an investigative interview training platform that uses a rule-based algorithm to select a pre-defined response based on user input. The user is presented with pre-recorded videos of children displaying different emotions, selected by a human operator. They conducted multiple studies with EIT system and investigated the impact of feedback on training effectiveness using their system and conducted multiple studies to analyze its learning effects[28–31]. Their findings suggest that incorporating feedback enhances learning effects, but the system has limited response generation capabilities.

While these systems have contributed to enhancing investigative interviewing abilities, they still possess certain limitations. Their response generation is inflexible or requires human intervention in their operation, leading to greater operational costs and the possibility of human errors. We have developed a training system for investigative interviews that incorporates lifelike avatars capable of dynamically responding to various queries, providing an immersive experience through virtual reality (VR) technology[32]. Our approach utilizes advanced natural language processing (NLP), which enables computers to understand and generate human language, and vision technologies, which allow machines to interpret and generate visual content, in combination to achieve this goal to create a virtual talking avatar. These technologies synergize to create a virtual talking avatar that simulates an abused child[33].

The utility of VR extends beyond mere immersion. Increasingly, research indicates that VR can be a powerful tool for learning and education. Compared to 2D, audio, chatbot, or other interactive environments, VR can create a more engaging environment. Pausch et al.[34] along with Bailenson et al[35]. discuss the use of immersive VR in educational settings and its impact on engagement and motivation. Both sets of authors concur that the immersive qualities of VR can elevate engagement and motivation, potentially leading to increased practice time. However, Pausch et al.[34] also caution that over-immersion can sidetrack users from learning goals, diminishing the effectiveness of learning. Similarly, Bailenson et al.[35] stress the significance of aligning VR experiences with learning objectives and balancing immersion and learning outcomes to optimize educational benefits. Numerous studies have been conducted to assess the impact of VR on learning in comparison to alternative methods. Lai et al.[36] examined a vocabulary learning game in both VR and 2D, concluding that VR resulted in a significantly higher mean Quality of Experience (QoE) than the 2D desktop for vocabulary learning. Krokos et al.[37] obtained similar outcomes for a memory learning task. In contrast, Madden et al.[38] conducted a comparative study on learning moon phases through VR, PC, and hands-on training. They reported no significant difference between the three environments.

Studies have also explored the effectiveness of VR interview training platforms in improving interview skills and vocational outcomes, primarily focusing on individuals with conditions like autism spectrum disorder (ASD), schizophrenia, and substance use disorders. These studies consistently reveal positive results, including enhanced interview skills, boosted self-confidence, and a higher rate of job offers for participants who underwent VR interview training[39–41]. However, it is important to note that the efficacy of VR interview training may vary depending on the specific population and context in which it is used. Using VR for investigative interviews with artificial children is a complex issue that requires investigation. Adaptability and acceptability of VR are also key issues, which are influenced by various social groups based on gender or profession. Raaen et al.[42] examine the level of acceptance and utilization of VR technology among the Norwegian population. The study reveals that merely 20% of the population has had any prior experience with VR, and only 0.6% of the individuals use it once per week. The feasibility of using VR for investigative interviews with artificial children is a complex issue that requires further investigation, especially in the context of the QoE for CPS workers.

In our earlier work[32], a pilot QoE study was conducted with child protective services (CPS) professionals, who are experts in interviewing abused children, where they interacted with a virtual child avatar in VR. QoE measures the degree of delight or annoyance of the user of an application or service[43]. The results showed that the perceived realism, encompassing aspects such as appearance, speech, and lip-sync, along with a sense of presence, feeling fully immersed and engaged in an environment[44,45], were crucial factors for the users. To further investigate these aspects of realism, another user study was conducted[46,47]. Realism study's findings reveal the presence of an uncanny valley, wherein participants favored interacting with animated avatars over highly realistic avatars generated by GAN technology[48]. Additionally, the research demonstrated that employing natural voices for virtual avatars did not enhance realism or improve user experiences when compared to computer-generated voices. Participants in the pilot study also expressed the importance of note-taking during long-duration interviews, and it remains a challenge to develop a VR environment where users can take notes.

Furthermore, previous research did not focus on domain experts, and as demonstrated by Sabet et al.[49], domain experts have different perceptions and expectations of a child avatar system than non-experts. Domain experts have specialized knowledge, training, and experience in dealing with real-world situations involving children, which may shape their expectations and responses when interacting with virtual child avatars in a

sensitive context. For instance, experts may have a more nuanced understanding of the non-verbal cues and behavioral patterns of children, which can influence their assessment of the realism and effectiveness of virtual child avatars for interviews. It is essential to conduct further research to explore the feasibility and potential benefits and risks of using VR for investigative interviews with artificial children. This motivated our comparative study to assess the QoE of expert participants in various interactive environments. The aim of our research is to investigate the possibility and potential advantages and drawbacks of utilizing VR as a means of conducting investigative interviews with artificial children, in comparison to other interactive mediums like 2D, audio, and text. Our study involved participation from CPS workers and child welfare psychology students, referred to as non-CPS in our paper.

The first part of the study design involves participants engaging with a child avatar for 90 seconds in various environments and sharing their feedback on the experience. ITU-T, which stands for the International Telecommunication Union's Telecommunication Standardization Sector, provides recommendations through ITU-T P.809 for conducting subjective experiments aimed at assessing the QoE for gaming services[50]. According to recommendations, short interactive sessions are suitable for measuring overall QoE, but may not be sufficient for assessing flow and immersion for sessions lasting less than 10 to 12 minutes. We argue, however, that it could be adequate for comparing different stimuli within the same setup. In the second part of the study, we also ask participants to interact with the child avatar just in VR environment without any time constraints. The motivation for second part came from the positive feedback we received from the pilot study[32], where participants gave high ratings for quality metrics and QoE in the VR environment. Our aim was to gather more data on the VR experience in investigative interview training and investigate any disparities in short and long interactive sessions in VR. Furthermore, we sought to confirm our hypothesis that brief interactive sessions are sufficient for comparing stimuli in different environments. Both groups, CPS and non-CPS, took part in both the first and second parts of the study, and their data were collectively analyzed.

The system architecture uses the dialogue model described in[32] with a new persona of a child. In addition to VR in current system architecture, for this study we also added three additional interactive environments (2D desktop, audio and text). We focused on identifying the optimal interactive environment to enhance the learning outcomes of the investigative interview training system. We do not evaluate the dialogue model or the participants' ability to conduct investigative interviews during the user study. The findings of this study have implications for the training of police and CPS personnel in investigative interviewing. We aim to provide a more engaging and realistic training experience, which can lead to better learning outcomes and higher motivation for practice.

In investigative interview training, it is crucial to provide trainees with feedback that can help them improve their questioning techniques. Accurately distinguishing between recommended and not recommended questions, the training program can provide trainees with targeted feedback, enabling them to learn from their mistakes and improve their questioning techniques[51]. So apart from conducting the user study, we introduce and validate our classification model that can assess the quality of interviews by categorizing the questions as either open-ended or closed-ended. This model will be integrated into the system and provide feedback to the participants on how did they perform. In this paper, we focus on validating the performance of our classification model and conducting a comparative analysis of the interview quality between two groups: CPS and non-CPS. To the best our knowledge, Haginoya et al.'s model[52] is the only relevant work with which we can compare our classification model's performance. They present an automated question classification system for simulated CSA interviews using avatar. The system was found to work well in classifying interviewer questions, and automated interventions (feedback and modeling) were provided to improve interview quality. They managed to achieve an accuracy of 70% for their classifier. In their study, the intervention groups showed improvement in their second interview while the no intervention group did not. In their research, the authors utilize N-gram features in combination with an XGBoost classifier. N-gram refers to a grouping of N consecutive characters, words, or phrases within a given text. For instance, if we apply bigrams to the question "tell me everything" we extract three distinct combinations: "tell me," "me everything," and "everything about," each of which represents a pair of adjacent words. In this case, we assign a frequency value of "1" to each of these bigrams to indicate their occurrence within the text and use it as features for training machine learning model. XGBoost, an abbreviation for Extreme Gradient Boosting, is a powerful machine learning algorithm employed in both classification and regression tasks. It operates by aggregating the predictions of numerous weak classifiers to make decisions.) Their methodology is constrained by its limited contextual window around a given word. This approach, however, falls short in its ability to comprehensively understand and contextualize entire sentences.

While addressing the limitation in previous work our paper aimed to address following questions:

- How do different environments influence the users' QoE, presence, flow, realism, responsiveness, and learning experience?
- Does different environments affect the quality of the interview?
- Are 90 seconds enough for the comparison of different stimuli and to capture user experience for different system modalities?

In this paper, we build upon our work earlier work[53] where we provided preliminary findings for the first part of the comparative study that we conducted. Specifically, we delve deeper into the qualitative feedback received from the study participants to gain a better understanding of their experiences. Furthermore, we introduce a feedback classification model and analyze its performance to better understand its potential applications. The main contributions of our work presented in this paper are:

- The development of a new dialogue model with a persona resembling a child that was exposed to sexual abuse.
- Extending the prototype of the child avatar in VR to three more environments of 2D, audio, and text.
- Conducting a comparative study for the four different interactive environments and evaluating them based on multiple quality experiences which captures user experience.
- A binary classification model, that can learn to recognize patterns in the interview questions and categorize them as either closed-ended or open-ended to assess the quality of interviews

Results

In this section, we present the results of part 1 of our study in which we compared the user experience in four different interactive environments, namely VR, 2D, audio, and text, over a 90-second period. We analyzed the participants' subjective feedback about each environment, as well as the quality of questions asked in each environment. Furthermore, as part two of our study, we evaluated the participants' experience in the VR environment during an unrestricted session. Additionally, we assessed a classification model's ability to distinguish between open-ended and close-ended questions.

Study Part-1

We use five quality assessment metrics namely, QoE, realism, responsiveness, presence and flow to compare the user experience in each environment. We employ the Mean Opinion Score (MOS) to gauge the comprehensive quality assessment of a system. MOS is derived by averaging the individual ratings provided by participants for each quality metric. Figure 1 shows that VR has an higher average ratings for each quality metrics except responsiveness. We discussed the detailed results for each metric in this section.

Quality of experience

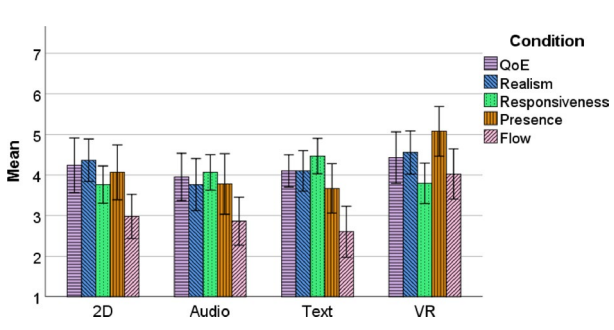
QoE is the degree of delight or annoyance of the user of an application or service, which results from the fulfillment of the users' expectations with respect to the utility and/or enjoyment of the service [43]. Figure 1a shows participants' overall QoE ratings were averaged across conditions. We observe that VR created a slightly higher overall QoE ($M = 4.50, SD = 1.40$), followed by 2D ($M = 4.25, SD = 1.52$), Text ($M = 4.10, SD = 0.85$) and Audio ($M = 3.95, SD = 1.32$). However, a one-way repeated measured Analysis of variance (ANOVA), with a significance level of $\alpha = 0.05$, finds no significant differences ($F(3.0, 17.0) = 1.115, p = 0.371$) between the MOS of overall QoE across different environments.

Presence

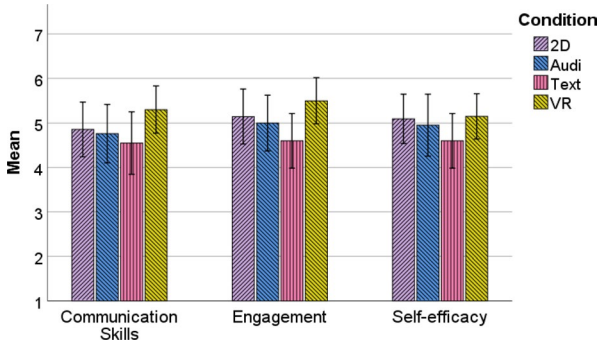
The quality of presence refers to a psychological state where an individual feels fully immersed and engaged in an environment, as if they are physically present in that space[44,45]. To evaluate any significant effects of the environments on the sense of presence, a one-way repeated measure ANOVA is used. The result of $F(3.0, 18.0) = 5.528, p = 0.007$ indicates a significant effect. The post-hoc tests followed by a Bonferroni correction indicate that, as we expected, VR ($M = 5.08, SD = 1.35$) creates a significantly higher presence than 2D ($M = 4.06, SD = 1.49, p = 0.048$) and Audio ($M = 3.78, SD = 1.64, p = 0.005$), but surprisingly not compared to Text ($M = 3.87, SD = 1.59, p = 0.070$). No other pairwise comparisons were found to be significant.

Realism

The realism score is the mean of three questions regarding realism in the *appearance*, *talking*, and *lip-sync* of the avatar, mentioned in methods section. Distinct questionnaires were used to assess the realism of VR and 2D in comparison to audio and text, primarily due to the visual element involved. The impact of different environments on the perceived realism of the avatar was investigated using a one-way repeated measure ANOVA. The result $F(3.0, 17.0) = 3.532, p = 0.037$ indicates a significant difference. The Text environment had a MOS of $(M) = 4.10, (SD) = 1.07$, Audio had a $M = 3.75, SD = 1.45$, 2D had a $M = 4.35, SD = 1.18$ and VR scored the



(a) Bar-plot (95% confidence interval) of means showing quality assessments for each environment.



(b) Bar-plot (95% confidence interval) of means showing learning experience assessments for each environment.

Figure 1. User experience.

highest ($M = 4.57, SD = 1.20$). The post-hoc test followed by a Bonferroni correction indicates that there is a significant difference for the VR-Audio pair ($p = 0.033$), but not for any other pairs.

Flow
Flow refers to a state of optimal experience in which a player is fully immersed and engaged in an interactive environment. It is a key indicator of a positive experience, as it indicates that the user is fully engaged and enjoying the experience[54]. For the flow, one-way repeated measure ANOVA indicates a significant impact of environments ($F(3.0, 17.0) = 9.020, p = 0.001$). Text had a MOS of $(M) = 2.60, (SD) = 1.33$, Audio had a $M = 2.78$, $SD = 1.28$, 2D had a $M = 2.88, SD = 1.13$ and VR the highest score of $M = 4.00, SD = 1.40$. The post-hoc tests followed by a Bonferroni correction indicate that VR was rated significantly higher than 2D ($p = 0.03$), Audio ($p = 0.001$), and Text ($p = 0.008$), but there were no significant differences between any other pairs.

Responsiveness
Responsiveness describes the temporal aspects of the feedback a player receives after performing an action[55]. The quicker the response time, the more responsive the environment is regarded. Moreover, the fluidity of the visual and audio feedback also contributes to the responsiveness, impacting the overall experience[56]. One-way repeated measure ANOVA of responsiveness for the avatar in each environment does not show a significant difference among them ($F(3.0, 17.0) = 2.893, p = 0.064$). The highest MOS was achieved by text ($(M) = 4.63, (SD) = 1.19$), followed by Audio ($M = 4.06, SD = 0.96$), VR ($M = 3.79, SD = 1.10$) and 2D ($M = 3.76, SD = 1.01$). This can be attributed to the STT and TTS synthesis in the environment with the audio element involved.

Learning experience
Apart from questions relating to user experience, we also incorporated three questions that pertained to the learning experience, which can be found in methods section. Figure 1b displays the average rating for these three questions. To investigate whether diverse environments could result in varying learning experiences, a one-way repeated measures ANOVA was conducted. Table 1 reveals that the different environments did not have any significant impact on communication and self-efficacy. Nevertheless, there was a noteworthy disparity between the perceived user engagement during learning across the various environments. A pairwise post-hoc comparison followed by a Bonferroni correction demonstrated that VR led to significantly higher engagement compared to Audio ($p = 0.022$) and Text ($p = 0.020$), but not when compared to 2D ($p = 0.85$). No significant differences were detected between any other pairs.

Sensitivity analysis and effect size comparisons
We further scrutinized our user study's data through sensitivity analysis to gauge the robustness of our findings. The sensitivity analysis underscored that, with our sample size of 21 and a desired power of 0.8, the smallest effect size (f) our study could reliably detect when comparing four different environments was approximately 0.81. Table 2 shows the observed effects for each metric, derived from the ANOVA test.

Comparing these observed effect sizes to our desired effect highlighted potential limitations of our study. The observed effect size for *QoE* and *self-efficay* was considerably below the desired effect. Metrics such as *presence*, *responsiveness*, and *engagement* displayed effect sizes that, while meaningful, were below the threshold of what our study was optimally powered to detect. In contrast, *flow* demonstrated stronger effects, suggesting that our sample size was more aligned with capturing such effects. An observed effect size that is less than the desired effect size does not invalidate your findings. It suggests that there is a higher risk of Type II errors which imply failing to detect an effect that is there. Conducting a future study with a larger sample size can reduce the risk of errors.

Study Part-2

The motivation behind this study was to assess the overall user experience of participants in the absence of time constraints. Additionally, the study aimed to validate our design decision that a 90-second duration is sufficient for users to engage with an interactive environment effectively and provide reliable feedback. Figure 2 shows that there is no significant difference in participants' ratings of the quality aspects between the 90-second VR session and the unconstrained session in VR. Table 3 shows the results of the repeated measure t-test between the long and short VR experience across different quality metrics, which included QoE, realism, responsiveness, presence, and flow.There were no significant differences observed in the assessment metrics, when comparing short stimulus tests to longer sessions for VR. This suggests that a 90-second short stimulus is sufficient for measuring these quality metrics for each interactive environment.

Metric	*VR*	*2D*	*F*-value	*p*-value
Communication skills	3	2.211	2.174	0.113
Engagement	3	1.99	4.422	0.015*
Self efficacy	3	2.311	2.174	0.179

Table 1. The results of one-way ANOVA tests comparing the means of learning experience aspects in different environments. The mean difference is significant at the .05 level.

Metric/source	SS_{effect}	SS_{error}	η_p^2	Cohen's f
QoE	3.3	55.7	0.056	0.059
Presence	22.619	93.881	0.194	0.241
Realism	7.383	39.283	0.158	0.188
Flow	24.213	52.038	0.318	0.465
Responsiveness	10.296	43.093	0.193	0.239
Comm.skills	6.789	6.211	0.108	0.121
Engagement	10.338	44.413	0.189	0.233
Self-efficacy	4.5	49.5	0.0833	0.0909

Table 2. Partial eta squared (η_p^2) and cohen's f score calculated using sum of squares of effects (SS_{effect}) and sum of squares of error (SS_{error}) from ANOVA test.

Metric	Environment	Paired Samples Statistics		Paired Samples Test	
		Mean	Std. Deviation	t	Sig. (2-tailed)
Overall Exp	Long	4.0952	1.57812	− 1.046	0.308
	Short	4.4286	1.39898		
Realism	Long	4.5079	1.25883	− 0.349	0.731
	Short	4.5556	1.16587		
Responsiveness	Long	3.6984	1.03228	− 0.573	0.573
	Short	3.7937	1.09786		
Presence	Long	4.9048	1.44585	− 0.757	0.458
	Short	5.0794	1.34951		
Flow	Long	3.9762	1.54496	− 0.197	0.846
	Short	4.0238	1.36452		
Comm.skills	Long	5.0000	1.48678	− 1.552	0.137
	Short	5.3000	1.12858		
Eng	Long	5.2857	1.34695	− 1.552	0.055
	Short	5.5714	1.12122		
Self-efficacy	Long	5.0000	1.54919	− 1.096	0.286
	Short	5.2381	1.13599		

Table 3. The results of one-way ANOVA tests comparing the means of learning experience aspects in different environments. The mean difference is significant at the .05 level.

Classification model

The assessment was conducted on 40 mock transcripts, containing 400 closed and 2,015 open-ended questions. Table 4a presents the performance evaluation of three classification models on test data. The models are random baseline classifier which generates predictions on assumption that classes are uniformly distributed, the model

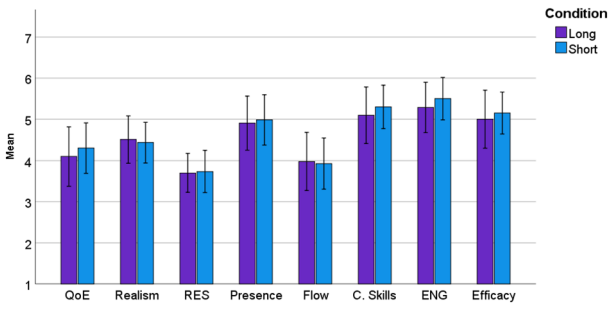
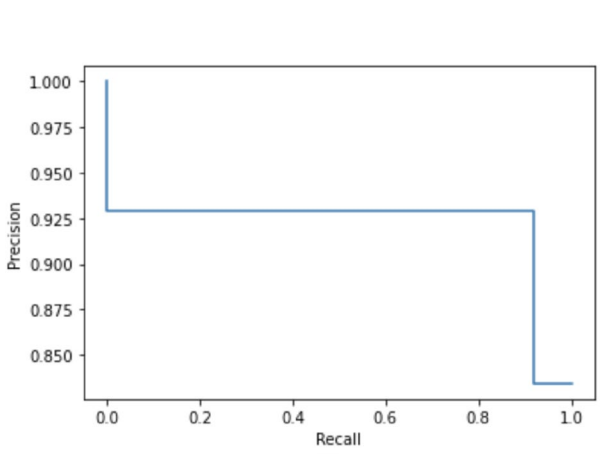


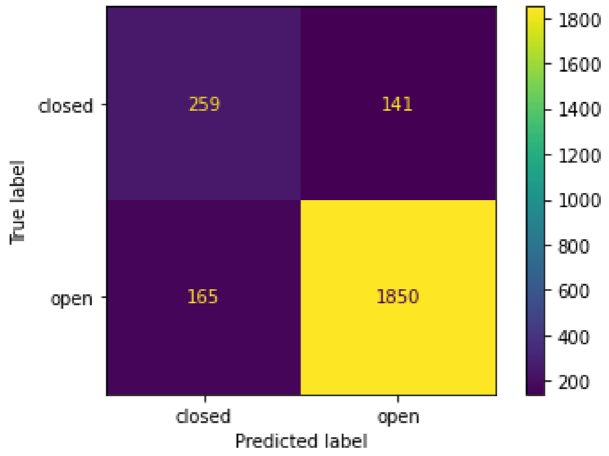
Figure 2. Bar-plot (95% confidence interval) of means showing quality assessments for VR session of 90-second VR termed short and without any time constrained termed as long. RES, C.Skills and ENG refers to responsiveness, communication skills and engagement respectivel.

(a) Our model performs better than the Haginoya's binary classifier in distinguishing between open-ended and close-ended questions			
Metrics	**Baseline**	**Haginoya[52]**	**Our model**
Macro F1-score	0.44	?	**0.78**
Weighted F1-score	0.57	?	**0.87**
Balanced accuracy	0.48	0.72 (Normal)	**0.87**
MCC	− 0.02	?	**0.55**
(b) Per-Class analysis of the model performance: Evaluating on test data			
Metrics	**Closed**	**Open**	
Precision	0.61	0.93	
Recall	0.65	0.92	
F1-score	0.63	0.92	

Table 4. Evaluation of the classification model: Test Data Classification Report. Best results are highlighted using bold fonts. "?" represents that information is not available.



(a) The Precision-Recall (PR) curve: Representation the trade-off between the precision and recall of the classification model.



(b) Confusion matrix: A visual representation of the performance of a classification model, showing the distribution of predicted and actual class labels.

Figure 3. Qualitative analysis of classification model performance.

by Haginoya et al.[52], and our model. The metrics used to evaluate the models include macro F1-score, weighted F1-score, balanced accuracy, and Matthews correlation coefficient (MCC).

From Table 4a, we can observe that the dummy model, which randomly predicts labels, has the lowest performance in all metrics. The Haginoya et al.'s model has better performance than the dummy model, but it is not clear from the reported data how it performs relative to the our proposed model since they only reported the accuracy score which we assume is not balanced accuracy and rest of the metrics are missing. Our model still outperforms the Haginoya's models in basic accuracy metric.

The performance of the proposed classification model was further evaluated for each class. Table 4b presents a per-class analysis of our model's performance on test data. The classes are closed and open questions, and the metrics used to evaluate the performance are precision, recall, and F1-score.

From table 4b, we can observe that the proposed model performs better for the open class than the closed class, as indicated by higher precision, recall, and F1-score values for the open class. The model's precision for the closed class is lower than that for the open class, which means that there are more false positive predictions for the closed class. The model's recall for the closed class is lower than that for the open class, which means that there are more false negative predictions for the closed class. Figure 3b illustrates that false negatives (FN) are more frequent for closed-ended questions than open-ended ones.

Overall, the proposed model's performance is good, but there is room for improvement in predicting the closed class. We are convinced that transforming it into a multi-label task will enhance generalization and reduce false predictions. According to Lamb et al.[51], the inter-coder reliability rate among experts for coding investigative interviews is approximately 85%. While our classification model outperforms other automated coding methodologies, as stated in Haginoya et al.'s research[52], our goal is to achieve expert-level precision.

Environment	Number of turns
Text	4.5
2D	7.82
VR	8.17
Audio	8.52
Long VR	29.70

Table 5. The average number of questions asked in the 90s interview for each environment.

The model's performance, as indicated by the metrics we used is quite promising, with a significant improvement over the Dummy model and potentially better performance than the Haginoya model. Despite the stepped appearance of the PRC, which suggests some limitations in the model's ability to distinguish between classes or a lack of diverse confidence score distributions, the overall metrics demonstrate good performance. The model has high precision, recall, and F1-scores for both classes, indicating a good balance between correctly identifying true positives and minimizing false positives. Additionally, balanced accuracy and MCC values further support the ability of the model in classifying both positive and negative classes. While the Precision-Recall curve's (PRC), in figure 3a, appearance suggests potential areas for improvement, the model's performance is quite effective based on the provided metrics and the fact that the dataset is biased.

Quality of interviews

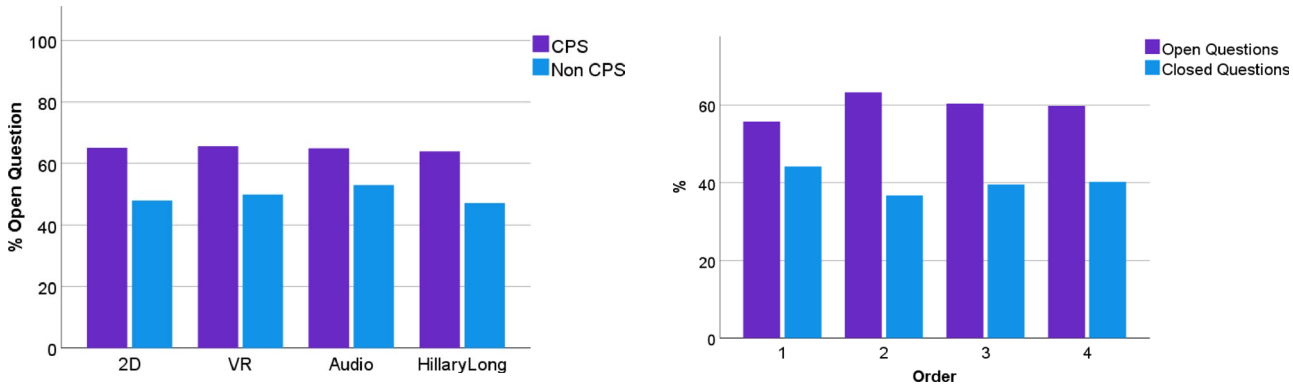
Table 5 presents the average number of questions asked in each environment. The average number of turns in the *text* environment was limited to 4.5 and Long VR was 29.7, due to being an outlier they are removed from the following data analyses, which focus on comparing the quality of the interviewers' questions between different environments. Figure 4a shows the percentage of open-ended questions asked between CPS and non-CPS participants in different environments. Overall, 64% of the questions asked by CPS participants were open-ended, while for non-CPS it was 48%. That shows, as expected, experts conduct higher-quality interviews. However, the number of open-ended questions asked could have been higher among CPS.

Furthermore, no difference between the environment and quality of the interview even when split by CPS versus non CPS, which shows the higher cognitive load created in environments with visual elements does not distract the users from conducting interviews. Figure 4b shows no significant impact of the order in which participants interact with environments over the quality of the interviews.

A chi-square test of independence was performed to evaluate the relationship between the training environment (2D, audio, VR short and VR long) and question type (open versus closed-ended questions). The relationship between these variables was not significant, $X^2(3) = 0.143, p = 0.986$. Thus, the training environment did not affect the quality of the interviews.

Qualitative results

A 4-item post-test questionnaire was utilised, as shown in Table 9, which was then followed by an open question asking participants to provide their reasons behind their selection. A top-down coding approach was utilised to analyse the post-test questionnaire, as shown in Table 6. Many of the participants selected VR as their preferred environment. 76% of the participants perceived VR as the most realistic environment, 66% preferred VR over other environments, and 61% agreed that VR created the best experience for them. Participants with previous VR experience showed a preference for VR – 80% rated VR favourable in the post-test questionnaire. However, also 75% of the participants with no VR experience preferred VR over the other environments, and 60% rated VR as the environment that they liked, preferred, and created the best experience.



(a) Quality of questions asked for each environment for two different groups.

(b) Distribution of open and closed questions asked

Figure 4. Quality of Interviews.

Items	VR	2D	Audio	Text
Q1-Which environment felt more realistic for you?	16	2	1	2
Q2-Which environment do you like the most?	14	3	1	3
Q3-Which environment do you prefer to use?	14	3	1	3
Q4-Which environment creates the best experience for you?	13	3	2	3

Table 6. The results of the user study to evaluate the number of votes in response to the questions.

In addition, a thematic analysis was conducted to examine the responses of the open question regarding the participants' experiences related to the different training environments. The following themes, *Sentiment towards environment* and *Realism* were explored.

Sentiment towards environment addressed the participants' feelings about using the training program in 2D, audio, text, and VR environments. Sentiments were categorized into positive or negative feelings and were examined by whether the participant worked within CPS or not. While discussing participants' feedback, we add extra text in '[ ]' to their responses, to adjust the tense and make participants' feedback grammatically correct. CPS participants only expressed negative sentiments regarding the audio environment. Participants noted that audio "felt strange as [they] mostly talk to children face-to-face", that "it [was] difficult to read body language" and that it "miss[es] the non-verbal communication". Regarding the text environment, CPS participants were generally positive, stating that it felt familiar and that it "can be a good way to practice, as there will be many times when you also work on chat". However, a non-CPS participant felt like "it [was] difficult to interact with the avatar in text". A few participants expressed that they preferred the 2D environment as it has both the visual and body language of the child, and they are used to working with PCs. CPS participants had positive experiences with 2D, with a participant expressing that "2D felt more familiar" and it was "even better than audio, because it's better to see the child [they were] talking to". The VR environment generated the most positive responses, with many participants expressing that the realism of the avatar made them feel like it was easier to talk to the child. However, one CPS participant stated that this environment "[did] not really reflect [their] experience with having these kinds of conversations...because it seem[ed] a bit scripted...like you have to have the right kind of question to get a reply".

Participants also reflected on the *Realism* of the interactive training program. Most participants stated that VR created the most realistic environment compared to the other environments. One participant stated that VR "made [them] feel like [they were] in the same room as Hillary" and that the realism made it "easier to talk to the child". However, one participant noted that despite the realism, they quote "didn't connect with Hillary". These views were similar in regards to the 2D environment, with participants agreeing that it felt realistic but not as realistic as VR. Regarding the audio and text environments, participants stated that these "environments did not feel as real" compared to the other environments. A few participants expressed that although the text is not as realistic as VR, it can be a better environment for training as they are used to this way of communication compared to VR. One participant noted that although "text only was not very realistic, [it] can work when training". Overall, the participants' descriptions gave the impression that realism is an important aspect of an interactive training program.

Overall, the participants responded quite positively towards their experiences in this study. One participant "believe[ed] this way to practice is a good option [for] training in the school setting". However, some participants had difficulties using the program as one participant stated that "it would have been much easier if everything was in Norwegian" and that "it took some time to formulate and write in English". Only one participant reported a negative overall experience, stating that it was "frustrating all together, the system did not give me any responses that [were] useful in VR, audio, and 2D".

Discussion

Our study aimed to compare the quality and learning aspects of experts in four different environments. Interactions lasted for 90 seconds in each environment. We observed that audio-based environments had more conversational turns than the text environments, as participants typed slower than they spoke. Despite a consistent dialogue model used in all environments, participants' experiences were influenced by the responses they received from the model. Environments with fewer responses limited participants' information, potentially affecting the richness and depth of interactions. As a consequence, the results of the study could have been impacted by this. Responses from the avatar were shaped by the questions asked, and because the questions themselves could vary, they also had the potential to influence participants' experiences across different environments. It is worth noting that although the conversation's quality may differ, we did not evaluate the dialogue model in this study.

Our findings indicate that VR was rated higher in four out of five quality assessments used for comparative analysis. However, there was no significant difference in the overall rated QoE for each environment. As anticipated, participants rated VR highest in terms of presence, indicating a greater sense of being part of the environment compared to the other options. VR also generated a significantly greater flow for users, creating an immersive environment that caused participants to lose track of time and forget their surroundings. These quality aspects may enhance users' willingness to engage, potentially leading to longer practice sessions with the child avatar system. When examining the quantitative analysis for realism, the only noteworthy variation was found between the VR-audio pair. The text environment was rated highest in responsiveness due to the lack of delay because of TTS and STT synthesis, and participants were in control of when to send their messages. In addition,

as shown in Figure 1b, all three metrics measuring the learning experience have a relatively higher mean for VR. Of particular significance is the higher user engagement for VR compared to audio and text environments. Mulqueeny et al.[57] highlight the importance of engagement in achieving long-term learning and outcomes. With significantly greater engagement and the highest average values across all learning experience metrics, VR may lead to better learning and knowledge transfer. The majority of participants preferred VR as their interactive environment for the interview training system, which was likely influenced by the fact that most users were using VR for the first time and found it fascinating.

Nevertheless, it is premature to assert that VR is the superior environment at this stage. While VR offers better levels of realism, presence, and flow, about 40% of the participants preferred other environments due to familiarity, accessibility, and regularity of use. Moreover, it is important to note the size of our study with the two different groups of our participants, CPS students and CPS professionals. We found both of these groups to be well-suited for participation in our study due to their shared knowledge, training, and experience in conducting interviews with children. Recruiting participants with relevant experience was a crucial and difficult job. When we looked closely at our results using sensitivity analysis, we found something crucial. While our findings are significant and provide valuable insights into the various metrics, the study might not have been optimally powered to detect certain nuanced effects, especially for specific metrics. This is why we aim to continue our research with more participants to get a clearer picture. Additionally, future studies should look more closely at each group. This could help us to see if different levels of expertise lead to different results. As we continue our research, these insights will be essential to help us improve our future studies even better.

The quality of the questions plays a crucial role in determining the effectiveness of an interview. Open-ended questions tend to elicit more informative responses, while closed-ended questions can yield potentially inaccurate answers. A system that accurately classifies questions can significantly impact the quality of interview training by providing effective feedback. Distinguishing between recommended and not recommended questions is essential for delivering high-quality feedback.

Achieving a balanced accuracy of 87% in classifying these two categories is a notable accomplishment, surpassing the state-of-the-art results of 72% reported by Haginoya et al.[52]. Although the model has a weighted average F1 score of 0.87, there is room for improvement in its ability to predict closed questions correctly. Reducing the original 15 question types to two fundamental categories may have affected the results when converting the multi-class task to a binary classification task. Future work could focus on refining the model's performance in classifying closed questions to further enhance the interview training experience. Furthermore, to accurately measure the learning impact of the system, a more extensive study is needed, where each participant conducts multiple interviews with a child. This approach would allow us to assess the quality of the interviews over time, similar to the pattern observed in the studies by Pomedda et al.[28,30].

When the classification model was applied to the conversational data from the study, the results revealed that the interviews conducted by both CPS professionals and non-CPS participants were of satisfactory quality using a rather high percentage of open-ended questions. However, the results also indicate that the interviewers rely on a high frequency of closed-ended questions, which is not good according to best practice standards. Closed-ended questions should be minimized as children often respond with little information. In addition, these questions tend to be suggestive, prone to response biases, and problematic to be used in investigative interviews[58]. Nevertheless, after examining the data, it was observed that the quality of interviews conducted by CPS professionals was comparatively better than those conducted by non-CPS participants. The participants seemed to perform consistently across different settings, indicating that the interview environment may not have a significant impact on their overall performance. Further research with more participants is necessary to determine the most suitable environment for interview training, particularly in terms of QoE and learning effect.

Future research should prioritize improving the various quality aspects of our child avatar system to provide a more effective learning platform for interview trainees. Our objective is to develop a flexible and efficient interview training system that is conducive to effective learning. Our future work will mainly focus on enhancing the dialogue model to generate more realistic responses by leveraging the advances made in large language models (LLMs), in conjunction with the use of more realistic-looking avatars generated using neural radiance fields (NeRF) or generative adversarial networks (GANs). In addition, we will also focus on improving the quality of feedback and exploring how feedback can be effectively utilized to enhance the overall learning experience.

Methods

System architecture

In this section, we discuss the overall architecture of the interactive child avatar system architecture for conducting a comparative study with CPS professionals and students shown in Figure 5. The system consists of three main components: (i) The language component, which comprises a dialogue model that we developed using the RASA[59] framework and a classification model that enables us to evaluate the quality of interviews; (ii) The speech synthesis component, which utilises IBM Watson services to provide speech-to-text (STT) and text-to-speech (TTS) synthesis[60]; and (iii) The front-end of the system, which offers four different interactive environments: a VR environment created using Unity and Oculus Quest2, a 2D environment and audio also developed on Unity. The text-based environment was created using the RasaX platform.

Language

The language component of the model is designed to analyze a child's conversation and produce realistic and suitable responses to an interviewer's questions. Furthermore, there is a classification model that can offer valuable feedback on the quality of interviews and assist in enhancing interview skills.

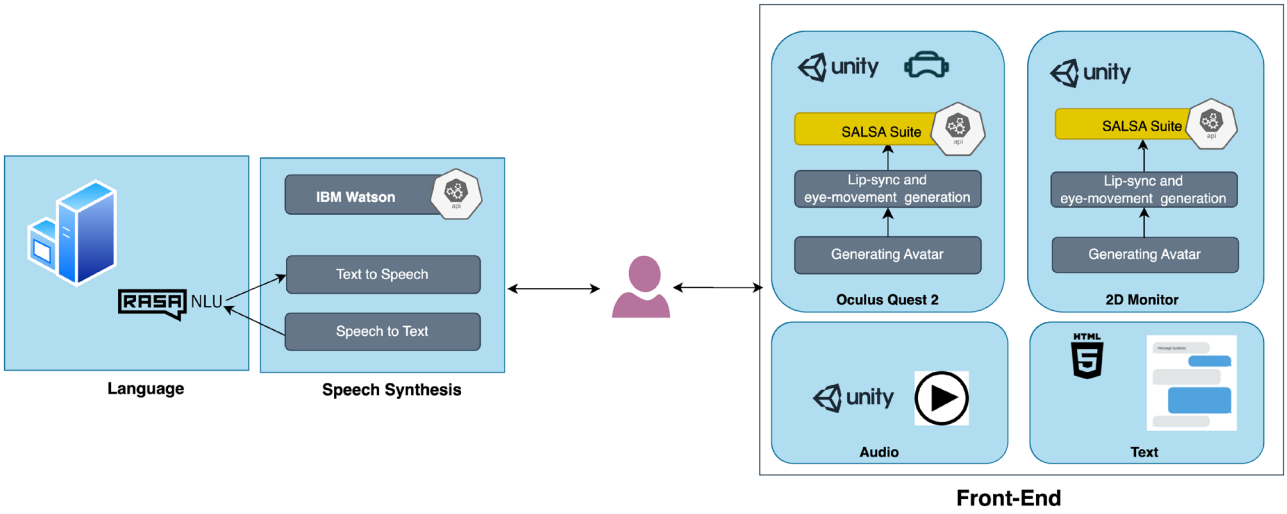


Figure 5. Architecture of the child avatar system for our comparative study with CPS professionals and students. (i) Language component incorporates a dialogue model that we developed using the RASA framework, which was trained on mock interviews, and a classification model for evaluating the quality of the interviews. (ii) Speech Synthesis component utilizes IBM Watson services for speech-to-text and text-to-speech synthesis. (iii) Front-end of the system provides four interactive environments: a VR environment created using Unity and Oculus Quest2, a 2D environment with audio also developed using Unity, and a text-based environment created using the RasaX platform.

Dataset

The dataset consists of transcripts from interviews conducted by professionals according to specific guidelines, as part of a training program at the Centre for Investigative Interviewing, which is situated in Griffith University, Australia[61]. The transcripts consist of simulated dialogues between professional actors portraying 5-7-year-old children and an interviewer who could be from CPS or law enforcement. These conversations aim to replicate an investigative interview with a child who may have been a victim of abuse.

Dialogue model

The current dialogue model is specifically designed to apply the dialogical communication method (DCM)[62], which is an investigative interviewing methodology currently employed in Norway. The Norwegian model is based on the research-based protocol developed by Michael Lamb and his colleagues at the National Institute of Child Health and Human Development (NICHD)[13]. The dialogue model was developed using an open-source framework of RASA which runs on tensorflow back-end. It has a modular structure with deep learning models, or rule-based techniques, which are employed to set up a complete dialogue model. Each module accepts the data in a predefined format, which requires quite a lot of pre-processing and data augmentation work for training. The Natural-language Understanding (NLU) module uses the Dual Intent and Entity Transformer (DIET)[63] architecture to model a joint task of intent and entity recognition. It processes the input text and predicts the best possible intent from the training data along with entities in that input. The rule module is used to model fallback strategies or control the behaviour of the dialogue model where necessary. According to the rules defined, a specific appropriate action is performed if a certain intent is recognized. These actions are defined in the action modules. The story module is used to control the evolution of the conversation and control the flow of conversation based on questions asked based on slot values defined in the NLU module or some specific intent. If a user asks a leading question, the child avatar answers with a vague or unproductive reply with a certain probability. This is designed to reinforce the need of asking open-ended questions and communicate negative feedback to the user.

Table 7 presents an excerpt of a participant's conversation during the study, with the participant's name removed for privacy purposes. The conversation shows that interviewer started with a bit of introduction and rapport building and then transition into substantive phase. Table 7 don't have the complete substantive as later child start talking about the alleged abuse.

Front-end

We used the system developed for our pilot study as the baseline[32]. The VR environment is running through an Oculus Quest 2 as a front-end. Three more interactive environments have been added to the current VR-based system. Thus, users can interact with the child's avatar in four different environments. Regardless of the environments (front-end), the system's back-end has the same dialogue model, which means that the responses generated by the virtual child remain consistent across different environments. The front-end of the child avatar system was developed in the Unity game engine in all environments except the text-only environment. RASA-X is used to develop a web interface for the text-only environment, where users have a textual conversation with the child. The 2D environment runs on a PC displayed using a 24-inch desktop monitor, and the audio-only environment was the same as the 2D with the difference that no visual elements were involved. The same child

Conversation Excerpt	
User	Hi Hillary
Bot	Hello.
User	What do you want to talk about today
Bot	About Janet's house.
User	Okay tell me about it
Bot	Janet's house is big and she has a bubble pool in the garden. I go there for play dates, but I don't want to go anymore.
User	Okay why don't you wanna go anymore.
Bot	Because I don't like her Pop.
User	Okay tell me why you don't like her pop.
Bot	He's Janet's grandpa, but I call him Pop.
User	Okay you call him pop. Tell me about pop.
Bot	He has grey hair and he lives with Janet in their house and he's a bit fat.
User	Okay he's a bit fat and what does he do that you don't like.
Bot	I just don't like him because he did a rude thing.
User	A rude thing what did you do?
Bot	I went to Janet's house, and it was really hot, so we were playing outside in the garden with the sprinklers and running through them and that was really fun.
User	...
Bot	...

Table 7. Example excerpt during the interaction between the user and a virtual child.



Figure 6. A snap-shot of the child avatar seen through the VR headset.

avatar character is used in both 2D and VR environments developed using the open-source project Unity Multipurpose Avatar (UMA)[64] which allows customization for combining meshes and textures of the characters. The unity asset Salsa Suit[65] was used to match the generated voices to the meshes and the texture of the avatar, to generate eye, head, and lip movements synced to the voices. In addition to facial movements created dynamically from the generated voice, a few pre-recorded animations were created by moving the meshes and the joints to create more natural movements and gestures in the avatar's hand and neck[66]. The snap-shot from a VR interface of the child avatar is shown in Figure 6.

Speech synthesis

We use the IBM Watson services for STT and TTS synthesis. Watson STT and TTS cloud service APIs play the role of a communication bridge between the dialogue model (back-end) and interactive environment (front-end) components. In our pilot study, participants gave us feedback that audio did not match the child's age[32]. IBM TTS services do not offer a child character option. However, users can utilize speech synthesis markup language (SSML) attributes and elements to manage text synthesis. To determine the most appropriate voice for a six-year-old girl, we conducted a small user study in collaboration with our colleagues in the social sciences, who are among authors of this paper. Nineteen distinct voices were developed, each featuring different female characters and prosody, which regulates pitch and speaking speed. The outcome of this study was the identification of the ideal audio representation of a six-year-old girl. The chosen voice's character, prosody rate, and prosody pitch were established as Allison, slow, and x-high, respectively. Except for the textual environment, the user interacts with the front-end system verbally. Questions asked by the users are synthesised by the IBM STT API and then processed by the dialogue model to generate appropriate responses. These responses are then

Subset	No of transcripts	No of turns	No of closed	No of closed
Training	58	2745	1191	1554
Test	40	2415	400	2015

Table 8. Training and test set class distribution.

synthesised by the IBM TTS API, which is played to the user through either speakers or VR headsets. Although multiple characters are available in IBM TTS API, all of them are adults. We adjusted the speed and pitch of one of the female characters to generate a child-like sound, as the pilot study showed that an adult voice over a child character is not taken well by the participants.

Classification model

To train the model, 58 transcripts were randomly selected and coded by a team of two psychologists and two research assistants. One of the authors, experienced with previous coding trained the other coders with best-practice guidelines and coding categories. Data was annotated with 14 different type of questions of open-ended (initial invitation, breadth prompt, depth prompt, descriptive, minimal encourager) and closed-ended (specific Yes-No, specific forced choice/option posing, leading, repeated question, leading, pressure, referring, indication, and visual). The coding process was conducted using the SIM and corresponding coding manual[67].The coders established inter-rater reliability on a separate set of transcripts until they reached on moderate agreement regarding the question types with Cohen's Kappa >0.70.

Training transcripts contained 2,745 turns of conversation. We focus on two higher-level classes: open-ended and closed questions. This was motivated by the the fact that we wanted to keep it simple for first model we train to do this task and also be able to compare our model performance in similar setting with Haginoya et al.'s model[52].

Table 8 shows the distribution of classes,the number of open-ended and closed questions,in the training and test sets used to train. In the training set, there were 1,191 closed questions and 1,554 open-ended questions. Using these annotations, we fine-tuned the GPT-3 davinci model to predict the question type in the conversations. To evaluate the performance of the classification model, we used an additional 40 transcripts, which were also annotated. In the test set, there were 400 closed questions and 2,015 open-ended questions.

When generating predictions from the fine-tuned GPT-3 model and evaluating a classification model that predicts the question types in conversations, we used a temperature of 0.0 and set the maximum tokens to one.

Experiment details

The study had two parts. The first part was designed according to the recommendations of ITU-T P.809[50] for short interactive test stimuli, which provides guidelines for conducting interactive subjective studies and assessing the QoE of a gaming system. We chose 90 seconds interactions to keep the study brief and for ease of recruiting participants who have domain knowledge. There are no strict recommendations for the duration of an ideal stimuli, it suggests that 30 seconds may not be sufficient to measure flow and immersion. For this, 10-15 minutes might be reasonable. We argue that 90 seconds could be enough for the comparison of different stimuli in the same setup and interactions can be enough to capture user experience for different system modalities. In order to validate this argument, part two was designed to evaluate the impact of the duration of the interactive session on the quality assessment metrics. Participants were asked to have an interactive session with a child avatar in VR without any time constraints.

In the first part of the study, we asked users to have a 90 second interaction in four different environments: text, audio, 2D, and VR. The order of these settings was randomized using the Latin squared design to avoid any potential sequencing bias. The primary objective of this research was to gauge and compare the QoE across various settings while minimizing any impact the sequence of settings may have on user experience. We also did not measure the quality of the dialogue model or the impact of the dialogues on QoE of each environment, for that purpose response generation from the dialogue model was kept constant across all settings. The second part of our study entailed participants interacting with a child avatar in VR with no time limit. It is referred as *long VR* in the result section to avoid ambiguity.

Ethical approval

- We affirm that our data collection process complies with the General Data Protection Regulation (GDPR). We have maintained data accuracy in accordance with Article 5(1)(d), ensured data integrity and confidentiality in accordance with Article 5(1)(f), and implemented robust security measures as required by Article 32. We have adhered to purpose limitation (Article 5.1 b), collecting data only for explicit, legitimate purposes, and not processing it for incompatible uses. We have practiced data minimization (Article 5.1 c), processing only adequate, relevant, and necessary data. Finally, we've enforced storage limitation (Article 5.1 e), not retaining data longer than necessary for its intended purpose.

- This study has received approval from the Norwegian Agency for Shared Services in Education and Research (SIKT) and the Data Protection Officer, OsloMet. No personal data was stored. SIKT Application number is 614272.
- We affirm our ethical commitment to ensure legality, fairness, and transparency (Article 5.1 a) by providing clear information and obtaining consent for data processing. Before beginning, participants were informed about the nature of their conversation with the virtual child and the potential adverse effects of prolonged VR use, such as dizziness or cyber-sickness. They signed a consent form allowing their feedback to be stored for research purposes, with the option to withdraw consent at any time in the future. The entire study took approximately 40 minutes for participants to complete. Participants were given a 200 Norwegian Kroner gift card for their voluntary participation.

Study participants were recruited in April 2022. The recruitment was specifically designed to target two groups: students specializing in child welfare psychology at Oslo Metropolitan University (Oslomet), and professionals currently employed in Child Protective Services (CPS). Information about the study was disseminated to these demographics to ensure a comprehensive representation of individuals with relevant knowledge and experience in child welfare. The study was executed during the time span from April to May of 2022, specifically during weeks 17 through 19. All participants were invited to attend at the SimulaMet research facilities, where they participated in the study. The data collected from the user study underwent an initial processing phase in July 2022, contributing to our preliminary work[53]. Subsequently, an additional round of data processing was performed in December 2022 and January 2023, purposed to facilitate the current extension of our work.

Storyline

To choose a storyline for the experiment, the interview transcripts were categorized into different personas. The persona selected for this study was determined by the abundance of transcripts and the extent of information that could be used to generate dialogues of sufficient length. The chosen persona was named Hillary and had the following background:

> *Hillary is a 6-year-old girl whose cognitive, emotional, social, and physical development appears to be typical for her age. She has several close friends at school, including one named Janet. Hillary lives with her biological parents and a younger brother. She has just started her first year of school and appears to have adapted well. Occasionally, she visits Janet at her home for play dates. Janet lives with her biological parents and her grandfather, whom the children refer to as "Pop." If Janet's parents are still at work, Pop usually takes care of the children. During these visits, Hillary and Janet usually play indoors and outdoors in the garden until Hillary's parents come to pick her up and take her home. A few days ago, while Hillary was in the car with her mother after a visit with Janet, she told her mother that Pop did something inappropriate to her and touched her private parts while they were in the pool in the garden. As a result, Hillary has come to the interview to discuss this incident.*

Demographic of the participants

In this research, a total of 21 participants took part, consisting of 20 females and one male. Their ages ranged from 20 to 59 years, with a mean age of 31.2 years and a standard deviation of 9.7. To begin with, they were requested to provide their ratings on their familiarity with interviewing children and VR. Approximately 60% of the participants had one or more years of experience as a CPS professional. Nonetheless, despite their expertise in CPS, most of them (75%) had no prior experience with VR. Each participant interacted with a child avatar in all environments, and none of them reported experiencing any form of motion sickness while utilizing the prototype.

Questionnaire

In addition to a question measuring the overall user QoE, we included three questions regarding *responsiveness* with GIQS[68], three questions regarding *realism* of the avatar inspired by the work conducted by Wilson et al.[69], two questions for measuring *flow* from GEQ[70], three questions for *presence* from IPQ[71], and three questions measuring the learning effects, engagement in learning, and self-efficacy as follows.

- *Communication skills* using this system aids me in acquiring knowledge and skills in interviewing maltreatment children.
- *Engagement* using this system would increase my engagement in the learning process.
- *Self-efficacy* using this system can improve my self-efficacy.

The questionnaire was answered on a 7-point Likert scale. The post-questionnaire is shown in Table 9 to measure the post-experience and user preference as the environment.

Post experience
Q1: Which environment felt more realistic for you?
Q2: Which environment do you like the most?
Q3: Which environment do you prefer to use?
Q4: Which environment creates the best experience for you?

Table 9. List of post-questionnaire items.

Data availability

All the subjective data collected during the user studies of this paper are available at https://docs.google.com/spreadsheets/d/e/2PACX-1vS8BzPnV7YbZxcQjcPjlN0EIYomA5E5qzmMzz3rNxUnqovoQiGEU0k1ZUPHwPsjSQ/pub?gid=2002569374 &single=true &output=csv. Although, the implementation details of the dialogue model on RASA and the UNITY 3D model of the visual avatar can be provided upon reasonable request. However, we regret to inform you that due to the ethical considerations and privacy concerns associated with the sensitive nature of the conversational data - specifically related to alleged sexual abuse - we cannot make the models and implementation of the child avatar trained on this data publicly available. Additionally, we would like to clarify that our classification model relies on a proprietary API key from OpenAI, which prevents us from sharing the model. Correspondence and requests for materials should be addressed to S.Z.H and S.S.S.

Received: 19 June 2023; Accepted: 13 November 2023
Published online: 21 November 2023

References

1. Kobulsky, J. M., Dubowitz, H. & Xu, Y. The global challenge of the neglect of children. *Child Abuse Neglect* **110**, 104296. https://doi.org/10.1016/j.chiabu.2019.104296 (2020).
2. Stoltenborgh, M., Van Ijzendoorn, M. H., Euser, E. M. & Bakermans-Kranenburg, M. J. A global perspective on child sexual abuse: Meta-analysis of prevalence around the world. *Child Maltreat.* **16**, 79–101. https://doi.org/10.1177/1077559511403920 (2011).
3. Stoltenborgh, M., Bakermans-Kranenburg, M. J., van IJzendoorn, M. H. & Alink, L. R. Cultural-geographical differences in the occurrence of child physical abuse? a meta-analysis of global prevalence. *Int. J. Psychol.* **48**, 81–94. https://doi.org/10.1080/00207594.2012.697165 (2013).
4. Pereda, N., Guilera, G., Forns, M. & Gómez-Benito, J. The prevalence of child sexual abuse in community and student samples: A meta-analysis. *Clin. Psychol. Rev.* **29**, 328–338. https://doi.org/10.1016/j.cpr.2009.02.007 (2009).
5. Andrews, G. *et al.* Child sexual abuse. *Comp. Quantif. Health Risks Global Reg. Burd. Dis. Attrib. Sel. Major Risk Factors* **2**, 1851–940 (2004).
6. Walsh, W. A., Jones, L. M., Cross, T. P. & Lippert, T. Prosecuting child sexual abuse: The importance of evidence type. *Crime Delinq.* **56**, 436–454. https://doi.org/10.1177/0011128708320484 (2010).
7. Elliott, D. M. & Briere, J. Forensic sexual abuse evaluations of older children: Disclosures and symptomatology. *Behav. Sci. Law* **12**, 261–277. https://doi.org/10.1002/bsl.2370120306 (1994).
8. Everson, M. D. & Sandoval, J. M. Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements. *Child Abuse Neglect* **35**, 287–298. https://doi.org/10.1016/j.chiabu.2011.01.001 (2011).
9. Brainerd, C. & Reyna, V. Reliability of children's testimony in the era of developmental reversals. *Dev. Rev.* **32**, 224–267. https://doi.org/10.1016/j.dr.2012.06.008 (2012).
10. Bala, N., Lee, J. & McNamara, E. Children as witnesses: Understanding their capacities, needs, and experiences. *J. Soc. Distress Homeless* **10**, 41–48. https://doi.org/10.1023/A:1009429602266 (2001).
11. Lyon, T. D., Malloy, L. C., Quas, J. A. & Talwar, V. A. Coaching, truth induction, and young maltreated childrens false allegations and false denials. *Child Dev.* **79**, 914–929. https://doi.org/10.1111/j.1467-8624.2008.01167.x (2008).
12. Gordon, B. N., Baker-Ward, L. & Ornstein, P. A. Children's testimony: A review of research on memory for past experiences. *Clin. Child. Fam. Psychol. Rev.* **4**, 157–181. https://doi.org/10.1023/A:1011333231621 (2001).
13. Lamb, M. E., Brown, D. A., Hershkowitz, I., Orbach, Y. & Esplin, P. W. *Tell me what happened: Questioning children about abuse* (Wiley, 2018).
14. Lamb, M. E., Orbach, Y., Hershkowitz, I., Esplin, P. W. & Horowitz, D. A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the nichd investigative interview protocol. *Child Abuse Neglect* **31**, 1201–1231 (2007).
15. Newlin, C. *et al.Child forensic interviewing: Best practices* (US Department of Justice, Office of Justice Programs, Office of Juvenile, 2015).
16. Brubacher, S. P., Benson, M. S., Powell, M. B., Goodman-Delahunty, J. & Westera, N. J. Chapter twenty-two: An overview of best practice investigative interviewing of child witnesses of sexual assault. In *Child Sexual Abuse*, (eds Bryce, I. & Petherick, W.) 445–466 (Academic Press, 2020). https://doi.org/10.1016/B978-0-12-819434-8.00022-2
17. Baugerud, G.-A., Johnson, M. S., Hansen, H. B., Magnussen, S. & Lamb, M. E. Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015–2017). *Appl. Cogn. Psychol.* **34**, 654–663. https://doi.org/10.1002/acp.3647 (2020).
18. Cederborg, A.-C., Orbach, Y., Sternberg, K. & Lamb, M. Investigative interviews of child witnesses in Sweden. *CAN* **24**, 1355–1361. https://doi.org/10.1016/S0145-2134(00)00183-6 (2000).
19. Korkman, J., Santtila, P., Westeråker, M. & Sandnabba, N. K. Interviewing techniques and follow-up questions in child sexual abuse interviews. *Eur. J. Dev. Psychol.* **5**, 108–128. https://doi.org/10.1080/17405620701210460 (2008).
20. Sternberg, K. J., Lamb, M. E., Davies, G. M. & Westcott, H. L. The memorandum of good practice: Theory versus application. *Child Abuse Neglect* **25**, 669–681. https://doi.org/10.1016/S0145-2134(01)00232-0 (2001).
21. Benson, M. S. & Powell, M. B. Evaluation of a comprehensive interactive training system for investigative interviewers of children. *Psychol. Public Policy Law* **21**, 309. https://doi.org/10.1037/law0000052 (2015).
22. Cederborg, A.-C., Alm, C., da Silva, Lima, Nises, D. & Lamb, M. E. Investigative interviewing of alleged child abuse victims: An evaluation of a new training programme for investigative interviewers. *Police Pract. Res.* **14**, 242–254. https://doi.org/10.1080/15614263.2012.712292 (2013).

23. Yi, M., Jo, E. & Lamb, M. E. Effects of the NICHD protocol training on child investigative interview quality in Korean police officers. *J. Police Crim. Psychol.* **31**, 155–163. https://doi.org/10.1007/s11896-015-9170-9 (2016).
24. Johnson, M. *et al.* Best practice recommendations still fail to result in action: A national 10-year follow-up study of investigative interviews in CSA cases. *Appl. Cogn. Psychol.* **29**, 661–668. https://doi.org/10.1002/acp.3147 (2015).
25. Dalli, K. C. Technological acceptance of an avatar based interview training application: The development and technological acceptance study of the AvBIT application. *Digitala Vetenskapliga Arkivet* (2021).
26. Dalli, K. C. Design and evaluation of an avatar-mediated system for child interview training. *Digitala Vetenskapliga Arkivet* (2015).
27. Pompedda, F., Zappalà, A. & Santtila, P. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychol. Crime Law* **21**, 28–52. https://doi.org/10.1080/1068316X.2014.915323 (2015).
28. Pompedda, F., Antfolk, J., Zappalà, A. & Santtila, P. A combination of outcome and process feedback enhances performance in simulations of child sexual abuse interviews using avatars. *Front. Psychol.* **8**, 1474. https://doi.org/10.3389/fpsyg.2017.01474 (2017).
29. Krause, N., Pompedda, F., Antfolk, J., Zappalá, A. & Santtila, P. The effects of feedback and reflection on the questioning style of untrained interviewers in simulated child sexual abuse interviews. *Appl. Cogn. Psychol.* **31**, 187–198. https://doi.org/10.1002/acp.3316 (2017).
30. Pompedda, F. *et al.* Transfer of simulated interview training effects into interviews with children exposed to a mock event. *Nord. Psychol.* **73**, 43–67. https://doi.org/10.1080/19012276.2020.1788417 (2020).
31. Pompedda, F., Zhang, Y., Haginoya, S. & Santtila, P. A mega-analysis of the effects of feedback on the quality of simulated child sexual abuse interviews with avatars. *J. Police Crim. Psychol.* **37**, 485–498. https://doi.org/10.1007/s11896-022-09509-7 (2022).
32. Hassan, S. Z. *et al.* Towards an AI-driven talking avatar in virtual reality for investigative interviews of children. In *Proceedings of the ACM Workshop on Games Systems (GameSys)*, 9–15 (2022). https://doi.org/10.1145/3534085.3534340
33. Baugerud, G. A. *et al.* Multimodal virtual avatars for investigative interviews with children. In *Proceedings of the Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR)*, 2–8 (2021). https://doi.org/10.1145/3463944.3469269.
34. Pausch, R., Proffitt, D. & Williams, G. Quantifying immersion in virtual reality. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, 13–18 (1997).
35. Bailenson, J. N. *et al.* The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. *The J. Learn. Sci.* **17**, 102–141. https://doi.org/10.1080/10508400701793141 (2008).
36. Lai, K.-W. K. & Chen, H.-J. H. A comparative study on the effects of a VR and PC visual novel game on vocabulary learning. *Computer Assisted Language Learning* 1–34 (2021). https://doi.org/10.1080/09588221.2021.1928226.
37. Krokos, E., Plaisant, C. & Varshney, A. Virtual memory palaces: Immersion aids recall. *Virtual Real.* **23**, 1–15. https://doi.org/10.1007/s10055-018-0346-3 (2019).
38. Madden, J. *et al.* Ready student one: Exploring the predictors of student learning in virtual reality. *PLoS One* **15**, e0229788. https://doi.org/10.1371/journal.pone.0229788 (2020).
39. Kumazaki, H. *et al.* Android robot-mediated mock job interview sessions for young adults with autism spectrum disorder: A pilot study. *Front. Psych.* **8**, 169. https://doi.org/10.3389/fpsyt.2017.00169 (2017).
40. Smith, M. J. *et al.* Virtual reality job interview training and 6-month employment outcomes for individuals with substance use disorders seeking employment. *J. Vocat. Rehabil.* **44**, 323–332. https://doi.org/10.3233/JVR-160802 (2016).
41. Smith, M. J. *et al.* Virtual reality job interview training and 6-month employment outcomes for individuals with schizophrenia seeking employment. *Schizophr. Res.* **166**, 86–91. https://doi.org/10.1016/j.schres.2015.05.022 (2015).
42. Raaen, K. & Eg, R. Who uses vr in norway? In *Norsk IKT-konferanse for forskning og utdanning*, Vol. 1, 150–153 (NIK Norsk informatikkonferanse, 2021).
43. Brunnström, K. *et al.* Qualinet white paper on definitions of quality of experience. *HAL open science* (2013).
44. Witmer, B. G., Jerome, C. J. & Singer, M. J. The factor structure of the presence questionnaire. *Presence Teleoper Virtual Environ* **14**, 298–312. https://doi.org/10.1162/105474605323384654 (2005).
45. Perkis, A. *et al.* Qualinet white paper on definitions of immersive media experience (IMEx). arXiv:2007.07032 (2020).
46. Salehi, P. *et al.* Is more realistic better? A comparison of game engine and gan-based avatars for investigative interviews of children. In *Proceedings of the ACM Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR)*, 41–49 (2022). https://doi.org/10.1145/3512731.3534209
47. Sabet, S. S. *et al.* Comparison of crowdsourced and remote subjective user studies: A case study of investigative child interviews. In *Proceedings of the International Conference on Quality of Multimedia Experience (QoMEX)*, 1–6 (2022). https://doi.org/10.1109/QoMEX55416.2022.9900900
48. Zhou, Y. *et al.* Makelttalk: Speaker-aware talking-head animation. *ACM Trans. Gr. (TOG)* **39**, 1–15. https://doi.org/10.1145/3414685.3417774 (2020).
49. Sabet, S. S. *et al.* Comparison of crowdsourced and remote subjective user studies: A case study of investigative child interviews. In *Proceedings of the IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, 1–6 (2022). https://doi.org/10.1109/QoMEX55416.2022.9900900
50. Union, I. T. P.809 : Subjective evaluation methods for gaming quality (201e).
51. Lamb, M. E. *et al.* The effects of intensive training and ongoing supervision on the quality of investigative interviews with alleged sex abuse victims. *Appl. Dev. Sci.* **6**, 114–125. https://doi.org/10.1207/S1532480XADS0603_2 (2002).
52. Haginoya, S. *et al.* AI avatar tells you what happened: The first test of using AI-operated children in simulated interviews to train investigative interviewers. *Front. Psychol.*https://doi.org/10.3389/fpsyg.2023.1133621 *(2023).*
53. Zohaib Hassan, S. *et al.* A comparative study of interactive environments for investigative interview of a virtual child avatar. In *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, 194–201 (2022). https://doi.org/10.1109/ISM55400.2022.00043
54. Michailidis, L., Balaguer-Ballester, E. & He, X. Flow and immersion in video games: The aftermath of a conceptual challenge. *Front. Psychol.* **9**, 1682. https://doi.org/10.3389/fpsyg.2018.01682 (2018).
55. Sabet, S. S. *The Influence of Delay on Cloud Gaming Quality of Experience* (Springer, 2022).
56. Jörg, S., Normoyle, A. & Safonova, A. How responsiveness affects players' perception in digital games. In *Proceedings of the ACM Symposium on Applied Perception*, 33–38 (2012). https://doi.org/10.1145/2338676.2338683
57. Mulqueeny, K., Kostyuk, V., Baker, R. S. & Ocumpaugh, J. Incorporating effective e-learning principles to improve student engagement in middle-school mathematics. *Int. J. STEM Educ.* **2**, 1–14. https://doi.org/10.1186/s40594-015-0028-6 (2015).
58. Myers, J. E., Saywitz, K. J. & Goodman, G. S. Psychological research on children as witnesses: Practical implications for forensic interviews and courtroom testimony. *Pac. LJ* **28**, 3 (1996).
59. Bocklisch, T., Faulkner, J., Pawlowski, N. & Nichol, A. Rasa: Open source language understanding and dialogue management. arXiv:1712.05181 (2017).
60. Santiago, F., Singh, P., Sri, L. *et al.Building cognitive applications with IBM Watson services: Volume 6 speech to text and text to speech* (IBM Redbooks, 2017).
61. Powell, M. B., Guadagno, B. & Benson, M. Improving child investigative interviewer performance through computer-based learning activities. *Polic. Soc.* **26**, 365–374. https://doi.org/10.1080/10439463.2014.942850 (2016).
62. Langballe, Å. & Davik, T. Sequential interviews with preschool children in norwegian barnahus. In *Collaborating Against Child Abuse: Exploring the Nordic Barnahus Model* (eds Johansson, S., Stefansen, K., Bakketeig, E. & Kaldal, A.) 165–183 (Springer International Publishing, Cham, 2017). https://doi.org/10.1007/978-3-319-58388-4_8

63. Bunk, T., Varshneya, D., Vlasov, V. & Nichol, A. Diet: Lightweight language understanding for dialogue systems. arXiv:2004.09936 (2020).
64. UMA Steering Group. UMA Repository. https://github.com/umasteeringgroup/UMA. Online repository (2022).
65. Unity Asset Store. SALSA LipSync Suite. https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442. Software package (2022).
66. Hassan, S. Z. *et al.* A virtual reality talking avatar for investigative interviews of maltreat children. In *Proceedings of the ACM International Conference on Content-Based Multimedia Indexing (CBMI)*, 201–204 (2022). https://doi.org/10.1145/3549555.3549572
67. Powell, M. B. & Brubacher, S. P. The origin, experimental basis, and application of the standard interview method: An information-gathering framework. *Aust. Psychol.* **55**, 645–659. https://doi.org/10.1111/ap.12468 (2020).
68. Schmidt, S. *Assessing the Quality of Experience of Cloud Gaming Services* (Springer, 2022).
69. Wilson, E., Hewett, D. G., Jolly, B. C., Janssens, S. & Beckmann, M. M. Is that realistic? The development of a realism assessment questionnaire and its application in appraising three simulators for a gynaecology procedure. *Adv. Simul.* **3**, 1–7. https://doi.org/10.1186/s41077-018-0080-7 (2018).
70. IJsselsteijn, W. A., de Kort, Y. A. & Poels, K. The game experience questionnaire. *Eindh. Tech. Univ. Eindh.***46** (2013).
71. Igroup Presence Questionnaire (IPQ) Overview: A multi-disciplinary project consortium addressing new interfaces between humans and the real and virtual environment. Available at: http://www.igroup.org/pq/ipq/index.php

Acknowledgements

This research is sponsored by the Research Council of Norway, project number 314690 ("Interview training of child-welfare and law-enforcement professionals interviewing maltreated children supported via artificial avatars") .

Author contributions

S.Z.H, S.S.S, M.A.R, G.A.B, R.K.R and P.H conceived the experiment(s), S.Z.H, S.S.S and H.K conducted the experiment(s) and analyzed the data. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.Z.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Article VI: Immersive Virtual Reality in Child Interview Skills Training: A Comparison of 2D and 3D Environments

Authors: *Pegah Salehi*, Syed Zohaib Hassan, Gunn Astrid Baugerud, Martine B Powell, M Cayetana López Cano, Miriam S Johnson, Ragnhild Klingenberg Røed, Dag Johansen, Saeed Shafiee Sabet, Michael A Riegler, Pål Halvorsen

Candidate's contributions: The candidate contributed to this study by designing the methodology, conducting data collection and analysis, and serving as the primary author of the manuscript. Her work focused on evaluating the comparative effectiveness of 2D and 3D virtual environments for investigative interview training with children. She also played a key role in interpreting the results and shaping the discussion on immersive training tools.

Published in: Proceedings of the 16th International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE) - Bari, Italy - April 15-18, 2024

DOI: 10.1145/3652212.3652219

Contributed to: This work contributed to the **RO-3**.

Immersive Virtual Reality in Child Interview Skills Training: A Comparison of 2D and 3D Environments

Pegah Salehi*
SimulaMet, Norway

Syed Zohaib Hassan†
SimulaMet, Norway

Gunn Astrid Baugerud
OsloMet, Norway

Martine Powell
Griffith University, Australia

M. Cayetana López Cano
OsloMet, Norway

Miriam S. Johnson
OsloMet, Norway

Ragnhild Klingenberg Røed
OsloMet, Norway

Dag Johansen
UiT The Arctic University of Norway

Saeed Shafiee Sabet
SimulaMet, Norway

Michael A. Riegler†
SimulaMet, Norway

Pål Halvorsen†
SimulaMet, Norway

ABSTRACT

The current study aims to evaluate and compare the subjective quality of an AI-based training system developed for conducting child interviews, focusing on the distinction between immersive 3D (using virtual reality) and 2D desktop environments. To this end, a structured user study was conducted, involving 36 participants who were exposed to these two distinct environments. The study evaluated various aspects of user experience, namely *presence*, *usability*, *visual fidelity*, *emotion*, *responsiveness*, *appropriateness*, and *training effectiveness*. The findings reveal significant differences in user experience between the 2D and 3D environments. Notably, the 3D environment enhanced *presence*, *visual fidelity*, *training effectiveness*, and *empathy*. In contrast, the 2D environment was favored for *usability*. The study highlights the potential of immersive VR while also pointing out the need to improve the system response and emotional expressiveness of the avatars.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Virtual reality**;
• **Computing methodologies** → *Machine learning*.

KEYWORDS

Virtual Reality (VR), Virtual Environments (VEs), Immersion, Quality of Experience (QoE), Large language model (LLM)

ACM Reference Format:
Pegah Salehi, Syed Zohaib Hassan, Gunn Astrid Baugerud, Martine Powell, M. Cayetana López Cano, Miriam S. Johnson, Ragnhild Klingenberg Røed, Dag Johansen, Saeed Shafiee Sabet, Michael A. Riegler, and Pål Halvorsen. 2024. Immersive Virtual Reality in Child Interview Skills Training: A Comparison of 2D and 3D Environments. In *Proceedings of the international workshop on immersive mixed and virtual environment systems (MMVE '24).* ACM, New York, NY, USA, 7 pages. https://doi.org/3652212.3652219

*Also affiliated with UiT The Arctic University of Norway
†Also affiliated with OsloMet, Norway

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MMVE '24, April 15-18, 2024, Bari, Italy
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0618-9/24/04…$15.00
https://doi.org/3652212.3652219

1 INTRODUCTION

Child abuse is a critical global issue that negatively affects children's development, mental and physical well-being. Meta-analysis studies have estimated that 22.6% of children experience physical abuse and 11.8% experience sexual abuse before reaching adulthood [36]. In particular, less than 15% of child sexual abuse (CSA) cases are corroborated by physical evidence [42], and in 70% of these cases, the child is the sole witness [11]. Properly conducted investigative interviews are crucial for prosecution, as children can be reliable witnesses when interviewed according to best-practice guidelines [4]. Such guidelines encourage communication with the child through open-ended questions to obtain elaborate and relevant evidence details [7]. Unfortunately, these guidelines are not often followed despite investments in training programs [18]; interviewers tend to use too many suggestive and closed questions, and directive questions generate short, rather than elaborate, responses [1, 2, 8]. This is due to training programs that contain inadequate practice opportunities (with feedback) that shape interviewer performance in the use of open questions [44]. Recent studies, including a decade-long Norwegian study, indicate that there is no significant advancement in interview quality despite training innovations. Mock interviews with trained respondents have been shown to be highly effective in shaping interviewer performance [27].

However, face-to-face training with a trained respondent is expensive and cumbersome to set up. It requires the availability of both parties and considerable investment to ensure that the trained respondent responds in a way that effectively shapes the performance of the trainee interviewer. To potentially improve the update of practice opportunities, we introduced an innovative AI-based training platform to conduct investigative interviews [14, 31]. This platform integrates VR technology with an advanced avatar system. The avatar is designed to emulate the responses of children in high-fidelity abuse scenarios, enhancing the realism of the digital training environment. The core of our system lies in the seamless

integration of state-of-the-art natural language processing (NLP) and vision technologies. NLP empowers the avatar with the ability to understand and express human language, while vision technologies enable them to present visually accurate and responsive representations of human-trained respondents. This dual-technology approach effectively simulates the complexities involved in interviewing child abuse victims, providing a crucial training tool for professionals in this field.

In our previous work [14], we conducted a user study that evaluated the efficacy of different interactive platforms, including VR, 2D desktop, audio, and text chat; however, it lacks crucial elucidation aspects of realism, such as *emotion* and *visual fidelity*. Also, the study may not have possessed ideal statistical power to discern certain subtle effects. This study focuses on a detailed assessment of qualitative feedback within both 2D and 3D visual environments, using a comprehensive questionnaire to investigate various evaluation aspects. In this paper, we present the following main contributions:

- Fine-tuning of the GPT-3 [6] using interview data from abused children and seamlessly integrating the dialogue model with a Unity3D framework.
- Demonstration of the potential of virtual reality (VR) for professional training in real-world scenarios.
- A comparison of trainee interviewers' perceptions across different evaluative aspects in 2D and 3D environments, including *presence*, *usability*, *visual fidelity*, *emotion*, *responsiveness*, *appropriateness*, *training effectiveness*, and *empathy*.

2 RELATED WORK

Child avatar training systems aimed at improving interview quality exist, but none are fully automated. One uses prerecorded child responses manually selected by an operator [9], another [26] employs a probabilistic rule-based algorithm for response selection after a person has categorized the question manually, that still requires human intervention. The existing systems do not utilize advanced large language models such as GPT-3 [6]. Instead, they are constrained and limited in generating responses, relying on predetermined sentences from which to select a response. Also, constrained by their dependency on human operators, these systems can have potential issues with flexibility, operational costs, and error rates.

Parallel to these developments, the emergence of VR as a dynamic educational tool is significant [25]. Several studies have shown that it enhances direct learning experiences and helps in memory retention [29, 30]. It increases the engagement and motivation of the learner [24], and improves the understanding of spatial and visual concepts [17]. VR also facilitates better decision-making in simulations [30], although its effectiveness depends on the application of sound pedagogical methods [12].

Furthermore, the adaptability and acceptability of VR, particularly in specialized training scenarios, are noteworthy [13]. Furthermore, the efficacy of VR in job interview training for individuals with serious mental illness underscores its wide applicability in various educational and training contexts [41].

The comparative analysis between 2D and 3D VR environments further emphasizes these benefits. Research indicates that the 3D environment generates stronger emotional reactions than the 2D interfaces when the same content is presented [19, 34]. This increased participation in VR is correlated with its proven efficacy in enhancing educational outcomes, as evidenced in studies focusing on vocabulary acquisition and memory retention [20, 21]. However, Madden et al. [22] did not identify significant differences in learning about moon phases, indicating that the effectiveness of VR may depend on the specific nature of the task.

Current research trends in this domain include the use of electroencephalogram (EEG) measurements to gain more profound insights. These studies reveal a greater sense of presence in 3D environments and a reduced cognitive load in comparison to 2D tasks [35]. Notably, cognitive load appears higher in 2D tasks, suggesting that 3D technologies might provide cognitive benefits in learning contexts due to their reduced cognitive demands [10]. In support of this notion, Tian et al. [38] experienced that VR's stereoscopic vision can lead to increased emotional arousal, further differentiating 3D experiences from 2D ones.

This study aims to obtain feedback on the effectiveness of our tools and to gain a deeper understanding of the experiences and perceptions of the interviewers associated with child abuse.

3 SYSTEM ARCHITECTURE

The architecture of our interactive child avatar system, which supports both 2D and 3D simulation environments, is shown in Figure 1. The system is structured into three main components: i) the language component, which features large language models (LLMs) based on OpenAI's GPT-3 [6]; ii) the speech synthesis component leveraging IBM Watson services for effective speech-to-text (STT) and text-to-speech (TTS) capabilities; and iii) a Unity-based user interface supporting two distinct interactive modalities: a 3D virtual environment using the Oculus Quest2 Head Mounted Display (HMD), and a 2D virtual environment.

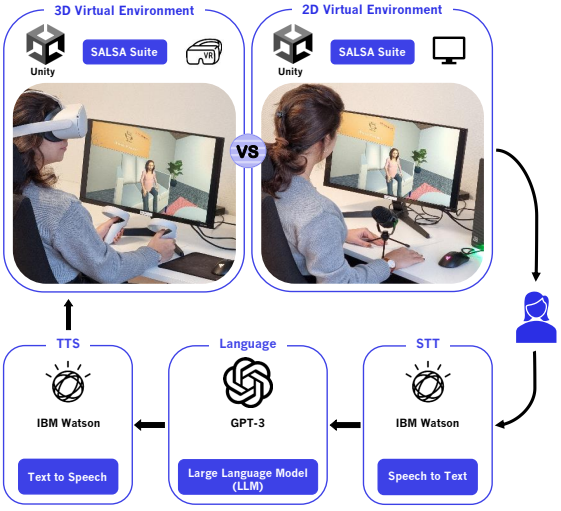


Figure 1: Architecture of the child avatar system for the comparative study.

3.1 Language

The model's language component is designed to process and respond to interviewers' inquiries, emulating a child's conversational patterns. This is based on a dataset of interview transcripts from the Centre for Investigative Interviewing at Griffith University, Australia [28]. The data set includes mock interviews, conducted by trained professionals, that simulate interactions between actors representing children and interviewers from Child Protection Services or law enforcement. In our earlier work [14], our dialogue model was developed using the RASA framework. Now we have fine-tuned the GPT-3 [6] Davinci model for two specific situations: sexual abuse and physical abuse. This was done using 10 simulated forensic interviews with children aged 6 to 8 years who were potential victims. The aim was to produce dynamic and appropriate responses to the questions asked by the interviewers in these contexts. The dialogue model was integrated with a Unity-based user interface using OpenAI API calls.

3.2 Speech Synthesis

IBM Watson's STT and TTS services are used to link the dialogue model with the interactive user interface. The IBM TTS API primarily offers adult voices; however, for a more realistic interaction, we altered the pitch and speed of a female voice to simulate a childlike sound. This modification was based on feedback from a pilot study that indicated that the voices of adults for the children were not well received by the participants [15, 32].

3.3 Visual Environment

This system allows users to interact with a virtual child avatar in two distinct environments. Despite the difference in environments, the system maintains a consistent dialogue model in its back-end, ensuring uniform responses from the avatar in both settings. The system's front-end, developed using the Unity game engine, offered two environments: a 3D environment accessible via an Oculus Quest 2 HMD and a 2D environment displayed on a 24-inch desktop monitor. Both 2D and 3D environments feature the same avatar, created using the Unity Multipurpose Avatar (UMA) open source project [39], which allows the customization of character meshes and textures. For realistic avatar movements, the Salsa Suit asset [40] is employed to synchronize lip, eye, and head movements with a generated voice. To enhance naturalism, prerecorded animations are used to animate the avatar's hands and neck.

4 EXPERIMENT DESIGN

This study involved two groups of participants: experienced investigative interviewers and individuals with backgrounds in psychology, criminology, or related fields but who lacked professional interviewing experience. Recruitment of participants was facilitated through the support of managerial personnel at their respective workplaces, guiding the researchers towards professionals best suited for participation. Eligibility for selection required experience in the criminal justice sector, with a preference for those who had previously worked as investigative interviewers. Out of those approached, all but a few individuals agreed to participate.

The methodology involved simulations in both 2D and 3D settings in random order. Each participant engaged in these environments for a duration ranging from five to ten minutes. Following the simulations, participants were asked to complete a questionnaire. This questionnaire focused on their experiences and observations while interacting with the child avatars in both the 2D and 3D environments. The study was conducted over six separate days without any specific arrangement to distinguish between participants with and without experience in interviewing children.

Our research has received ethical approval from the Norwegian Agency for Shared Services in Education and Research (SIKT) (project number #614272), titled "Interview training of child-welfare and law-enforcement professionals interviewing maltreated children supported via artificial avatars." Additionally, this study was approved by Griffith University, Australia (project number #2023/501), titled "Evaluation of Child Avatar Interview Simulation Learning Activity."

4.1 Participant Demographics

Our study initially included 39 individuals, but with the exclusion of three participants who only interacted in the 2D environment, the number of participants considered for the final analysis was reduced to 36. Within this group, 24 identified as female, 11 as male, and one participant preferred not to disclose gender. The age distribution revealed that most of the participants, 26 participants, fell within the 30-49 age range, 9 were over 50 years old, and one was under 29 years old. Regarding their professional background in child interviewing, 14 participants had no prior experience, 11 had more than ten years of experience, and the remaining 11 had less than ten years of experience. Regarding their exposure to VR, 16 participants had previous VR experience, while 20 had not used VR technology before.

4.2 Questionnaires

The questionnaire comprised 30 items, including 28 questions on a 5-point Likert scale and two open-ended questions. These questions were categorized into eight different evaluation keys: *presence* [16], *usability* [5], *visual Fidelity* [3, 43], *emotion* [3], *responsiveness* [33], *appropriateness* [45], *training effectiveness*, and *empathy*. Each question was designed to provide specific insights relevant to its category, ensuring comprehensive coverage of the user experience. For the full text of these questions, please refer to Figure 2.

5 RESULTS

In this section, we analyze the subjective feedback of participants for both 2D and 3D environments. The results are organized into three segments: Descriptive Analysis, Comparative Analysis, and Qualitative Feedback.

5.1 Descriptive Analysis

We began our analysis by calculating the central tendencies and dispersion of each evaluation aspect. In Figure 3, we present a visual comparison of mean scores and standard deviations between 2D and 3D environments. In the 2D environment, the *usability* scored the highest ($Mean = 4.09$, $SD = 0.67$), and in the 3D environment, *presence* scored the highest ($Mean = 4.12$, $SD = 0.72$).

Evaluation Aspects	NUM	Questions	Shorthand	Response Type
Presence	1	I felt engaged during the simulation.	Engagement	Likert scale
	2	I felt immersed in the computer-generated world.	Immersion	Likert scale
	3	I was able to concentrate on the simulation without being distracted by my surroundings.	Concentration	Likert scale
	4	I forgot about the real world during the interaction.	World Forgetfulness	Likert scale
Usability	5	The equipment was comfortable to use.	Equipment Comfort	Likert scale
	6	I felt comfortable interacting with the child Avatar.	Avatar Interaction Comfort	Likert scale
	7	The interface of the tool was easy to understand and use.	Interface Usability	Likert scale
	8	I did not experience technical difficulties while interacting with the child Avatar.	Technical Difficulty	Likert scale
	9	I would feel very comfortable using this tool on my own next time.	Ease of Future Use	Likert scale
Visual Fidelity	10	The appearance of the child Avatar was realistic.	Appearance Fidelity	Likert scale
	11	The virtual environment where the child Avatar was located felt real and contributed to my overall immersive experience.	Environment Fidelity	Likert scale
	12	I perceived hand-movements/gestures from the child Avatar.	Hand Movement Perception	Likert scale
	13	The quality of the child Avatar's movements was satisfactory (naturalness, realism, …).	Movement Quality	Likert scale
	14	The child Avatar's lip movements were well synchronized with the speech.	Lip Sync Accuracy	Likert scale
	15	The child Avatar's face expressions/movements felt realistic and were well synchronized with the speech.	Facial Expression Fidelity	Likert scale
	16	The overall perception was realistic and pleasant.	Overall Realism Perception	Likert scale
Emotion	17	I felt emotionally engaged during the interaction with the child Avatar.	Emotional Engagement	Likert scale
	18	I perceived emotions in the child Avatar's responses.	Emotional Response Perception	Likert scale
	19	The child Avatar's emotional reactions (e.g. body language, facial expressions and behaviour) looked realistic.	Emotional Reaction Realism	Likert scale
	20	The child Avatar's emotional reactions (e.g. body language, facial expressions and behaviour) consistently matched the content of the interview.	Emotion-Content Match	Likert scale
Responsiveness	21	The responsiveness of the system to my inputs felt right, natural and smooth (e.g. the system's reaction time, the consequent responses/actions from the child Avatar).	System Responsiveness	Likert scale
	22	I noticed a delay between my questions and the child Avatar's responses/reactions.	Response Delay Notice	Likert scale
	23	The pace was the usual for a conversation with a child in such circumstances.	Conversation Pace Normalcy	Likert scale
Appropriateness	24	The child Avatar's responses felt age appropriate.	Age-Appropriate Response	Likert scale
	25	The child Avatar's responses were consistent with respect to the general story.	Story Consistency	Likert scale
	26	The child Avatar's responses were appropriate and on-topic with my questions.	Response Relevance	Likert scale
Training Effectiveness	27	From a learning perspective, my interaction with the child Avatar felt as effective as interacting with a human actor/trainer.	Training Comparability	Likert scale
	28	I think this tool should be included in investigative interviewing training programs.	Tool Inclusion Recommendation	Likert scale
Empathy	29	Please provide one or more examples of the aspects of the child Avatar that felt particularly effective in eliciting your empathy and understanding.	Effective Empathy Elicitation Examples	Open-ended
	30	Please provide one or more examples of the aspects of the child Avatar that felt particularly ineffective in eliciting your empathy and understanding.	Ineffective Empathy Elicitation Examples	Open-ended

Figure 2: Categorization of questionnaire items according to Evaluation Aspects.

Conversely, the *responsiveness* scored the lowest in both the 2D and 3D environments, indicating concerns regarding interaction delays. Specifically, the mean score was 2.61 ($SD = 0.61$) in the 2D environment and 2.75 ($SD = 0.71$) in the 3D environment.

5.2 Comparative Analysis

In this section, we investigate the differential impact of the 2D versus 3D environment on QoE. This investigation utilized a paired sample T-test to analyze the variations between the two environments. Additionally, we calculated Cohen's d to assess the effect sizes. The results of these analyses are presented in Table 1.

In terms of *presence*, the results reveal that the 3D environment enhances the sense of presence more significantly than a 2D environment due to its sensory-rich and interactive dynamics. Specifically, participants experienced a higher sense of 'immersion', 'engagement' and 'world forgetfulness' in a 3D environment. However, 'concentration' levels did not show a significant difference, suggesting a complex relationship between immersive features and user focus. It is possible that the use of HMDs captures and diverts the participant's attention, thus impacting their concentration levels.

From the perspective of *usability*, this study claims that due to the absence of physically demanding equipment such as HMD, the 2D environment offers a more user-friendly experience. Specifically, participants experienced higher 'equipment comfort' in the 2D environment, suggesting that the absence of additional gear may contribute to increased comfort. Similarly, 'ease of future use' was

Table 1: Paired Sample T-Test and Cohen's d Effect Size for Learning Experience Variables. The gray areas show significant differences noted at p < .05.

	T-Statistic	P-Value	Cohen's d Effect Size
Engagement	2.798	0.008	0.466
Immersion	5.940	0.000	0.990
Concentration	2.023	0.050	0.337
World Forgetfulness	2.704	0.010	0.450
Equipment Comfort	-4.159	0.000	-0.693
Avatar Interaction Comfort	1.000	0.324	0.166
Interface Usability	-0.867	0.391	-0.144
Technical Difficulty	0.000	1.000	0.000
Ease of Future Use	-2.256	0.030	-0.376
Appearance Fidelity	1.190	0.242	0.198
Environment Fidelity	3.944	0.000	0.657
Hand Movement Perception	2.841	0.007	0.473
Movement Quality	1.847	0.073	0.307
Lip Sync Accuracy	0.849	0.401	0.141
Facial Expression Fidelity	2.142	0.039	0.357
Overall Realism Perception	1.661	0.105	0.276
Emotional Engagement	1.454	0.154	0.242
Emotion Response Perception	1.000	0.324	0.166
Emotional Reaction Realism	1.357	0.183	0.226
Emotion-Content Match	0.452	0.653	0.075
System Responsiveness	0.122	0.903	0.020
Response Delay Notice	1.183	0.244	0.197
Conversation Pace Normalcy	1.540	0.132	0.256
Age-Appropriate Response	1.152	0.257	0.192
Story Consistency	0.273	0.785	0.045
Response Relevance	-1.177	0.246	-0.196
Training Comparability	3.365	0.001	0.560
Tool Inclusion Recommendation	1.715	0.095	0.285

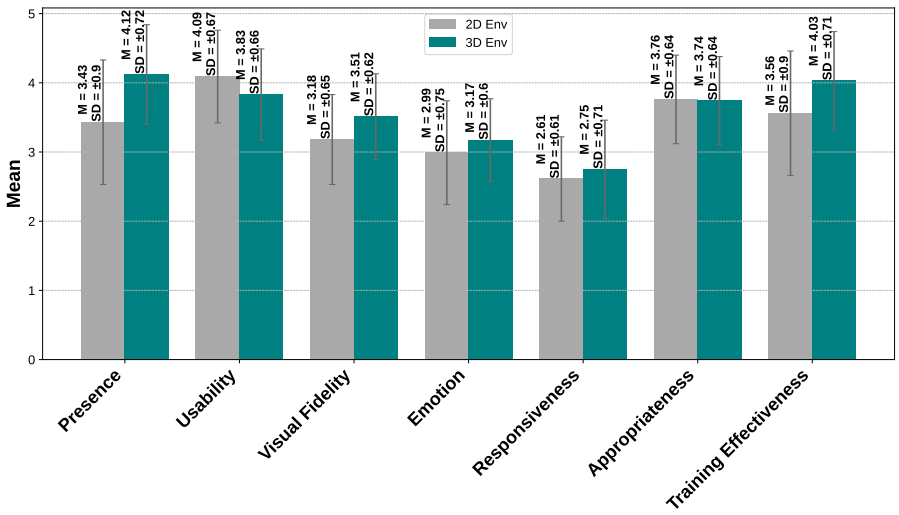


Figure 3: Bar-plot (95% confidence interval) of mean scores across the evaluative aspects in 2D and 3D environments.

significantly better in the 2D environment, which implies that users might prefer the 2D environment for future interactions, possibly due to its less physically exhausting nature. Conversely, 'avatar interaction comfort', 'interface usability', and 'technical difficulty' showed no significant differences, indicating that these aspects of *usability* are perceived similarly across both 2D and 3D environments.

Regarding *visual fidelity*, the study examined several key aspects, which indicated that there is an improved quality of visual experience in the 3D environment. In particular, 'environment fidelity' shows a substantial increase in 3D environment, suggesting that participants perceived a 3D environment as more lifelike. Similarly, 'hand movement perception' is significantly enhanced in 3D, indicating that the fluidity of VR technology provides a more lifelike representation of hand movement. Further, 'facial expression fidelity' is rated significantly higher in the 3D environment, reflecting that the immersive nature of the 3D environment may inherently strengthen the user's perception of facial expression naturalness.

However, 'movement quality', while not reaching statistical significance, shows a trend towards enhanced perception in 3D, implying a possible positive impact of VR technology on the naturalness and consistency of avatar movement. The 'overall realism perception' also trended towards a better experience in a 3D environment, but did not reach conventional significance levels. In addition, according to our expectations, there are no significant differences between 2D and 3D environments regarding 'avatar appearance fidelity' and 'lip sync accuracy'.

In terms of *emotion*, the study investigated how users perceive emotional dynamics with an avatar lacking emotional expressions in 2D and 3D environments. Contrary to the belief that a 3D immersive environment may compensate for the absence of emotional expressions [23, 37], the results indicate that there are no statistically significant differences between the 2D and 3D environments. This finding highlights the importance of explicit emotional signals in improving emotional dynamics during avatar-user interactions. However, a trend is observed to favor 'emotional engagement' and 'emotional reaction realism' in the 3D environment, although this trend is not statistically significant, which implies that immersion provided by the 3D environment leads the user to perceive certain behaviors and facial expressions of the avatar as an emotional reaction.

Regarding two conceptually close aspects, *responsiveness* and *appropriateness*, the results show that there is no significant difference in user experience between the 2D and 3D environments.

Furthermore, from the perspective of *training effectivenes*, we examined participants' perceptions of the training with the child avatar relative to human actors ('training comparability') and their approval of incorporating this tool into investigative interview training programs ('tool inclusion recommendation'). The results demonstrate a statistically significant preference for 3D over 2D in terms of 'training comparability', suggesting that participants found the 3D interaction to be nearly as effective as a real-life training experience. Additionally, there is a trend of preferring 'tool inclusion recommendation' in the 3D environment, although not statistically significant, suggesting that immersion provided by the 3D environment enhances the effectiveness of training by simulating real-world interactions more closely than in the 2D environment.

5.3 Qualitative Feedback

In the open-ended section of the questionnaire survey, participants were asked to clarify the aspects that contributed to the effective and ineffective elicitation of empathy within two simulated environments, as presented in Table 2. This table quantifies the feedback, offering a clear count of positive and negative comments for each evaluated aspect of the avatar's performance.

Notably, the 3D environment was dependent on its *visual fidelity*, with participants expressing mixed reviews, one stating: *'The hand*

Table 2: Distribution of Participant Comments on Effective and Ineffective Aspects of Empathy Elicitation in 2D and 3D Environments.

	Effective Aspects		Ineffective Aspect	
	2D	3D	2D	3D
Presence	0	1	2	0
Visual Fidelity	8	19	9	15
Emotion	2	0	4	5
Voice	6	7	4	5
Responsiveness	16	10	13	9
Response Speed	0	0	7	5

movements and eye contact were great!' In contrast, another participant stated: *'The avatar's body was not that of a child but more of a woman, which was distracting'*, underscoring the importance of appropriate visual cues in engendering empathy.

On the contrary, the participant responses indicated that the efficacy of the 2D environment was largely dependent on the avatar's *responsiveness*. One participant highlighted this as a strength, stating, *'When she spoke of harm perpetrated by her father, that tugged the heartstrings.'* Nonetheless, certain comments, like the one regarding the avatar's limited response of *'YEAH'*, underscore the necessity for enhanced emotional expressiveness and sophisticated interaction within virtual training platforms.

Furthermore, the participant feedback in both environments revealed aspects that hindered empathy engagement, indicating potential areas for improvement. A repeated critique related to the delay in the avatar's reactions; as one respondent explicitly observed, *'There was a large delay in her responses that made it hard to engage with the avatar and affected rapport building significantly.'* Emotional reaction was another aspect that received critical attention, with comments such as *'No emotion shown when disclosing sexual abuse'*, implying that more emotional expression could be crucial in augmenting the avatar's realism.

6 DISCUSSION AND FUTURE WORK

We explored the comparative effectiveness of 2D and 3D training environments in the context of CPS interview training. Differing from our earlier research [14], where VR's novelty significantly influenced participants, the current study provided a more comprehensive understanding of VR's practical utility in training scenarios. A noteworthy finding is the pronounced preference for the 3D environment in most of the questionnaire responses, suggesting that immersive experiences are important to further investigate in relation to improving training efficacy. This aligns with previous research suggesting that immersive learning contexts can substantially improve skill acquisition by providing more realistic and engaging experiences, as they closely mimic real-world interactions and scenarios [17, 20, 24]. This is demonstrated by a significant difference in 'facial expression fidelity', a crucial aspect of non-verbal communication, even without explicit emotion settings in the system. Participant feedback such as *'In the 2D version non-verbal behavior of the avatar was more difficult to perceive than in the VR version of the program'* supports this observation. Recognizing the importance of emotional dynamics in virtual environments, future iterations of the system will integrate Unreal Engine technology with NVIDIA Omniverse[1] to enhance emotional reactions. This integration is expected to result in a more sophisticated simulation of human-like emotional responses, thereby enriching the user's immersive experience by providing a nuanced and context-sensitive rendering of non-verbal cues. These advancements will likely contribute to the development of more empathetic and interactive virtual environments.

Another significant outcome is the advantage of visual interaction over text-based methods in developing effective communication skills. User comments like *'making eye contact but then also looking away when talking about negative moments (assault) seemed like a normal response in a human interview'* highlight the system's potential in simulating realistic human interactions. Interestingly, the immersive nature of the visual environment sometimes leads to overlooked responses, highlighting its realism. This was demonstrated in some participant's observations. For instance, one noted, *'When the child paused and then said the age of her brother very quietly. It seemed like a realistic moment'*, suggesting that in the visual environment, users think the delay is part of the child's hesitation. However, there is still a concern raised regarding the delay observed in the avatar's responses, primarily due to the current necessity for manual input to trigger these responses. Recognizing the criticality of fluid interaction in the context of child investigative interview interviews, future iterations will focus on automating the response mechanism, thereby eliminating the need for manual intervention.

Despite these advantages in VR technology, no significant differences were observed between the 2D and 3D environments regarding the 'tool inclusion recommendation'. This lack of distinction could be partly attributed to the discomfort associated with using HMDs, as indicated by participant comments such as *'Using headphones did not cause me pain, (…) with the VR I was in pain.'*

Additionally, it should be noted that the 3D environment requires additional equipment and a special setup, which is intrusive and uncomfortable for some individuals. In contrast, the 2D environment offers greater ease of use, which makes it ideal for 'anytime, anywhere' training sessions, making it a more accessible option for a broader audience.

7 CONCLUSION

This paper has presented a comprehensive analysis contrasting 2D and 3D virtual environments in their application to child interview training for CPS and law enforcement personnel. The findings reveal that the immersive 3D environment significantly enhanced aspects such as *presence*, *visual fidelity*, and *training effectiveness*, providing a more realistic and engaging interaction that greatly benefits training scenarios. Conversely, the 2D environment, while less immersive, was favored for its higher *usability* due to its fewer equipment requirements. However, this research also identifies critical challenges within these virtual training environments, particularly the need for emotional expressiveness and improved system responsiveness. Based on these findings, it is evident that future developments in virtual training tools must adopt a balanced approach. Prioritizing the refinement of the 3D environment is crucial, with a specific focus on integrating emotional expressiveness in avatars and improving avatar response speed. These improvements are essential for increasing the efficacy and accessibility of these virtual training environments for CPS and law enforcement personnel.

[1] https://www.nvidia.com/en-us/omniverse/apps/audio2face/

REFERENCES

[1] Gunn Astrid Baugerud, Miriam S Johnson, Helle BG Hansen, Svein Magnussen, and Michael E Lamb. 2020. Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015–2017). *Applied cognitive psychology* 34, 3 (2020), 654–663.

[2] Gunn Astrid Baugerud, Ragnhild Klingenberg Røed, Helle BG Hansen, Julie Schøning Poulsen, and Miriam S Johnson. 2023. Evaluating child interviews conducted by child protective services workers and police investigators. *The British Journal of Social Work* (2023), bcac245.

[3] Frank Biocca and Mark R Levy. 2013. *Communication in the age of virtual reality*. Routledge.

[4] C.J. Brainerd and V.F. Reyna. 2012. Reliability of children's testimony in the era of developmental reversals. *Developmental review* 32, 3 (2012), 224–267.

[5] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Sonja P. Brubacher, Mairi S. Benson, Martine B. Powell, Jane Goodman-Delahunty, and Nina J. Westera. 2020. Chapter Twenty-Two - An overview of best practice investigative interviewing of child witnesses of sexual assault. In *Child Sexual Abuse*, India Bryce and Wayne Petherick (Eds.). Academic Press, 445–466.

[8] Ann-Christin Cederborg, Yael Orbach, Kathleen J Sternberg, and Michael E Lamb. 2000. Investigative interviews of child witnesses in Sweden. *Child abuse & neglect* 24, 10 (2000), 1355–1361.

[9] Kevin Charles Dalli. 2021. Technological Acceptance of an Avatar Based Interview Training Application: The development and technological acceptance study of the AvBIT application. *Digitala Vetenskapliga Arkivet* (2021).

[10] Alex Dan and Miriam Reiner. 2017. EEG-based cognitive load of processing events in 3D virtual worlds is lower than processing events in 2D displays. *International Journal of Psychophysiology* 122 (2017), 75–84.

[11] Mark D. Everson and Jose Miguel Sandoval. 2011. Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements. *Child Abuse & Neglect* 35, 4 (2011), 287–298.

[12] Chris Fowler. 2015. Virtual reality and learning: Where is the pedagogy? *British journal of educational technology* 46, 2 (2015), 412–422.

[13] Stephanie G Fussell and Dothang Truong. 2023. Accepting virtual reality for dynamic learning: An extension of the technology acceptance model. *Interactive Learning Environments* 31, 9 (2023), 5442–5459.

[14] Syed Zohaib Hassan, Saeed Shafiee Sabet, Michael Alexander Riegler, Gunn Astrid Baugerud, Hayley Ko, Pegah Salehi, Ragnhild Klingenberg Røed, Miriam Johnson, and Pål Halvorsen. 2023. Enhancing investigative interview training using a child avatar system: a comparative study of interactive environments. *Scientific Reports* 13, 1 (2023), 20403.

[15] Syed Zohaib Hassan, Pegah Salehi, Ragnhild Klingenberg Røed, Pål Halvorsen, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pierre Lison, Michael Riegler, Michael E. Lamb, Carsten Griwodz, and Saeed Shafiee Sabet. 2022. Towards an AI-Driven Talking Avatar in Virtual Reality for Investigative Interviews of Children. In *Proceedings of the ACM Workshop on Games Systems (GameSys)* (Athlone, Ireland). 9–15.

[16] Igroup Presence Questionnaire (IPQ) Overview. . A Multi-Disciplinary Project Consortium Addressing New Interfaces Between Humans and the Real and Virtual Environment. Available at: http://www.igroup.org/pq/ipq/index.php.

[17] Lasse Jensen and Flemming Konradsen. 2018. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies* 23 (2018), 1515–1529.

[18] Miriam Johnson, Svein Magnussen, Christian Thoresen, Kyrre Lønnum, Lisa Victoria Burrell, and Annika Melinder. 2015. Best practice recommendations still fail to result in action: A national 10-year follow-up study of investigative interviews in CSA cases. *Applied Cognitive Psychology* 29, 5 (2015), 661–668.

[19] Maria Kandaurova and Seung Hwan Mark Lee. 2019. The effects of Virtual Reality (VR) on charitable giving: The role of empathy, guilt, responsibility, and social exclusion. *Journal of Business Research* 100 (2019), 571–580.

[20] Eric Krokos, Catherine Plaisant, and Amitabh Varshney. 2019. Virtual memory palaces: immersion aids recall. *Virtual reality* 23, 1 (2019), 1–15.

[21] Kuo-Wei Kyle Lai and Hao-Jan Howard Chen. 2021. A comparative study on the effects of a VR and PC visual novel game on vocabulary learning. *Computer Assisted Language Learning* 0, 0 (2021), 1–34.

[22] J Madden, S Pandita, JP Schuldt, B Kim, A S. Won, and NG Holmes. 2020. Ready student one: Exploring the predictors of student learning in virtual reality. *PloS one* 15, 3 (2020), e0229788.

[23] Valentina Mancuso, Francesca Bruni, Chiara Stramba-Badiale, Giuseppe Riva, Pietro Cipresso, and Elisa Pedroli. 2023. How do emotions elicited in virtual reality affect our memory? A systematic review. *Computers in Human Behavior* (2023), 107812.

[24] Thomas K Metzinger. 2018. Why is virtual reality interesting for philosophers? *Frontiers in Robotics and AI* 5 (2018), 101.

[25] Kathy A Mills and Alinta Brown. 2022. Immersive virtual reality (VR) for digital media making: transmediation is key. *Learning, Media and Technology* 47, 2 (2022), 179–200.

[26] Francesco Pompedda, Angelo Zappalà, and Pekka Santtila. 2015. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law* 21, 1 (2015), 28–52.

[27] Martine B Powell, Ronald P Fisher, and Carolyn H Hughes-Scholes. 2008. The effect of intra-versus post-interview feedback during simulated practice interviews about child abuse. *Child Abuse & Neglect* 32, 2 (2008), 213–227.

[28] Martine B Powell, Belinda Guadagno, and Mairi Benson. 2016. Improving child investigative interviewer performance through computer-based learning activities. *Policing and Society* 26, 4 (2016), 365–374.

[29] Jaziar Radianti, Tim A Majchrzak, Jennifer Fromm, and Isabell Wohlgenannt. 2020. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education* 147 (2020), 103778.

[30] Fabin Rasheed, Prasad Onkar, and Marisha Narula. 2015. Immersive virtual reality to enhance the spatial awareness of students. In *Proceedings of the 7th Indian Conference on Human-Computer Interaction*. 154–160.

[31] Pegah Salehi, Syed Zohaib Hassan, Myrthe Lammerse, Saeed Shafiee Sabet, Ingvild Riiser, Ragnhild Klingenberg Røed, Miriam S Johnson, Vajira Thambawita, Steven A Hicks, Martine Powell, et al. 2022. Synthesizing a talking child avatar to train interviewers working with maltreated children. *Big Data and Cognitive Computing* 6, 2 (2022), 62.

[32] Pegah Salehi, Syed Zohaib Hassan, Saeed Shafiee Sabet, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pål Halvorsen, and Michael A. Riegler. 2022. Is More Realistic Better? A Comparison of Game Engine and GAN-Based Avatars for Investigative Interviews of Children. In *Proceedings of the ACM Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR)* (Newark, NJ, USA). 41–49.

[33] Steven Schmidt. 2022. *Assessing the quality of experience of cloud gaming services*. Springer Nature.

[34] Donghee Shin. 2018. Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience? *Computers in human behavior* 78 (2018), 64–73.

[35] Semyon M Slobounov, William Ray, Brian Johnson, Elena Slobounov, and Karl M Newell. 2015. Modulation of cortical activity in 2D versus 3D virtual reality environments: an EEG study. *International Journal of Psychophysiology* 95, 3 (2015), 254–260.

[36] Marije Stoltenborgh, Marian J. Bakermans-Kranenburg, Marinus H. van IJzendoorn, and Lenneke R. Alink. 2013. Cultural–geographical differences in the occurrence of child physical abuse? A meta-analysis of global prevalence. *International Journal of Psychology* 48, 2 (2013), 81–94.

[37] Feng Tian, Minlei Hua, Wenrui Zhang, Yingjie Li, and Xiaoli Yang. 2021. Emotional arousal in 2D versus 3D virtual reality environments. *PloS one* 16, 9 (2021), e0256211.

[38] Feng Tian, Xuefei Wang, Wanqiu Cheng, Mingxuan Lee, and Yuanyuan Jin. 2022. A Comparative Study on the Temporal Effects of 2D and VR Emotional Arousal. *Sensors* 22, 21 (2022), 8491.

[39] UMA Steering Group. 2022. UMA Repository. https://github.com/umasteeringgroup/UMA. Online repository.

[40] Unity Asset Store. 2022. SALSA LipSync Suite. https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442. Software package.

[41] Pınar Üstel, Matthew J Smith, Shannon Blajeski, Jeffery M Johnson, Valerie G Butler, Johanna Nicolia-Adkins, Monica J Ortquist, Lisa A Razzano, and Adrienne Lapidos. 2021. Acceptability and feasibility of peer specialist-delivered virtual reality job interview training for individuals with serious mental illness: A qualitative study. *Journal of technology in human services* 39, 3 (2021), 219–231.

[42] Wendy A Walsh, Lisa M Jones, Theodore P Cross, and Tonya Lippert. 2010. Prosecuting child sexual abuse: The importance of evidence type. *Crime & Delinquency* 56, 3 (2010), 436–454.

[43] Erin Wilson, David G Hewett, Brian C Jolly, Sarah Janssens, and Michael M Beckmann. 2018. Is that realistic? The development of a realism assessment questionnaire and its application in appraising three simulators for a gynaecology procedure. *Advances in Simulation* 3, 1 (2018), 1–7.

[44] Misun Yi, Eunkyung Jo, and Michael E Lamb. 2016. Effects of the NICHD protocol training on child investigative interview quality in Korean police officers. *Journal of police and criminal psychology* 31, 2 (2016), 155–163.

[45] Qingfu Zhu, Lei Cui, Weinan Zhang, Furu Wei, and Ting Liu. 2018. Retrieval-enhanced adversarial training for neural response generation. *arXiv preprint arXiv:1809.04276* (2018).

Article VII: A Theoretical and Empirical Analysis of 2D and 3D Virtual Environments in Training for Child Interview Skills

Authors: *Pegah Salehi*, Syed Zohaib Hassan, Gunn Astrid Baugerud, Martine Powell, Miriam S Johnson, Dag Johansen, Saeed Shafiee Sabet, Michael A Riegler, Pål Halvorsen

Candidate's contributions: The candidate was the primary contributor to this study, leading all aspects of the research. She designed and developed the comparative evaluation framework for 2D and 3D virtual environments, conceptualized the study methodology, and implemented the technical infrastructure. Her work included analyzing both quantitative and qualitative data, and performing in-depth thematic analysis of participant interviews to assess key aspects such as visual fidelity, emotional engagement, and training effectiveness. As the first author, she drafted the manuscript and synthesized the findings.

Published in: IEEE Access, vol. 12, pp. 131842-131864 (2024)

DOI: 10.1109/ACCESS.2024.3442297

Contributed to: This work contributed to the **RO-3** and **RO-5**.

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A theoretical and empirical analysis of 2D and 3D Virtual Environments in Training for Child Interview Skills

PEGAH SALEHI[1,4], SYED ZOHAIB HASSAN[1,2], GUNN ASTRID BAUGERUD[2], MARTINE POWELL[3], MIRIAM S. JOHNSON[2], DAG JOHANSEN[4], SAEED SHAFIEE SABET[1], MICHAEL A. RIEGLER[1,2], PÅL HALVORSEN[1,2]

[1]SimulaMet, Oslo, Norway
[2]OsloMet, Oslo, Norway
[3]Griffith University, Australia
[4]UiT The Arctic University of Norway, Tromsø, Norway

Corresponding author: Pegah Salehi (e-mail: pegah@simula.no).

Our research has received ethical approval from the Norwegian Agency for Shared Services in Education and Research (SIKT) (project number #614272), titled "Interview training of child-welfare and law-enforcement professionals interviewing maltreated children supported via artificial avatars." Additionally, this study was approved by Griffith University, Australia (project number #2023/501), titled "Evaluation of Child Avatar Interview Simulation Learning Activity."

ABSTRACT This paper presents a detailed study of an AI-driven platform designed for the training of child welfare and law enforcement professionals in conducting investigative interviews with maltreated children. It achieves a realistic simulation of interview situation through the integration of fine-tuned GPT-3 models within the Unity framework. The study recruited participants from a range of backgrounds, including professionals experienced in conducting investigative interviews and individuals with academic qualifications in psychology, criminology, or related disciplines. To assess the effectiveness of this tool, a multi-method evaluation approach was utilized, incorporating both quantitative analyses and qualitative interviews. The quantitative methods included mixed-effects models, which provided insights into how effects such as the type of virtual environment (2D vs. 3D), scenario variations, virtual reality (VR) familiarity, and professional expertise influence the user experience. Additionally, structural equation modeling (SEM) provided deeper insights into the relationships between variables, offering a comprehensive understanding of how they collectively impact the user experience. The qualitative method included a detailed semi-structured interview that provided a deeper understanding of user experiences and perceptions.

The findings indicate significant advantages of the 3D environment in enhancing 'flow' and 'virtual fidelity'; however, the 2D environment was favored for 'usability'. Despite the 3D environment's potential for greater immersion, the discomfort associated with VR head-mounted displays (HMDs) led some users to prefer the 2D setup. Familiarity with VR technology positively influenced user perception, indicating that prior exposure can mitigate some of the 'avatar interaction comfort' issues. Additionally, the 'hand movement perception' was better understood in scenarios with sensitive themes. As user experience increased, participants had a more positive view of the 'age-appropriateness responses'. Furthermore, the dialog system's effectiveness, particularly 'responses relevance' and 'detail responses', played a significant role in 'empathy elicitation', often outweighing 'visual fidelity'. However, 'emotion' in facial expressions and 'responsiveness' were two factors that negatively impacted the effectiveness of the tool, indicating areas that need improvement in the future.

INDEX TERMS Virtual Reality (VR), Virtual Environments (VEs), Immersion, Usability, Quality of Experience (QoE), Large language model (LLM)

I. INTRODUCTION

CHILD abuse is a pervasive global issue that detrimentally affects children's psychological, developmental, and physical health. Meta-analyses reveal that before reaching adulthood, 22.6% of children face physical abuse and 11.8% encounter sexual abuse [1]. Specifically, child sexual abuse (CSA) cases rarely have corroborative physical evidence, with less than 15% of cases supported by such data [2], and in 70% of cases, the child is the sole witness [3]. The lack of corroborative evidence underscores the importance of conducting effective investigative interviews, as these can harness reliable testimonies from children when aligned with best-practice guidelines [4]. These guidelines promote the use of open-ended questions to elicit detailed, accurate, and relevant evidence [5].

Despite the critical role of proper interviewing techniques and the availability of extensive training programs, compliance with best practices remains low [6]. Interviewers frequently resort to using too many suggestive, closed, and directive questions, which hampers the collection of comprehensive and accurate information from child witnesses [7]–[9]. The reasons for this include training programs that lack effective practice opportunities and do not adequately adjust interviewers' behavior towards using open questions [10]. The persistence of these issues even after professional training highlights significant gaps in current training methodologies [11]. Innovative methods such as mock interviews with trained actors have proven beneficial in enhancing interviewer skills; however, these face-to-face training sessions are costly and logistically challenging, requiring significant resources and the availability of both trainee and trainer [12].

To address these challenges and improve training accessibility and effectiveness, we have developed an AI-based training platform that utilizes VR technology with artificial intelligence (AI) to develop a novel training platform that simulates realistic interview scenarios with high-fidelity avatars mimicking child behavior in these situations [13], [14].

The core of this system combines natural language processing (NLP), computer vision, and audio technologies. NLP allows the avatar to comprehend and articulate human language, while computer vision technologies enable it to present visually accurate and responsive representations of children. The audio component leverages IBM Watson's speech synthesis and recognition services for effective speech-to-text (STT) and text-to-speech (TTS) capabilities, ensuring the avatar can respond with childlike voices by modifying pitch and speed. This triad of technologies effectively mimics the complexities of interviewing child abuse victims, thereby providing a vital, scalable tool for training professionals in this critical field.

In [14], we undertook a user study to evaluate the effectiveness of various interactive platforms, including VR, 2D desktop environments, audio, and text chat. However, this study did not fully address different dimensions of visual fidelity and the importance of emotional facial expression. Moreover, the statistical power of the study may have been insufficient to detect some nuanced effects.

Building on these preliminary insights, we conducted a compressive study to explore deeper into the integration of AI technologies with VR platforms to augment the realism and efficacy of child interview sessions. We utilized the advanced features of the GPT-3 language model, which was specially fine-tuned using mock interview datasets. This model was then seamlessly integrated into a Unity3D framework, creating an audio-visual training environment that closely simulates the complexities of real-world situations. The study recruited participants from diverse backgrounds, including professionals with experience in conducting investigative interviews and individuals with academic qualifications in psychology, criminology, or related fields.

As a result, this paper is an extended version of our previously published work [15]. We have expanded our research with additional experiments, deeper analysis, extended discussions, and new findings. Specifically, these extensions aim to probe deeper into the dynamics between virtual environment design and user experience. The core advancements presented in this paper involve a detailed examination of the interactions within these simulated environments through mixed-effects models. These models assess how effects such as the type of environment (2D vs. 3D), scenario variation, VR familiarity, and the user's professional expertise influence the user experience. Additionally, we used SEM to analyze how different variables are interconnected and to assess their combined impact on factors such as 'training effectiveness' and 'empathy elicitation'. In addition to the quantitative findings, our study incorporates in-depth semi-structured interviews with all participants who interacted in the environments. These interviews were designed to explore participants' detailed experiences, perceptions, and emotional responses while engaging with both 2D and 3D virtual environments, particularly focusing on aspects not fully captured by quantitative measures.

II. RELATED WORK

The following section reviews pertinent literature in two primary areas: investigative interview training and the application of VR in education. By examining existing research in these fields, we aim to establish a foundation for understanding the advancements and current challenges in these domains.

A. INVESTIGATIVE INTERVIEW TRAINING

Existing child avatar training systems have significantly contributed to the field of investigative interviews, yet they show varied levels of automation and effectiveness. Predominantly semi-automated, these systems require varying degrees of human intervention. The system developed by Linnæus University in collaboration with AvBIT Labs exemplifies an early approach, where prerecorded child responses are manually selected by operators to simulate interactions during interviews [16]. This method, while innovative, restricts dy-

namic interaction due to its reliance on predefined responses. Furthermore, another research team developed LiveSimulation [17], [18] which allows participants to interact with a videotaped five-year-old child discussing alleged sexual abuse. Participants select from four predetermined questions, with the child's video responses reflecting typical response patterns of five-year-olds. Research shows that this training leads to a significant increase in the use of open-ended questions, with the effects persisting in follow-ups up to 12 months later [19], [20].

In parallel, Empowering Interviewer Training (EIT) [21], which utilizes a virtual child created from animated, morphed images of real children. Unlike LiveSimulation, EIT allows for more dynamic interactions where participants can pose verbal questions freely. The child's responses are video clips with predefined answers, selected either manually or through a probabilistic rule-based algorithm that activates after manual question categorization [22], [23]. Feedback on both process and outcome is provided after each session, enhancing learning effectiveness [21], [22], [24].

Building on these foundations, ViContact system [25] represents an advancement in VR training for child interviewing. Through the ViContact, participants are able to engage in verbal interactions with virtual children, simulating real-life scenarios of suspected abuse conversations. The system combines VR-based simulated conversations with automated, personalized feedback and classical seminar training, enhancing both open-ended questioning skills and socio-emotional support among trainees.

These systems, however, do not yet leverage advanced language processing technologies like GPT-3, which could potentially address many of the current systems' limitations by improving response dynamism and relevance [26]. Furthermore, the integration of VR in educational tools like these could significantly enhance immersive learning experiences, as VR has been shown to increase engagement, memory retention, and decision-making capabilities depending on the applied pedagogical methods [27].

B. VIRTUAL REALITY FOR EDUCATION

The transformative potential of VR as an educational tool has been increasingly recognized in recent literature [27]. VR not only augments direct learning experiences by improving memory retention but also significantly boosts learner engagement and motivation [28]–[30]. These attributes are important in creating an interactive learning environment where spatial and visual concepts are better understood and learners are more immersed in the educational content [31]. Moreover, VR has shown promise in simulation-based training, enhancing decision-making skills through realistic scenarios [28]. However, the success of these simulations depends on integrating sound pedagogical practices that tailor the VR experience to educational objectives [32].

In specialized training scenarios, particularly in fields requiring high levels of practical skill, VR's adaptability and acceptance have been remarkable [33]. For instance, its application in job interview training for individuals with serious mental health conditions illustrates VR's broad utility across various educational and training frameworks [34].

A critical aspect of VR's educational efficacy is its ability to induce stronger emotional responses compared to traditional 2D learning environments [35], [36]. Studies suggest that 3D VR environments, by presenting the same content, engage users more deeply, thus enhancing learning outcomes related to vocabulary acquisition and memory retention [37], [38]. Nevertheless, the technology's effectiveness can vary depending on the nature of the learning task, as some studies, like those on learning moon phases, show minimal differences between VR and traditional methods [39].

Recent trends in this field involve employing electroencephalogram (EEG) to delve deeper into cognitive processes, uncovering that 3D VR may reduce cognitive load compared to 2D experiences [40]. This reduction is thought to facilitate learning by minimizing mental strain [41]. Research by Tian et al. supports this, noting that VR's stereoscopic vision not only reduces cognitive load but also increases emotional engagement, offering a distinct advantage over traditional 2D approaches [42].

This study aims to obtain feedback on the effectiveness of the 2D and 3D virtual environments and to gain a deeper understanding of the experiences and perceptions of the interviewers when interviewing an alleged abused child avatar.

III. SYSTEM ARCHITECTURE

Figure 1 illustrates the design of our interactive child avatar system, which is compatible with both 2D and 3D interactive environments. The system is divided into three key components: 1) the language module, which utilizes OpenAI's large language models (LLM), GPT-3 [26]. It is fine tunned to converse like a child with various personas; 2) the speech synthesis module, which employs IBM Watson services for efficient STT and TTS functions; and 3) a Unity-based user interface that offers two interactive modes: a 3D virtual environment accessible via the Oculus Quest2 HMD and a 2D virtual environment.

A. LANGUAGE

The language component of the model is designed to emulate a child's conversational style, responding to interviewers' questions. This is achieved by utilizing a dataset of interview transcripts from the Centre for Investigative Interviewing at Griffith University, Australia [20]. The dataset features mock interviews conducted by trained professionals, simulating interactions between actors portraying children and interviewers from Child Protection Services or law enforcement.

In our previous work [14], we utilized the RASA framework to develop our dialogue model. Currently, we have fine-tuned GPT-3 [26] Davinci model for two specific scenarios: sexual abuse and physical abuse. This fine-tuning involved using 10 simulated forensic interviews with children aged 6 to 8 years who were potential victims. The objective was to generate dynamic and contextually appropriate responses

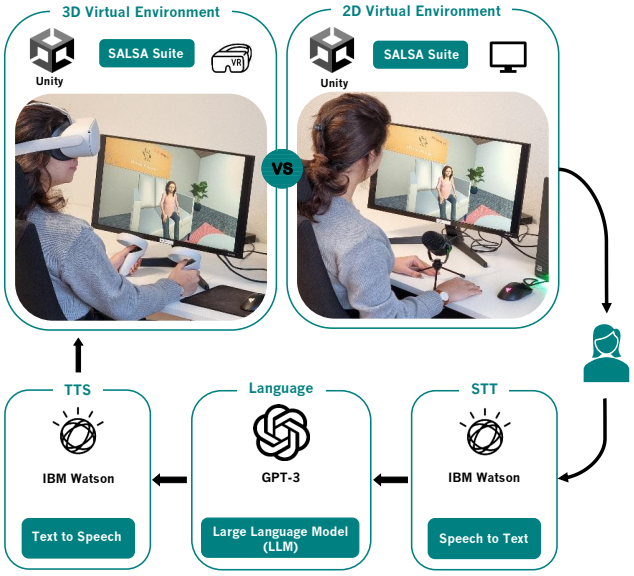


FIGURE 1. Architecture of the child avatar system used in the comparative study.

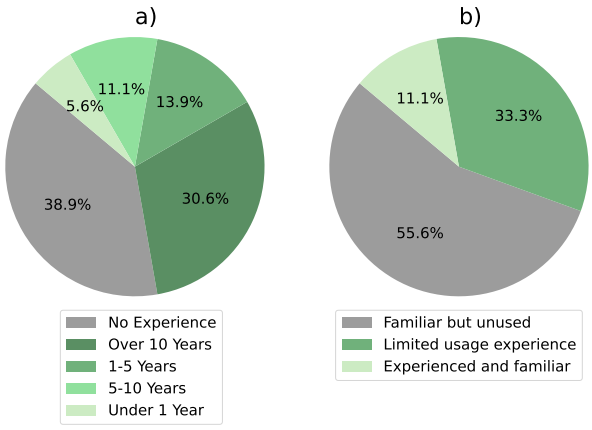


FIGURE 2. Distribution of experience levels among participants, comparing professional years (a) and familiarity with VR (b).

to the interviewers' questions. The process of fine-tuning involved tweaking crucial hyperparameters such as a batch size to 1, setting learning rate multiplier to 0.1, a total of four epochs of fine-tuning, and a prompt loss weight of 0.01. We used 46 examples of prompts and completions, which enhanced the model's ability to comprehend and adjust to the different conversation scenarios. The dialogue model has been integrated with a Unity-based user interface via OpenAI API calls.

B. SCENARIO DESIGN

To select a scenario for the experiment, the interview transcripts were organized into various personas. The personas chosen for this research were identified based on the volume of available transcripts and the depth of information suitable for creating sufficiently long dialogues. The selected personas were named Hillary and Rebecca, each with distinct backgrounds:

a: Rebecca Scenario

This scenario involves a child named Rebecca who is designed to simulate a case of sexual abuse. The dialogue model has been fine-tuned to generate appropriate responses that reflect the experiences and emotional state typical of a child who has faced such trauma. This scenario helps trainees understand the nuances of interviewing a child in such delicate situations, emphasizing the need for empathy and careful questioning techniques.

b: Hilary Scenario

In this scenario, the child avatar named Hilary portrays a victim of physical abuse. The responses and behaviors of Hilary are crafted to illustrate the challenges faced by children undergoing physical harm. This scenario is crucial for training interviewers to recognize signs of physical abuse and to handle the interview process with the sensitivity and support needed to ensure the child's safety and well-being.

C. SPEECH SYNTHESIS

IBM Watson's STT and TTS services facilitate the connection between the dialogue model and the user interface. Although the IBM TTS API typically provides adult voices, we adjusted the pitch and speed of a female voice to create a childlike tone. This adjustment was informed by a pilot study, which revealed that the participants did not respond well to adult voices intended for children [43].

D. VISUAL INTERFACE

This system enables users to interact with a virtual child avatar within two distinct environments. The front-end, built using the Unity game engine, features two environments: a 3D environment accessible via an Oculus Quest 2 headset and a 2D environment displayed on a 24-inch desktop monitor. The same avatar is present in both environments, created with the Unity Multipurpose Avatar (UMA) open-source project [44], allowing for customization of character meshes and textures. For realistic avatar movements, the Salsa Suit asset [45] is used to synchronize lip, eye, and head movements with a generated voice. Additionally, prerecorded animations animate the avatar's hands and neck to enhance naturalism.

IV. EXPERIMENT DESIGN

The experimental protocol was structured into two main components following each interactive session: an interview and a survey. Participants experienced both 2D and 3D environments in a random sequence. They interacted with the child avatar in these environments for an average duration of 8:27 minutes, with a standard deviation of 1:59 minutes. Based on a pre-study, we determined that 10 minutes is the optimal duration for these interactions. This duration was found to strike an effective balance, allowing participants

to fully engage with the avatar while preventing fatigue and maintaining focus throughout the session. Immediately following each interactive session, participants underwent a semi-structured interview lasting between 10 and 15 minutes to collect qualitative feedback. Subsequently, they completed a questionnaire survey, which further contributed to the study's comprehensive data collection. Prior to commencing the experimental activities, participants read and signed an informed consent form to confirm their understanding of the study's objectives and to ensure ethical compliance. Afterwards, a demographic survey was conducted. Throughout the experiment, the processes were meticulously monitored to ensure participant comfort and to gather extensive data on the efficacy of the tool and the overall user experience.

a: Participant Demographics

This study recruited two distinct groups: experienced investigative interviewers and individuals with academic backgrounds in psychology, criminology, or related fields but without professional interviewing experience. Recruitment was facilitated by managerial staff at their workplaces who identified employees matching the study's criteria. All participants were required to have relevant criminal justice sector experience, with a preference for a background in investigative interviewing.

The initial cohort of our study consisted of 39 individuals. After excluding three participants who engaged solely in the 2D environment, the final sample size for analysis was reduced to 36 individuals. Demographically, 24 participants identified as female, 11 as male, and one preferred not to disclose their gender. The age distribution was primarily within the 30-49 year range, with 26 participants, nine over 50 years, and one under 29 years. Concerning professional experience in child interviewing, 14 participants had no prior experience, and 22 had experience in investigative interviewing. Regarding prior exposure to VR, 16 participants had previous experience with VR technologies, whereas 20 had not previously used VR. The levels of professional expertise and VR familiarity are illustrated in the pie chart presented in Figure 2, which are specifically referenced in the results section for further analysis.

A. INTERVIEW STUDY

Following the interview sessions with the child avatar, participants were engaged in a semi-structured interview designed to elicit detailed feedback on their experiences. The interview consisted of a series of open-ended questions as outlined in Table 1, which probed various aspects of their interaction with the avatars. These aspects included their overall experience, ability to engage with the avatars from a visual/interactive perspective, the appropriateness of the avatars' responses, and the usefulness of the activity for training purposes. Additionally, participants were asked to comment on areas for improvement and their preferred version of the environment. These questions were instrumental in capturing comprehensive insights into the user experience.

To ensure accuracy in data collection, the interviews were recorded and subsequently transcribed using Otter transcription software, adhering to the guidelines established in the Informed Consent document. To analyze the transcripts, we first categorized the data based on questionnaire items and then clustered them into different topics. We assessed the importance of each topic by counting the number of participants who mentioned it and visualized this data in the form of graphs (see Figure **??**). The entire review process of the interview transcripts was conducted with meticulous attention to detail, ensuring thorough representation and analysis of participant feedback.

TABLE 1. Open-ended questions in post-test interview.

	Items
Q1	what was your overall experience with this avatar?
Q2	Please reflect solely on your ability to engage with the Avatar from a visual/interactive perspective.
Q3	Please reflect on the appropriateness of her responses.
Q4	Please reflect on the usefulness of this activity from a training perspective.
Q5	Please provide your perspective on the things that went well and those that need to improve.
Q6	Which was your preferred version? Please elaborate.
Q7	What haven't we asked you today that you think would be valuable for us to know?

B. SURVEY STUDY

Following each interview, the participant was asked to complete a detailed questionnaire designed to gather insights into their experiences and observations while interacting with child avatars in both 2D and 3D environments. The study was conducted across six distinct sessions on separate days. There was no predefined grouping based on the participants' prior experience in interviewing children, which allowed for a diverse range of responses. The questionnaire features 31 questions rated on a 5-point Likert scale. These questions are organized into seven distinct evaluative factors: 'Flow' [46], 'Usability' [47], 'visual Fidelity' [48], [49], 'emotion' [48], 'responsiveness' [50], 'response relevance' [51], and 'training effectiveness'. Additionally, there are four independent questions: 'empathy elicitation', 'age-appropriate response', 'detailed responses', and 'response suggestibility'. Each question was crafted to elicit detailed information pertinent to its respective category, thereby ensuring a thorough evaluation of the user experience. For the complete list of these questions, refer to Table 2.

V. QUANTITATIVE RESULTS

In this section, we analyze the subjective feedback of participants for both the 2D and 3D environment. The results are organized into four segments: Reliability Analysis, Descriptive Analysis, Comparative Analysis, and Structural Equation Modeling Analysis.

A. RELIABILITY ANALYSIS

We assessed the reliability of the evaluation factors using Cronbach's alpha to determine the internal consistency of

TABLE 2. Categorization of questionnaire items by evaluation factors. Items that are not grouped into predefined categories are noted for their distinctive attributes and the absence of a reliable category, based on Cronbach's alpha (shown in Table 3). The columns labeled 'Shorthand' and 'Tag' refer to the concise labels assigned to each item, which facilitate referencing them efficiently throughout the text and figures of the paper.

Evaluation Factors	NUM	Questions	Shorthand	Tag
Flow (Flw)	1	I felt engaged during the simulation.	Engagement	Eng
	2	I felt immersed in the computer-generated world.	Immersion	Imm
	3	I was able to concentrate on the simulation without being distracted by my surroundings.	Concentration	Cnc
	4	I forgot about the real world during the interaction.	World Forgetfulness	WrF
Usability (Usb)	5	The equipment was comfortable to use.	Equipment Comfort	EqC
	6	I felt comfortable interacting with the child Avatar.	Avatar Interaction Comfort	AIC
	7	The interface of the tool was easy to understand and use.	Interface Usability	InU
	8	I did not experience technical difficulties while interacting with the child Avatar.	Technical Difficulty	TcD
	9	I would feel very comfortable using this tool on my own next time.	Ease of Future Use	EFU
Visual Fidelity (VsF)	10	The appearance of the child Avatar was realistic.	Appearance Fidelity	ApF
	11	The virtual environment where the child Avatar was located felt real and contributed to my overall immersive experience.	Environment Fidelity	EnF
	12	I perceived hand-movements/gestures from the child Avatar.	Hand Movements Perception	HMP
	13	The quality of the child Avatar's movements was satisfactory (naturalness, realism, …).	Movement Quality	MvQ
	14	The child Avatar's lip movements were well synchronized with the speech.	Lip Sync Accuracy	LSA
	15	The child Avatar's face expressions/movements felt realistic and were well synchronized with the speech.	Facial Expressions Fidelity	FEF
	16	The overall perception was realistic and pleasant.	Overall Realism Perception	ORP
Emotion (Emt)	17	I perceived emotions in the child Avatar's responses.	Emotional Response Perception	ERP
	18	The child Avatar's emotional reactions (e.g. body language, facial expressions and behaviour) looked realistic.	Emotional Reaction Realism	ERR
	19	The child Avatar's emotional reactions (e.g. body language, facial expressions and behaviour) consistently matched the content of the interview.	Emotion-Content Match	ECM
_	20	Interacting with the child Avatar evoked my empathy towards their situation.	Empathy Elicitation	EmE
Responsiveness (Rsp)	21	The responsiveness of the system to my inputs felt right, natural and smooth (e.g. the system's reaction time, the consequent responses/actions from the child Avatar).	System Responsiveness	SyR
	22	I noticed a delay between my questions and the child Avatar's responses/reactions.	Response Delay Notice	RDN
	23	The pace was the usual for a conversation with a child in such circumstances.	Conversation Pace Normalcy	CPN
Response Relevance (RsR)	24	The child Avatar's responses were consistent with respect to the general story.	Story Consistency	StC
	25	The child Avatar was able to understand my questions and statements.	Avatar Comprehension	AvC
	26	The child Avatar's responses were appropriate and on-topic with my questions.	Response On-Topic	ROT
_	27	The child Avatar's responses felt age appropriate.	Age-Appropriate Response	AAR
_	28	The child Avatar's responses were highly detailed when asked specific questions (i.e. Y/N, forced choice or cued recall questions).	Detailed Responses	DtR
_	29	The child Avatar was suggestible in response to my questions or statements.	Response Suggestibility	RdS
Training Effectiveness (TrE)	30	From a learning perspective, my interaction with the child Avatar felt as effective as interacting with a human actor/trainer.	Training Comparability	TrC
	31	I think this tool should be included in investigative interviewing training programs.	Tool Inclusion Recommendation	TIR

TABLE 3. Cronbach's Alpha values for assessing reliability of the different factors in 2D and 3D environments.

	Iteam	2D Env	3D Env
Presence	4	0.89	0.77
Comfort	5	0.77	0.67
Visual Fidelity	7	0.83	0.80
Emotion	3	0.81	0.78
Responsiveness	3	0.61	0.73
Response Relevance	3	0.76	0.66
Training Effectiveness	2	0.70	0.61

the items in 2D and 3D environments. The results, as detailed in Table 3, highlight the reliability of the different factors across environments. The Cronbach's alpha values indicate generally good reliability for most factors in both environments. For the 2D environment, the factor of 'flow' shows the highest reliability ($\alpha = 0.89$), suggesting a very consistent internal structure. However, the factor of 'responsiveness' displays relatively lower reliability ($\alpha = 0.61$), which might reflect variability in the participants' responses concerning interaction delays. In the 3D environment, the reliability scores generally decreased compared to the 2D environment, with the highest reliability observed in 'visual fidelity' ($\alpha = 0.80$). The lowest reliability was seen in 'training effectiveness' ($\alpha = 0.61$), indicating potential inconsistencies in how participants perceived the usefulness of this tool in a 3D environment.

B. DESCRIPTIVE ANALYSIS

In this section, we calculated the central tendencies and measures of dispersion for each evaluation factor. Figure 3 displays a graphical comparison of the mean scores and standard deviations across 2D and 3D environments. In the

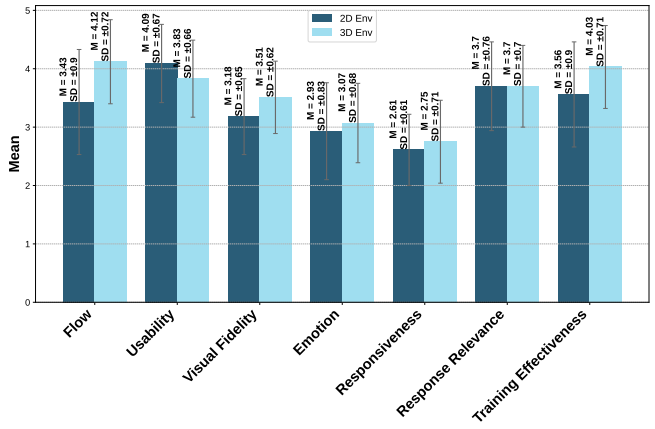


FIGURE 3. Bar-plot (95% confidence interval) of mean scores across seven evaluative factors in 2D and 3D environments.

2D environment, the highest mean score was observed for 'usability' ($Mean = 4.09$, $SD = 0.67$), while in the 3D environment, 'flow' received the highest mean score ($Mean = 4.12$, $SD = 0.72$). In contrast, 'responsiveness' had the lowest scores in both environments, indicating issues with interaction latency. Specifically, the mean responsiveness score was 2.61 ($SD = 0.61$) in the 2D setting and 2.75 ($SD = 0.71$) in the 3D setting. This low 'responsiveness' score can be attributed to several factors, primarily the delays introduced by the speech synthesis process, which involves both STT and TTS conversions via cloud API calls. Additionally, the dialogue model's reliance on calls to the OpenAI API further contributes to these delays. In contrast, our previous work [52], which utilized a locally run dialogue model, achieved a mean responsiveness score of 3.18 ($SD = 0.45$). We believe that the cumulative delays from these two API calls are the primary cause of the observed low score.

C. COMPARATIVE ANALYSIS

In this section, we investigate the influence of various effects on the QoE across different experimental conditions. Our analysis differentiates between experiences in 2D and 3D environments and examines the impact of two distinct scenarios. We also consider participants' expertise as investigative interviewers, categorized into three groups: those without experience, those with over ten years of experience, and those with under ten years of experience, as depicted in Figure 2 (a). We also analyze participants' familiarity with VR, dividing them into two groups: those who have used VR so far and those who have not, as shown in Figure 2 (b).

A potential downside of having multiple conditions in our study is the risk of low statistical power due to the small sample size. To mitigate this risk, we employed mixed-effects models, which are well-suited for studies with small sample sizes and multiple conditions as they account for within-subject correlations and allow for the inclusion of random effects, thereby enhancing the reliability of our findings. In this model, we incorporated a random intercept for each participant to account for within-subject correlation, acknowledging that each participant experienced both environments. This statistical approach allows us to robustly estimate and compare the effects of each variable on perceived quality. We decided to conduct this analysis on isolated questions rather than groups to achieve a more detailed examination. The detailed results, including the coefficients and p-values for each assessed variable, are systematically presented in Table 4. Following this, the results are discussed.

1) 2D vs. 3D Environment Effects

To assess the distinct impacts of 2D and 3D virtual environments on the user experience, we discuss all evaluation factors separately.

a: Flow:

The 3D environments substantially enhance 'flow' compared to 2D environments. Metrics like 'immersion' ($p < 0.001$), 'engagement' ($p = 0.008$), 'concentration' ($p = 0.050$), and 'world forgetfulness' ($p = 0.017$) were significantly higher in the 3D settings, indicating that participants were more deeply involved and engaged and more capable of forgetting the real world. These outcomes suggest that the immersive properties of 3D environments, with their interactive and sensory-rich dynamics, are crucial for fostering a sense of flow.

b: Usability:

In contrast, the 2D environment was preferred for its 'usability', particularly in terms of 'equipment comfort' ($p < 0.001$) and 'ease of future Use' ($p = 0.022$). The less physically demanding nature of 2D interfaces, without the need for complex equipment like HMDs, contributes to higher comfort levels and a preference for future use. This highlights the trade-offs between ease of use and immersion.

c: Visual Fidelity:

'Visual fidelity' is notably enhanced in the 3D environment, especially in aspects such as 'environment fidelity' ($p < 0.001$) and 'hand movement perception' ($p = 0.022$), which are perceived as more realistic. This improvement in visual fidelity suggests that 3D technology provides a more authentic representation of space and movement, thereby improving the overall visual experience.

d: Emotion:

In fact, the game engine utilized did not incorporate 'emotion' components. We aim to find out if this limitation could have been perceived in the 3D environment through immersion, even if emotions were not explicitly present. Despite the potential for enhanced sensory inputs in 3D environment, the results did not demonstrate any substantial differences, and emotion factors remained similar across both environments.

TABLE 4. Results of Mixed Effects Analysis Examining the Impact of Environment (2D vs. 3D VR), Story Variation, Expertise as investigative interviewers, and Familiarity with VR on User Experience. Cells with blue highlights indicate statistically significant results (p-value ≤ 0.05).

	Environment		**Story**		**Expertise**		**VR Familiarity**	
Questions	Coef	P-Value	Coef	P-Value	Coef	P-Value	Coef	P-Value
Engagement	0.574	0.008	-0.049	0.837	-0.075	0.641	0.219	0.413
Immersion	1.142	0.000	0.018	0.938	-0.047	0.773	0.288	0.288
Concentration	0.387	0.050	-0.01	0.962	-0.162	0.197	0.117	0.571
World Forgetfulness	0.586	0.017	-0.272	0.317	-0.314	0.078	-0.085	0.772
Equipment Comfort	-0.912	0.000	-0.117	0.612	0.066	0.645	0.193	0.414
Avatar Interaction Comfort	0.121	0.473	-0.233	0.213	0.089	0.482	0.416	0.046
Interface Usability	-0.200	0.223	-0.315	0.099	-0.138	0.322	-0.051	0.826
Technical Difficulty	-0.003	0.992	-0.014	0.963	-0.196	0.300	0.482	0.123
Ease of Future Use	-0.466	0.022	-0.11	0.641	-0.139	0.419	0.253	0.374
Appearance Fidelity	0.206	0.225	0.058	0.786	-0.109	0.556	0.267	0.386
Environment Fidelity	0.685	0.000	0.092	0.661	0.190	0.242	-0.136	0.615
Hand Movement Perception	0.416	0.022	-0.431	0.050	-0.135	0.429	-0.104	0.714
Movement Quality	0.238	0.057	0.081	0.618	-0.024	0.896	0.220	0.478
Lip Sync Accuracy	0.06	0.654	-0.265	0.106	0.18	0.212	-0.096	0.689
Facial Expression Fidelity	0.272	0.065	-0.175	0.334	-0.066	0.680	0.051	0.846
Overall Realism Perception	0.263	0.128	-0.078	0.681	0.054	0.671	0.247	0.241
Emotional Response Perception	0.230	0.256	0.185	0.412	-0.027	0.850	0.067	0.779
Emotional Reaction Realism	0.161	0.208	-0.028	0.862	0.048	0.760	0.337	0.199
Emotion-Content Match	0.106	0.576	0.119	0.587	0.067	0.672	-0.084	0.747
Empathy Elicitation	0.104	0.596	-0.034	0.883	-0.218	0.222	-0.337	0.254
System Responsiveness	0.018	0.939	-0.052	0.825	0.006	0.965	0.216	0.344
Response Delay Notice	0.164	0.262	-0.013	0.938	0.023	0.856	-0.05	0.809
Conversation Pace Normalcy	0.156	0.304	-0.342	0.071	0.098	0.508	-0.226	0.356
Story Consistency	0.006	0.977	-0.254	0.283	0.182	0.188	-0.340	0.136
Avatar Comprehension	0.18	0.384	-0.076	0.74	-0.230	0.106	0.028	0.906
Response On-Topic	-0.26	0.228	-0.051	0.819	0.107	0.413	-0.248	0.250
Age-Appropriate Response	0.117	0.349	-0.113	0.462	0.290	0.037	-0.135	0.561
Detailed Responses	0.199	0.259	0.311	0.135	-0.075	0.609	-0.362	0.136
Response Suggestibility	0.137	0.434	-0.007	0.971	-0.155	0.265	-0.041	0.861
Training Comparability	0.687	0.001	-0.038	0.879	-0.211	0.247	0.005	0.988
Tool Inclusion Recommendation	0.227	0.135	-0.118	0.508	-0.015	0.910	0.070	0.749

e: Empathy:

There were no significant differences in 'empathy elicitation' between 2D and 3D environments, as it appears to be influenced more by the content and emotional depth of the dialogues than the visual component in which they are presented. This conclusion is further supported by findings from SEM, which emphasize the critical role of 'response relevance' and 'detailed responses' in enhancing empathy.

f: Dialog System:

Dialog system includes the factors of 'responsiveness' and 'response relevance', along with independent questions of 'age-appropriate response', 'detailed responses', and 'response suggestibility', the results indicate no significant difference in user experience between the 2D and 3D environments. This can be attributed to the nature of the dialog system, which is primarily based on verbal interactions rather than visual elements. Since the effectiveness of the dialogue system relies on the quality of the verbal exchanges and the contextual appropriateness of responses, the transition from a 2D to a 3D environment does not impact its performance. Therefore, the user experience remains consistent across both environments.

g: Training Effectiveness:

The 3D environment showed significantly higher scores in 'training comparability' ($p = 0.001$), indicating that participants perceived the 3D interactions to be almost as effective as those with human actors. However, there was no significant difference in 'tool inclusion recommendation' ($p = 0.135$), despite a tendency to favor the 3D environment. This conclusion is further supported by findings from SEM, which indicate that 'usability' has a greater impact on 'tool inclusion recommendations' than flow. To better understand the rationale behind these results, the additional analyses are detailed in the subsequent part of the results.

2) Varied Story Effect

In our analysis of scenario variations, we observed significant effects on the 'hand movements perception' ($coef = -0.431$, $p = 0.05$). This finding suggests that different scenario contexts may prime users to focus more intently on specific types of visual information, such as hand movements, which are critical for conveying non-verbal cues and enhancing the realism of interactions within virtual environments. Hand movements were particularly noticeable in the scenario where a child avatar discussed experiences of sexual abuse, making these movements significant as the child described their feelings and actions. This context likely heightened participants' awareness of hand movements, enhancing their

perception of these non-verbal cues.

In VR applications, narrative-driven design is crucial, especially where accurate and detailed perceptions of movement are crucial for user immersion and realism [53]. Additionally, we noted a trend in the perceived normalcy of conversation pacing, although this was not statistically significant ($p = 0.071$). This trend indicates that the story context might subtly influence participants' expectations and perceptions of the interaction flow, further emphasizing the narrative's role in shaping the user experience in immersive environments.

3) Expertise Effects

Our investigation examined how specialized professional expertise, particularly in conducting interviews with children, influences user perceptions within VR settings. A key observation from our data analysis revealed a significant correlation between such expertise and the assessment of 'age-appropriate response' ($p - value = 0.037$). Participants with a background in child interactions demonstrated a distinct pattern in their evaluation of the appropriateness of responses when dealing with child avatars. This finding indicates that professionals with extensive experience tend to perceive responses as more age-appropriate. However, it does not necessarily imply that the experts' assessments are more accurate or that they are inherently better at evaluating these interactions compared to non-experts. Instead, it shows the influence of professional training and experience on perception and evaluation criteria in the virtual interaction session.

Despite this significant finding in the domain of 'age-appropriate response', our study revealed that other aspects of the virtual interaction were not markedly influenced by the participants' expertise. This indicates that while specialized knowledge significantly enhances specific facets of the virtual experience, the general accessibility and efficacy of VR environments remain robust across users with varying levels of expertise. This broad usability suggests that VR platforms can serve as effective tools for a wide audience, including those new to professional practices involving children, thereby expanding the potential applications of VR in training and educational settings.

These results align with our previous findings [54], where non-experts could reliably evaluate general aspects of interaction, but domain experts identified subtle and critical elements that non-experts often overlooked. Specifically, the study underscores how expertise is essential for recognizing and interpreting finer details and complexities that may be critical for effective training but are not as apparent to non-experts.

4) VR Familiarity Effect

In analyzing the influence of participants' prior familiarity with VR technology on their experiences, a notable difference emerged in 'avatar interaction comfort'. Specifically, participants familiar with VR reported significantly higher comfort levels when interacting with avatars compared to those without prior VR experience ($p < 0.05$). This familiarity effect suggests that users accustomed to VR are more at ease when engaging with virtual entities, possibly because they have fewer cognitive and psychological barriers. This familiarity may facilitate a more immersive and comfortable interaction within VR environments.

Despite the significant effect on avatar comfort, no other major aspects demonstrated substantial differences based on VR familiarity, suggesting that VR environments are accessible and can be effectively used by novice users as well. However, this finding underscores the benefits of prior VR exposure, which appears to enhance comfort levels by enabling users to better anticipate virtual behaviors and reducing the novelty effects. This highlights the importance of considering user backgrounds in VR interface design to optimize engagement and satisfaction.

D. STRUCTURAL EQUATION MODELING ANALYSIS

In continuation of our evaluation, we utilized SEM to examine the complex interrelations between user experience variables identified in our study. SEM allowed us to specify and estimate multiple and interrelated dependencies simultaneously. Following an initial exploratory factor analysis (EFA), we refined our model by identifying key factors and eliminating variables with insufficient factor loadings, thereby enhancing both the precision and interpretability of our model. For a detailed representation of these relationships, please refer to Figure 4 and Table 5, which illustrate the SEM path diagram and the corresponding statistical analysis results.

1) Impact on Empathy Elicitation

'Empathy elicitation' was selected as a key variable in our SEM analysis due to its foundational importance in the context of child interview training. Our own earlier work [15] has shown the significance of both 'visual fidelity' and 'dialogue systems' in fostering 'empathy elicitation'. Therefore, we aim to identify which of these components has a more substantial impact on the user's empathetic connection with the avatar.

In examining the factors that significantly enhance 'empathy elicitation' during interactions with the child avatar, three components stand out: 'flow', 'response relevance', and 'detailed responses'. 'flow' ($\beta = 0.252$, $p = 0.024$) effectively evokes empathy, indicating that when users feel more engaged, immersed, and less distracted by their real-world surroundings, their empathetic engagement with the avatar intensifies.

'Response relevance' ($\beta = 0.302$, $p = 0.012$) also plays a critical role in eliciting empathy. Accurate and relevant responses from the avatar, which align with the narrative consistency and user's understanding, enhance the realism of the interaction, making the training experience feel more genuine and emotionally impactful. This alignment ensures that the interactions are not only contextually appropriate but

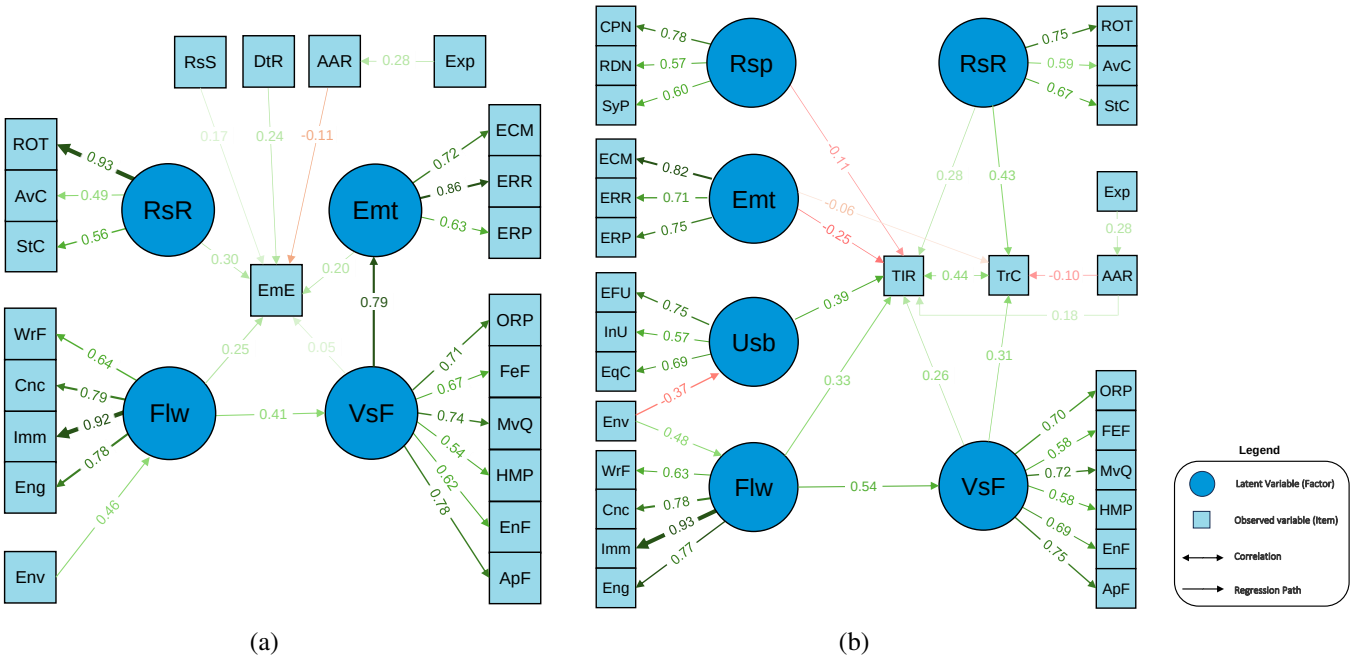


FIGURE 4. SEM Path Diagram: This diagram shows relationships among user experience constructs and their outcomes in the model. Significant paths are in bolder lines; green indicates positive relationships and red indicates negative ones. a) Influences on Empathy Elicitation. b) Influences on Tool Inclusion Recommendation and Training Comparability.

TABLE 5. SEM Analysis of User Experience Predictors: This table details the impact of various predictors on outcomes such as Empathy Elicitation, Tool Inclusion Recommendation, and Training Comparability. Results significant at $p < 0.05$ are indicated. For a graphical representation of these relationships, refer to Figure 4.

Predictor	Outcome	Coefficient	Std. Error	t-value	p-value	Standardized Coeff. (β)
Env	Flow	1.047	0.361	2.898	0.004	0.464
Flow	VisualFidelity	0.399	0.181	2.210	0.027	0.411
VisualFidelity	Emotion	1.187	0.450	2.641	0.008	0.793
Exp	AgeAppropriateResponse	0.268	0.094	2.846	0.004	0.284
Flow	EmpathyElicitation	0.228	0.101	2.252	0.024	0.252
VisualFidelity	EmpathyElicitation	0.046	0.313	0.146	0.884	0.049
ResponseRelevance	EmpathyElicitation	0.308	0.123	2.500	0.012	0.302
ResponseSuggestibility	EmpathyElicitation	0.214	0.127	1.686	0.092	0.175
AgeAppropriateResponse	EmpathyElicitation	-0.147	0.136	-1.084	0.278	-0.113
Emotion	EmpathyElicitation	0.127	0.201	0.633	0.527	0.205
DetailedResponses	EmpathyElicitation	0.274	0.112	2.444	0.015	0.235
Env	Flow	1.089	0.351	3.100	0.002	0.478
Env	Usability	-0.809	0.342	-2.366	0.018	-0.375
Flow	VisualFidelity	0.556	0.204	2.729	0.006	0.535
Exp	AgeAppropriateResponse	0.268	0.094	2.846	0.004	0.284
Flow	ToolInclusionRecommendation	0.208	0.105	1.981	0.048	0.332
Usability	ToolInclusionRecommendation	0.256	0.073	3.514	0.000	0.387
VisualFidelity	ToolInclusionRecommendation	0.155	0.090	1.723	0.085	0.257
Emotion	ToolInclusionRecommendation	-0.182	0.101	-1.805	0.071	-0.254
Responsiveness	ToolInclusionRecommendation	-0.076	0.092	-0.827	0.408	-0.107
ResponseRelevance	ToolInclusionRecommendation	0.197	0.151	1.300	0.194	0.276
AgeAppropriateResponse	ToolInclusionRecommendation	0.165	0.091	1.818	0.069	0.180
VisualFidelity	TrainingComparability	0.289	0.139	2.076	0.038	0.311
Emotion	TrainingComparability	-0.063	0.200	-0.314	0.753	-0.057
ResponseRelevance	TrainingComparability	0.471	0.162	2.896	0.004	0.428
AgeAppropriateResponse	TrainingComparability	-0.137	0.127	-1.079	0.281	-0.097

also substantively meaningful, thereby deepening the user's emotional involvement.

Interestingly, 'detailed responses' ($\beta = 0.235$, $p = 0.015$) significantly contribute to empathy elicitation. When the avatar provides responses that are rich in detail and nuance, it likely enhances the user's perception of the avatar as a realistic and responsive entity. This level of detail in communication can make the virtual interaction more relatable and emotionally resonant, prompting a stronger empathetic response from the user.

2) User Experience Impact on Training Effectiveness

In this section, we assess the items of the 'training effectiveness' factor—'tool inclusion recommendation' and 'training comparability'—individually. Each item is analyzed in detail to determine how different user experience factors impact the overall effectiveness of training.

a: Impact on Tool Inclusion Recommendation

In this analysis, we examine how user experience factors influence the likelihood of recommending using the tool for training purposes. The results indicate significant influences on 'flow' ($p = 0.048$) and 'usability' ($p < 0.001$). While the significance of 'flow' and 'usability' were anticipated based on mixed-effect results, a notable finding from the current analysis is the predominance of 'usability' over 'flow' in influencing the likelihood of recommending the tool. This outcome highlights the importance of both 'flow' and 'usability' in the adoption of this tool for training applications. The more substantial impact of 'usability' can be attributed to the intrinsic requirements of VR environments, where physical ease and interface usability are essential for prolonged engagement and effective learning. These environments necessitate that users feel physically comfortable to fully engage in and extract the educational value offered. Furthermore, the significant influence of 'usability' suggests that discomfort could detract from an otherwise immersive VR experience, thereby reducing the overall recommendation of the tool. The analysis also highlighted two notable areas of concern: the integration of 'emotion' and system 'responsiveness'. Although these coefficients were not statistically significant, they highlight important areas for potential improvement within the tool. The negative coefficient for 'emotion' ($\beta = -0.254$, $p = 0.071$) suggests that the emotional dynamics might not align with user expectations in professional training settings. This misalignment could detract from the overall engagement and educational effectiveness of the tool, indicating a need for enhanced emotional realism and responsiveness. Regarding system responsiveness, the analysis indicated a potential negative impact on the user experience due to perceptible delays in system interaction ($\beta = -0.107$). While these delays did not adversely impact tool inclusion recommendations, improving this aspect could significantly enhance user satisfaction and the perceived utility of the tool [55].

The factors of visual fidelity and response relevance show trends but not reaching statistical significance. 'visual fidelity' ($\beta = 0.257, p = 0.085$) suggests a moderately positive effect indicating that more realistic representations in avatar's appearance and movement might enhance the effectiveness. Similarly,'response relevance' ($\beta = 0.276$, $p = 0.194$) points to the importance of accurate and contextually appropriate responses, which may improve user trust and reliance on the tool.

b: Impact on Training Comparability

This analysis assesses the effectiveness of interactions with a child avatar relative to human actors ('training comparability') and the result emphasizes the roles of 'visual fidelity' and 'response relevance' in this aspect. The significant impact of 'visual fidelity' ($\beta = 0.311$, $p = 0.038$) shows that the realism of the avatar, including aspects like detailed facial expressions and environment authenticity, markedly enhances the learning experience by mimicking real-life interactions. More impactful, however, is the factor of 'response relevance' ($\beta = 0.428$, $p = 0.004$), which stood out as a stronger predictor of training effectiveness. The result underscores the importance of the accuracy and relevance of the avatar's responses in ensuring that the training is perceived as being on par with human-facilitated training. High fidelity in responses ensures that the trainees receive meaningful feedback on their interviewing techniques, which is crucial for learning and improving their skills. The significant influence of response relevance highlights its essential role in simulating realistic and productive training scenarios that can rival interactions with human trainers. However, the non-significant results concerning 'age-appropriate responses' ($\beta = -0.097$, $p = 0.281$) may reflect that the appropriateness of the avatar's responses might not meet the expected standards. In training environments, it is particularly important that the avatar's responses reflect age-appropriate cognition and emotional maturity. This alignment not only adds to the realism but also better prepares trainees for real-life interactions with children. Enhancing this aspect could help simulate more realistic and educationally valuable scenarios, thereby improving the overall training comparability to human actors. The emotion factor ($\beta = -0.057$, $p = 0.753$) did not significantly influence the perceived 'training comparability.' This suggests that the current emotional modeling within the avatar system falls short of simulating the complex emotional cues exhibited by human actors. This consistency with previous findings highlights a critical area for improvement in avatar design. Enhancing emotional intelligence in virtual systems is essential to supporting the empathetic and contextually sensitive responses expected in human interactions.

VI. INTERVIEW ANALYSIS

We examined user feedback from the interviews to identify recurring themes and patterns in participant responses. The respondents' responses to the questionnaire (see Table 1) were categorized into four main themes: user preferences, visual fidelity, dialog systems, and training usefulness. Each theme was further broken down into sub-themes to capture specific aspects of the user experience. As illustrated in Figure 5, these themes and sub-themes were identified based on the frequency of mentions and their significance in the overall user experience, offering a clear overview of the areas most emphasized by the respondents during the interviews. Each of these themes and sub-themes is reviewed in detail below.

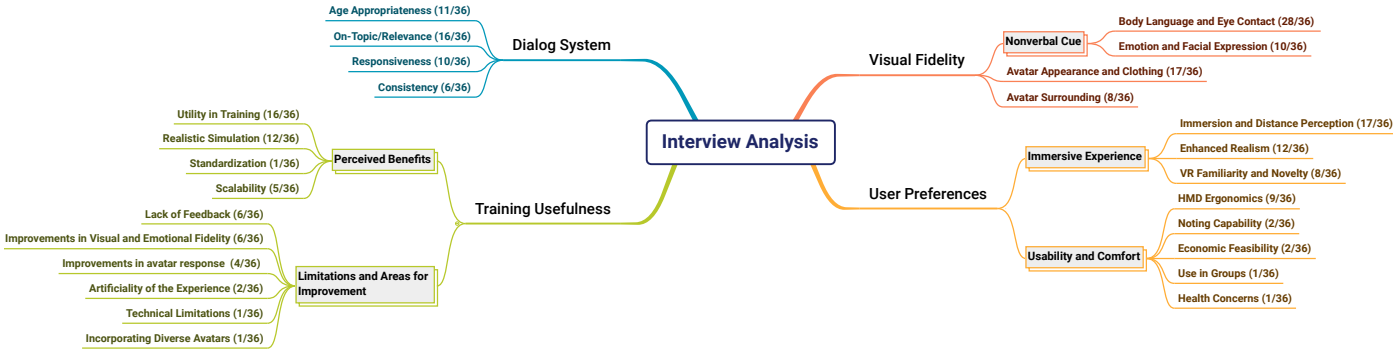


FIGURE 5. Overview of Interview Analysis: Breakdown of key themes with participant counts (x/36), where 'x' represents the number of participants who emphasized each theme out of a total of 36.

A. VISUAL FIDELITY

Visual fidelity is an important component in the effectiveness of training tools. It encompasses various aspects such as the realism of the movement and expression, the avatar's appearance, and the fidelity of the virtual environment. Following this, each of these aspects is analyzed based on the user's viewpoint.

1) Nonverbal Cue

Nonverbal cues are essential not only for improving realism but also for fulfilling the psychological and educational goals of the training process. Participants expressed varied perspectives on the nonverbal cues of the virtual avatar, particularly focusing on body language, eye contact, and facial expression, which significantly impact their experience when interacting with the avatar.

a: Body Language and Eye Contact

According to participants, effective body language and eye contact serve not just as signals of attention but as essential tools for understanding and interpreting behavioral cues.

Positive feedback highlighted instances where these aspects enhanced the realism and engagement in the interactive sessions. One participant noted, *"the body language was really good, like the way she moves her little hands, and like, looks directly at you and stuff like that. That was super cool"*. Another participant appreciated the interactivity, stating, *"I also like in some parts, the avatar kind of look at you when with their responses, then there was also like, it's more interactive"*.

However, several participants identified areas for improvement in the avatar's movement. The most common critique involved the unnatural or repetitive movements of the avatar's hands, which some found distracting. One participant remarked, *"I felt like I wanted her body to shift around"*. Another echoed this sentiment, *"her hands didn't rest upon her knees, they just floated on the button all the time. And I think that's something that was constantly kind of, I think, plant the hands, technically animation was hands on"*. Another participant observed, *"The hands just float awkwardly, almost as if playing a piano"*. Additionally, some participants noted the robotic nature of the avatar's movements, describing them as *"a bit surprised"* and *"not too bad,"* but still noticeably artificial.

Suggestions for enhancing the avatar's movement included more natural hand placements, dynamic movements, and consistent eye contact. One participant recommended, *"add some movement to the lower body, some actual displacement in the chair, you know, crossing legs turning away a bit more whole bodily, rather than just the torso"*. This feedback aligns with the need for the avatar to exhibit more realistic and relatable gestures to improve the overall training experience.

Furthermore, the absence of adequate eye contact was frequently highlighted, with participants expressing dissatisfaction and a feeling of disconnection. For instance, one participant mentioned, *"I just thought I felt frustrated that she wasn't looking at me."* This comment reflects the widespread view that eye contact is important to fostering an engaging, responsive interaction environment and establishing a meaningful communicative connection.

Despite these critiques, some participants perceived meaning from the avatar's movements and body language, interpreting them as a sign of the avatar's emotional state or comfort level. For instance, one participant noted, *"it seems like she was not comfortable... trying to say something... it was hard to know her expression... she was a bit scared, trying to hide something"*. Another participant mentioned that *"kids when they're nervous and not quite engaging, rather than look straight ahead of a wall, they'll focus on themselves fiddling with their fingers looking down until they're ready to respond"*.. Another participant noted that *"when you talk to him, she was looking over that side. That's normal for kids."*. This observation suggests that even imperfect body language can convey significant information, potentially adding depth to the interaction by prompting interviewers to consider the emotional and psychological state of the avatar. Another participant noted that such nonverbal cues, even when not perfectly executed, align with real-life scenarios where children might exhibit nervous or guarded behavior, enhancing the authenticity of the training sessions.

The feedback suggests that consistent and realistic natural body movements, along with eye contact, are crucial for maintaining user engagement and enhancing the perceived realism of the avatar. Improvements in more natural body language, as well as the avatar's eye-tracking and gaze-following capabilities, could significantly enhance the training experience by making interactions feel more natural and responsive.

b: Emotion and Facial Expression

In the analysis of the avatar's emotion and facial expression, several participants identified areas for improvement in the avatar's emotional expressions. A common critique was the lack of strong facial expressions, which some participants found diminished the immersive experience. One participant noted, *"the expressions weren't very strong,"* while another mentioned, *"It would be better if she showed a bit more emotion."* The absence of adequate emotional expression often led to a sense of disconnection and frustration. Additionally, some participants felt that the avatar's emotional responses were too generic and lacked dynamism, making the interaction feel less spontaneous. One participant remarked that the visuals were *"a little generic,"* implying that the avatar's facial expressions did not adequately convey nuanced emotions. Suggestions for enhancement included better synchronization of facial movements with speech and a wider range of emotions.

2) Avatar Appearance and Clothing

In the analysis of the avatar's appearance and clothing, participants raised significant concerns, crucially influencing the authenticity and effectiveness of the interactions. These centered around the avatar's physical proportions and attire, impacting the realism expected in a child avatar designed for training purposes. The avatar's proportions were frequently cited as inappropriate for its intended age, particularly noting features suggestive of maturity, such as developed breasts, which are atypical for a child aged six to eight. One participant pointed out, *"So, I just felt it ended up she's not a six-year-old, just got boobs. Right. So, it's really not relatable at all in terms of who the character is in front of me."* This discrepancy could lead trainees to interact with the avatar as they would with an older individual, thus skewing the training dynamics intended for child interviews.

Furthermore, the avatar's clothing also drew criticism for not aligning with typical children's attire, skewing more towards teenage or adult styles. *"She looks to me more like a teenager than an eight-year-old girl... Kids just don't wear jeans like that, and the V-neck is not appropriate,"* remarked a participant. To address this, suggested improvements include adopting more age-appropriate, neutral clothing that avoids misinterpretation and maintains focus on the training content. As one participant recommended, *"Considering more playful and colorful clothing, akin to what actual eight-year-olds might wear, such as graphic tees or simple patterns, could help make the avatar appear more childlike and less mature."*

The spatial arrangement and the avatar's placement within the virtual environment also diminished the realism of interactions. Descriptions such as *"like a dwarf sitting on a huge chair"* and feedback like *"The big discrepancy in size... makes a difference. Like I feel like I'm just looking at something quite little... it kind of takes away from the feeling of interacting with a person,"* underscored how scale and positioning critically impacted the immersive experience.

These insights from participants underscore the importance of designing avatars that accurately reflect the physical and stylistic attributes of the age group they intend to represent.

3) Avatar Surrounding

In the analysis of the virtual surroundings of the avatar system, most participants described the setting as resembling *"a police interview room,"* terms that reflect the sterile and impersonal nature of the environment. One participant described the setting as overly formal and intimidating for children, stating, *"The environment feels very clinical, like something you'd expect in a therapy session for adults rather than children."* Another confirmed this sentiment, stating, *"It looks like a police interview room."*

Conversely, few participants expressed satisfaction with the simplicity and minimalism of the environment, which contributed to maintaining focus during interviews. As one participant described, *"The uncluttered nature of the environment helps maintain focus on the interview content, which is beneficial for gathering accurate information."* The setting was designed based on the pictures we received from actual interviewing rooms to ensure realism. However, the feedback suggests a pressing need to redesign the setting to foster a more welcoming and less daunting atmosphere. Recommendations for such a redesign include incorporating elements that are child-friendly and less formal, potentially enhancing a more open and productive dialogue.

B. DIALOG SYSTEM

In evaluating the avatar's dialogue model, we focused on several key criteria critical to its effectiveness and utility: age appropriateness, on-topic relevance, responsiveness/delay and consistency. User feedback provided valuable insights into how well the avatar performed across these dimensions.

1) Age Appropriateness

Eight of the participants who talked about this topic in their interviews believed that the responses were suitable for a child's level of understanding and articulation. There was a consensus that the child seemed to speak and respond in ways typical for her age, though sometimes there were miscommunications or misunderstandings, attributed to age-appropriate cognitive and language development. Some participants noted the child's answers were not always directly on-topic, reflecting a natural conversational pattern for a young child. This indirect response style was mostly seen as appropriate, considering the child's developmental stage.

When discussing more sensitive or complex topics (like body parts), participants noted the responses as generally appropriate but sometimes lacking in detail, which could be reflective of a typical child's knowledge or comfort discussing such topics. One of the participants responded, *"Exactly what I would expect a little girl saying"*. On contrast, one of the participant responded, *"I mean, you know, again, I'm not sure how old she is. So, but she could articulate really well what happened and how she felt about it. So that, to me is saying that she must be a little bit older, because a younger child might just be, you know, not be able to articulate what happened as easily as she did."* The negative feedback from two highlights concerns about the authenticity and emotional depth of the responses given by the child in the user study. One of the participants said, *"... emotional or the emotive language, I think, is some what lacking .. and she could articulate that normally would have like, a bit more emotive language ... you know, really upsets me and makes me feel sick in my tummy when dad yells at me and sends me to my room."* Lammerse et al. [56] has proposed the framework for modelling the child's emotion in investigative interviews for potential abuse, this work has not been integrated into avatar tool yet.

The participants' feedback highlights that while the child's responses were generally age-appropriate, there are areas for improvement in how informative the responses should be for certain young ages, vocabulary and adding the emotional depth to these responses.

2) On-Topic/Relevance

In the context of user experience interviews, a on-topic and relevant response is defined as an answer that directly addresses the interviewer's question.The participants' feedback indicates a mix of positive and negative experiences with the on-topic and relevance of responses. 11 participants generally found the responses appropriate and informative, appreciating the detail and alignment with their questions. They noted that with repeated questioning, responses became more accurate and provided valuable insights. One participant stated, *"They were appropriate. They were related to what I was asking. [...] You need to repeatedly ask them, so I feel it was like, real."* While five participants reported issues with relevance, clarity, and highlighting areas where the responses did not meet their expectations or needs. One of participants stated, *"I think they were unappropriate. Like, there's definitely some really close kinds of answers that she would expect. Okay. But it was just more. Yeah, I don't think it answered some of the times like, particularly thing around the family,.."*

This highlights the need for refining dialogue model to better match the expectations and requirements of different interview contexts.

3) Responsiveness

Responsiveness refers to delay between a user's question and the avatar's response. The interviews with participants regarding their experience with responsiveness and delay in the interaction highlighted several key insights. The feedback was mixed, with some participants finding the delay acceptable and even realistic, while others found it frustrating and disruptive. Some participants found the delay realistic, reflecting real-life scenarios where children do not always respond immediately. This was seen as a positive aspect by a few, as it added to the authenticity of the interaction. A quote from one of the participant's interview, *"I enjoy that the Avatar had, like some sort of delay with the responses. So but that's okay. Because that simulates as well, like children. Not all children, like responding immediately to your questions, some of them take some time."*

Conversely, many participants experienced frustration due to the delays, which they felt interrupted the flow of conversation and made it challenging to maintain engagement or determine if the system had understood their input. They compared the interaction with the system to real-life conversations with children, noting that while children do take time to respond, but the delays in the system felt longer and more artificial, which affected their overall experience. One of the participant stated, *"Sometimes it was too late. Where I wasn't sure, you know whether I spoke too softly. Or, you know, like, sometimes no response came for quite a while. So not sure whether it's, you know, it's what the child would know or not."* and *A little awkward. Yeah, it's delayed. And asking the questions and getting the response so that I know whether it's wanting more from me or if it's, perhaps that's the standard response that the avatar is going to give."* .

Addressing the delay in responsiveness is crucial for enhancing user experience in avatar interactions. Most delays stem from speech synthesis processes like TTS and STT, often resulting in frustration and interruptions. Transitioning to local speech synthesis and recognition models instead of relying solely on cloud services could potentially mitigate these delays, offering quicker responses and smoother interactions for users.

4) Consistency

Consistency in refers to the coherence and reliability of avatar's responses generated by the model across different questions. A consistent dialogue model produces responses that align with its previous statements, are logical and plausible within the given conversational context, without contradicting itself. In the post-experience interviews, six participants highlighted inconsistencies in the responses given by the avatar. These inconsistencies ranged from contradictory answers to the same question to responses that did not logically follow from the questions asked. One participant mentioned a case where a child included her dad when asked about safe people, but later mentioned her mom when the participant asked a different question and another Another participant noted a contradictory response regarding a special occasion gift. The child initially denied it was for a birthday but later confirmed it was. One participants observed that the child avatar initially denied having siblings but later

mentioned living with a little brother. Participants associated this inconsistency with the lack of ability to understand the term "siblings" by the avatar. Participant stated, *"And then she also made the mistake when I asked her, Do you have siblings? And she said, No. And I said, Who do you live with? Our mum and his little brother? Yeah."*

The study revealed several inconsistencies in the system's responses, ranging from contradictory answers to different handling of similar questions.Although frequency of these inconsistencies is not high but addressing these inconsistencies is crucial for improving the system's reliability and user trust. Thorough analysis of transcripts of participants' interactions with the avatar will provide deeper insights into these issues, guiding targeted improvements to enhance the system's performance and ensure more consistent and accurate responses.

C. USER PREFERENCES

In evaluating user preferences between 2D and 3D virtual environments, participants were asked to elaborate on their preferred environment. 16.7% of participants valued both environments. One participant noted, *"I think it depends on the situation. I would prefer both gradually. The screen one is just an introduction, and then the VR would be like my final test."* Another participant confirmed this sentiment of equivalence: *"The experience wasn't terribly different between the two. I could easily screen out and focus on the child. To me, it's quite similar; I really don't have a sense that the VR was super different or super better."*

However, a majority of 61.1% favored the 3D environment, citing enhanced realism and immersion as key factors. Meanwhile, 22.2% opted for the 2D environments, preferring their simplicity and ease of use. Following is a detailed analysis of the reasons behind users' preferences across different themes.

1) Immersive Experience

a: Immersion and Distance Perception

Participants in 3D environments reported a profound sense of flow, often becoming so engaged that they lost awareness of their actual surroundings, as one explained: *"I was totally focused on what's happening; it felt like everything else in the room faded away."* This level of immersion might be beneficial in scenarios that demand high concentration and empathetic engagement, such as interviews or complex training sessions. Another participant reflected on the realism of the experience: *"It felt like I was actually in the room, dealing with a real situation, which heightened the perceived stakes and intensified the experience."* Another participant vividly described the immersive quality of 3D environments, saying, *"It was really incredible because I sort of forgot you guys were there, and I felt like I was just in the room with this girl."* In stark contrast, participants interacting with 2D setups reported frequent distractions and a diminished sense of presence, highlighting the limitations of less immersive technologies.

b: Enhanced Realism

The enhanced realism extends beyond mere visual fidelity; it includes spatial awareness and the dynamic nature of interactions, where the movements and gestures of the avatars are perceived with greater clarity and immediacy. For example, one participant remarked, *"It just felt a lot more like I was interacting with a person in a more realistic way, the way I would in a therapeutic space. 3D is a lot more realistic; it feels like you're actually in a space with the avatar, rather than just looking at a computer screen."* Such feedback highlights the capability of 3D VR to closely mimic real-life interactions, which is essential for effective training in fields requiring nuanced human interaction. Additionally, the ability of 3D settings to convey subtle non-verbal cues like eye contact and body language enhances the communication process, a feature often diminished in 2D environments. This enhancement was valued by participants, with one stating, *"I could really pick up on a lot more of the hand movements and eye contact and things that I didn't notice as much with the other ones."*

c: VR Familiarity and Novelty

Some participants clearly favored the 2D environment due to its resemblance to traditional computer interfaces, which many are already accustomed to from daily use at work or home. They found it more approachable and less intimidating compared to the immersive 3D VR settings. This familiarity reduces the initial barrier to entry, allowing users to engage with the training material more quickly without the distraction of navigating a new technological interface. This tendency underscores a significant aspect of technology adoption—familiarity often enhances comfort and reduces the cognitive load associated with learning new systems. One participant explicitly highlighted this by stating, *"To be honest, I think I'm probably supposed to say that the 3D was better because it's real and everything, but this one 2D just felt more like things I've done before."*. This statement reflects a reluctance to embrace newer technologies and reveals a generational divide in technology adoption, as further emphasized by their comment, *"Maybe I'm just too old for it."* Another participant reinforced this sentiment by preferring the known comforts of existing technology: *"You probably need to make sure we have a good mic and headphones. I kind of just prefer this computer experience where it's familiar territory for me."*

Conversely, the novelty of the 3D virtual environments markedly enhanced participant engagement. Those new to VR often described the immersive experience as *"extraordinary"* and *"exciting,"* distinctly different from conventional computer interfaces. One participant noted *"how the real world went away,"* indicating a deep level of engagement. Another vividly expressed, *"It's like stepping into another world where everything you know is replaced by what you see and feel here."* This novelty appeared to enhance receptiveness to learning, with participants stating that the innovative aspect made the training feel more promising, de-

spite any initial hesitations. However, the novelty diminishes after repeated exposure, and subsequent experiences may not evoke the same *"extraordinary"* reaction. Such enthusiasm for new technology underlines how the novel elements of VR can impact user motivation and openness, particularly when engaging with learning experiences.

2) Usability and Comfort

This subsection examines the factors of usability and comfort that influence participant preferences in 2D and 3D training environments. Key aspects such as the ergonomics of HMDs, user familiarity with technology, the capability for note-taking, economic feasibility, suitability for group use, and health concerns are discussed in detail.

a: HMD Ergonomics

The physical comfort and ergonomics of HMD significantly impact user acceptance and the effectiveness of learning technologies. Physical attributes such as headset weight, the fit around the head and over the eyes, and the integration with other personal accessories like glasses can deeply affect the user experience [57]. In our study, participants frequently expressed concerns over the discomfort caused by VR headsets, noting the awkwardness of the fit. The weight and bulkiness of the headsets often led to fatigue and distraction, undermining the immersive potential of VR. One participant explicitly highlighted this issue, stating, *"I was aware all the time that I had this heavy thing hanging on my head over glasses. It just wasn't comfortable."*. Another participant lamented the poor fit, remarking, *"The headset didn't fit well. It's not just about the VR; it's about making sure it can be worn comfortably by everyone,"* pointing to the need for designs that accommodate diverse users. Additionally, the concern about physical strain was expressed by another, who mentioned, *"It just the equipment is actually heavy and puts a lot of pressure on my forehead. So I started getting headaches after a while"* illustrating the discomfort that can detract from the learning experience.

b: Noting Capability

The capability to take notes during training sessions was highlighted as a limitation in the 3D VR environment. Participants felt that the immersive setting, while beneficial for certain aspects of learning, hindered their ability to perform simple but crucial tasks such as jotting down notes, a common practice in traditional learning and interview scenarios. One participant clearly articulated this concern, stating, *"While I loved the immersiveness of the 3D version, it made taking notes impossible."* Participants suggested that the ability to easily take notes as they would in traditional learning environments remains a critical feature, particularly in professional training contexts where details matter and are often revisited post-session. This feedback points to the need for integrating functionalities within VR platforms that can accommodate note-taking, possibly through voice-to-text features or virtual notepads.

c: Economic Feasibility

Furthermore, economic feasibility has been raised as a another considerable concern regarding the use of VR technology for training purposes, particularly when compared to more conventional computer-based systems. Participants noted the higher costs and logistical challenges associated with VR setups. One participant emphasized this point by stating, *"It's a lot more economically doable to just have a software program that people can work from the computer, rather than being like, we've got this program via your Oculus headset."*. They further expressed concerns about the procurement and maintenance of VR equipment, mentioning, *"The VR would obviously be more expensive, more cumbersome in training."*. This feedback underscores the need for a cost-effective balance in training tools, where the benefits of advanced VR technology must be weighed against their economic and operational implications to ensure broad and sustainable adoption.

d: Use in Groups

In addition, the use of VR in group settings elicited a particular viewpoint about its effectiveness for collaborative or observational learning. While the immersive nature of VR provides a unique personal experience, it may not suit group interactions as effectively as 2D environments. One participant noted the potential of the 3D setting for individual use but questioned its utility in a group context, stating, *"3D is very individual, whereas I think this could be done with someone observing or providing cues, or it could be done even in a group setting."* While VR can indeed be interactive and allows both users to wear headsets and enter the same scene, or one to observe the other, this feedback highlights the perception that VR might still feel more isolating compared to the more open and easily shared 2D environments. This implies that although VR can facilitate interaction, the inclusive and collective dynamics of group learning environments might be better supported by 2D settings, where immediate visual and verbal feedback is more effectively integrated for all users.

e: Health Concerns

Health concerns were raised as a consideration for the adoption of VR headsets in training environments. Participants expressed worries about the cleanliness of shared equipment, a concern that has become more pronounced in the wake of recent global health events such as the COVID-19 pandemic. One participant mentioned, *"I was worried about the possibility of catching germs from the VR gear. I noticed the person who used it before me had a bad cold."* Additionally, the potential for cosmetic residues, such as sunscreen and makeup, to be left on the VR goggles was highlighted. These comments confirm the importance of maintaining strict hygiene protocols. To address these concerns, we implemented thorough cleaning procedures between uses and provided individual sanitizing wipes, ensuring the safety and comfort of all users when utilizing VR technology for training purposes.

D. TRAINING USEFULNESS

Main purpose of this user study aimed to evaluate the effectiveness of an avatar training tool designed to simulate realistic interview experiences with avatars. The feedback gathered from participants on the use of a training tool was primarily centered around its potential utility, realism, and areas needing improvement. 85% of participants opinioned that this avatar training tool can be useful. 15% of raised their concerns and did not think it is ready to be used for training.

1) Perceived Benefits

Participants found the avatar extremely useful for training purposes and highlighted various aspect which they liked.

a: Utility in Training

The tool was seen as particularly valuable for practicing interview techniques in a controlled environment. One participant stated, *"I can see a lot of applications for this, particularly with the reputation coppers or whoever had to interview kids, but also like, psychologists would be great."*. Many participants mentioned the value of the tool as a starting guide to get used to asking questions and as a way to practice before encountering real-life situations. This hands-on practice was seen as essential for building confidence and competence. *"It would be useful as kind of like a starting guide to get used to asking or phrasing questions, because that's always one of the things that is a bit challenging"*, remarked one participant. Also, it was noted by a participant for providing a less pressurized environment for practice, which could be particularly beneficial for students and new trainees. Participants stated, *"Yeah, it certainly allows you space to investigate. So yeah. Okay. Less pressurized environment."*

Overall, most participants thought the avatar training tool could really help with practicing interviews, build confidence and competence in interviewing techniques, especially for students and new trainees.

b: Realistic Simulation

Participants appreciated its ability to simulate realistic interview scenarios without the risk of causing trauma to real children or vulnerable individuals. One participant stated,*"You can create, you know, someone that looks and acts and answers as a child, and you can simulate that without the risk of actually, you know, messing up and traumatizing a child."* The tool was praised for simulating realistic scenarios where interviewees might not respond as expected, which helped users practice and refine their questioning techniques. A participant noted,*"You have to come up with more questions to try and get the child to answer. So that was pretty realistic for training."* Also, One of the participant highlighted the fact that it is hard to find a competent actor to play the role of a child for the training purposes and *"It just simulates a realistic experience without it's hard to find someone that can kind of that has the time or is willing to do an acting role and can actually really embody that role."*

The positive feedback underscores the tool's utility in enhancing the interview skills of users, making it a valuable asset in training programs that require nuanced and sensitive interaction techniques.

c: Standardization

The ability to deliver a consistent training package was highlighted as a significant advantage, ensuring all trainees are assessed on the same standards. A participant remarked,*"You have a consistent, consistent training package that can be delivered."*. Avatar provides provides a consistent training package that can be delivered repeatedly, ensuring all trainees receive the same level of training, and it does not have to depend on the quality of instructor.

d: Scalability

Scalability in the context of this user study refers to how easily this avatar training program can be deployed and scaled to a large number of users. The feedback from participants highlights various aspects of the scalability, including its usefulness and cost-effectiveness. Participants appreciated the flexibility of avatar training tool, allowing them to practice at any time without needing an instructor present. This aspect enhances the scalability as it removes the constraint of scheduling and availability of trainers. A participant stated that,*"I could practice this at two o'clock in the morning, like, I don't need someone right there in front of me."* The potential to reduce costs by not needing actors was seen as a major benefit. Once the system is set up, it can be used by multiple people without additional recurring costs. A participant stated that, *"... there's a cost to setting it up at the beginning, but then not having to hire an actor every time and that sort of thing. And being able to do it with multiple people at once."* The positive feedback suggests strong support for the adoption and deployment of avatar training on a larger scale.

2) Limitations and Areas for Improvement

Although participants generally found the tool valuable for training, they also provided constructive feedback and suggestions to enhance its effectiveness and user experience.

a: Lack of Feedback

Several participants noted that the lack of adequate feedback on their performance. This absence of feedback makes it challenging for them to evaluate and improve their interviewing techniques effectively. Feedback, at the end of practice session, is an important part of learning in developing these interviewing skills [58]. Quoting one of the participants, *"Well, for the training, I'm getting no feedback on my interview techniques. So from that perspective, what have I actually done? "*

b: Improvements in Visual and Emotional Fidelity

Participants pointed out several areas where the visual and emotional fidelity of the avatar could be enhanced to improve

the training experience. They emphasized that while the avatars performed adequately in basic interactions, their lack of emotional expressions and realistic in the visual elements detracted from the overall immersion and effectiveness of the training.

One participant suggested that *"improving the virtual reality experience with higher graphics and more realistic environments"* would significantly enhance the sense of presence during interactions. Similarly, another participant noted that the avatar's emotional responses were lacking, stating, *"If there was a way to make her start to cry, or get angry, or something, that would be good to test the person's skills to calm them down."* This sentiment was echoed by another participant, who proposed the addition of visual cues to guide interviewers during the training, which could further enhance the learning experience by providing real-time feedback on their performance. These suggestions collectively demonstrate the need for enhancing both the visual realism and emotional responsiveness of the avatar to create a more effective and engaging training tool.

c: Improvements in avatar response

Several participants mentioned specific areas where the avatar's responses could be enhanced to improve the overall training experience. One participant suggested the importance of measuring the proportion of information provided by the avatar to better assess the interaction's completeness and realism. They noted, *"When we interact with people, we start making up things and we have an infinite number of things that we can make. But obviously, when you program an avatar, you have some limited number of details."* Another participant emphasized the need to address lag time, which affected the flow of conversation and the perceived responsiveness of the avatar. They expressed frustration, stating, *"That lag time was frustrating. I'm waiting for acknowledgment that they need to hear what I've said."* Similarly, a participant recommended incorporating a variety of responses, including random questions from the avatar, to better mimic real child behavior. They observed, *"Often kids will ask you questions, or they'll throw out a random question that has nothing to do with your interview."* Another suggestion was that the avatar should include more personal and innocuous statements to build rapport before discussing serious topics. The feedback included, *"Having a good mix of some statements that you know, because people will always ask that with a student."* As a result, these observations suggest the need for more dynamic, responsive, and contextually rich interactions to enhance the training tool's effectiveness.

d: Artificiality of the Experience

A few participants expressed a preference for training with real human beings, citing the lack of emotional connection and the immersive experience was not convincing enough to replace real human interactions. Participants stated that *"I'd probably rather a real human being okay. Personally, I wouldn't find it useful."*

e: Technical Limitations

Participant raised issues with the tool's ability to accurately pick up what has been asked which ultimately affect the quality of the responses and training quality. It will also affect the quality of the feedback a user will get at the end of the training. Participant stated, *"If you're not picking that [words] up correctly [...] I don't see the value on this to be honest.* This issue can be associated with tool's inability to transcribe certain accents. While the avatar training tool shows promise as a consistent and practical training aid, participants highlighted several areas for improvement. Enhancing the emotional realism of interactions, addressing technical limitations, and providing more robust feedback mechanisms will be crucial for maximizing the tool's effectiveness.

f: Incorporating Diverse Avatars

One participant commented on the potential benefit of incorporating a wider range of avatars, including both male and female characters, to enhance the realism and applicability of the training tool. They commented, *"Are you only using girl avatars? It would be good to also have a boy avatar, just you know, for the variety of perhaps for a future product."* Additionally, incorporating both girl and boy avatars helps to avoid reinforcing gender stereotypes by ensuring that the training scenarios do not inadvertently suggest that certain behaviors or situations are gender-specific [59]. By integrating diverse avatars, the training tool can offer better preparing users for the range of situations they might encounter in actual investigative interviews and promoting a more balanced and unbiased approach to child interviewing.

VII. DISCUSSION AND FUTURE DIRECTIONS

This study aimed to evaluate the effectiveness of an AI-driven training platform that integrates 2D and 3D virtual environments to train child welfare and law enforcement professionals in conducting investigative interviews with maltreated children. To enhance the robustness of our findings, we incorporated qualitative data from semi-structured interviews alongside our quantitative analyses. These qualitative insights provided a deeper understanding of participants' experiences and perceptions, complementing the statistical results. Despite the ethical and logistical constraints that limited our sample size, we ensured a diverse participant pool, including both experienced professionals and individuals with academic qualifications in relevant disciplines. This diversity enhances the generalizability of our findings, ensuring that the results are reflective of a broad range of perspectives within the field.

The findings demonstrated that the 3D virtual environment significantly heightened user immersion and sense of flow compared to its 2D counterpart. Interestingly, in terms of recommending the inclusion of this tool in training, there was no meaningful difference between the two environments, though there was a tendency towards favoring the 3D environment because some users preferred the 2D environment

due to greater usability and the discomfort associated with VR HMDs. This preference was reflected in the interviews, where 61% of participants expressed a preference for the 3D environment.

Notably, SEM analysis revealed that both 'flow' and 'usability' significantly influenced the decision to include the tool in training, with usability exerting a stronger influence. The usability challenges were so significant that they outweighed the immersive experience of the 3D environment. Despite many users remarking on the impressive sense of immersion, the discomfort and complexity of the VR setup highlighted the critical importance of user comfort and simple interfaces in the adoption of new training technologies.

Additionally, participants with previous VR experience reported enhanced 'avatar comfort interaction' levels, suggesting that familiarity with VR technology can alleviate some of the initial usability challenges and discomfort linked to using VR headsets. Qualitative feedback further illuminated how the novelty of VR contributes positively to user comfort and engagement. Additionally, most of the criticism regarding the discomfort of the headset came from older users who were less familiar with this new technology. This highlights the importance of considering user demographics when implementing new training tools. The detailed and contextually appropriate responses of the avatars played a crucial role in empathy elicitation, often outweighing the importance of visual fidelity. This finding underscores the importance of continuous improvements in the AI-driven dialogue system to ensure realistic and contextually appropriate interactions.

Furthermore, both qualitative and quantitative results highlighted significant areas for improvement in the avatars' emotional reaction and visual fidelity, two key components of the training tool's overall effectiveness. Participants' feedback revealed that the avatars' facial expressions lacked the emotional depth necessary to fully engage and connect with users, particularly in situations requiring empathetic responses. While the avatars could exhibit basic emotional cues, these were often too generic and failed to convey appropriate emotional states for various scenarios. This misalignment between the avatars' expressions and the conversational context often led to a disconnect, reducing the overall realism and impact of the training. Quantitative analysis supported these observations, indicating that 'emotion' in facial expressions and 'visual fidelity' were factors that negatively impacted the tool's effectiveness. Enhancing these expressions to include a broader range of emotions and more fluid, lifelike movements, along with age-appropriate representation and realistic environments, would significantly improve the perceived realism and effectiveness of the training tool.

Despite these criticisms, some user expectations about the technology may not be entirely realistic. For instance, one participant expected *"far more furtive glances"* and acknowledged having preconceived notions about abuse survivors. Another user noted, *"if it was a real child, they might sit forward and maybe roll their eyes,"* highlighting the challenge of replicating complex human behaviors in virtual avatars. Recognizing the importance of emotional dynamics and visual fidelity in virtual environments, future iterations of the system will integrate Unreal Engine technology with NVIDIA Omniverse[1] to enhance emotional reactions. This integration is expected to result in more sophisticated simulations of human-like emotional responses, enriching the user's immersive experience by providing a nuanced and context-sensitive rendering of non-verbal cues. These advancements will significantly enhance the development of more empathetic and interactive virtual environments, ultimately providing a more effective training tool for professionals in child welfare and law enforcement.

VIII. CONCLUSION

This study provides an in-depth evaluation of the effectiveness of AI-driven virtual environments for training professionals in child interview skills. Through comprehensive analysis, we have demonstrated the significant benefits and limitations of both 2D and 3D environments in enhancing the training experience. The findings indicate that while 3D environments offer superior 'flow' and 'visual fidelity', they are also associated with increased 'usability' challenges due to the hardware requirements. In contrast, the 2D environment was found to be more user-friendly and less physically demanding, making it preferable for certain users.

Furthermore, the study highlights that familiarity with VR technology positively influences user perception, mitigating some of the discomfort associated with VR head-mounted displays. The dialog system's effectiveness, particularly the relevance and detail of responses, plays a crucial role in empathy elicitation, often outweighing visual fidelity. However, areas such as 'emotion' in facial expression and system 'responsiveness' need improvement to enhance the tool's effectiveness. As user experience increased, participants had a more positive view of the age-appropriateness of responses. This research contributes valuable insights into the development and application of AI-driven training tools, ultimately aiming to improve the effectiveness of investigative interviews with maltreated children by leveraging the strengths of both 2D and 3D virtual environments.

IX. ACKNOWLEDGEMENT

We would like to thank M. Cayetana López Cano for her assistance in conducting our studies in Australia.

REFERENCES

[1] Marije Stoltenborgh, Marian J. Bakermans-Kranenburg, Marinus H. van IJzendoorn, and Lenneke R. Alink. Cultural–geographical differences in the occurrence of child physical abuse? a meta-analysis of global prevalence. International Journal of Psychology, 48(2):81–94, 2013.

[2] Wendy A Walsh, Lisa M Jones, Theodore P Cross, and Tonya Lippert. Prosecuting child sexual abuse: The importance of evidence type. Crime & Delinquency, 56(3):436–454, 2010.

[3] Mark D. Everson and Jose Miguel Sandoval. Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements. Child Abuse & Neglect, 35(4):287–298, 2011.

[1] https://www.nvidia.com/en-us/omniverse/apps/audio2face/

[4] C.J. Brainerd and V.F. Reyna. Reliability of children's testimony in the era of developmental reversals. Developmental review, 32(3):224–267, 2012.

[5] Sonja P. Brubacher, Mairi S. Benson, Martine B. Powell, Jane Goodman-Delahunty, and Nina J. Westera. Chapter twenty-two - an overview of best practice investigative interviewing of child witnesses of sexual assault. In India Bryce and Wayne Petherick, editors, Child Sexual Abuse, pages 445–466. Academic Press, 2020.

[6] Miriam Johnson, Svein Magnussen, Christian Thoresen, Kyrre Lønnum, Lisa Victoria Burrell, and Annika Melinder. Best practice recommendations still fail to result in action: A national 10-year follow-up study of investigative interviews in csa cases. Applied Cognitive Psychology, 29(5):661–668, 2015.

[7] Gunn Astrid Baugerud, Miriam S Johnson, Helle BG Hansen, Svein Magnussen, and Michael E Lamb. Forensic interviews with preschool children: An analysis of extended interviews in norway (2015–2017). Applied cognitive psychology, 34(3):654–663, 2020.

[8] Gunn Astrid Baugerud, Ragnhild Klingenberg Røed, Helle BG Hansen, Julie Schøning Poulsen, and Miriam S Johnson. Evaluating child interviews conducted by child protective services workers and police investigators. The British Journal of Social Work, page bcac245, 2023.

[9] Ann-Christin Cederborg, Yael Orbach, Kathleen J Sternberg, and Michael E Lamb. Investigative interviews of child witnesses in sweden. Child abuse & neglect, 24(10):1355–1361, 2000.

[10] Misun Yi, Eunkyung Jo, and Michael E Lamb. Effects of the nichd protocol training on child investigative interview quality in korean police officers. Journal of police and criminal psychology, 31(2):155–163, 2016.

[11] Kathleen Coulborn Faller. Forty years of forensic interviewing of children suspected of sexual abuse, 1974–2014: Historical benchmarks. Social Sciences, 4(1):34–65, 2014.

[12] Martine B Powell, Ronald P Fisher, and Carolyn H Hughes-Scholes. The effect of intra-versus post-interview feedback during simulated practice interviews about child abuse. Child Abuse & Neglect, 32(2):213–227, 2008.

[13] Pegah Salehi, Syed Zohaib Hassan, Myrthe Lammerse, Saeed Shafiee Sabet, Ingvild Riiser, Ragnhild Klingenberg Røed, Miriam S Johnson, Vajira Thambawita, Steven A Hicks, Martine Powell, et al. Synthesizing a talking child avatar to train interviewers working with maltreated children. Big Data and Cognitive Computing, 6(2):62, 2022.

[14] Syed Zohaib Hassan, Saeed Shafiee Sabet, Michael Alexander Riegler, Gunn Astrid Baugerud, Hayley Ko, Pegah Salehi, Ragnhild Klingenberg Røed, Miriam Johnson, and Pål Halvorsen. Enhancing investigative interview training using a child avatar system: a comparative study of interactive environments. Scientific Reports, 13(1):20403, 2023.

[15] Pegah Salehi, Syed Zohaib Hassan, Gunn Astrid Baugerud, Martine Powell, M Cayetana López Cano, Miriam S Johnson, Ragnhild Klingenberg Røed, Dag Johansen, Saeed Shafiee Sabet, Michael A Riegler, et al. Immersive virtual reality in child interview skills training: A comparison of 2d and 3d environments. In Proceedings of the 16th International Workshop on Immersive Mixed and Virtual Environment Systems, pages 1–7, 2024.

[16] Kevin Charles Dalli. Technological acceptance of an avatar based interview training application: The development and technological acceptance study of the avbit application. Digitala Vetenskapliga Arkivet, 2021.

[17] Belinda Guadagno and Martine Powell. E-simulations for the purpose of training forensic (investigative) interviewers. In Professional education using e-simulations: Benefits of blended learning design, pages 71–86. IGI Global, 2012.

[18] Ragnhild Klingenberg Røed, Martine B Powell, Michael A Riegler, and Gunn Astrid Baugerud. A field assessment of child abuse investigators' engagement with a child-avatar to develop interviewing skills. Child Abuse & Neglect, 143:106324, 2023.

[19] Mairi S Benson and Martine B Powell. Evaluation of a comprehensive interactive training system for investigative interviewers of children. Psychology, Public Policy, and Law, 21(3):309, 2015.

[20] Martine B Powell, Belinda Guadagno, and Mairi Benson. Improving child investigative interviewer performance through computer-based learning activities. Policing and Society, 26(4):365–374, 2016.

[21] Francesco Pompedda. Training in investigative interviews of children: Serious gaming paired with feedback improves interview quality. 2018.

[22] Niels Krause, Francesco Pompedda, Jan Antfolk, Angelo Zappalá, and Pekka Santtila. The effects of feedback and reflection on the questioning style of untrained interviewers in simulated child sexual abuse interviews. Applied Cognitive Psychology, 31(2):187–198, 2017.

[23] Francesco Pompedda, Angelo Zappalà, and Pekka Santtila. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. Psychology, Crime & Law, 21(1):28–52, 2015.

[24] Shumpei Haginoya, Shota Yamamoto, Francesco Pompedda, Makiko Naka, Jan Antfolk, and Pekka Santtila. Online simulation training of child sexual abuse interviews with feedback improves interview quality in japanese university students. Frontiers in psychology, 11:998, 2020.

[25] Niels Krause, Elsa Gewehr, Hermann Barbe, Marie Merschhemke, Frieda Mensing, Bruno Siegel, Jürgen L Müller, Renate Volbert, Peter Fromberger, Anett Tamm, et al. How to prepare for conversations with children about suspicions of sexual abuse? evaluation of an interactive virtual reality training for student teachers. Child Abuse & Neglect, 149:106677, 2024.

[26] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[27] Kathy A Mills and Alinta Brown. Immersive virtual reality (vr) for digital media making: transmediation is key. Learning, Media and Technology, 47(2):179–200, 2022.

[28] Fabin Rasheed, Prasad Onkar, and Marisha Narula. Immersive virtual reality to enhance the spatial awareness of students. In Proceedings of the 7th Indian Conference on Human-Computer Interaction, pages 154–160, 2015.

[29] Jaziar Radianti, Tim A Majchrzak, Jennifer Fromm, and Isabell Wohlgenannt. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. Computers & Education, 147:103778, 2020.

[30] Thomas K Metzinger. Why is virtual reality interesting for philosophers? Frontiers in Robotics and AI, 5:101, 2018.

[31] Lasse Jensen and Flemming Konradsen. A review of the use of virtual reality head-mounted displays in education and training. Education and Information Technologies, 23:1515–1529, 2018.

[32] Chris Fowler. Virtual reality and learning: Where is the pedagogy? British journal of educational technology, 46(2):412–422, 2015.

[33] Stephanie G Fussell and Dothang Truong. Accepting virtual reality for dynamic learning: An extension of the technology acceptance model. Interactive Learning Environments, 31(9):5442–5459, 2023.

[34] Pınar Üstel, Matthew J Smith, Shannon Blajeski, Jeffery M Johnson, Valerie G Butler, Johanna Nicolia-Adkins, Monica J Ortquist, Lisa A Razzano, and Adrienne Lapidos. Acceptability and feasibility of peer specialist-delivered virtual reality job interview training for individuals with serious mental illness: A qualitative study. Journal of technology in human services, 39(3):219–231, 2021.

[35] Maria Kandaurova and Seung Hwan Mark Lee. The effects of virtual reality (vr) on charitable giving: The role of empathy, guilt, responsibility, and social exclusion. Journal of Business Research, 100:571–580, 2019.

[36] Donghee Shin. Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience? Computers in human behavior, 78:64–73, 2018.

[37] Kuo-Wei Kyle Lai and Hao-Jan Howard Chen. A comparative study on the effects of a vr and pc visual novel game on vocabulary learning. Computer Assisted Language Learning, 0(0):1–34, 2021.

[38] Eric Krokos, Catherine Plaisant, and Amitabh Varshney. Virtual memory palaces: immersion aids recall. Virtual reality, 23(1):1–15, 2019.

[39] J Madden, S Pandita, JP Schuldt, B Kim, A S. Won, and NG Holmes. Ready student one: Exploring the predictors of student learning in virtual reality. PloS one, 15(3):e0229788, 2020.

[40] Semyon M Slobounov, William Ray, Brian Johnson, Elena Slobounov, and Karl M Newell. Modulation of cortical activity in 2d versus 3d virtual reality environments: an eeg study. International Journal of Psychophysiology, 95(3):254–260, 2015.

[41] Alex Dan and Miriam Reiner. Eeg-based cognitive load of processing events in 3d virtual worlds is lower than processing events in 2d displays. International Journal of Psychophysiology, 122:75–84, 2017.

[42] Feng Tian, Xuefei Wang, Wanqiu Cheng, Mingxuan Lee, and Yuanyuan Jin. A comparative study on the temporal effects of 2d and vr emotional arousal. Sensors, 22(21):8491, 2022.

[43] Pegah Salehi, Syed Zohaib Hassan, Saeed Shafiee Sabet, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pål Halvorsen, and Michael A. Riegler. Is more realistic better? a comparison of game engine and gan-based avatars for investigative interviews of children. In Proceedings of the ACM Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR), page 41–49, 2022.

[44] UMA Steering Group. UMA Repository. https://github.com/umasteeringgroup/UMA, 2022. Online repository.
[45] Unity Asset Store. SALSA LipSync Suite. https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442, 2022. Software package.
[46] Igroup Presence Questionnaire (IPQ) Overview. A Multi-Disciplinary Project Consortium Addressing New Interfaces Between Humans and the Real and Virtual Environment. Available at: http://www.igroup.org/pq/ipq/index.php, .
[47] John Brooke. Sus: a "quick and dirty'usability. Usability evaluation in industry, 189(3):189–194, 1996.
[48] Frank Biocca and Mark R Levy. Communication in the age of virtual reality. Routledge, 2013.
[49] Erin Wilson, David G Hewett, Brian C Jolly, Sarah Janssens, and Michael M Beckmann. Is that realistic? the development of a realism assessment questionnaire and its application in appraising three simulators for a gynaecology procedure. Advances in Simulation, 3(1):1–7, 2018.
[50] Steven Schmidt. Assessing the quality of experience of cloud gaming services. Springer Nature, 2022.
[51] Qingfu Zhu, Lei Cui, Weinan Zhang, Furu Wei, and Ting Liu. Retrieval-enhanced adversarial training for neural response generation. arXiv preprint arXiv:1809.04276, 2018.
[52] Syed Zohaib Hassan, Pegah Salehi, Ragnhild Klingenberg Røed, Pål Halvorsen, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pierre Lison, Michael Riegler, Michael E. Lamb, Carsten Griwodz, and Saeed Shafiee Sabet. Towards an ai-driven talking avatar in virtual reality for investigative interviews of children. In Proceedings of the ACM Workshop on Games Systems (GameSys), page 9–15, 2022.
[53] Richard Skarbez, Frederick P Brooks, Jr, and Mary C Whitton. A survey of presence and related concepts. ACM computing surveys (CSUR), 50(6):1–39, 2017.
[54] Saeed Shafiee Sabet, Cise Midoglu, Syed Zohaib Hassan, Pegah Salehi, Gunn Astrid Baugerud, Carsten Griwodz, Miriam Johnson, Michael Alexander Riegler, and Pål Halvorsen. Comparison of crowd-sourced and remote subjective user studies: A case study of investigative child interviews. In 2022 14th International Conference on Quality of Multimedia Experience (QoMEX), pages 1–6. IEEE, 2022.
[55] Saeed Shafiee Sabet. The influence of delay on cloud gaming quality of experience. Springer, 2023.
[56] Myrthe Lammerse, Syed Zohaib Hassan, Saeed Shafiee Sabet, Michael Alexander Riegler, and Pål Halvorsen. Human vs. gpt-3: The challenges of extracting emotions from child responses. In 2022 14th International conference on quality of multimedia experience (QoMEX), pages 1–4. IEEE, 2022.
[57] Arian Mehrfard, Javad Fotouhi, Giacomo Taylor, Tess Forster, Nassir Navab, and Bernhard Fuerst. A comparative analysis of virtual reality head-mounted display systems. arXiv preprint arXiv:1912.02913, 2019.
[58] Michael E Lamb, Kathleen J Sternberg, Yael Orbach, Irit Hershkowitz, Dvora Horowitz, and Phillip W Esplin. The effects of intensive training and ongoing supervision on the quality of investigative interviews with alleged sex abuse victims. Applied developmental science, 6(3):114–125, 2002.
[59] Naomi Ellemers. Gender stereotypes. Annual review of psychology, 69:275–298, 2018.



PEGAH SALEHI is currently a PhD candidate at Simulamet Research Laboratory, Oslo, Norway, and UiT—The Arctic University of Norway. Her research interests include deep learning, generative models, and computer vision. She is interested in the potential applications of computer vision in various domains, such as healthcare, and is also keen on improving user experience (UX) through intuitive and responsive visual interfaces and enhanced interaction with AI systems.



SYED ZOHAIB HASSAN



GUNN ASTRID BAUGERUD is an associate professor at Oslo Metropolitan University (OsloMet) holds a PhD in cognitive developmental psychology; she is an expert in child welfare and forensic psychology and has published extensively on these topics. She has an extensive research network as well as strong leadership experience as head of several large projects in collaboration with the CPS, and has lead of a large study, funded Research Council of Norway as ground-breaking research (FRIPRO).



MARTINE POWELL is a Professor at Griffith University and Founding Director of the Centre for Investigative Interviewing (Griffith Criminology Institute). She is a world-leading expert in investigative interviewing - her work focuses on the 'how to' of obtaining accurate and detailed information from people about events to assist decision-making.



MIRIAM S. JOHNSON is a clinical psychologist and an associate professor of psychology at Oslo Metropolitan University (OsloMet). She holds a PhD in cognitive developmental psychology and is an expert in forensic psychology and investigative interviewing of children. Dr. Johnson has published extensively on these topics and she has significant leadership experience, having headed several national research projects focused on the quality of investigative interviews with allegedly abused children.



DAG JOHANSEN is currently a Full Professor with the Department of Computer Science, UiT—The Arctic of Norway. He is exploring interdisciplinary research problems at the intersection of sport science, medicine, and computer science. A usecase receiving special attention is elite soccer performance development and quantification technologies as basis for evidence-based decisions. His research interests include intervention technologies where privacy is a first-order concern and design principle.



SAEED SHAFIEE SABET is a developer and machine learning engineer at Forzasys AS. He earned his PhD from TU Berlin, where he delved deeply into his research interests, which include multimedia quality of experience, gaming, and virtual/augmented reality technologies. His work focuses on improving the user overall quality of experience in multimedia applications.



MICHAEL A. RIEGLER received the Ph.D. degree from the Department of Informatics, University of Oslo, Oslo, Norway, in 2015. He is currently working as a Chief Research Scientist at SimulaMet, Oslo, Norway. His research interests include machine learning, video analysis and understanding, image processing, image retrieval, crowdsourcing, social computing, and user intentions.



PÅL HALVORSEN is currently a Chief Research Scientist at SimulaMet, Oslo, Norway, a Full Professor with the Department of Computer Science, Oslo Metropolitan University, and an Adjunct Professor at the Department of Informatics, University of Oslo, Norway. His research interest includes distributed multimedia systems, including operating systems, processing, storage and retrieval, communication, and distribution.

Article VIII: Comparative Analysis of Audio Feature Extraction for Real-Time Talking Portrait Synthesis

Authors: *Pegah Salehi*, Sajad Amouei Sheshkal, Vajira Thambawita, Sushant Gautam, Saeed S Sabet, Dag Johansen, Michael A Riegler, Pål Halvorsen

Candidate's contributions: The candidate was the primary contributor to this study, leading all aspects of the research. She conceptualized the scientific problem, designed the methodology, and developed the comparative framework for evaluating AFE models in real-time talking portrait synthesis. Her work included implementing the Whisper model to optimize processing efficiency and reduce latency, conducting experiments across multiple datasets, and analyzing both quantitative metrics and qualitative outcomes. As the first author, she drafted the manuscript, prepared visualizations, and synthesized the findings, highlighting Whisper's advantages in speed and synchronization for real-time applications. Her contributions also encompassed software implementation, data curation, and interpretation of results, ensuring a comprehensive evaluation of the system's performance. Additionally, she coordinated feedback and revisions from co-authors.

Published in: Big data and cognitive computing (BDCC), vol. 9, pp. 59 (2025)

DOI: 10.3390/bdcc9030059

Contributed to: This work contributed to the **RO-4**.

big data and cognitive computing

MDPI

Article

Comparative Analysis of Audio Feature Extraction for Real-Time Talking Portrait Synthesis

Pegah Salehi [1,2,*], Sajad Amouei Sheshkal [1], Vajira Thambawita [1], Sushant Gautam [1], Saeed S. Sabet [3], Dag Johansen [2], Michael A. Riegler [4] and Pål Halvorsen [1,5]

1 SimulaMet, 0167 Oslo, Norway; sajad.amouei@gmail.com (S.A.S.); vajira@simula.no (V.T.); sushant@simula.no (S.G.); paalh@simula.no (P.H.)
2 Department of Computer Science, UiT The Arctic University of Norway, 9019 Tromsø, Norway; dag.johansen@uit.no
3 Forzasys AS, 0840 Oslo, Norway; saeed.sh.sabet@gmail.com
4 Simula Research Laboratory, 0164 Oslo, Norway; michael@simula.no
5 Faculty of Technology, Art, and Design, Department of Computer Science (IT), Oslo Metropolitan University (OsloMet), 0130 Oslo, Norway
* Correspondence: pegah@simula.no

Abstract: This paper explores advancements in real-time talking-head generation, focusing on overcoming challenges in Audio Feature Extraction (AFE), which often introduces latency and limits responsiveness in real-time applications. To address these issues, we propose and implement a fully integrated system that replaces conventional AFE models with OpenAI's Whisper, leveraging its encoder to optimize processing and improve overall system efficiency. Our evaluation of two open-source real-time models across three different datasets shows that Whisper not only accelerates processing but also improves specific aspects of rendering quality, resulting in more realistic and responsive talking-head interactions. Although interviewer training systems are considered a potential application, the primary contribution of this work is the improvement of the technical foundations necessary for creating responsive AI avatars. These advancements enable more immersive interactions and expand the scope of AI-driven applications, including educational tools and simulated training environments.

Keywords: talking portrait synthesis; interactive avatar; Whisper; neural radiance fields (NeRFs); child protective services (CPS)

check for updates

Academic Editor: Jun Wang

Received: 15 November 2024
Revised: 29 December 2024
Accepted: 18 February 2025
Published: 4 March 2025

Citation: Salehi, P.; Sheshkal, S.A.; Thambawita, V.; Gautam, S.; Sabet, S.S.; Johansen, D.; Riegler, M.A.; Halvorsen, P. Comparative Analysis of Audio Feature Extraction for Real-Time Talking Portrait Synthesis. *Big Data Cogn. Comput.* **2025**, *9*, 59. https://doi.org/10.3390/bdcc9030059

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The application of AI in education has gained widespread attention for its potential to enhance learning experiences across disciplines, including psychology [1,2]. In the context of investigative interviewing, especially when questioning suspected child victims, AI offers a promising alternative to traditional training approaches. These conventional methods, often delivered through short workshops, fail to provide the hands-on practice, feedback, and continuous engagement needed for interviewers to master best practices in questioning child victims [3,4]. Research has shown that while best practices recommend open-ended questions and discourage leading or suggestive queries [5,6], many interviewers still struggle to implement these techniques effectively during real-world investigations [7]. The adoption of AI-powered child avatars provides a valuable solution, enabling child protective services (CPS) workers to engage in realistic practice sessions without the ethical dilemmas associated with using real children, while simultaneously offering personalized feedback on their performance [8].

Our current system leverages advanced AI techniques within a structured virtual environment to train professionals in investigative interviewing. Specifically, this system integrates the Unity Engine to generate virtual avatars. Despite the potential advantages of our AI-based training system, its effectiveness largely depends on the perceived realism and fidelity of the virtual avatars used in these simulations [9]. Based on our findings, we observe that avatars generated using Generative Adversarial Networks (GANs) demonstrate higher levels of realism compared to those created with the Unity Engine in several key aspects [10].

Accordingly, in this paper, we propose leveraging existing real-time talking portrait generation techniques to create an interactive avatar and evaluate its potential for improving perceived realism and interaction quality.

Audio-driven talking portrait synthesis enables the animation of a specific person based on arbitrary speech input, using deep learning-based methods such as neural radiance fields (NeRF) [11] and 3D Morphable Models (3DMMs) to generate high-quality 3D head models that offer flexibility in head poses and superior visual fidelity. Despite advancements in the development of real-time talking-head systems, their broad availability remains constrained by the inherent complexity of integrating multimodal data inputs, including audio and visual cues such as facial landmarks. This complexity poses challenges in achieving both synchronization and processing efficiency, which are imperative for the smooth operation of these systems in real-time scenarios. Furthermore, the field of real-time talking portrait systems is still in its nascent stages, with limited publications and, to the best of our knowledge, no standardized benchmarks or evaluation metrics currently available.

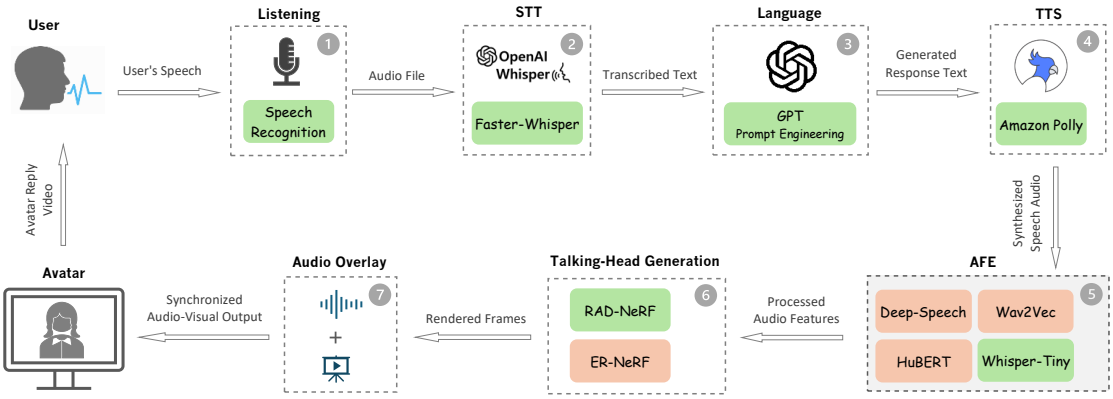
In audio-driven talking-head synthesis, the nuances of audio features directly influence the realism and synchronization of the generated visuals. Modern Audio Feature Extraction (AFE) models such as Wav2Vec [12] and HuBERT [13] have demonstrated high accuracy in tasks like Speech Emotion Recognition (SER) and Automatic Speech Recognition (ASR).

However, their reliance on transformer architectures poses significant challenges for real-time applications. For instance, Wav2Vec's transformer-based structure exhibits poor memory and computational scalability, leading to increased inference time with larger input sizes as observed in edge deployment scenarios [14–16]. Additionally, these models require substantial computational resources even when using optimized configurations, which further exacerbate latency issues in energy-constrained or low-power environments [15,16]. Moreover, through our exploration of open-source real-time talking-head models, we observe that latency in AFE can hinder real-time performance, impacting the overall responsiveness and realism of the avatar. Addressing these limitations necessitates the development of lightweight, efficient AFE systems that balance accuracy and speed for real-time use cases [14].

To mitigate the latency issues associated with existing AFE models, we leverage Whisper [17], an advanced ASR system that has been adapted for AFE tasks. Whisper's encoder–decoder Transformer architecture, trained on extensive multilingual and multitask datasets, offers an efficient and accelerated solution for AFE. Compared to earlier models like Deep-Speech 2 [18], which relies on computationally intensive recurrent networks, or Wav2Vec 2.0 [12] and HuBERT [13], which demand task-specific fine-tuning, Whisper delivers high performance across diverse conditions without additional optimization. Performance benchmarks (see Section 4.3) reveal that Whisper achieves an 80–90% reduction in processing latency relative to Deep-Speech, Wav2Vec, and HuBERT, without compromising quality. In summary, our main contributions are as follows:

- Integrating a complete AI-based avatar system combining GPT-3 for conversations, Speech-To-Text (STT), Text-To-Speech (TTS), AFE, and talking portrait synthesis (Figure 1) for full end-to-end experiments.
- Evaluating and comparing talking-head synthesis models.
- Evaluation and comparison of four AFE models using two open-source talking-head frameworks across three datasets.
- Modifying Whisper for efficient and accelerated AFE in talking portrait systems.
- Assessing and discussing the best combinations of talking portrait synthesis systems and AFE systems.

The code and resources related to this work are available on GitHub [19].



(a) Child avatar system architecture



(b) User interacting with the avatar

Figure 1. (**a**) System architecture of the interactive child avatar, detailing the integration of key modules: (1) Listening, (2) STT, (3) Language Processing, (4) TTS, (5) AFE, (6) Frames Rendering, and (7) Audio Overlay. This setup simulates natural conversation, allowing the user to interact with the avatar as if communicating with a real person. (**b**) User interaction with the child avatar system.

2. Related Work

The development and research of our avatar is touching upon two main areas: virtual avatars for interviewer training and real-time talking portrait synthesis. Notably, avatar systems provide the interactive framework necessary for dynamic user engagement, while talking portrait synthesis improves visual and auditory realism. Together, these technologies address both interaction logic and expressive fidelity, forming a synergistic foundation for our solution. This integration is vital for achieving a immersive user experience.

2.1. Virtual Avatar for Interview Training

The progression of child avatar training systems has significantly advanced investigative interviewing techniques, introducing varying degrees of automation and efficacy. Initial systems, such as those by Linnæus University and AvBIT Labs [20], relied on pre-

recorded responses, limiting interaction dynamics and necessitating human intervention for response selection. This dependency increased operational costs and introduced potential for human error. The LiveSimulation [21] addressed some limitations by enabling interactions with videotaped child avatars, improving trainees' open-ended questioning skills [22]. However, its reliance on predefined question options restricted the ability to practice generating spontaneous, context-specific queries crucial for real-world interviews.

The Empowering Interviewer Training (EIT) platform [23] advanced these methods with rule-based algorithms that allowed more dynamic interactions, though the fixed-response architecture limited adaptability. ViContact [24] further evolved the landscape by integrating virtual reality (VR) with automated feedback to refine both questioning techniques and socio-emotional support skills. Despite these advances, ViContact's rule-based interaction system hindered its ability to handle diverse inputs, and the gamification effect it introduced risked overemphasizing certain question patterns, potentially reducing skill transfer to real scenarios.

Building on these foundations, our previous platform [9] integrated GPT-3 into a Unity-based framework to create a dynamic child interview simulation. While this approach improved the system's responsiveness, criticisms of unrealistic avatar appearances pointed to a need for greater engagement during interactions. To address this, our current work incorporates lifelike talking portrait generation to improve visual realism and improve the overall efficacy of child interview training systems. This advancement aims to bridge the gap between simulation fidelity and real-world applicability, fostering more effective skill development.

2.2. Talking Portrait Synthesis

In recent years, real-time audio-driven talking portrait synthesis has gained significant attention, driven by applications in digital humans, virtual avatars, and video conferencing. Several approaches have been proposed to balance visual quality, synchronization, and computational efficiency.

Talking-head synthesis employs diverse methodologies, each with unique strengths and limitations. GANs generate high-quality static images with controllable attributes but face challenges in temporal consistency and complex dynamics. NeRF excels in photo-realistic 3D rendering and offers flexible viewpoints, but it is computationally intensive for real-time use. The 3D Gaussian models balance efficiency and identity preservation, though they may lack fine detail [25].

In the following, we discuss recent advancements in real-time audio-driven talking portrait synthesis, which build on these foundational methods to improve computational efficiency while maintaining synchronization and visual quality.

Live Speech Portrait [26] uses auto-regressive predictive coding (APC) [27] for extracting speech information, predicts 3D lip landmarks from audio, and synthesizes video frames through an image-to-image translation network (U-Net). Similarly, RealTalk [28] utilizes 3D facial priors and efficient expression rendering modules to achieve precise lip–speech synchronization while preserving facial identity.

Furthermore, 3D Gaussian Splatting (3DGS) [29] introduces a point-based rendering technique that uses ellipsoidal, anisotropic Gaussians to represent scenes with high accuracy. GSTalker [30] builds on this by incorporating deformable Gaussian splatting, significantly reducing training times and boosting rendering speeds compared to earlier NeRF-based models. Gaussian Talker [31] further advances the field by using a Gaussian-based model to generate talking faces with high-quality lip synchronization while reducing computational complexity.

Moreover, NeRF [11] has recently gained attention for generating talking portraits, given their capacity to capture intricate visual scenes. To enhance system efficiency, RAD-NeRF [32] incorporates discrete learnable grids in AD-NeRF [33], resulting in faster training and inference processes. Building upon this, ER-NeRF [34] utilizes tri-Plane hash representation to minimize hash collisions, leading to faster convergence, while a cross-modal fusion mechanism has been developed to improve lip–speech synchronization. To further refine lip synchronization, GeneFace++ [35] introduces a dedicated audio-to-motion module within the NeRF-based rendering framework. Additionally, R2-Talker [36] employs a progressive multilayer conditioning approach, improving performance and visual fidelity by integrating hash-grid landmark encoding.

3. Methodology

This section introduces the AFE models Deep-Speech 2, Wav2Vec 2.0, HuBERT, and Whisper, which we compare in this study, and outlines the interactive avatar's system architecture.

3.1. Audio Feature Extraction

A key aspect of talking-head generation is the ability to capture distinguishing speech features, as this directly affects the synchronization and quality of the audiovisual output. In the proposed framework, feature extraction is conducted using four ASR models: Deep-Speech 2 [18], Wav2Vec 2.0 [12], HuBERT [13], and Whisper [17]. Each of these models extracts both acoustic features and language representations from raw audio signals as part of their architecture. The following sections provide further details on these four models.

3.1.1. Deep-Speech 2

Deep-Speech 2, developed by Baidu, utilizes bidirectional recurrent neural networks (BRNNs) [37] alongside convolutional layers to capture context from both past and future frames, enhancing speech recognition accuracy. Key techniques like Batch Normalization and SortaGrad stabilize training, making the model effective across different acoustic conditions, including noisy environments [18].

In this study, building on prior methodologies [33,38], the DeepSpeech is applied to generate a 29-dimensional feature representation for every 20 ms audio segment. Consecutive frames of these audio features are processed through a temporal convolutional network to mitigate the noise present in raw audio signals. Specifically, a feature vector $\mathbf{a} \in \mathbb{R}^{16\times 29}$, derived from 16 contiguous audio frames, characterizes the current state of the audio modality and facilitates synchronization and performance in the proposed system.

3.1.2. Wav2Vec 2.0

Wav2Vec 2.0 is a transformer-based model developed for self-supervised feature extraction directly from raw audio signals [12]. Initially, the model processes audio waveforms into representations using a convolutional neural network (CNN) paired with a Gaussian Error Linear Unit (GELU) activation function [39], which captures latent speech features across temporal frames $z_1, z_2, \ldots, z_T$. These features are then fed into a transformer network [40], which is trained through a contrastive loss objective. This loss function enables the model to differentiate between correct and incorrect quantized representations of the audio signal [41].

This self-supervised training allows Wav2Vec 2.0 to learn rich, contextualized embeddings from unlabeled speech data, effectively building contextual representations across continuous speech and capturing dependencies over the entire sequence of latent representations. This approach reduces the need for hand-engineered features, while leveraging the powerful representations learned by Wav2Vec 2.0, resulting in improved performance

for various speech processing applications. In this study, we integrate Wav2Vec 2.0 as utilized in [32], leveraging its ability to maintain consistent classification logits when audio signals are accurately recognized. To further refine these features, they apply an audio attention mechanism, which functions as a temporal smoothing layer to generate the final audio conditions.

3.1.3. HuBERT

Hidden-Unit BERT (HuBERT) [13] introduces a self-supervised approach that addresses key challenges in speech processing: handling multiple sound units per utterance, the lack of lexicon during pre-training, and the absence of the explicit segmentation of sound units. By applying prediction loss to masked regions, HuBERT learns a combined acoustic and language model from unmasked inputs. Pre-trained on Librispeech [42] (960 h) and Libri-light [43] (60,000 h), it outperforms previous methods and comes in three sizes: BASE (90 M parameters), LARGE (300 M), and X-LARGE (1 B). HuBERT employs masking similar to SpanBERT [44] and wav2vec 2.0 [12], using cross-entropy loss on masked and unmasked time steps, encouraging the model to capture both acoustic features and long-range speech patterns.

In this study, we apply HuBERT, as utilized in GeneFace [45], to facilitate high-fidelity audio-to-motion transformation. HuBERT extracts semantic features from the input audio. The audio is segmented for memory-efficient processing, and each segment is analyzed to produce high-dimensional hidden states. These features provide semantically rich and temporally consistent representations, optimized for precise mapping to expressive motion outputs.

3.1.4. Whisper

In this paper, we propose the use of Whisper Tiny [17], designed for lightweight applications, which offers efficient processing and broad applicability in low-resource or edge environments. With approximately 39 million parameters, it is tailored for real-time applications on less powerful devices while maintaining the core Whisper architecture. This model follows an encoder–decoder Transformer structure, allowing it to perform versatile multilingual transcription, translation, and voice activity detection within a compact and efficient design, making it particularly suitable for talking-head systems and real-time audio–visual synchronization tasks.

Whisper Tiny uses log-Mel spectrograms as input features, derived from 25 ms windows with a 10 ms stride, which are scaled to a near-zero mean. This spectrogram is processed through Transformer encoder blocks containing convolutional layers with GELU activations [39]. These layers capture critical acoustic and linguistic features across languages and environments, leveraging Whisper's extensive pre-training on multilingual, multitask datasets. This approach provides the model with robust noise resilience and the ability to maintain high accuracy without dataset-specific fine-tuning.

The Whisper model's design has demonstrated a substantial 80–90% reduction in processing times compared to alternative models like Deep-Speech, Wav2Vec, and HuBERT, particularly for longer audio clips.

The Whisper model processes raw audio $A(t)$ by transforming it into a log-Mel spectrogram $S(f,t)$ as follows:

$$S(f,t) = \log\left(\sum_{k=0}^{N-1} A(t) \cdot e^{-j2\pi ft}\right)$$

This spectrogram, capturing core frequency components, is then passed through Whisper's encoder to generate high-dimensional audio embeddings E:

$$E = \text{WhisperEncoder}(S(f,t))$$

where E has shape (T_{initial}, C), with T representing the number of time steps aligned to the visual frames, and $C = 384$ as the dimensionality of the feature embedding space. Synchronization with a 25 FPS visual frame rate is achieved by applying a sliding window with parameters $w = 16$, stride $s = 2$, and padding $p = 7$, yielding a final feature matrix with shape $(750, 16, 384)$. This setup ensures precise temporal alignment across 750 frames over 30 s, enhancing the real-time accuracy and fluidity of interactions in talking-head applications.

3.2. System Architecture

The system architecture of our interactive child avatar, as depicted in Figure 1, is composed of several modules: Listening, STT, Language, TTS, AFE, Frames Rendering, and Audio Overlay.

The Listening module is the entry point of the system, where user speech is captured through a button-based recording process. Users click to start recording their voice input and click again to stop, saving the recorded audio file. During this recording phase, the avatar remains in a listening state, utilizing a pre-rendered video based on an empty audio file where it exhibits natural, non-verbal behaviors such as blinking and subtle movements while remaining silent. This module employs a speech recognition system to continuously listen for spoken input from the user, initiating the conversion process by passing the audio data to the speech-to-text module.

The STT module utilizes OpenAI's Whisper model [17] to perform the real-time conversion of spoken input into text, transcribing the user's speech for further processing. The language module is responsible for generating contextually appropriate and dynamic responses. It leverages GPT for prompt engineering, simulating a child's conversational style. The transcribed text from the STT is processed here, where GPT generates the relevant responses tailored to the interaction context. Once the response text is generated, it is passed to the TTS, which uses Amazon Polly [46] to convert the text into speech. This module maintains the consistency of the avatar's voice with a child's persona.

AFE processes the incoming audio from the TTS module to extract relevant features. To achieve rapid and efficient audio processing that enhances the system's responsiveness, four models are evaluated: Deep-Speech [18], Wav2Vec [12], HuBERT [13], and Whisper-Tiny [17].

The Frames Rendering module manages the visual representation of the avatar. It is responsible for rendering frames in real-time based on the processed audio and text input, allowing the avatar to exhibit natural behaviors such as lip-syncing and facial expressions synchronized with the spoken output. We also compare two frameworks, RAD-NeRF [32] and ER-NeRF [34], to identify the most effective solution for this purpose. The final component is the Audio Overlay module, which combines the rendered frames with the audio output.

4. Experiments

This section rigorously evaluates various AFE models across two frameworks, focusing on model efficiency, synchronization accuracy, and responsiveness.

4.1. Experimental Setup

In this section, our datasets, hardware, and configurations are outlined for evaluating AFE model performance in real-time talking portrait synthesis.

4.1.1. Dataset

The dataset used in our experiments comprises a combination of publicly available video datasets and a privately sourced video. We select three high-definition speech video clips, each with an average duration of approximately 6700 frames (around 4.5 min, as recorded at 25 FPS). The raw videos, originally recorded at their native resolutions (which can also be used in the experiments), are cropped and resized to 512×512. However, the Obama video from AD-NeRF [33], which was originally processed at 450×450 pixels, is used in that resolution without further resizing.

To ensure fairness and reproducibility, two of the video clips used in our experiments are sourced from the Internet. Specifically, we utilize the "Obama" video from the publicly released data of AD-NeRF [33] and the "Shaheen" video. These videos are chosen for their recognition and frequent use as benchmarks in talking portrait synthesis research. The third video clip, featuring a young girl's speech, is privately sourced to align with the application objectives of the investigative interview. Explicit permission was obtained from the individual for the use of this video in this research. Our selection includes a range of speaking styles, expressions, and conditions, ensuring a balanced and comprehensive dataset for training and evaluation.

For each video, the first 91% frames, along with the corresponding audio, are used as training data, while the final 9% of the data are reserved for subsequent evaluation, in accordance with previous studies [32–34].

4.1.2. System Configuration

Experiments are conducted on a machine with a 12th Gen Intel(R) Core(TM) i9-12900F CPU, 31 GiB of RAM, and an NVIDIA RTX 4090 GPU with 24 GiB of VRAM, running CUDA 12.4 on an Ubuntu operating system.

4.2. Real-Time Talking-Head Speed Analysis

Despite progress in real-time talking-head systems, their widespread use is limited by the complexity of integrating multimodal inputs like audio and facial landmarks. This complexity challenges both synchronization and processing efficiency, which are required for smooth real-time performance. Given these challenges, we deemed it necessary to evaluate the real-time capabilities of existing models to substantiate performance claims. To accurately assess the efficiency of these models, we conduct an experiment using the Obama dataset [33]. The experiment is carried out under identical hardware and CUDA environments across all applicable open-source methods, ensuring a consistent basis for comparison.

Our findings confirm that a major challenge identified in previous studies is the latency introduced during the AFE process. To better understand the impact of AFE on overall system performance, we measure the execution time using two different approaches: one excluding AFE and the other including it. GeneFace++ [35] is excluded from this comparison, as its audio features are deeply interwoven within the core of the model, making it impossible to measure its performance without AFE. The results, presented in Figure 2, illustrate the comparative processing efficiency of these models across different audio durations.

4.3. AFE Analysis

Based on the results presented in Figure 2, we select RAD-NeRF [32] and ER-NeRF [34] as the frameworks for further experimentation due to their better performance compared to other models. To conduct a systematic comparison of AFE models, we employ DeepSpeech, Wav2Vec [12], HuBERT [13], and Whisper-Tiny [17] within both RAD-NeRF [32]

and ER-NeRF [34]. The models are trained from scratch with each AFE configuration to assess their impact on system performance, focusing on factors such as lip synchronization accuracy, visual quality, and overall execution time. In the following, we divide the analysis into two key aspects—speed and quality—to provide a more detailed evaluation of each AFE configuration's impact on system performance.

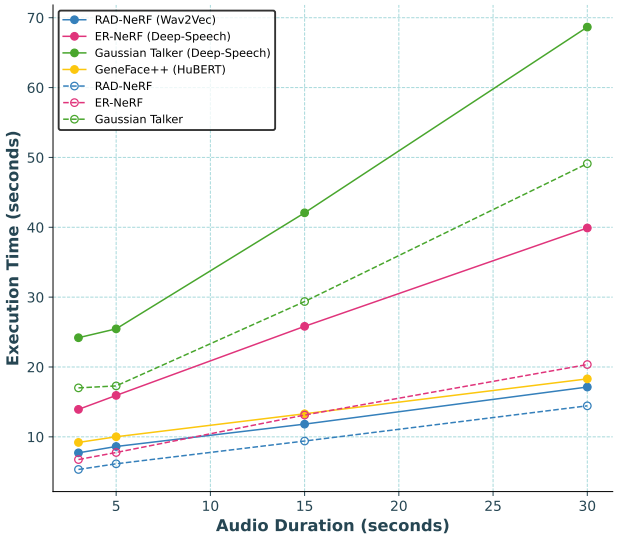


Figure 2. Execution time comparison of open-source real-time talking-head generation models, including RAD-NeRF [32], ER-NeRF [34], Gaussian Talker [31] and GeneFace++ [35]. The solid lines represent execution times excluding AFE, while the dashed lines indicate execution times that include AFE.

4.3.1. AFE Speed Analysis

In evaluating the speed of AFEs, we conduct an analysis between Whisper [17] and other well-known AFE models, including Deep-Speech [18], Wav2Vec [12], and HuBERT [13]. The results, as shown in Figure 3, reveal that Whisper greatly outperforms the other models across varying audio durations, especially compared to Deep-Speech, which shows increasing execution times as the audio duration grows. Whisper consistently achieves notably lower execution times, making it particularly advantageous for conversational systems.

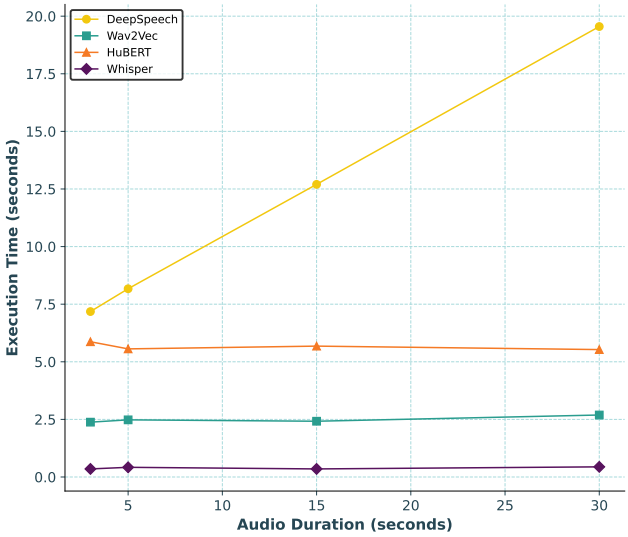


Figure 3. Execution time comparison of different AFE models, including Deep-Speech [18], Wav2Vec [12], HuBERT [13], and Whisper [17].

Additionally, Figure 4 compares the execution times of RAD-NeRF [32] and ER-NeRF [34] using various AFE models, including Whisper, Deep-Speech [18], Wav2Vec [12], and HuBERT [13]. The results show the time-saving advantages of Whisper across both models. By integrating Whisper, we achieve a reduction in processing time, which is important for real-time applications such as talking-head generation and interactive avatars. These findings reveal the potential of Whisper as an effective solution for accelerating AFE processes in interactive avatar systems, making them more efficient and responsive.

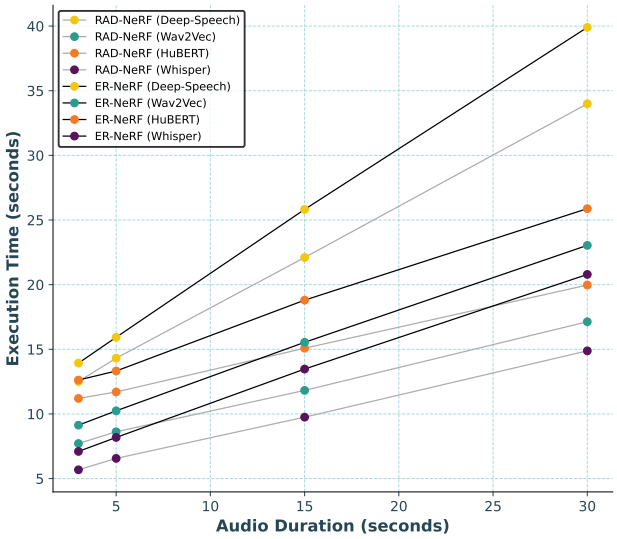


Figure 4. Execution time comparison of RAD-NeRF [32] and ER-NeRF [34] across different AFE models.

4.3.2. AFE Quality Analysis

To further validate utilizing Whisper as AFE, we conduct an evaluation of the system's rendering quality across various settings. In the self-driven setting, where the ground truth data correspond to the same identity as the generated output, we utilize the widely recognized quantitative metrics to assess the quality of portrait reconstruction. These metrics are standard in the field and provide a comprehensive assessment of both visual fidelity and synchronization, ensuring alignment with the established evaluation practices:

- Peak Signal-to-Noise Ratio (PSNR): This metric measures the fidelity of the reconstructed image relative to the ground truth. The PSNR is calculated as

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}_\text{I}^2}{\text{MSE}}\right)$$

  where MAX_I is the maximum possible pixel value of the image, and MSE is the Mean Squared Error between the reconstructed and ground truth images.
- Structural Similarity Index Measure (SSIM): SSIM evaluates the structural similarity by considering the luminance, contrast, and structure. The formula is

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

  where μ_x and μ_y are the average intensities, σ_x and σ_y are variances, and σ_{xy} is the covariance between images x and y. C_1 and C_2 are constants for stability.
- Learned Perceptual Image Patch Similarity (LPIPS) [47]: Measures the perceptual similarity between the generated and ground truth images by calculating the distance between feature representations extracted from a deep neural network. This metric

captures differences in visual features that align more closely with human perception than simple pixel-wise comparisons, making it useful for assessing image quality in terms of perceptual fidelity:

$$\text{LPIPS} = \frac{1}{N}\sum_{i=1}^{N} \|f(x_i^{\text{pred}}) - f(x_i^{\text{truth}})\|_2$$

where $f(x)$ represents the feature representation of image x extracted by a neural network (e.g., AlexNet), x_i^{pred} and x_i^{truth} are the predicted and ground truth images respectively, and N is the number of patches or feature points compared.

- Landmark Distance (LMD) [48]: This metric measures the geometric accuracy of facial landmarks by calculating the Euclidean distance between corresponding landmark points in the generated and ground truth images:

$$\text{LMD} = \frac{1}{N}\sum_{i=1}^{N} \sqrt{(x_i^{\text{pred}} - x_i^{\text{truth}})^2 + (y_i^{\text{pred}} - y_i^{\text{truth}})^2}$$

where $(x_i^{\text{pred}}, y_i^{\text{pred}})$ and $(x_i^{\text{truth}}, y_i^{\text{truth}})$ are the coordinates of the i-th landmark in the predicted and ground truth images, respectively, and N is the total number of landmarks.

- Fréchet Inception Distance (FID) [49]: FID assesses the similarity between distributions of real and generated images. The formula is

$$\text{FID} = \mu_r - \mu_g^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r\Sigma_g})$$

where μ_r and Σ_r represent the mean and covariance of features for real images, and μ_g and Σ_g represent those for generated (images).

- Action Units Error (AUE) [50]: AUE measures the accuracy of lower facial muscle movements by calculating the squared differences in action unit intensities between the generated and ground truth images. In this study, we specifically evaluate the lower face region, which is relevant for expressions related to speech and emotion:

$$\text{AUE}_{\text{lower}} = \frac{1}{N}\sum_{i=1}^{N}(AU_i^{\text{pred}} - AU_i^{\text{truth}})^2$$

where AU_i^{pred} and AU_i^{truth} are the intensities of the i-th action unit in the lower face region for the predicted and ground truth data, respectively, and N is the total number of lower face action units evaluated.

- SyncNet Confidence Score (Sync) [51]: Measures the lip-sync accuracy by evaluating the alignment between audio and lip movements in generated talking-head videos. This metric utilizes SyncNet, which calculates a confidence score based on the similarity between the embeddings of audio and video frames:

$$\text{Sync} = \frac{1}{N}\sum_{i=1}^{N} \frac{v_i \cdot s_i}{\max(\|v_i\|_2 \cdot \|s_i\|_2, \epsilon)}$$

where v_i and s_i are the embeddings for the i-th video and audio frames, respectively, computed by SyncNet. This formula calculates the cosine similarity between the embeddings, giving a score between 0 and 1 for each frame pair. N is the total number of frame pairs evaluated, and higher average values indicate better lip-sync quality.

The self-driven evaluation results are shown in Table 1, where the utilization of Whisper for AFE leads to a slight improvement in the PSNR and LPIPS metrics. Additionally,

the FID scores reflect a minor but consistent improvement in the perceptual similarity between generated and real images. However, the most notable improvement is observed in the Sync, which shows better lip synchronization.

Table 1. Quantitative comparison of face reconstruction quality under self-driven synthesis on the same identity's test set. The arrows in the metric names indicate the preferred trend for each metric: ↑ denotes that higher values are better, ↓ signifies that lower values are preferable.

Methods	AFE	Dataset	PSNR ↑	SSIM ↑	LPIPS ↓	LMD ↓	FID ↓	AUE ↓	Sync_{conf} ↑
RAD-NeRF [32]	Deep-Speech	Obama	27.14	0.9304	0.0738	2.675	31.29	1.995	7.171
		Donya	27.79	0.9045	0.0917	2.750	12.82	1.911	4.720
		Shaheen	30.13	0.9314	0.0697	3.199	33.05	2.837	7.330
		Mean	28.35	0.9221	0.0784	2.874	25.72	2.247	6.407
	HuBERT	Obama	26.58	0.9261	0.0769	2.762	28.78	2.006	0.563
		Donya	28.05	0.9071	0.0868	2.518	14.25	2.511	0.365
		Shaheen	30.45	0.9332	0.0729	3.050	35.96	3.229	0.494
		Mean	**28.36**	0.9221	0.0788	**2.776**	26.33	2.582	0.474
	Wav2Vec	Obama	26.59	0.9268	0.0785	2.696	15.15	1.707	6.744
		Donya	27.12	0.8972	0.0845	2.726	24.58	1.531	4.820
		Shaheen	30.08	0.9306	0.0698	3.221	34.77	2.966	7.946
		Mean	27.93	0.9182	0.0776	2.881	24.83	2.068	6.503
	Whisper	Obama	26.10	0.9231	0.0723	2.573	12.67	1.693	7.143
		Donya	28.65	0.9138	0.0844	2.640	26.85	1.504	5.269
		Shaheen	30.05	0.9303	0.0660	3.045	29.44	2.696	8.488
		Mean	28.07	**0.9224**	**0.0761**	2.826	**24.04**	**1.964**	**6.966**
ER-NeRF [34]	Deep-Speech	Obama	26.44	0.9339	0.0441	2.561	7.14	1.923	7.201
		Donya	28.91	0.9165	0.0605	2.647	14.59	1.874	4.722
		Shaheen	29.92	0.9267	0.0450	2.900	16.10	2.668	8.215
		Mean	**28.16**	**0.9257**	0.0689	2.7932	20.92	2.155	6.712
	HuBERT	Obama	26.30	0.9297	0.0473	2.758	8.33	1.711	0.300
		Donya	24.20	0.7826	0.1255	2.545	49.81	2.284	0.408
		Shaheen	30.45	0.9322	0.0420	2.852	16.56	3.172	0.434
		Mean	26.98	0.8815	0.0716	**2.718**	24.90	2.389	0.380
	Wav2Vec	Obama	25.59	0.9268	0.0497	2.645	8.83	1.704	6.616
		Donya	24.21	0.7777	0.1509	2.754	68.20	1.730	4.403
		Shaheen	29.81	0.9245	0.0470	3.003	15.59	2.948	7.917
		Mean	26.53	0.8763	0.0825	2.800	30.87	**2.127**	6.312
	Whisper	Obama	26.30	0.9314	0.0462	2.501	8.06	1.797	7.647
		Donya	27.36	0.9020	0.0641	2.516	14.67	1.852	5.704
		Shaheen	30.20	0.9305	0.0434	2.935	15.61	3.030	8.575
		Mean	28.12	0.9213	**0.0654**	2.7640	**19.29**	2.226	**7.308**

Note: The Mean row, highlighted in gray, represents the average value obtained from three independent experiments, providing a summarized performance metric. Furthermore, the best-performing result for each criterion is highlighted in bold to indicate the best outcome.

In the cross-driven setting, the model generates a talking portrait based on audio clips that do not match the identity of the visual data used during training. To create variation, two different audio clips are selected for each dataset, corresponding to the gender and age of the character. One audio clip is synthetic, generated using Amazon Polly [46], while the other is a natural recording from a real human, allowing for a fair comparison between synthetic and natural speech.

Because ground truth images corresponding to the same identity are absent, identity-specific metrics are not applicable. Therefore, consistent with prior studies [32,34], we use

the identity-agnostic Sync score [51] as the primary metric to evaluate synchronization between audio and lip movements. This approach allows us to effectively measure the model's performance when direct comparison with ground truth is not feasible.

The cross-driven results, as shown in Table 2, indicate that models utilizing Whisper generally outperform others, particularly in terms of SyncNet scores. However, Whisper's superior performance is not consistent across all scenarios. While it shows particular strength in handling human voices, its effectiveness diminishes when dealing with bot-generated voices. Specifically, in our experiments, the bot's voice is generated at a slower pace and interspersed with more pauses. This slower speech setting may unintentionally benefit other models, which makes them perform better relative to Whisper in this context.

Table 2. Quantitative comparison of audio–lip synchronization under the cross-driven setting. The arrows in the metric names indicate the preferred trend for each metric: ↑ denotes that higher values are better, ↓ signifies that lower values are preferable.

Methods	AFE	Dataset	Sync_{conf} ↑ Synthetic	Sync_{conf} ↑ Natural
RAD-NeRF [32]	Deep-Speech	Obama	6.581	6.489
		Donya	4.208	3.576
		Shaheen	6.319	5.840
		Mean	5.702	5.301
	HuBERT	Obama	5.109	0.548
		Donya	4.750	0.680
		Shaheen	6.670	0.741
		Mean	5.509	0.656
	Wav2Vec	Obama	6.851	7.388
		Donya	4.593	5.017
		Shaheen	6.837	7.303
		Mean	**6.093**	6.569
	Whisper	Obama	6.884	7.137
		Donya	4.628	5.742
		Shaheen	6.347	7.052
		Mean	5.953	**6.643**
ER-NeRF [34]	Deep-Speech	Obama	7.306	7.224
		Donya	5.054	4.851
		Shaheen	6.505	6.074
		Mean	**6.323**	6.050
	HuBERT	Obama	5.542	0.614
		Donya	4.822	0.480
		Shaheen	7.068	0.581
		Mean	5.810	0.558
	Wav2Vec	Obama	7.219	7.428
		Donya	4.115	4.155
		Shaheen	7.145	7.242
		Mean	6.159	6.275
	Whisper	Obama	7.399	8.247
		Donya	4.651	6.486
		Shaheen	6.148	7.093
		Mean	6.066	**7.275**

Note: The Mean row, highlighted in gray, represents the average value obtained from three independent experiments, providing a summarized performance metric. Furthermore, the best-performing result for each criterion is highlighted in bold to indicate the best outcome.

Our experimental results reveal that HuBERT consistently underperforms in lip synchronization, with the lowest Sync scores among all the models tested. Given these results, particularly the poor performance in cross-driven settings (e.g., Sync scores as low as 0.564), HuBERT appears inadequate for applications requiring precise lip-sync. Future work should consider incorporating subjective evaluations to further assess the perceptual quality and user satisfaction, which may provide a more thorough understanding of its limitations.

Furthermore, we compare the subjective quality and lip-sync performance of two models, RAD-NeRF [32] and ER-NeRF [34], shown in Figure 5 and Figure 6, respectively, using four distinct AFEs under self-driven settings. Each model's output is evaluated against ground truth to assess its accuracy in audio-driven 3D reconstruction and lip-sync fidelity. For a more comprehensive view of the results, please refer to the supplemental video available on our GitHub repository [19].



Figure 5. Quality comparison: Examples of visualizations of RAD-NeRF [32] under the self-driven setting, based on two frames extracted from each video illustrating typical challenges. Yellow boxes highlight areas of noisy image quality, while red boxes indicate regions with inaccurate lip synchronization.

ER-NeRF and RAD-NeRF demonstrate varying degrees of fidelity across the AFEs, particularly in achieving accurate lip-sync with the audio input. Generally, Whisper achieves the closest approximation to ground truth across both ER-NeRF and RAD-NeRF, exhibiting strong performance in synchronizing lip movements with audio, especially during pronounced lip articulations, such as the wide-open mouth movements observed when pronouncing sounds like "wa." This precise alignment enhances the realism of the output, particularly in sequences requiring dynamic mouth shapes.



Figure 6. Quality comparison: Examples of visualizations of ER-NeRF [34] methods under the self-driven setting, based on two frames extracted from each video illustrating typical challenges. Yellow boxes highlight areas of noisy image quality, while red boxes indicate regions with inaccurate lip synchronization.

Wav2vec and HuBERT provide reasonable approximations but show slight misalignments in lip-sync. Deep-Speech, while effective, displays the greatest variance from ground truth, with notable lip-sync discrepancies, indicating less robust performance in these NeRF-based reconstructions.

4.4. System Responsiveness Analysis

Table 3 outlines the execution times per component, with the Listening component excluded due to its dependence on user input duration. The analysis is segmented into raw execution time and the percentage contribution of each stage to the overall system latency, providing a granular view of where time is consumed in the process of generating the interactive avatar's responses.

Table 3. Execution time analysis of each component in the system architecture as depicted in Figure 1. AA Tokens: number of tokens in the avatar's answer. AA Duration (s): length of the avatar's answer in seconds.

AA Tokens	AA Duration (s)		STT	Language	TTS	AFE	Frame Rendering	Audio Overlay	SUM
1	0.41	Exe. Time (s)	0.06	0.80	0.22	0.29	4.05	0.14	5.56
		% of Total	1.08%	14.39%	3.96%	5.22%	72.84%	2.52%	100%
8	1.69	Exe. Time (s)	0.07	0.96	0.33	0.28	4.75	0.16	6.55
		% of Total	1.07%	14.66%	5.04%	4.27%	72.52%	2.44%	100%
14	3.63	Exe. Time (s)	0.07	2.45	0.44	0.28	5.45	0.14	8.83
		% of Total	0.79%	27.74%	4.98%	3.17%	61.71%	1.59%	100%
21	5.08	Exe. Time (s)	0.1	2.27	0.44	0.28	5.88	0.17	9.14
		% of Total	1.09%	24.83%	4.81%	3.06%	64.36%	1.86%	100%
30	6.55	Exe. Time (s)	0.06	2.76	0.49	0.27	6.58	0.15	10.31
		% of Total	0.58%	26.77%	4.75%	2.62%	63.84%	1.45%	100%
39	9.55	Exe. Time (s)	0.09	1.95	0.78	0.28	7.52	0.18	10.8
		% of Total	0.83%	18.06%	7.22%	2.59%	69.63%	1.67%	100%
50	12.16	Exe. Time (s)	0.07	2.08	0.55	0.28	8.65	0.18	11.81
		% of Total	0.59%	17.61%	4.66%	2.37%	73.24%	1.52%	100%

To evaluate the system's responsiveness, we test various lengths of avatar answers. This approach allows us to assess how the system handles different interaction complexities, ranging from brief exchanges like *"Hi"* and *"I'm OK"* to more extended dialogues, such as *"Um... at Jenny's house, we play a lot. It's nice there... but, well, I don't really like talking about the pool. It wasn't very fun last time"*. The results indicate that while the overall system performs efficiently, the frame rendering stages are identified as the most time-consuming, due to the computational demands of generating high-quality, real-time visual output that matches the synchronized audio input.

Conversely, the integration of the Whisper model as the AFE component proves to be much faster than the conventional AFE models. Whisper's improved processing speed has substantially reduced the time required to extract and synchronize audio features, contributing to a faster overall system performance.

The execution times presented in Table 3 indicate delays significantly beyond the commonly recommended thresholds for interactive systems, typically under 100 ms [52,53]. However, our project does not fall into the category of high-speed interaction tasks. Given the sensitive and often emotionally charged nature of the scenarios—such as simulating a child avatar discussing distressing topics—longer response times might even be perceived as reflective of the avatar's emotional state, enhancing realism [9]. To better understand the impact of these delays on user experience and engagement, conducting a user-centered study is essential. This will allow us to evaluate whether the observed response times align with the expectations and needs of users in this specific context.

5. Discussion and Future Work

The AFE execution time plays a critical role in the performance of interactive avatar systems. Whisper's ability to reduce latency by up to 90% compared to alternative models is a decisive factor in its selection. To ensure that this acceleration does not compromise

rendering quality or synchronization, a thorough evaluation of image quality metrics is conducted. While these metrics are secondary to the execution time, the results confirm that Whisper's integration preserves the system's visual and functional integrity and, in some cases, even enhances it. Whisper's streamlined architecture significantly reduces system latency compared to resource-intensive models like Deep-Speech [18], which slow down as the audio duration increases. This reduction in processing latency is especially important for real-time applications, where delays can disrupt user engagement. By enabling faster responses, Whisper facilitates smoother interactions in applications like interactive avatars, thereby enhancing both user experience and system reliability [9].

However, despite Whisper's efficiency, the overall system still experiences latency due to other computationally intensive components, such as frame rendering. This issue becomes more pronounced when changes in hardware specifications significantly affect the frame rendering time of talking-head generation models. Addressing these bottlenecks and further exploring system scalability across diverse hardware configurations are essential steps toward improving real-world applicability and robustness.

Future work could focus on optimizing frame rendering by incorporating lightweight techniques such as Gaussian Splatting [29], which are well suited for real-time applications. These methods could reduce the computational overhead while maintaining high-quality output, addressing bottlenecks and increasing the system's scalability and responsiveness across diverse hardware configurations. This aligns with GSTalker's demonstration of the Gaussian Splatting's efficiency in fast training and real-time rendering [30].

Additionally, we can utilize advanced solutions like NVIDIA's Avatar Cloud Engine (ACE) (https://developer.nvidia.com/ace, accessed on 20 January 2025), a suite of technologies that combines generative AI and hardware acceleration specifically for creating realistic digital humans. By leveraging tools like ACE, which use powerful NVIDIA GPUs to accelerate processes like speech recognition, natural language processing, and real-time facial animation, system responsiveness can be further improved.

Moreover, although Whisper has demonstrated significant advantages in two real-time talking-head synthesis models, its adoption should not be presumed universally effective across all models or contexts. Each talking-head generation model leverages unique strategies and architectures tailored to specific applications, which may necessitate alternative feature extraction approaches. The diversity in model designs underscores the importance of contextual considerations when selecting feature extraction methods. Nevertheless, Whisper's robust real-time capabilities suggest potential for future integration as a preprocessing step in emerging systems, particularly those prioritizing real-time performance. Such integration could streamline audio processing pipelines, enhancing efficiency without compromising synchronization or quality. This outlook provides a promising direction for optimizing real-time systems while acknowledging the necessity for model-specific customization.

Another limitation of the current work is the absence of a robust framework for handling errors or implementing fallback mechanisms across the modular pipeline components. In modular systems, individual component failures can propagate through the pipeline, potentially compromising both system reliability and real-time performance. To address these challenges, future research should prioritize the development of dynamic error detection and mitigation strategies.

Furthermore, conducting broader user studies with professionals in the field could provide a deeper understanding of the system's impact on practical outcomes, such as user training and engagement. Expanding this research to include real-world evaluations with subjective user feedback will offer valuable insights into perceived realism, usability, and overall user satisfaction. Incorporating user-centric metrics will provide

a more comprehensive view of the system's effectiveness in high-engagement applications, such as investigative interview training. Through these advancements, AI-driven avatars may achieve greater utility and realism, enhancing their role as effective tools in training applications.

Ethical Considerations

In integrating AI technologies, such as deepfake-powered avatars, into sensitive domains like CPS, ethical considerations are paramount. Deepfake technology [54], while advancing digital realism, raises significant ethical concerns. Misuse could lead to the creation of nonconsensual or harmful content, infringing on privacy, consent, and psychological safety [55]. In CPS training, the misuse of AI avatars may manifest in inappropriate or unprofessional interactions, such as engaging avatars in casual or sexualized dialogues. Such behaviors compromise the seriousness of the educational context, disrespect the simulated victims, and violate professional ethics.

To mitigate these risks, stringent ethical oversight is necessary. This includes the implementation of privacy-preserving frameworks, content ownership protection for generative models, and robust guidelines on acceptable avatar interactions [56]. For example, CPS workers should engage with avatars solely for training purposes, adhering strictly to professional and ethical boundaries.

Despite these challenges, AI avatars offer unique opportunities for societal good. They provide CPS workers with realistic, interactive training tools that avoid involving real children, thus eliminating ethical dilemmas and ensuring victim privacy. Additionally, they can anonymize sensitive victim data, preserving dignity and confidentiality. To balance the benefits and risks, a multidisciplinary approach is important. Collaboration among technologists, legal experts, child protection professionals, and ethicists can ensure AI technologies are leveraged responsibly. Developing robust detection systems to identify misuse, alongside legal frameworks to regulate the deployment of such technology, is essential for protecting vulnerable populations.

6. Conclusions

This study addressed the latency challenges associated with AFE, which has hindered the practical deployment of real-time talking portrait systems in real-world applications. By integrating the Whisper model—a high-performance ASR system—into our framework, we achieved notable reductions in processing delays. These optimizations not only increased the overall responsiveness of the interactive avatars but also improved the accuracy of lip-syncing, making them more applicable for immersive training applications.

Our findings affirm Whisper's capability to meet real-time demands, particularly in applications requiring responsive interactions and minimal delay. This efficiency is important for training environments such as CPS, where timely and realistic interactions can greatly impact training efficacy. By achieving these improvements, Whisper-integrated systems emerge as promising solutions for a variety of real-time applications, including virtual assistants, remote education, and digital customer service platforms.

Author Contributions: Conceptualization, P.S., S.A.S., V.T., S.S.S., M.A.R. and P.H.; methodology, P.S., S.A.S., V.T. and P.H.; software, P.S., S.A.S. and S.G.; validation, P.S., S.A.S. and S.G.; formal analysis, P.S.; investigation, P.S.; resources, P.H. and M.A.R.; data curation, P.S. and S.A.S.; writing—original draft preparation, P.S., S.A.S. and P.H.; writing—review and editing, P.S., S.A.S., V.T., S.G., S.S.S., D.J., M.A.R. and P.H.; visualization, P.S. and S.A.S.; supervision, V.T., S.S.S., D.J., M.A.R. and P.H.; project administration, P.S. and P.H.; funding acquisition, M.A.R. and P.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research Council of Norway grant number #314690.

Institutional Review Board Statement: Ethical approval has been obtained from The Norwegian Agency for Shared Services in Education and Research.

Informed Consent Statement: Informed consent for participation was obtained from all subjects involved in the study.

Data Availability Statement: The code and data for this study are available in the project's GitHub repository [19].

Conflicts of Interest: Author Saeed S. Sabet was employed by the company Forzasys. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Chernikova, O.; Heitzmann, N.; Stadler, M.; Holzberger, D.; Seidel, T.; Fischer, F. Simulation-based learning in higher education: A meta-analysis. *Rev. Educ. Res.* **2020**, *90*, 499–541. [CrossRef]
2. Crompton, H.; Bernacki, M.; Greene, J.A. Psychological foundations of emerging technologies for teaching and learning in higher education. *Curr. Opin. Psychol.* **2020**, *36*, 101–105. [CrossRef] [PubMed]
3. Lamb, M.E. Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink? *Am. Psychol.* **2016**, *71*, 710. [CrossRef] [PubMed]
4. Powell, M.B. Designing effective training programs for investigative interviewers of children. *Curr. Issues Crim. Justice* **2008**, *20*, 189–208. [CrossRef]
5. Lamb, M.E.; Orbach, Y.; Hershkowitz, I.; Esplin, P.W.; Horowitz, D. A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child Abuse Negl.* **2007**, *31*, 1201–1231. [CrossRef] [PubMed]
6. Lyon, T.D. Interviewing children. *Annu. Rev. Law Soc. Sci.* **2014**, *10*, 73–89. [CrossRef]
7. Lamb, M.E.; Brown, D.A.; Hershkowitz, I.; Orbach, Y.; Esplin, P.W. *Tell Me What Happened: Questioning Children About Abuse*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
8. Powell, M.B.; Brubacher, S.P.; Baugerud, G.A. An overview of mock interviews as a training tool for interviewers of children. *Child Abuse Negl.* **2022**, *129*, 105685. [CrossRef] [PubMed]
9. Salehi, P.; Hassan, S.Z.; Baugerud, G.A.; Powell, M.; Johnson, M.S.; Johansen, D.; Sabet, S.S.; Riegler, M.A.; Halvorsen, P. A theoretical and empirical analysis of 2D and 3D Virtual Environments in Training for Child Interview Skills. *IEEE Access* **2024**, *12*, 131842–131864. [CrossRef]
10. Salehi, P.; Hassan, S.Z.; Shafiee Sabet, S.; Astrid Baugerud, G.; Sinkerud Johnson, M.; Halvorsen, P.; Riegler, M.A. Is more realistic better? A comparison of game engine and gan-based avatars for investigative interviews of children. In Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval, Newark, NJ, USA, 27–30 June 2022; pp. 41–49.
11. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
12. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
13. Hsu, W.N.; Tsai, Y.H.H.; Bolte, B.; Salakhutdinov, R.; Mohamed, A. HuBERT: How much can a bad teacher benefit ASR pre-training? In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6533–6537.
14. Chakhtouna, A.; Sekkate, S.; Abdellah, A. Unveiling embedded features in Wav2vec2 and HuBERT msodels for Speech Emotion Recognition. *Procedia Comput. Sci.* **2024**, *232*, 2560–2569. [CrossRef]
15. Schubert, M.E.; Langerman, D.; George, A.D. Benchmarking Inference of Transformer-Based Transcription Models with Clustering on Embedded GPUs. *IEEE Access* **2024**, *12*, 123276–123293. [CrossRef]
16. Chakravarty, A. Deep Learning Models in Speech Recognition: Measuring GPU Energy Consumption, Impact of Noise and Model Quantization for Edge Deployment. *arXiv* **2024**, arXiv:2405.01004.
17. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the International Conference on Machine Learning. PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
18. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In Proceedings of the International Conference on Machine Learning. PMLR, New York, NY, USA, 19–24 June 2016; pp. 173–182.

19. Salehi, P. Whisper AFE for Talking Heads Generation, Version 1.0.0. 2024. Available online: https://github.com/pegahs1993/Whisper-AFE-TalkingHeadsGen (accessed on 20 February 2025).
20. Dalli, K.C. Technological Acceptance of an Avatar Based Interview Training Application: The Development and Technological Acceptance Study of the AvBIT Application. 2021. Available online: http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-107108 (accessed on 20 January 2025).
21. Røed, R.K.; Powell, M.B.; Riegler, M.A.; Baugerud, G.A. A field assessment of child abuse investigators' engagement with a child-avatar to develop interviewing skills. *Child Abuse Negl.* **2023**, *143*, 106324. [CrossRef]
22. Benson, M.S.; Powell, M.B. Evaluation of a comprehensive interactive training system for investigative interviewers of children. *Psychol. Public Policy Law* **2015**, *21*, 309. [CrossRef]
23. Pompedda, F. Training in Investigative Interviews of Children: Serious Gaming Paired with Feedback Improves Interview Quality. Doctoral Dissertation, Åbo Akademi University, Turku, Finland, 2018.
24. Krause, N.; Gewehr, E.; Barbe, H.; Merschhemke, M.; Mensing, F.; Siegel, B.; Müller, J.L.; Volbert, R.; Fromberger, P.; Tamm, A.; et al. How to prepare for conversations with children about suspicions of sexual abuse? Evaluation of an interactive virtual reality training for student teachers. *Child Abuse Negl.* **2024**, *149*, 106677. [CrossRef] [PubMed]
25. Meng, M.; Zhao, Y.; Zhang, B.; Zhu, Y.; Shi, W.; Wen, M.; Fan, Z. A Comprehensive Taxonomy and Analysis of Talking Head Synthesis: Techniques for Portrait Generation, Driving Mechanisms, and Editing. *arXiv* **2024**, arXiv:2406.10553.
26. Lu, Y.; Chai, J.; Cao, X. Live speech portraits: Real-time photorealistic talking-head animation. *ACM Trans. Graph. (ToG)* **2021**, *40*, 1–17. [CrossRef]
27. Chung, Y.A.; Glass, J. Generative pre-training for speech with autoregressive predictive coding. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3497–3501.
28. Ji, X.; Lin, C.; Ding, Z.; Tai, Y.; Yang, J.; Zhu, J.; Hu, X.; Zhang, J.; Luo, D.; Wang, C. RealTalk: Real-time and Realistic Audio-driven Face Generation with 3D Facial Prior-guided Identity Alignment Network. *arXiv* **2024**, arXiv:2406.18284.
29. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* **2023**, *42*, 1–25. [CrossRef]
30. Chen, B.; Hu, S.; Chen, Q.; Du, C.; Yi, R.; Qian, Y.; Chen, X. GSTalker: Real-time Audio-Driven Talking Face Generation via Deformable Gaussian Splatting. *arXiv* **2024**, arXiv:2404.19040.
31. Cho, K.; Lee, J.; Yoon, H.; Hong, Y.; Ko, J.; Ahn, S.; Kim, S. GaussianTalker: Real-Time Talking Head Synthesis with 3D Gaussian Splatting. In Proceedings of the ACM Multimedia 2024, Melbourne, Australia, 28 October–1 November 2024.
32. Tang, J.; Wang, K.; Zhou, H.; Chen, X.; He, D.; Hu, T.; Liu, J.; Zeng, G.; Wang, J. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv* **2022**, arXiv:2211.12368.
33. Guo, Y.; Chen, K.; Liang, S.; Liu, Y.J.; Bao, H.; Zhang, J. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5784–5794.
34. Li, J.; Zhang, J.; Bai, X.; Zhou, J.; Gu, L. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 7568–7578.
35. Ye, Z.; He, J.; Jiang, Z.; Huang, R.; Huang, J.; Liu, J.; Ren, Y.; Yin, X.; Ma, Z.; Zhao, Z. Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. *arXiv* **2023**, arXiv:2305.00787.
36. Ye, Z.; Zhang, L.; Zeng, D.; Lu, Q.; Jiang, N. R2-Talker: Realistic Real-Time Talking Head Synthesis with Hash Grid Landmarks Encoding and Progressive Multilayer Conditioning. *arXiv* **2023**, arXiv:2312.05572.
37. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
38. Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; Nießner, M. Neural voice puppetry: Audio-driven facial reenactment. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 716–731.
39. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
40. Baevski, A.; Schneider, S.; Auli, M. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv* **2019**, arXiv:1910.05453.
41. Jegou, H.; Douze, M.; Schmid, C. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 117–128. [CrossRef]
42. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.
43. Kahn, J.; Riviere, M.; Zheng, W.; Kharitonov, E.; Xu, Q.; Mazaré, P.E.; Karadayi, J.; Liptchinsky, V.; Collobert, R.; Fuegen, C.; et al. Libri-light: A benchmark for asr with limited or no supervision. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7669–7673.

44. Joshi, M.; Chen, D.; Liu, Y.; Weld, D.S.; Zettlemoyer, L.; Levy, O. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 64–77. [CrossRef]
45. Ye, Z.; Jiang, Z.; Ren, Y.; Liu, J.; He, J.; Zhao, Z. Geneface: Generalized and high-fidelity audio-driven 3D talking face synthesis. *arXiv* **2023**, arXiv:2301.13430.
46. Amazon Web Services. AWS Polly. Web Services. Amazon. 2024. Available online: https://aws.amazon.com/polly/ (accessed on 28 September 2024).
47. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595.
48. Chen, L.; Li, Z.; Maddox, R.K.; Duan, Z.; Xu, C. Lip movements generation at a glance. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 520–535.
49. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6626–6637.
50. Baltrušaitis, T.; Robinson, P.; Morency, L.P. Openface: An open source facial behavior analysis toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
51. Chung, J.S.; Zisserman, A. Out of time: Automated lip sync in the wild. In Proceedings of the Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, 20–24 November 2016; pp. 251–263.
52. Attig, C.; Rauh, N.; Franke, T.; Krems, J.F. System latency guidelines then and now–is zero latency really considered necessary? In Proceedings of the Engineering Psychology and Cognitive Ergonomics: Cognition and Design: 14th International Conference, EPCE 2017, Vancouver, BC, Canada, 9–14 July 2017; pp. 3–14.
53. Forch, V.; Franke, T.; Rauh, N.; Krems, J.F. Are 100 ms fast enough? Characterizing latency perception thresholds in mouse-based interaction. In Proceedings of the Engineering Psychology and Cognitive Ergonomics: Cognition and Design: 14th International Conference, EPCE 2017, Vancouver, BC, Canada, 9–14 July 2017; pp. 45–56.
54. Meskys, E.; Kalpokiene, J.; Jurcys, P.; Liaudanskas, A. Regulating deep fakes: Legal and ethical considerations. *J. Intellect. Prop. Law Pract.* **2020**, *15*, 24–31. [CrossRef]
55. Hu, H. Privacy Attacks and Protection in Generative Models. 2023. Available online: https://hdl.handle.net/10993/58928 (accessed on 20 February 2025).
56. Kugler, M.B.; Pace, C. Deepfake privacy: Attitudes and regulation. *Northwestern Univ. Law Rev.* **2021**, *116*, 611. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article IX: Multimodal Integration Challenges in Emotionally Expressive Child Avatars for Training Applications

Authors: *Pegah Salehi*, Sajad Amouei Sheshkal, Vajira Thambawita, Michael A Riegler, Pål Halvorsen

Candidate's contributions: The candidate was the primary contributor to this study, overseeing all stages of the research. She initiated the research questions, designed the system architecture, and led the integration of Unreal Engine 5 MetaHuman with NVIDIA Omniverse Audio2Face for real-time facial emotion synthesis in child avatars. Her work involved implementing the distributed two-PC framework for synchronized audio-driven facial animation, selecting TTS models for perceptual alignment, and optimizing the system for both VR and desktop environments. She designed and conducted the user study with 70 participants, analyzing quantitative outcomes using cumulative link mixed-effects models and interpreting qualitative feedback to assess perceived realism, emotional clarity, and empathy. As first author, she wrote the manuscript, developed the figures, and synthesized the results to highlight key design trade-offs in multimodal avatar systems. She also coordinated iterative feedback from co-authors and ensured that the system's software and data were curated and publicly released to support reproducibility.

submitted in: Nordic Machine Intelligence (NMI) Journal

Contributed to: This work contributed to the **RO-5**.

NORDIC MACHINE INTELLIGENCE

Nordic Machine Intelligence, vol. x, pp. 1–18, 20xx
Received xx Jan 20xx / published xx Jan 20xx
https://doi.org/10.5617/nmi.some-doi

Multimodal Integration Challenges in Emotionally Expressive Child Avatars for Training Applications

Pegah Salehi[1,2], Sajad Amouei Sheshkal[1], Vajira Thambawita[1], Michael A. Riegler[3], Pål Halvorsen[1,4]
[1]SimulaMet, Oslo, Norway [2]UiT, Tromsø, Norway [3]Simula, Oslo, Norway [4]OsloMet, Oslo, Norway
E-mail correspondence: pegah@simula.no

Abstract

Dynamic facial emotion is essential for believable AI-generated avatars; however, most systems remain visually inert, limiting their utility in simulations such as virtual training for investigative interviews with abused children. We introduce and evaluate a real-time architecture fusing Unreal Engine 5 MetaHuman rendering with NVIDIA Omniverse Audio2Face to translate vocal prosody into high-fidelity facial expressions on photorealistic child avatars. Due to limitations in synthetic voice options, both avatars were voiced using a young adult female TTS model, selected from two different systems in an attempt to better match each character. This compromise introduces a confounding factor. Voice-age mismatches and prosodic or emotional differences may disrupt audiovisual alignment. We implemented a two-PC setup decoupling language and speech synthesis from GPU-intensive rendering, designed to support low-latency interaction in desktop and VR environments. A between-subjects study ($N = 70$) using audio+visual and visual-only conditions assessed perceptual impacts as participants rated emotional clarity, facial realism, and empathy for two avatars expressing joy, sadness, and anger.

Our results show that while avatars could express emotions recognizably, particularly sadness and joy, recognition of anger dropped markedly without audio, highlighting the role of vocal cues in conveying high-arousal states. Interestingly, silencing the clips improved perceived realism by removing mismatches between facial animation and voice, especially where age or emotional tone were incongruent. These findings highlight that perceived believability hinges on the interplay between audiovisual congruence and facial geometry: a mismatched voice can undermine even well-crafted expressions, while a congruent one can enhance weaker visuals. This trade-off presents an ongoing challenge for designing emotionally coherent avatars in sensitive training contexts.

Keywords: Human-Computer Interaction; Interactive Avatars; Emotionally Expressive Avatars; Perceived Realism, Unreal Engine MetaHuman; NVIDIA Audio2Face

Introduction

AI-generated avatars are rapidly becoming integral components of digital interaction across diverse domains, with applications spanning virtual, augmented, and mixed reality (VR/AR/XR) platforms, virtual telepresence, embodied digital assistants, e-commerce, entertainment, education, and potential therapeutic interventions. The efficacy and adoption of these avatars critically depend on their perceived believability and their capacity to engage users in naturalistic interactions. For an avatar to "feel alive" and effectively bridge the human-computer interaction gap, it must convincingly replicate key facets of human social signaling [1]. Among these, the dynamic expression of facial emotions stands out as particularly vital. Facial expressions serve as a primary, often subconscious, channel for non-verbal communication, conveying a wealth of affective information that shapes interpersonal understanding, relationship dynamics, and the interpretation of intent [2]. Consequently, an avatar's ability to display appropriate, recognizable, and dynamically evolving facial emotions is paramount for creating engaging, trustworthy, and effective virtual agents. This paper investigates the technical feasibility of creating such avatars and evaluates their perceptual impact, particularly in training scenarios.

perceived believability hinges on the interplay between audiovisual congruence and facial geometry

While the importance of such affective realism is broadly recognized, its consistent and effective implementation is foundational and challenging in high-stakes domains. One such domain is simulation-based training for professionals who conduct forensic interviews with children, often concerning experiences of abuse or trauma. In this context, previous virtual-child systems have primarily focused on dialogue flow and question typology. However, empirical evaluations have repeatedly highlighted two significant shortcomings [4]: firstly, the avatars themselves



Figure 1: Example frames from the two avatars used in the user study: "Emory" and "Amelia". The top row presents their neutral facial expressions, while the bottom row illustrates expressions of joy (left), sadness (middle), and anger (right). High-resolution videos of all expressions are available in the project's GitHub repository [3].

are frequently emotionally inert, their faces remaining neutral regardless of the conversational context, thereby offering no micro-expressions that might signal anxiety, relief, or distress. Secondly, practitioners often find the overall visual fidelity of these avatars inadequate, with subtle inconsistencies in proportions, shading, or motion disrupting immersion and undermining trust in the simulation. Collectively, these deficits impede core learning objectives; interviewers are unable to accurately judge whether a question soothes or unsettles the virtual child, cannot adequately fine-tune their empathic responses, and consequently struggle to transfer refined interviewing skills to live cases where non-verbal cues are indispensable. Furthermore, the capacity to render these digital characters in real time is essential for such interactive applications, facilitating immediate responsiveness and a heightened sense of presence.

Addressing these limitations requires a focused research endeavor into developing AI avatar systems capable of rich, dynamic facial emotion expression with real-time responsiveness. This research directly tackles this objective by introducing and assessing a novel, distributed architecture engineered for high-fidelity emotional expressiveness. This system represents a significant advancement by integrating the sophisticated rendering capabilities of Unreal Engine 5 (UE5) MetaHuman software [5] with the intelligent, audio-driven facial animation provided by NVIDIA's Omniverse Audio2Face [6]. The system's distributed architecture employs two synchronized computational nodes to effectively balance processing demands by decoupling language understanding and speech generation from GPU-intensive rendering, thereby enabling low-latency interaction. Figure 1 displays representative screenshots of the two photorealistic MetaHuman avatars, "Amelia" and "Emory," utilized in this study.

In summary, this study offers the following contributions:

- Development of a modular, real-time pipeline integrating Unreal Engine MetaHuman and NVIDIA Audio2Face to generate emotionally expressive child avatars for interactive applications.
- Empirical investigation ($N = 70$) comparing audio+visual and visual-only conditions, revealing how mismatched modalities can reduce perceived realism, and in some cases, visual-only presentations may be preferred.
- Identification of a perceptual asymmetry across emotional arousal levels: while low-arousal emotions (e.g., sadness, joy) were recognizable from facial cues alone, high-arousal emotions (e.g., anger) were more dependent on audiovisual congruence.
- Observation that avatar facial morphology (e.g., soft vs. angular features) influences emotion interpretation, suggesting the need for morphology-

emotion alignment in sensitive contexts such as empathy training.

- Extraction of practical design lessons from participant feedback, including the impact of audiovisual desynchrony, the trade-off between expression clarity and naturalism, and the importance of age-appropriate voice synthesis.

Related Work

The pursuit of emotionally expressive AI avatars lies at the confluence of affective computing, human-computer interaction (HCI), and advancements in multimodal systems, all striving to endow machines with a semblance of human emotional intelligence [7, 8, 9, 10]. While foundational work in affective computing centered on the detection and modeling of emotional cues, the emphasis has progressively shifted towards the dynamic and adaptive expression of these emotions in avatars to foster user engagement and perceived believability. This aligns with broader HCI principles where anthropomorphism in digital agents consistently correlates with enhanced user engagement, perceived credibility, and overall satisfaction [11, 12, 13, 14]. The significance of expressive capabilities is underscored by research in applied domains like virtual therapy and education, where emotionally inert or insufficiently expressive avatars can lead to misunderstandings, diminished rapport, and ultimately, reduced system effectiveness [15, 16, 17, 18].

This need for dynamic emotional expressiveness becomes particularly acute in high-stakes interaction scenarios, such as the virtual training environments for child forensic interviews. In these contexts, an avatar's ability to display contextually appropriate and recognizable emotions can profoundly influence user perception and training outcomes. However, existing virtual-child systems tailored for such training have predominantly concentrated on dialogue management and question typologies. Empirical evaluations of these systems have consistently identified critical shortcomings: specifically, avatars often lack dynamic facial affect, remaining emotionally static, and their overall visual fidelity is frequently judged as inadequate [4]. These deficiencies impair training efficacy, as trainees cannot accurately gauge a child's non-verbal reactions to their questioning or refine their empathic responses, which are skills paramount in real-world interviews.

The technological evolution towards more expressive avatars has been considerable. Initial efforts often involved 2D video-based avatars that, while capable of displaying emotions, tended to produce rigid outputs through direct mappings from input signals [19, 20, 21]. The advent of 3D parametric models, such as FLAME [22], offered more granular control over facial geometry and pose. Subsequently, sophisticated speech-driven animation systems like EMOTE [23] and EmoTalk [24] have emerged, aiming to disentangle and map speech content and prosody to facial expressions. However, as these systems push the boundaries of realism, they often encounter challenges in achieving both high-fidelity emotional rendering and the stringent real-time performance demanded by interactive, immersive simulations. Many existing solutions struggle to seamlessly integrate state-of-the-art rendering quality with low-latency, audio-driven emotional animation, particularly for photorealistic child avatars where subtlety and believability are crucial.

Collectively, the literature highlights a persistent gap: the need for AI avatar systems that not only generate rich and dynamic facial emotions but do so with high visual fidelity and in real-time, especially for demanding applications like child interview simulations. While advancements in speech-driven animation and realistic rendering are notable, their integration into a cohesive, responsive system that meets these multifaceted requirements remains a significant hurdle. Our research directly confronts this challenge by proposing and evaluating a novel, distributed architecture. This system uniquely fuses the high-fidelity rendering capabilities of Unreal Engine 5's MetaHuman with NVIDIA Omniverse Audio2Face for real-time, prosody-driven facial emotion on photorealistic child avatars. By decoupling language processing and speech generation from GPU-intensive rendering via a two-PC setup, our approach is specifically designed to achieve the low-latency interaction essential for effective and immersive training simulations.

Methodology

System Design and Implementation

To overcome the limitations of previous systems [4] for emotional expression in AI avatars, this research introduces a novel architecture that integrates Unreal Engine 5 Metahuman [5] for high-fidelity facial rendering with NVIDIA's Omniverse Audio2Face [6] for generating emotionally expressive facial animations from audio input. The source code and implementation details for the system described in this study are publicly available [3]. Below, we provide a brief overview of these two core platforms:

- **Unreal Engine 5 MetaHuman** integrates advanced real-time graphics technologies such as Nanite geometry, the Lumen lighting system, and optimized post-processing to deliver film-quality visuals without compromising interactivity [25, 26]. Using the MetaHuman Creator, researchers can import head scans or digital sculpts and generate fully rigged characters that are customizable in every anatomical detail and exportable at multiple levels of detail (LOD) for deployment across platforms ranging from mobile devices to virtual reality.

  The system includes a preconfigured facial rig built on UE5's Control Rig, which offers hundreds of easily retargetable controls. This allows expressions

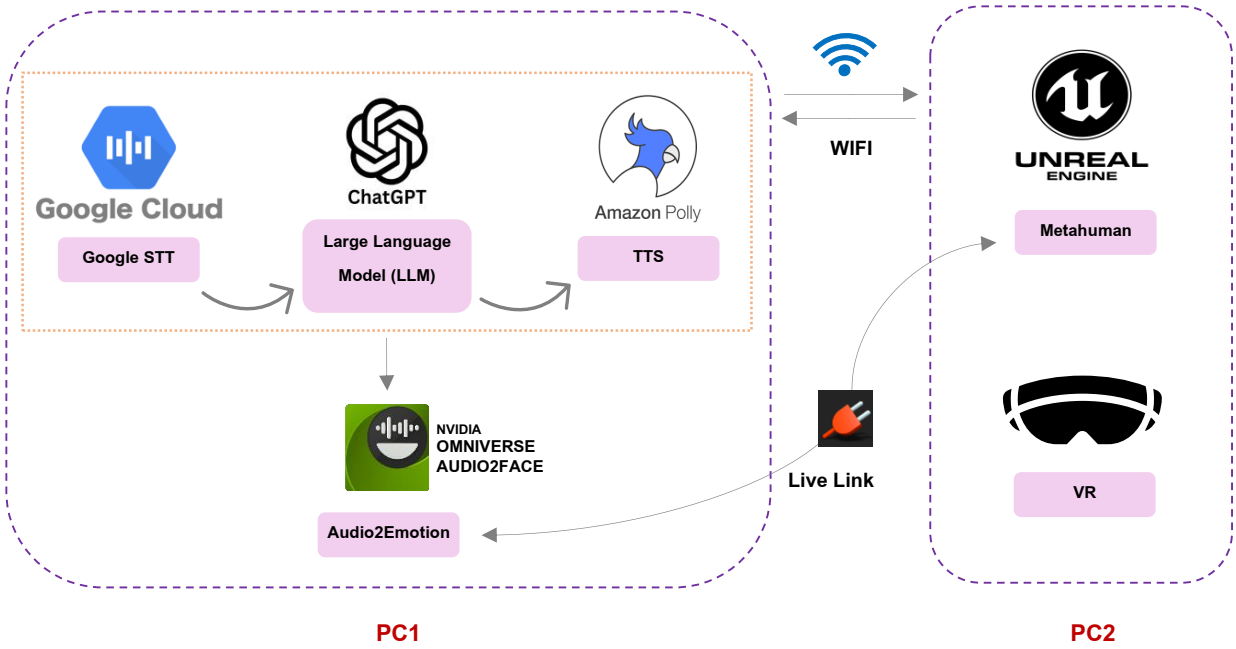


Figure 2: Architecture for real-time emotionally expressive child avatars. PC1 processes speech input (Google Cloud STT), generates a text response via a large language model (LLM), and synthesizes speech using Amazon Polly TTS. The resulting audio is analyzed by NVIDIA Audio2Face to extract emotional parameters for facial animation. PC2 renders a customizable Metahuman avatar (Unreal Engine) driven by this emotional data in real time. Synchronization is achieved using the Live Link plugin, and the system supports deployment in both desktop and VR environments.

to be driven by live capture, audio input, or keyframe animation. To achieve photorealistic skin, which is essential for avoiding the uncanny valley, MetaHuman incorporates subsurface scattering (SSS) [27], a rendering method that simulates the way light penetrates and diffuses through real skin. Additional realism is provided through groom-based hair and physically accurate materials. Together, these features enable a fast end-to-end pipeline for producing responsive and emotionally expressive avatars designed for AI-driven training scenarios involving child interactions.

- **NVIDIA Omniverse Audio2Face** is the core component responsible for generating lifelike facial animations in response to audio input. This generative AI tool analyzes speech to automatically produce synchronized lip movements and expressive facial animations in real time, thereby enhancing the emotional realism of virtual characters. It includes an intuitive character retargeting pipeline that enables users to apply animations to custom 3D characters, including seamless integration with Unreal Engine MetaHumans.

  The system provides a Blendshape Generation widget that converts audio into facial expressions using templates such as NV 46 Pose or ARKit 51 Pose. Although the tool primarily supports full-face animation, it can also transfer movements to individual facial features such as the eyes, jaw, and tongue.

  A key advantage of Audio2Face is its real-time performance, which makes it particularly well-suited for interactive applications including virtual reality, digital assistants, and serious games. It also supports traditional animation workflows via live playback or baked outputs. Additional features include multilingual input support, simultaneous animation of multiple characters, and manual adjustment of emotion intensity using facial action unit (FAU) controls. The interface also allows users to fine-tune automatic emotion extraction from audio using settings such as emotion detection range, strength, contrast, smoothing, and the number of concurrently active emotions. These capabilities enable fine-grained, context-aware emotional expression, which is especially important in sensitive domains such as child interview simulations.

  In this study, the "exaggerated" mode in Audio2Face was used to enhance the clarity and distinctiveness of facial expressions in short-duration clips, facilitating consistent emotion recognition among participants. This setting was chosen based on preliminary

observations during development, as the standard and automatic modes often produced expressions that were too subtle to be reliably perceived within the limited duration of the stimuli.

The proposed architecture, illustrated in Figure 2, enables real-time interactivity through the coordinated operation of two synchronized computational nodes. PC1 handles all speech and language processing. The user's spoken input is first transcribed using the Google Cloud Speech-to-Text (STT) engine [28], which leverages advanced models such as Google's foundation model Chirp to provide high accuracy across diverse accents and acoustic environments, with support for multilingual input. The resulting transcription is processed by OpenAI's ChatGPT Large Language Model (LLM) [29], which generates a context-appropriate textual response. This response is then synthesized into speech using Amazon Polly's Text-to-Speech (TTS) system [30], which offers lifelike multilingual voices and allows customization of intonation, pronunciation, and speaking style.

In parallel, NVIDIA Audio2Face analyzes the prosodic and spectral features of the synthesized speech [31]. It extracts frame-level emotional parameters, such as indicators of joy, fear, or sadness, based on vocal characteristics. These parameters are typically represented as blendshape weights or facial action unit (FAU) activations and are transmitted via Wi-Fi to PC2, the rendering node.

PC2 is responsible for avatar rendering and real-time facial animation. It runs Unreal Engine 5's Metahuman framework, which produces high-fidelity child avatars with anatomically accurate facial rigs. The emotional parameters received from PC1 are used to drive the avatar's expressions in real time by mapping the blendshape weights or FAU activations to the MetaHuman Control Rig via the Live Link plugin [32]. This ensures low-latency synchronization between facial animation and the synthesized voice.

This distributed system design separates the computationally intensive processes of natural language understanding and speech generation from the demanding task of real-time, high-resolution avatar rendering. While this architecture is designed for responsiveness, potential sources of audiovisual desynchrony can still arise from factors such as network latency or processing delays within either computational node. Nevertheless, this separation enables the system to maintain both overall responsiveness and visual realism, supporting immersive and emotionally expressive interactions in applications such as child interview training.

User Study

To examine the feasibility of incorporating emotional expressiveness into AI-driven avatars through dynamic facial emotion expression and to assess its perceptual impact, we conducted a user study focused on three key dimensions: emotional clarity, facial realism, and empathic connection. These dimensions were chosen as they are key indicators of avatar believability and effectiveness in interactive simulations. Two photorealistic MetaHuman avatars, named Emory and Amelia, were used to assess perceived emotional expressiveness under different conditions. Representative screenshots of the avatars are shown in Figure 1.

The study followed a between-subjects design with two experimental conditions: one condition included both voice and facial expression (audio+visual condition), while the other featured facial expression only (visual-only condition). This design was informed by prior findings suggesting that audio and visual modalities contribute differently to emotion perception, specifically that audio cues are more salient for detecting changes in emotional arousal, while visual cues are more informative for identifying emotional valence [33].

A total of 77 participants were initially recruited via the Prolific platform. Although initial eligibility was filtered using Prolific's pre-screening tools (e.g., age, language fluency, and device requirements), an additional quality control step was applied during data cleaning. Based on clearly invalid or inconsistent responses (e.g., repeated ratings across stimuli or contradictory answers), 7 participants were excluded from further analysis. The final sample included 70 participants whose responses met the criteria for attentiveness and internal consistency. Participants were randomly assigned to one of two experimental conditions ($N = 35$ per condition) using Prolific's built-in randomization tool. To minimize order effects, we manually randomized the sequence of avatar stimuli within each condition, varying both avatar identity and emotional expression. All participants in a given condition viewed the same randomized stimulus order. The sample exhibited diversity in gender, age, academic/professional background, and familiarity with virtual environments.

The participant pool ($N = 70$) was evenly distributed by gender, comprising 35 women (50.0%) and 35 men (50.0%). The predominant age group was 25–34 years (40 participants, 57.1%), followed by 35–44 years (16 participants, 22.9%). Smaller segments included participants aged 18–24 years (10 participants, 14.3%) and 45–54 years (4 participants, 5.7%). The academic and professional backgrounds of the participants were highly varied, with numerous unique responses.

Regarding familiarity with virtual environments and avatars, a significant portion of participants reported interacting "Occasionally" (32 participants, 45.7%). Other interaction frequencies included "Rarely" (15 participants, 21.4%), "Frequently" (12 participants, 17.1%), and "Very frequently" (9 participants, 12.9%). A small fraction (2 participants, 2.9%) reported "Never" interacting with avatars. This demographic profile indicates a sample primarily composed of young to middle-aged adults from diverse professional domains, possessing

a general, though not necessarily expert-level, familiarity with avatar interactions.

Participants viewed a series of six high-definition video clips, the presentation order of which was counterbalanced to avoid order effects. In these clips, the avatars expressed three emotional states: sadness, anger, and joy (see Figure 3). These specific emotions were selected as they represent a range of valence and arousal levels and are relevant emotions that children might express in sensitive interview contexts. In the audio+visual condition, speech was synthesized using two different TTS systems: OpenAI's TTS[1] for the *Emory* avatar and NaturalReaders[2] for the *Amelia* avatar. We evaluated several available text-to-speech options and selected these two voices based on their perceived suitability, as judged by two of the researchers. Although child-like voices were technically available, a young adult female voice was ultimately preferred for both avatars due to its better perceptual match with the visual and emotional characteristics of the characters. This choice was supported by findings from our prior studies [34], which showed that female voices can be effectively used to represent child characters in training scenarios, regardless of gender. Additionally, we found the child-like voices to be overly juvenile in tone and pitch, which created a mismatch with the avatars' appearance and demeanor, both of which were designed to resemble pre-teens or young adolescents rather than very young children.

The six video clips were generated using the described system architecture. For the audio+visual condition, the scripts were synthesized into audio, and this audio was processed by NVIDIA Audio2Face to generate facial animation data. This animation data was then applied to the MetaHuman avatars in Unreal Engine 5 and rendered into video clips. For the visual-only condition, the same facial animation data was applied to the avatars and rendered into video clips without the synthesized audio track.

To manage participant expectations and establish a perceptual anchor, we began with a deliberately low-fidelity stimulus, following evidence that anchoring effects require the presence of both semantic and numerical cues [35]. The inclusion of this initial low-fidelity stimulus was intended to provide a baseline or reference point for participants, potentially making their subsequent ratings of the higher-fidelity main stimuli more calibrated and less susceptible to being skewed by initial unfamiliarity or overly high expectations for avatar realism. An example frame from this anchoring stimulus is shown in Figure 4. By first exposing them to a clearly less sophisticated avatar, we aimed to help participants better contextualize and evaluate the nuances of the emotionally expressive avatars that formed the core of the study.

After viewing the main stimuli, participants completed a questionnaire comprising three main sections (see Table 1). items were rated on a five-point Likert scale. Emotion recognition and perceived facial realism were assessed for each individual video, while perceived realism across different contributing aspects (e.g., facial expressions, voice tone) and overall empathy toward the avatars were measured in a final, post-video section of the questionnaire. The final section included an open-ended question inviting participants to reflect on what stood out most about the avatars.

Ethical Considerations

This study involved the evaluation of emotional expressiveness in *synthetic* avatars by adult participants via anonymous online surveys. All visual and auditory stimuli were fully computer-generated and did not include any real individuals or personal data. This study did not require formal ethical approval, as it involved no interaction with vulnerable groups, no collection of sensitive personal data, and no deception. Participants were recruited through the Prolific platform and were informed about the nature and duration of the study prior to participation. No formal consent form was presented beyond Prolific's standard participation agreement. While the study featured simulated emotional expressions (e.g., sadness and anger) in photorealistic child avatars, no content warnings were provided prior to exposure. Some participants reported feeling uncomfortable or uneasy, particularly in response to expressions of sadness or anger. This is noted as a potential ethical consideration. Future implementations should consider including content advisories to reduce the risk of discomfort for sensitive viewers.

Results and Analysis

To comprehensively assess how participants interpreted the emotional expressions of the avatars, we conducted both quantitative and qualitative analyses.

Quantitative Analysis

The quantitative analysis in this study focused on how participants perceived and interpreted the emotional expressions of AI-generated child avatars across different experimental conditions. Specifically, we examined three main response categories:

1. **Emotion recognition**, defined as how accurately participants identified the intended emotional expression (anger, sadness, joy) after each video,
2. **Perceived facial expression realism**, referring to the extent to which facial animations appeared realistic in each video, and
3. **Perceived realism across contributing aspects** (e.g., facial expressions, visual appearance, voice tone, dialogue content), along with the **empathic connection** participants felt toward the avatars, assessed in a post-exposure questionnaire.

[1] https://ttsopenai.com/
[2] https://www.naturalreaders.com/

Sadness	*"I thought today would be different... but it wasn't. No one talked to me at school. I just sat alone... again. I don't know why they don't like me. Maybe I'm just not good enough."*
Anger	*"It's not fair! I told them it wasn't my fault, but no one listened. They just blamed me like always! I'm so tired of this. I try to do the right thing, but it never matters!"*
Joy	*"Today was amazing! My teacher said I did a great job on my homework, and my friends invited me to play during lunch. Everything just felt really good ... like things are finally going my way!"*

Figure 3: Scripted utterances used for emotional avatar expressions in the user study.

Table 1: Structure and content of the user study questionnaire, including measures of perceived realism, emotional expressiveness, and overall impressions. The full questionnaire is available in the project's GitHub repository [3].

Section	Question Type	Content
Emotion Recognition	Likert scale	Extent to which the avatar expressed: Anger, Sadness, Joy, Fear, Disgust
Facial Realism	Likert scale	Realism of facial expressions in each video
Realism Contributors	Likert scale	Contribution of each factor to perceived realism: - Facial expressions - Visual appearance - Voice tone* - Dialogue content*
Empathy/Connection	Likert scale	Emotional connection or empathy felt toward the avatars overall
Open-ended Reflection	Free text	One sentence describing what stood out most about the avatars
Demographics	Multiple choice	Gender, age group, academic/professional background, frequency of avatar interaction

** Voice tone and dialogue content were excluded from the visual-only condition.*



Figure 4: Low-fidelity anchoring stimulus used at the beginning of the user study to calibrate participants' expectations regarding avatar realism.

We first computed descriptive statistics (mean and standard deviation) for each category. To assess effects of condition, avatar identity, and emotion scenario on these outcomes, we employed cumulative link mixed-effects models (CLMMs) using the `ordinal` package in R. These models incorporated fixed effects for *Condition* (audio+visual vs. visual-only), *Avatar* (Amelia vs. Emory), and *Scenario* (anger, sadness, joy), along with their interaction terms. Random intercepts were included for participants and, where applicable, for emotion or video stimuli, to account for within-subject and item-level variability. Model fitting was performed using a logit link and adaptive Gauss–Hermite quadrature (`nAGQ = 10`) to ensure robust convergence and estimation accuracy.

Before examining each response category in detail, we begin by validating the robustness and goodness-of-fit of our statistical models. The results are then presented across the following dimensions: emotion recognition, perceived facial realism, and perceived realism/empathy. For completeness and cross-method validation, supplementary t-test analyses are reported in Appendix Table 8 and Table 9.

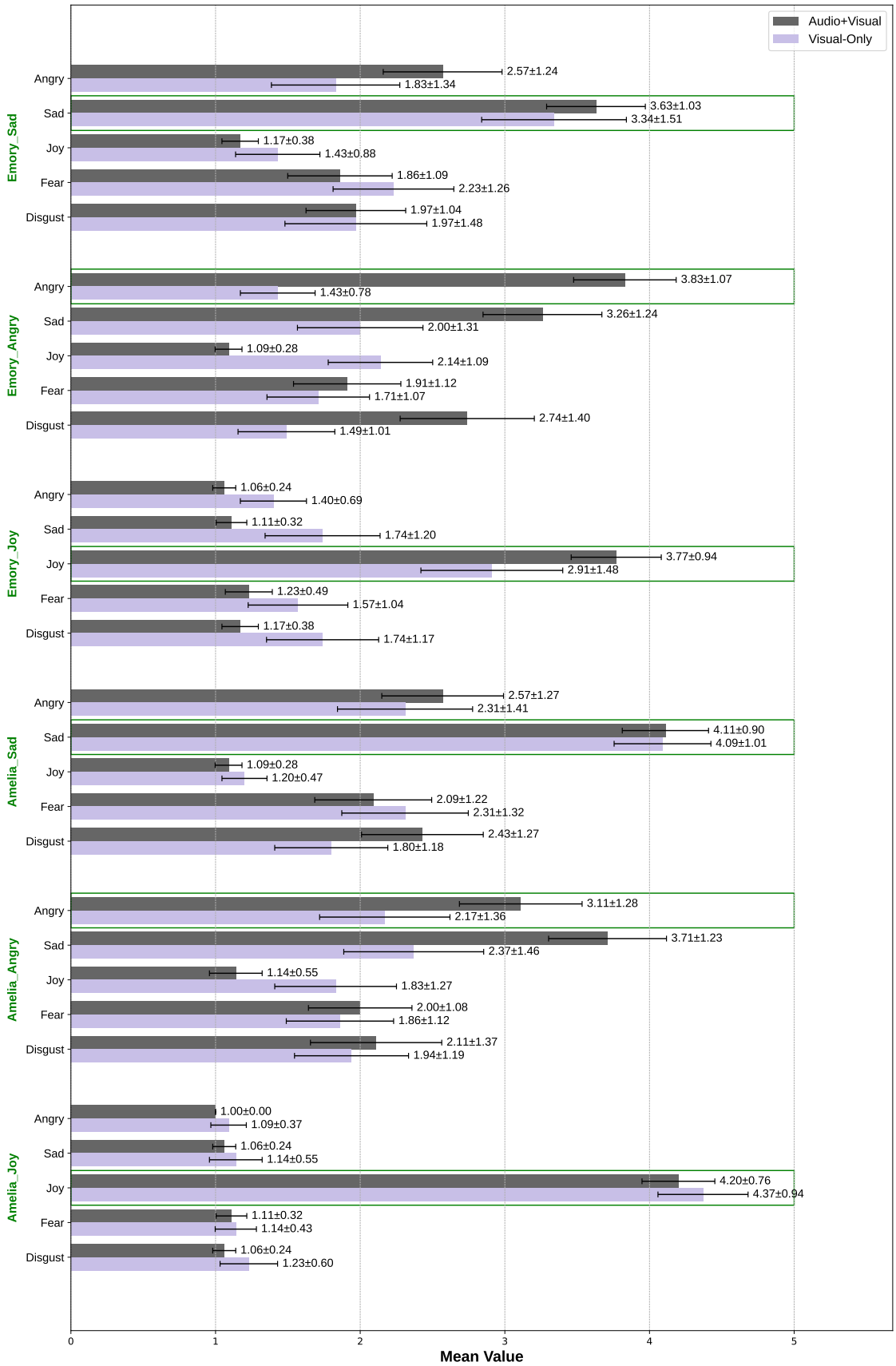


Figure 5: Mean ratings ($M \pm SD$) of emotional expressiveness across five emotions (Angry, Sad, Joy, Fear, Disgust) for the avatars *Emory* and *Amelia*, under both audio+visual and visual-only conditions. Bars represent group means; values next to each bar report the mean and standard deviation, while horizontal error bars indicate 95% confidence intervals (CI). Each video clip is labeled on the left, with color-coded bars distinguishing the experimental conditions. Green-highlighted boxes indicate target emotions critical for the study's recognition task.

Statistical Validation and Model Robustness

To evaluate the reliability and robustness of the statistical modeling framework, we performed an extensive set of diagnostic procedures, model comparisons, and fit assessments across all three primary response categories. These included emotion recognition, perceived facial realism, and post-exposure ratings of realism contributors and empathic engagement.

For the emotion recognition analysis, model selection was guided by information-theoretic criteria, specifically the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), along with likelihood ratio tests (LRTs) to assess violations of the proportional-odds (PO) assumption. As shown in Table 2, relaxing the Proportional Odds (PO) constraint significantly improved model fit ($p < 0.001$), justifying the use of more flexible specifications in the CLMMs. To ensure sufficient power, an a priori simulation-based power analysis was conducted, confirming that a sample size of $N = 70$ was adequate to detect medium-sized effects with a power of at least 0.80.

Similarly, Table 3 summarizes model comparisons for perceived facial expression realism. Again, significant improvements were observed when PO constraints were relaxed (e.g., LRT = 33.35, df = 15, $p = 0.004$), indicating that ordinal responses were not evenly distributed across thresholds. The best-performing CLMMs, detailed in Table 4, achieved a conditional R^2 of 0.640, which indicates strong explanatory power when both fixed and random effects were considered. The marginal R^2 of 0.099 reflects the variance attributable to fixed effects alone. Residual diagnostics revealed an RMSE of 3.176 and residual standard deviation ($\sigma = 3.238$), confirming acceptable levels of predictive error and model stability.

In addition to these analyses, we assessed model robustness for the third response category, which included perceived realism across contributing aspects (e.g., facial expressions, visual appearance, voice tone, dialogue content) and overall empathy toward the avatars. Internal consistency for the multi-item realism contributors scale was found to be marginal, with a Cronbach's alpha of $\alpha = 0.652$, suggesting moderate internal reliability. Furthermore, a logistic regression model predicting empathy responses yielded a Nagelkerke R^2 of just 0.011, indicating minimal variance explained. These findings suggest high individual variability and underscore the challenge of modeling subjective constructs like empathic connection using fixed predictors alone.

Across all three domains, model coefficients were exponentiated to derive interpretable odds ratios (OR) and visualized using forest plots. Where applicable, multiple comparisons were corrected using Holm's method. To contextualize statistical findings, open-ended responses were triangulated with quantitative patterns, particularly regarding visual fidelity, emotional clarity, and audiovisual congruence. These effects were especially pronounced in high-arousal conditions, such as anger, where audiovisual alignment proved critical for accurate emotion recognition.

Taken together, these results validate the overall modeling approach and underscore its suitability for capturing the complex interplay between avatar design, experimental conditions, and emotional perception. The convergence of statistical rigor, diagnostic confirmation, and qualitative insight provides a robust foundation for interpreting user responses to emotionally expressive virtual avatars.

Table 2: Model comparison for emotion recognition accuracy. Relaxing the PO assumption significantly improved model fit, as confirmed by the likelihood ratio test ($p < .001$).

Df	logLik	AIC	LRT	Pr(>Chisq)	no.par
6	-2521.47	5076.95	30.57	0.00	17

Table 3: Model comparison for perceived facial expression realism. Relaxing the PO assumption improved model fit significantly across multiple model specifications.

Df	logLik	AIC	LRT	Pr(>Chi)
3.00	-382.84	937.69	9.34	0.02
6.00	-380.78	939.57	13.46	0.03
15.00	-370.84	937.69	33.34	0.004

Emotion Recognition

To initiate our analysis of emotion recognition, we present the results of the ordinal mixed-effects model (Table 5), supported by two key visualizations: a forest plot illustrating the influence of model predictors (Figure 6) and a bar graph comparing perceived emotional expressiveness across experimental conditions and avatars identity (Figure 5). Together, these results offer a comprehensive overview of participants' accuracy in identifying the target emotions conveyed by the avatars.

Figure 5 shows that *sadness* and *joy* were recognized with high accuracy across both conditions. Sadness recognition remained stable for both avatars (Amelia: $M = 4.11$ vs. $M = 4.09$; Emory: $M = 3.63$ vs. $M = 3.34$), consistent with its strong reliance on static visual cues such as brow depression (AU1+4) and lip-corner droop (AU15) as defined in FACS [36].

In contrast, recognition of joy varied by avatar identity and condition. Amelia's joy remained consistently high, slightly improving in the visual-only condition ($M = 4.37$ vs. $M = 4.20$). Emory's joy ratings, however, dropped 22% in visual-only condition ($M = 3.77$ to $M = 2.91$). This disparity may reflect both visual ambiguity and a potential methodological limitation, namely the use of two different TTS applications, which may have introduced variations in emotional prosody, and the use of adult female voices for both avatars, which might not have fully matched

Table 4: Fit statistics for the best-performing CLMM in the facial realism analysis. Conditional and marginal R^2 values indicate explained variance; RMSE and σ reflect prediction error and residual variability.

AIC	BIC	R2_conditional	R2_marginal	RMSE	σ
1035.65	1100.30	0.64	0.09	3.17	3.23

their child-like appearance. Emory's voice, generated with OpenAI's TTS, exhibited strong emotional expressiveness, particularly in terms of pitch variation and intensity. However, this high affective prosody may not have been fully matched by the corresponding facial expressions, which might have contributed to a perceived mismatch between voice and visuals. In contrast, Amelia's voice, synthesized using NaturalReaders, seemed to align more closely with the visual and emotional tone of the avatar, although this could also be influenced by other factors such as facial morphology or expression timing. Nevertheless, we cannot attribute the entire effect to the voice alone, as the study also included visual-only conditions in which Amelia's avatar consistently performed better in terms of emotion recognition and perceived realism. This suggests that additional factors, such as facial morphology and expression clarity, also played a substantial role. Amelia's advantage may be partly explained by her softer facial features and the presence of clear Duchenne markers (e.g., AU6+12), which enhanced visual clarity even in the absence of audio. In contrast, Emory's more angular facial structure may have contributed to perceptual ambiguity, with features such as brow tension being more easily misinterpreted as negative affect.

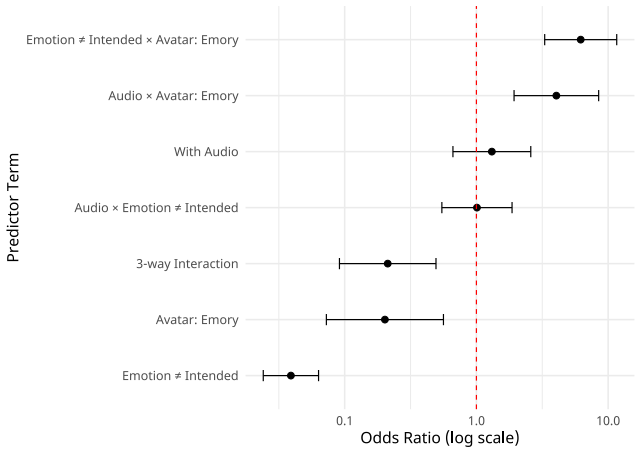


Figure 6: Forest plot showing OR and 95% CI for predictors of emotion recognition accuracy. The dashed vertical line indicates no effect (OR = 1). Statistically significant predictors are clearly indicated. *Reference levels:* Avatar = Amelia, Condition = visual-only, Emotion = intended.

These findings reinforce the idea that while sadness can be conveyed effectively through visual cues alone, accurate recognition of emotions like joy often depends on audiovisual congruence, particularly for avatars with less prototypical or more ambiguous facial expressions [2].

Among all emotions, anger recognition was the most negatively affected in the visual-only condition. Ratings for Emory fell from $M = 3.83$ to $M = 1.43$, and for Amelia from $M = 3.11$ to $M = 2.17$. Silent anger was often misinterpreted as sadness, reflecting the high reliance of anger on auditory cues such as spectral roughness and vocal intensity [37, 38]. Visual signals alone, particularly AU4, were insufficient for conveying high-arousal affect.

Even in the audio+visual condition, Amelia's anger ($M = 3.11$) was frequently overshadowed by sadness ($M = 3.71$), suggesting perceptual interference from her softer features and higher-pitched voice. Emory's more angular structure and pronounced brow tension better conveyed anger ($M = 3.83$ vs. $M = 3.26$ for sadness), underscoring the role of morphology in affect salience. This may have been further influenced by his highly emotive voice, which could have reinforced the expression of anger by providing strong auditory cues.

The statistical analysis (Table 5, Figure 6) confirmed a significant three-way interaction among experimental conditions (audio+visual vs. visual-only), avatar identity, and emotion congruency ($p < .001$, OR = 0.212). This finding highlights the complex and interdependent nature of emotion recognition accuracy, suggesting that participants' ability to correctly interpret emotional expressions was jointly shaped by the presentation mode, the specific avatar shown, and whether the expressed emotion aligned with the intended target.

In addition, the model identified a highly significant main effect for emotional incongruency. When the emotion recognized by the participant did not match the intended expression, the odds of accurate recognition dropped dramatically ($p < .001$, OR = 0.039, 95% CI [0.024, 0.063]). This indicates that congruency between intended and perceived emotion was a critical factor in successful emotion interpretation.

Notably, the main effect of audio+visual condition alone was not statistically significant ($p = .437$, OR = 1.310), suggesting that the presence of auditory cues did not independently enhance recognition accuracy across conditions. Rather, the influence of audio appeared to depend on interactions with other variables, particularly avatar identity.

Avatar identity emerged as a significant factor in recognition performance. Specifically, participants were substantially less accurate in identifying emotions displayed by the avatar *Emory* compared to *Amelia* ($p = .002$, OR = 0.202, 95% CI [0.073, 0.562]). Moreover, a significant

Table 5: Ordinal mixed-effects model results predicting emotion recognition accuracy. Odds ratios (OR), 95% confidence intervals (CI), and p-values are reported. Significant effects ($p < .05$) are highlighted. *Reference levels:* Avatar = Amelia, Condition = visual-only, Emotion = intended. All model terms are interpreted as contrasts relative to these reference levels.

term	p.value	OR	OR_low	OR_high
Emotion ≠ Intended	0.000	0.039	0.024	0.063
Audio+Visual	0.437	1.310	0.663	2.586
Avatar: Emory	0.002	0.202	0.073	0.562
Audio+Visual × Emotion ≠ Intended	0.976	1.010	0.547	1.864
Emotion ≠ Intended × Avatar: Emory	0.000	6.191	3.299	11.618
Audio+Visual × Avatar: Emory	0.000	4.045	1.931	8.472
3-way Interaction	0.000	0.212	0.091	0.493

interaction between experimental condition and avatar identity (*Audio+Visual* × *Avatar: Emory*; $p < .001$, OR = 4.045, 95% CI [1.931, 8.472]) indicated that the effect of audio cues seemed to vary depending on other factors—especially which avatar was shown.

Importantly, the interaction between incongruent emotional expressions and avatar identity (*Emotion ≠ Intended* × *Avatar: Emory*; $p < .001$, OR = 6.191, 95% CI [3.299, 11.618]) revealed that the negative impact of mismatched emotional cues was especially pronounced for Emory. This suggests that when Emory's facial expression did not match the target emotion, participants were particularly prone to misinterpretation, underscoring both the sensitivity of emotion recognition to the alignment between vocal and facial emotional cues, and the role of avatar-specific morphology in modulating perceptual accuracy.

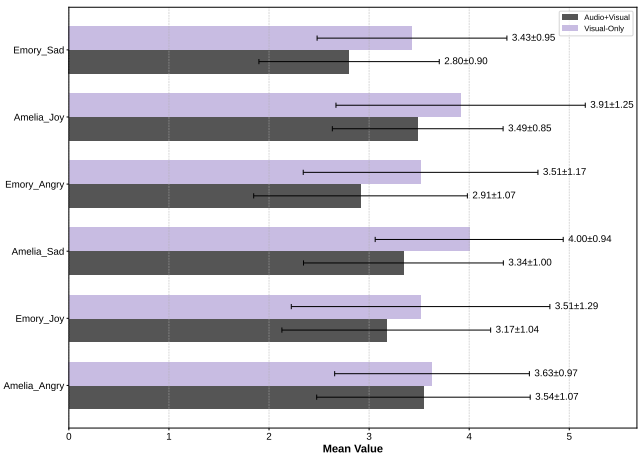


Figure 7: Mean ratings ($M \pm SD$) perceived Facial Expressions realism for each video clips under both audio+visual and visual-only conditions.

One plausible explanation for Amelia's superior emotion recognition lies in her softer facial morphology, which aligns more closely with users' expectations of emotional authenticity [39]. Her rounded features and gentle brow dynamics enhanced the perceived realism of low-arousal emotions like sadness and joy. In contrast, Emory's angular face and more pronounced brow movements made high-arousal expressions like anger more salient [38, 37]. However, beyond facial structure, the alignment between voice characteristics and facial appearance may also have played a role. Amelia's relatively soft and emotionally neutral voice appeared to match her visual design, potentially reinforcing the perception of coherence. In contrast, Emory's highly expressive voice may have occasionally clashed with his facial expressions, leading to moments of perceptual mismatch and ambiguity. This highlights a key design trade-off: while subtle facial geometry can enhance believability for certain emotions, it may reduce clarity for others, especially when not fully synchronized with vocal affect.

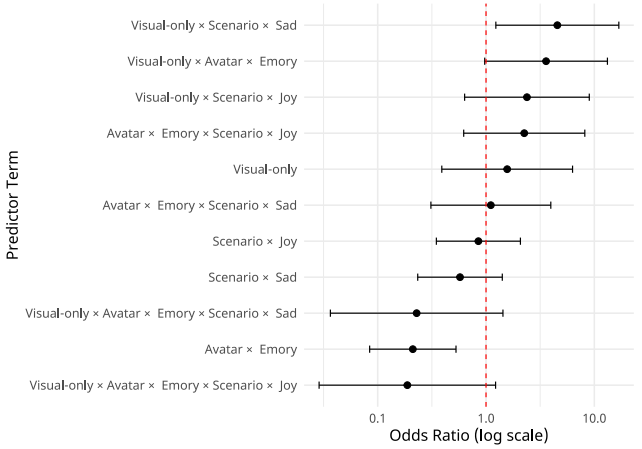


Figure 8: Forest plot showing OR with 95% CI for the model predictors of perceived realism of facial expressions. The vertical dashed line indicates no effect (OR = 1). Statistically significant predictors are clearly indicated. *Reference levels:* Avatar = Amelia, Condition = Audio+Visual, Scenario = Angry.

Table 6: Results from the ordinal mixed-effects model predicting perceived realism of avatars' facial expressions. Odds ratios (OR), 95% confidence intervals (CI), and p-values are presented. Significant effects ($p < .05$) are highlighted. *Reference levels:* Avatar = Amelia, Condition = Audio+Visual, Scenario = Angry. All contrasts are interpreted relative to these reference levels.

term	p.value	OR	OR.lower	OR.upper
Visual-only	0.526	1.568	0.391	6.294
Avatar: Emory	0.001	0.211	0.084	0.528
Scenario: Joy	0.722	0.850	0.348	2.077
Scenario: Sad	0.228	0.575	0.234	1.413
Visual-only × Avatar: Emory	0.055	3.584	0.972	13.208
Visual-only × Scenario: Joy	0.198	2.387	0.635	8.981
Visual-only × Scenario: Sad	0.023	4.554	1.232	16.829
Avatar: Emory × Scenario: Joy	0.216	2.254	0.622	8.165
Avatar: Emory × Scenario: Sad	0.875	1.108	0.310	3.963
Visual-only × Avatar: Emory × Scenario: Joy	0.081	0.188	0.029	1.225
Visual-only × Avatar: Emory × Scenario: Sad	0.115	0.229	0.037	1.433

Perceived Facial Expressions Realism

To investigate the perceived realism of facial expressions, we analyzed participants' ratings across two experimental conditions (*audio+visual* and *visual-only*) and two avatar identities.

As shown in Figure 7, facial expressions were consistently rated as more realistic in the *visual-only* condition across all scenarios. This trend was further supported by the ordinal mixed-effects model (Table 6; Figure 8). While the main effect of experimental condition was not statistically significant ($p = .526$), a significant interaction emerged between the visual-only condition and sad emotional content ($p = .023$, OR = 4.554, 95% CI [1.232, 16.829]). This indicates that without audio, low-energy emotions like sadness seemed more realistic to viewers.

This enhancement likely stems from audiovisual interference. In the audio+visual condition, even slight mismatches in lip-sync or emotional prosody can disrupt the illusion of authenticity. The visual-only condition eliminates such dissonance, allowing facial cues to be evaluated in isolation, which may reduce perceptual scrutiny and heighten tolerance for animation artifacts.

A strong main effect of avatar identity was also observed: *Emory* was rated significantly less realistic than *Amelia* ($p = .001$, OR = 0.211, 95% CI [0.084, 0.528]), regardless of experimental condition or emotion. As previously concluded in Section , Emory's angular features and more rigid expressions may have clashed with participants' implicit expectations of emotional expressiveness, whereas Amelia's softer, more rounded features aligned better with perceived emotional authenticity. In addition, a good match between the voice and the avatar's face likely made the avatar feel more realistic. Amelia's voice was perceived as more consistent with her visual design, while Emory's highly expressive vocal delivery may have introduced moments of mismatch when not fully supported by corresponding facial expressions.

Although some interaction effects (e.g., Visual-only × Avatar: Emory, $p = .055$) did not reach statistical significance, the directional trend suggests meaningful interplay between avatar morphology and presentation condition, which merits further investigation.

In summary, low-arousal emotions like sadness appear more effectively conveyed through visuals alone, as unsynchronized audio can introduce perceptual dissonance that reduces realism. These findings underscore the importance of condition-sensitive avatar design, especially in emotionally nuanced applications such as child-simulation training, where voice-age congruence and expressive fidelity are crucial to achieving believable interaction.

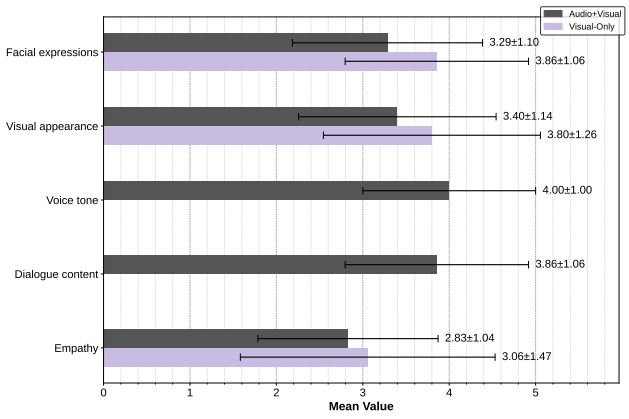


Figure 9: Mean ratings ($M \pm SD$) of perceived realism across different aspects (facial expressions, visual appearance, voice tone, dialogue content) and empathy, compared between audio+visual and visual-only conditions.

Perceived Realism and Empathy

The results summarized in Table 7 and Figure 9 illustrate how multimodal cues (audio and visual) influenced participants' perception of realism and empathy towards emotionally expressive avatars.

In the audio+visual condition, auditory elements played a major role in shaping perceptions. Voice tone emerged as

Table 7: Results of mixed-effects cumulative-logit analyses comparing perceived realism and empathy between audio+visual and visual-only conditions. Significant results ($p < .05$) are highlighted.

	term	p.value	OR	OR_low	OR_high
Realism	Audio+Visual vs. Visual-Only	0.038	0.258	0.072	0.925
	Visual Appearance vs. Facial Expressions	0.765	0.865	0.333	2.246
	Interaction: Audio+Visual × Visual Appearance	0.587	1.434	0.391	5.264
Empathy	Audio+Visual vs. Visual-Only	0.397	0.693	0.295	1.614

the strongest contributor, with an average realism rating of $M = 4.00$, followed by dialogue content ($M = 3.86$). However, in the visual-only condition, visual components, particularly facial expression realism and visual appearance, received higher ratings ($M = 3.86$ and $M = 3.80$, respectively) compared to the audio+visual condition ($M = 3.29$ and $M = 3.40$, respectively). This difference may partly reflect the fact that participants in the visual-only condition evaluated fewer realism contributors, with their attention concentrated on the available visual aspects. In contrast, participants in the audio+visual condition distributed their judgments across four factors, including voice tone and dialogue content, which may have diluted individual ratings for visual features.

A mixed-effects cumulative-logit analysis confirmed these findings. Including audio led to a significant 74% reduction in the likelihood of higher realism ratings (OR = 0.26, 95% CI [0.07, 0.93], $p = .038$). This suggests that audio cues may introduce inconsistencies, such as minor lip-sync issues or mismatches between tone and visual emotion, which reduce perceived authenticity. No significant differences were found between the two realism items ("Facial Expressions" vs. "Visual Appearance," $p = .765$), nor was the interaction between condition and realism item significant ($p = .587$), indicating that audio affected both visual aspects similarly.

Empathy ratings were not significantly influenced by the presence of audio. Although slightly higher empathy was reported in the visual-only condition ($M = 3.06$) than in the audio+visual condition ($M = 2.83$), the difference was not statistically significant (OR = 0.69, 95% CI [0.30, 1.61], $p = .397$). This suggests a possible trend toward enhanced emotional connection with visual-only stimuli, though not conclusive.

Collectively, these findings highlight a critical trade-off in avatar design: while audio cues enhance realism for high-arousal emotions, subtle audiovisual mismatches can negatively impact perceived authenticity. Under the conditions examined, removing audio increased visual credibility without substantially affecting emotional engagement.

Qualitative Feedback

Participants provided reflective qualitative feedback on what stood out the most about the avatars. These reflections revealed distinct and important observations regarding emotional expressiveness, realism, and immersion, providing directions for enhancing avatar technologies in interactive simulations. Two independent coders analyzed the responses. Discrepancies in coding were discussed and resolved collaboratively to ensure consistency and analytical rigor.

Audio and Visual Conditions

In the audio+visual condition, participants' qualitative feedback highlighted both the strengths and limitations of the avatars' emotional expressiveness. On the positive side, many participants recognized and appreciated the avatars' capacity to convey emotions dynamically, noting that *"They were all expressing emotions"* and that they *"expressed their feelings through facials"*. The quality of the synthesized voice was frequently mentioned as a significant contributor to engagement and realism, with comments such as *"The tone of voice was the most convincing"* and *"The dialogue content... did not sound like an avatar"*, suggesting successful prosody design and auditory fidelity. Some participants also felt a sense of connection, stating, *"You can feel them though facial expressions"*. The visual appearance was also praised, with feedback indicating that *"Their visual appearances were very good and looked human like"* and *"The faces were well done"*.

However, a recurring theme in the feedback was the presence of multimodal incongruities. Participants frequently observed a mismatch between the auditory and visual channels, as captured by comments like *"The facial expression does not match what the avatars are saying"*. Issues with lip synchronization were also noted as detrimental to realism, with feedback stating, *"When they speak, their lips do not correspond to the dialogue"* and *"The mouth moves but it doesn't seam connected to the rest of the face"*. These observations align with the quantitative findings regarding reduced perceived facial realism when audio was present (Table 6), suggesting that the inclusion of voice amplified scrutiny of temporal and spatial alignment.

Participants also reported experiencing uncanny valley effects, describing the avatars with phrases such as *"weird mix of looking real but still obviously fake"* and comparing their movement to *"NPC characters in video games"* or *"a rigid model in video games"*. Some felt the facial expressions were not consistently on point or that the faces *"didnt move all that much"*. Negative emotional responses were also reported by some participants, with comments like *"They all looked very scary"* and *"I feel*

uncomfortable and uneasy after this".

Collectively, the feedback from the audio+visual condition reveals a tension between the successful implementation of some aspects, such as voice quality and general emotional expression, and persistent challenges related to audiovisual synchronization and the subtlety and naturalness of facial movements. These points underscore the complexity of creating truly believable and comfortable AI-generated avatars.

Visual-Only Conditions

In the visual-only condition, where avatars were presented without synthesized speech, participant feedback predominantly focused on visual realism and the effectiveness of facial cues in conveying emotion. Positive comments highlighted the avatars' ability to express emotions through isolated facial dynamics, with participants noting that *"Facial expressions stood out the most from the avatars"* and that the avatars showed emotion *"like real human"*. Some feedback specifically praised the clarity of certain expressions, such as joy and sadness, with one participant observing a *"genuine smile, often characterized by crinkles around the eyes"* for happiness and another emphasizing the effectiveness of *"The emotional that was in the eyes of avatars... specially the emotion of sadness and disgust,"* adding *"good job!"*. The visual design was also acknowledged positively, with comments like *"They have human-like features"* and *"Amelia and Emory looked the most realistic to me"*. Some participants reported feeling *"Very connected with the avatar"* and that *"You can feel them though facial expressions"* even in the absence of audio.

However, critiques were also present, centering on perceived exaggeration or inconsistency in the expressions. Participants reported that *"facial expressions were quite exaggerated and did not look completely natural"* and that *"They almost all seem to have the same facial expression if I did not look closely"*. Technical limitations in animation fidelity were also noted, with comments such as *"I think they lack polygons to make them more expresive"* and comparisons to *"a rigid model in video games"*. The absence of audio seemed to shift attention to subtle visual flaws, including repetitive mouth movements (*"the movement of their mouth"*) and limited emotional range or gradation (*"I would see a slight smile here and there but the avatar still kept a straight face"*). Some feedback also simply stated, *"nothing"* stood out, or described the experience as *"interesting but not very realistic"*.

Collectively, the feedback from the visual-only condition suggests that while removing audio can enhance the focus on and perceived realism of facial expressions by eliminating cross-modal interference, it also exposes limitations in the subtlety, naturalness, and fidelity of the facial animations themselves. These points underscore the need for further refinement in visual-only emotional expression to achieve greater believability.

Discussion

This study offers several important lessons regarding the perceptual and design challenges of building emotionally expressive child avatars. Rather than emphasizing the system's technical capabilities, our findings point to deeper tensions in achieving believable emotional portrayals through multimodal integration. One key insight is that perceived realism depends not on modality fidelity in isolation, but on the coherence across modalities. Moreover, participant responses varied considerably, revealing the inherently subjective nature of emotional perception. What one viewer perceives as lifelike or relatable, another may find unsettling or unnatural. This variability underscores the importance of designing emotionally expressive avatars that can accommodate a range of user sensitivities. When auditory and visual channels are not well aligned, user experience suffers. In some cases, visual-only presentations were even rated as more realistic than audio-visual ones. Silencing the audio removes a major source of incongruence, but it can also weaken the communication of certain emotions, particularly those that involve high arousal.

The study highlights a recurring trade-off. Auditory cues are necessary for disambiguating intense emotional states such as anger, yet their inclusion risks drawing attention to imperfections in lip-sync or mismatched vocal tone. Synchronizing prosody with facial animation is not a small technical issue. It plays a key role in making avatars believable and emotionally clear. Participant feedback supports this, with several comments describing distracting or unsettling effects when lip movements did not match the dialogue or when the voice tone conflicted with the avatar's appearance.

A second lesson concerns the role of avatar morphology in shaping emotional interpretation. The "Amelia" avatar, with softer facial features, was generally perceived as more believable in low-arousal contexts such as sadness or joy. In contrast, the more angular "Emory" avatar was more effective in conveying anger. These outcomes suggest that certain facial structures are better suited to subtle, empathic expressions, while others may more effectively convey direct, high-intensity emotions. Designers of virtual agents may benefit from selecting or customizing facial geometry based on the emotional demands of the specific use case.

Additionally, the study raises questions about the optimal configuration of animation parameters. Our use of the "exaggerated" mode in Audio2Face, intended to increase clarity, some participants in the visual-only condition perceived the resulting facial movements as overly intense or unnatural. This suggests a potential trade-off between expressiveness and realism, especially in training contexts where exaggerated expressions may break immersion or reduce credibility. Exploring "automatic" or more

customized settings could help address this issue.

Overall, this work reinforces that perceived believability is not simply a function of system complexity or realism in individual channels, but rather the result of coherent integration across modalities and sensitivity to user perception. By focusing on these lessons, future systems can be better tailored to the emotional and communicative needs of users, especially in sensitive and high-stakes environments.

Limitations and Future Work

While this study offers several valuable insights into the design and perception of emotionally expressive child avatars, it also presents a number of limitations that inform directions for future research. The findings are based on two specific MetaHuman avatars, and user perceptions may differ with avatars of varying visual design, age representation, or artistic style. Moreover, although prior work supported the use of a young adult female voice for child avatars, this choice may have influenced participants' perceptions of emotional authenticity and audiovisual coherence. Integrating high-quality, age-appropriate synthesized voices that align more naturally with the avatars' appearance and emotional tone remains a significant technical and perceptual challenge.

The emotional scope of the study was also limited to three basic emotions: joy, sadness, and anger. As a result, the system's ability to portray a broader range of complex or subtle emotional states (such as fear, surprise, or confusion) remains untested. Additionally, participant recruitment via the Prolific platform, while offering demographic diversity, may not fully capture the perspectives of domain-specific end-users such as Child Protection Services (CPS) professionals. The controlled, pre-rendered nature of the video stimuli further limits ecological validity. Responses observed in dynamic, real-time interactions may differ substantially due to the role of responsiveness and timing in social perception.

These limitations point to several promising avenues for future work. Longitudinal studies are needed to explore how user perceptions evolve with repeated or sustained exposure to emotionally expressive avatars, potentially revealing learning effects and long-term engagement dynamics. Future research should also expand the emotional repertoire of avatars to better reflect the range of expressions relevant in sensitive contexts such as forensic interviews. Critically, upcoming evaluations should involve domain experts such as CPS or law enforcement professionals to assess the system's usability, realism, and training efficacy in practical, high-stakes environments. Such studies will help refine both the technical and perceptual aspects of the system, ensuring its relevance and effectiveness in its intended application context.

Conclusion

This study highlights a central challenge in the development of emotionally expressive child avatars: effective multimodal integration remains difficult to achieve. While our system demonstrates the technical feasibility of real-time emotional animation using Unreal Engine 5 MetaHuman and NVIDIA Audio2Face, the results reveal that audiovisual misalignment can undermine perceived authenticity. Notably, participants often rated visual-only clips as more realistic than audio-visual ones, suggesting that incomplete or poorly matched modalities may detract from overall believability.

This is especially relevant in sensitive applications such as child forensic interview training, where avatars must evoke trust and empathy. Our findings show that high-arousal emotions like anger depend heavily on synchronized vocal and facial cues, while lower-arousal states like sadness and joy can be conveyed more effectively through visual expressions alone. Additionally, factors such as avatar morphology and voice-age congruence significantly shaped emotional perception.

Rather than pursuing realism within individual modalities, this study underscores the importance of holistic and context-sensitive design. Emotional credibility arises only when voice, appearance, and facial behavior are cohesively aligned. These insights offer practical guidance for improving the design of virtual agents in emotionally demanding settings. Ultimately, consistent integration across modalities, rather than fidelity in isolation, is essential for achieving believable human–AI interaction.

References

1. Daněček R, Schmitt C, Polikovsky S, and Black MJ. Supervising 3D Talking Head Avatars with Analysis-by-Audio-Synthesis. arXiv preprint arXiv:2504.13386 2025
2. Pei G, Li H, Lu Y, Wang Y, Hua S, and Li T. Affective computing: Recent advances, challenges, and future trends. Intelligent Computing 2024; 3:0076
3. Salehi P. Emotionally_Expressive_Child_Avatars. https://github.com/pegahsalehi/Emotionally_Expressive_Child_Avatars. GitHub repository, accessed July 7, 2025. 2025
4. Salehi P, Hassan SZ, Baugerud GA, Powell M, Johnson MS, Johansen D, Sabet SS, Riegler MA, and Halvorsen P. A theoretical and empirical analysis of 2D and 3D Virtual Environments in Training for Child Interview Skills. IEEE Access 2024
5. Pan Y, Kim K, Lee J, Sang Y, and Cheon J. Research on the application of digital human production based on photoscan realistic head 3D scanning and unreal engine metahuman technology in the metaverse. International journal of advanced smart convergence 2022; 11:102–18
6. NVIDIA. NVIDIA Audio2Face 3D. https://build.nvidia.com/nvidia/audio2face-3d. Accessed: 2025-04-04. 2025
7. Aranha RV, Corrêa CG, and Nunes FL. Adapting software with affective computing: a systematic review. IEEE Transactions on Affective Computing 2019; 12:883–99

8. Maas A, Daly RE, Pham PT, Huang D, Ng AY, and Potts C. Learning word vectors for sentiment analysis. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011 :142–50
9. Tian YI, Kanade T, and Cohn JF. Recognizing action units for facial expression analysis. IEEE Transactions on pattern analysis and machine intelligence 2001; 23:97–115
10. Pereira ET, Gomes HM, Veloso LR, and Mota MRA. Empirical evidence relating EEG signal duration to emotion classification performance. IEEE Transactions on Affective Computing 2018; 12:154–64
11. McDuff D and Czerwinski M. Designing emotionally sentient agents. Communications of the ACM 2018; 61:74–83
12. Lalwani T, Bhalotia S, Pal A, Rathod V, and Bisen S. Implementation of a Chatbot System using AI and NLP. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Volume-6, Issue-3 2018
13. Nass C and Moon Y. Machines and mindlessness: Social responses to computers. Journal of social issues 2000; 56:81–103
14. Jack RE and Schyns PG. The human face as a dynamic tool for social communication. Current Biology 2015; 25:R621–R634
15. Cabanac M. What is emotion? Behavioural processes 2002; 60:69–83
16. Schachter S and Singer J. Cognitive, social, and physiological determinants of emotional state. Psychological review 1962; 69:379
17. Gratch J and Marsella S. Evaluating the modeling and use of emotion in virtual humans. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.* Vol. 1. IEEE Computer Society. 2004 :320–7
18. Gratch J and Marsella S. Lessons from emotion psychology for the design of lifelike characters. Applied Artificial Intelligence 2005; 19:215–33
19. Ji X, Zhou H, Wang K, Wu W, Loy CC, Cao X, and Xu F. Audio-driven emotional video portraits. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021 :14080–9
20. Papantoniou FP, Filntisis PP, Maragos P, and Roussos A. Neural emotion director: Speech-preserving semantic control of facial expressions in" in-the-wild" videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022 :18781–90
21. Thies J, Elgharib M, Tewari A, Theobalt C, and Nießner M. Neural voice puppetry: Audio-driven facial reenactment. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer. 2020 :716–31
22. Li T, Bolkart T, Black MJ, Li H, and Romero J. Learning a model of facial shape and expression from 4D scans. ACM Trans. Graph. 2017; 36:194–1
23. Daněček R, Chhatre K, Tripathi S, Wen Y, Black M, and Bolkart T. Emotional speech-driven animation with content-emotion disentanglement. *SIGGRAPH Asia 2023 Conference Papers*. 2023 :1–13
24. Peng Z, Wu H, Song Z, Xu H, Zhu X, He J, Liu H, and Fan Z. Emotalk: Speech-driven emotional disentanglement for 3d face animation. *Proceedings of the IEEE/CVF international conference on computer vision*. 2023 :20687–97
25. Epic Games. MetaHuman. https://www.unrealengine.com/en-US/metahuman. Accessed 30 April 2025. 2025
26. Epic Games. Rig Logic White Paper v2. https://cdn2.unrealengine.com/rig-logic-whitepaper-v2-5c9f23f7e210.pdf. Accessed 30 April 2025. 2021
27. Epic Games Documentation Team. Using Subsurface Scattering in Unreal Engine Materials. https://dev.epicgames.com/documentation/en-us/unreal-engine/using-subsurface-scattering-in-unreal-engine-materials. Accessed 30 April 2025. 2025
28. Cloud G. Google Cloud Speech-to-Text. https://cloud.google.com/speech-to-text. Accessed 5 May 2025. 2023
29. OpenAI. ChatGPT. https://chat.openai.com. Accessed 5 May 2025. 2024
30. Services AW. Amazon Polly Neural Voices. https://docs.aws.amazon.com/polly/latest/dg/neural-voices.html. Accessed 5 May 2025. 2025
31. Daly S. NVIDIA Audio2Face: Revolutionizing Facial Animation in Real-Time. https://9meters.com/technology/ai/nvidia-audio2face-revolutionizing-facial-animation-in-real-time. Accessed 5 May 2025. 2024
32. Corporation N. Audio2Face Live Link UE Plugin — User Manual. https://docs.omniverse.nvidia.com/audio2face/latest/user-manual/livelink-ue-plugin.html. Accessed 5 May 2025. 2025
33. Wu J, Dang T, Sethu V, and Ambikairajah E. Multimodal affect models: An investigation of relative salience of audio and visual cues for emotion prediction. Frontiers in Computer Science 2021; 3:767767
34. Salehi P, Hassan SZ, Shafiee Sabet S, Astrid Baugerud G, Sinkerud Johnson M, Halvorsen P, and Riegler MA. Is more realistic better? a comparison of game engine and gan-based avatars for investigative interviews of children. *Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*. 2022 :41–9
35. Onuki Y, Honda H, and Ueda K. What stimuli are necessary for anchoring effects to occur? Frontiers in Psychology 2021; 12:602372
36. Ekman P and Friesen WV. Facial action coding system. Environmental Psychology & Nonverbal Behavior 1978
37. Gao C, Weber CE, Wedell DH, and Shinkareva SV. An fMRI study of affective congruence across visual and auditory modalities. Journal of cognitive neuroscience 2020; 32:1251–62
38. Marian V, Hayakawa S, and Schroeder SR. Cross-modal interaction between auditory and visual input impacts memory retrieval. Frontiers in Neuroscience 2021; 15:661477

39. Demeure V, Niewiadomski R, and Pelachaud C. How is believability of a virtual agent related to warmth, competence, personification, and embodiment? Presence 2011; 20:431–48

40. Seymour M. Omniverse & Audio2Face. https : / / www.fxguide.com/fxfeatured/omniverse-audio2face/. Accessed 30 April 2025. 2021

41. Goudbeek M and Scherer K. Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. The Journal of the Acoustical Society of America 2010; 128:1322–36

42. Field A, Field Z, and Miles J. Discovering statistics using R. Sage, 2012

43. Hussain T, Wang W, Bouaynaya N, Fathallah-Shaykh H, and Mihaylova L. Deep learning for audio visual emotion recognition. *2022 25th International Conference on Information Fusion (FUSION)*. IEEE. 2022 :1–8

44. Brooks JA, Kim L, Opara M, Keltner D, Fang X, Monroy M, Corona R, Tzirakis P, Baird A, Metrick J, et al. Deep learning reveals what facial expressions mean to people in different cultures. Iscience 2024; 27

Appendix: Supplementary Material

This appendix includes supplementary statistical tables (Table 8 and Table 9) that report model comparisons and detailed regression results supporting the analyses presented in Section .

Between-condition evaluations in Table 8 reveal a robust modality effect: when the audio track was muted, participants rated the avatars' facial expressions as more realistic overall, with significantly higher ratings in the sad scenarios for both characters (Emory-Sad: $M_{\text{no-audio}} = 3.43$ vs $M_{\text{audio}} = 2.80$, $d = 0.68$, FDR-corrected $p = .021$; Amelia-Sad: $M_{\text{no-audio}} = 4.00$ vs $M_{\text{audio}} = 3.34$, $d = 0.67$, FDR-corrected $p = .021$). This baseline shift suggests that any mismatch between vocal prosody and facial timing, or the use of an ill-suited voice, draws attention and diminishes perceived realism.

Conversely, in scenarios depicting high-arousal emotions (Emory-Angry, Amelia-Angry, and Emory-Joy), the presence of audio improved emotion recognition and overall expressiveness. This pattern indicates that synchronized vocal cues can enhance the clarity of emotional intent, particularly when facial expressions alone provide insufficient affective information.

Table 9 provides a direct comparison between the two avatars. Across both conditions, Amelia was consistently rated as more realistic, with higher accuracy in emotion identification for sadness and joy. Notably, in the audio-visual condition, Emory was rated more believable in the angry scenario, suggesting that his angular features may better support high-arousal affect when complemented by vocal intensity. In contrast, Amelia's rounded, soft-featured face reliably conveyed emotional states across modalities, enabling consistent perception even without audio.

Taken together, these two tables reinforce our central claim: perceived believability depends critically on the interplay between audiovisual congruence and facial geometry.

Table 8: Detailed results of independent-sample t-tests (including Welch's corrections where needed) comparing visual-only and audio+visual conditions across emotional expressions and avatars. Equality of variances was assessed using Levene's test. Cohen's d reflects effect size magnitude. FDR-corrected p-values account for multiple comparisons. Statistically significant results (FDR-adjusted $p < .05$) are highlighted in gray.

	Variable	Mean visual-only	Mean audio+visual	p-value	Cohen's d	FDR-corrected p-value
Emory_Sad	Facial exp.	3.429	2.800	0.006	0.680	0.021
	Angry	1.829	2.571	0.019	-0.575	0.053
	Sad	3.343	3.629	0.360	-0.221	0.520
	Joy	1.429	1.171	0.119	0.378	0.222
	Fear	2.229	1.857	0.192	0.315	0.312
	Disgust	1.971	1.971	1.000	0.000	1.000
Emory_Angry	Facial exp.	3.514	2.914	0.028	0.535	0.074
	Angry	1.429	3.829	0.000	-2.565	0.000
	Sad	2.000	3.257	0.000	-0.985	0.001
	Joy	2.143	1.086	0.000	1.329	0.000
	Fear	1.714	1.914	0.448	-0.182	0.540
	Disgust	1.486	2.743	0.000	-1.029	0.001
Emory_Joy	Facial exp.	3.514	3.171	0.226	0.292	0.339
	Angry	1.400	1.057	0.008	0.661	0.027
	Sad	1.743	1.114	0.005	0.717	0.021
	Joy	2.914	3.771	0.005	-0.690	0.021
	Fear	1.571	1.229	0.082	0.423	0.177
	Disgust	1.743	1.171	0.009	0.656	0.027
Amelia_Sad	Facial exp.	4.000	3.343	0.006	0.678	0.021
	Angry	2.314	2.571	0.425	-0.192	0.540
	Sad	4.086	4.114	0.901	-0.030	0.925
	Joy	1.200	1.086	0.224	0.293	0.339
	Fear	2.314	2.086	0.455	0.179	0.540
	Disgust	1.800	2.429	0.036	-0.513	0.081
Amelia_Angry	Facial exp.	3.629	3.543	0.726	0.084	0.787
	Angry	2.171	3.114	0.004	-0.714	0.021
	Sad	2.371	3.714	0.000	-0.997	0.001
	Joy	1.829	1.143	0.005	0.700	0.021
	Fear	1.857	2.000	0.589	-0.130	0.656
	Disgust	1.943	2.114	0.577	-0.134	0.656
Amelia_Joy	Facial exp.	3.914	3.486	0.098	0.401	0.200
	Angry	1.086	1.000	0.179	0.325	0.304
	Sad	1.143	1.057	0.400	0.203	0.540
	Joy	4.371	4.200	0.405	0.200	0.540
	Fear	1.143	1.114	0.754	0.075	0.795
	Disgust	1.229	1.057	0.119	0.377	0.222

Table 9: Statistical comparison using paired-sample t-tests of emotional expression ratings between the Amelia and Emory avatars using paired-sample t-tests, under both Audio+Visual and Visual-Only conditions. Ratings include facial expressiveness and perceived intensity across five target emotions (Angry, Sad, Joy, Fear, Disgust). Cohen's d reflects effect size. FDR-corrected p-values account for multiple comparisons. Statistically significant differences (FDR-adjusted $p < .05$) are highlighted in gray. For each significant result, the avatar with higher mean ratings is highlighted in green to indicate user preference.

		Audio+Visual					Visual-Only				
	Variable	p-value	Mean Amelia	Mean Emory	Cohen's d	FDR-corrected p-value	p-value	Mean Amelia	Mean Emory	Cohen's d	FDR-corrected p-value
Sad	Facial exp.	0.001	3.343	2.800	0.591	0.010	0.001	4.000	3.429	0.646	0.005
	Angry	1.000	2.571	2.571	0.000	1.000	0.088	2.314	1.829	0.297	0.143
	Sad	0.020	4.114	3.629	0.414	0.045	0.018	4.086	3.343	0.419	0.041
	Joy	0.263	1.086	1.171	-0.192	0.338	0.118	1.200	1.429	-0.271	0.163
	Fear	0.244	2.086	1.857	0.201	0.338	0.619	2.314	2.229	0.085	0.619
	Disgust	0.014	2.429	1.971	0.440	0.041	0.280	1.800	1.971	-0.186	0.336
Angry	Facial exp.	0.002	3.543	2.914	0.578	0.010	0.579	3.629	3.514	0.095	0.613
	Angry	0.002	3.114	3.829	-0.582	0.010	0.006	2.171	1.429	0.495	0.016
	Sad	0.009	3.714	3.257	0.466	0.041	0.113	2.371	2.000	0.275	0.163
	Joy	0.571	1.143	1.086	0.097	0.643	0.162	1.829	2.143	-0.242	0.208
	Fear	0.609	2.000	1.914	0.087	0.644	0.483	1.857	1.714	0.120	0.543
	Disgust	0.012	2.114	2.743	-0.451	0.041	0.004	1.943	1.486	0.516	0.016
Joy	Facial exp.	0.054	3.486	3.171	0.337	0.108	0.041	3.914	3.514	0.358	0.075
	Angry	0.160	1.000	1.057	-0.243	0.262	0.006	1.086	1.400	-0.498	0.016
	Sad	0.324	1.057	1.114	-0.169	0.389	0.002	1.143	1.743	-0.565	0.012
	Joy	0.020	4.200	3.771	0.413	0.045	0.000	4.371	2.914	1.025	0.000
	Fear	0.254	1.114	1.229	-0.196	0.338	0.026	1.143	1.571	-0.392	0.053
	Disgust	0.103	1.057	1.171	-0.283	0.186	0.006	1.229	1.743	-0.495	0.016

