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]Oslo Metropolitan University, 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 subjective 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-Appropriate Response*. Furthermore, the dialog system's effectiveness, particularly *Response Relevance* and *Detailed Responses*, played a significant role in *Empathy Elicitation*, often outweighing *Virtual 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 Immersion, Large Language Model (LLM), Quality of Experience (QoE), Usability, Virtual Environments (VEs), Virtual Reality (VR)

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 pivotal role of proper interviewing techniques and the availability of extensive training programs, compliance with best practices remains low [6]. Interviewers frequently use 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 developed an AI-based platform that presents interview scenarios using high-fidelity avatars to mimic child behavior [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. Computer vision technologies enable it to present visually similar 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 highly important 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 *Virtual 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 extends our previously published work [15] in which the findings revealed significant differences in user experience between 2D and 3D environments. The 3D environment provided greater sense of presence and visual fidelity, while the 2D environment was favored for usability. We have expanded this research with additional experiments, deeper analysis, extended discussions, and new findings. Specifically, we 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 structural equation modeling (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 three primary areas: investigative interview training, the application of VR in education and visual fidelity in virtual environments. 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, despite their 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 dynamic 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 facilitates 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 following 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 utilize 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 enhances 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 major 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.

C. VISUAL FIDELITY IN VIRTUAL ENVIRONMENTS

Visual Fidelity in virtual environments contributes significantly to user engagement and interaction quality. Research has shown that avatars with realistic facial expressions significantly improve the conveyance of emotions, thereby providing a more immersive and authentic user experience [43]. Studies comparing different technologies for conveying emotions through avatars indicate that dynamic facial expressions are crucial for effective non-verbal communication, trust-building, and user satisfaction in virtual interactions [44]–[46]. Additionally, the psychological impact of realistic avatar representations reinforces the importance of incorporating detailed facial and eye movements to achieve higher levels of social presence and emotional resonance with users [47]. Thus, the inclusion of realistic facial expressions in virtual environments is essential for creating compelling and emotionally engaging experiences.

III. SYSTEM ARCHITECTURE

Figure 1 illustrates the design of our interactive child avatar system, which is compatible with both 2D and 3D

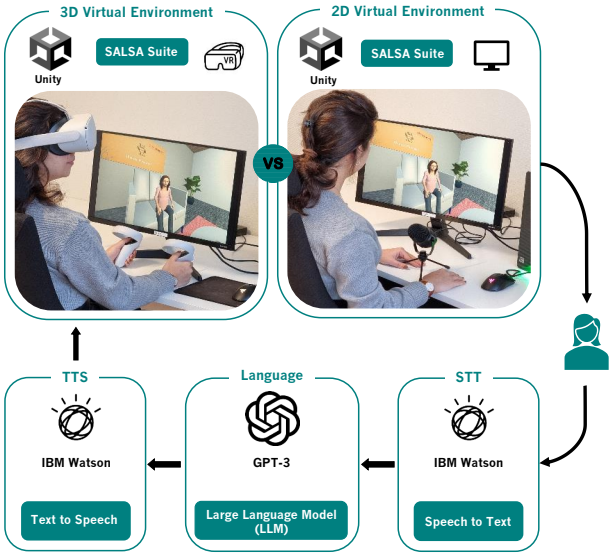


FIGURE 1. The high-level architecture of the child avatar system used in this comparative study also showing the two alternative interaction environments (2D vs. 3D).

interactive environments. The system is divided into three key components:

1) A language module, which utilizes OpenAI's Large Language Model (LLM), GPT-3 [26]. It is fine tuned to converse like a child with various personas.
2) A speech synthesis module, which employs IBM Watson services for efficient STT and TTS functions.
3) A Unity-based user interface that offers two interactive modes: a 3D virtual environment accessible via an Oculus Quest2 HMD and a 2D virtual environment using a regular screen.

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 the GPT-3 [26] Davinci model for two specific case 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 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:

1) **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 selection of questions.
2) **Hilary Scenario**: In this scenario, a 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 important 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.

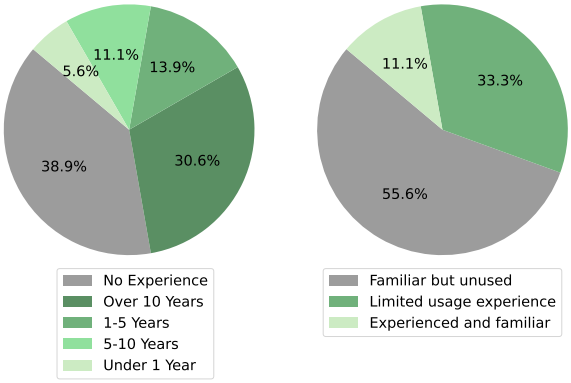
C. SPEECH SYNTHESIS

IBM Watson's STT and TTS services[1] 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 [48].

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[2], allowing character meshes and

[1]https://www.ibm.com/cloud/watson-text-to-speech
[2]https://github.com/umasteeringgroup/UMA



a) Professional years b) Familiarity with VR

FIGURE 2. Distribution of experience levels among participants.

textures to be customized. For realistic avatar movements, the Salsa Suit asset[3] 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.

E. FEEDBACK

Feedback is an important feature in training systems to improve learning efficiency. In our system, a feedback module is integrated where we analyze and classify the person's questions using AI techniques, as discussed in our previous work [49] This module aims to provide users with insights into their performance, highlighting strengths and areas for improvement. The feedback is generated based on various metrics, including question relevance, response accuracy, and the overall interaction quality. This automatic feedback function, which classified and provided feedback on the types of questions asked, was found to be highly reliable ($Cohen's\ kappa\ =\ 0.80$). While the feedback module was not evaluated within the scope of this study, preliminary results from our previous research suggest that users find the feedback valuable in understanding their performance and making necessary adjustments.

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. Afterward, 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 gender ratio mimics the gender balance we see among professionals who work as trainers and in child protection. 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.

B. 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 prompts, 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

[3]https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442

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. For ease of presentation, the quotes have been corrected for grammatical errors.

To ensure accuracy in data collection, the interviews were recorded and subsequently transcribed using the Otter transcription software, adhering to the guidelines established in the informed consent document. To analyze the transcripts, we first categorized the data based on the 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 5). 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 prompts in post-test interview.

	Items
Q1	what was your overall experience with this avatar?
Q2	Reflect solely on your ability to engage with the avatar from a visual/interactive perspective.
Q3	Reflect on the appropriateness of her responses.
Q4	Reflect on the usefulness of this activity from a training perspective.
Q5	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 not we asked you today that you think would be valuable for us to know?

C. 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, allowing 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*[4], *Usability* [50], *Virtual Fidelity* [5], [51], *Emotion* [51], *Responsiveness* [53], *Response Relevance* [54], 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 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 *Virtual 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 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 [55], 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 investigated 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,

[4]https://www.igroup.org/pq/ipq/index.php

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

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 (coef) and p-values (p) for each assessed variable, are systematically presented in Table 4, cells with blue highlights indicate statistically significant results

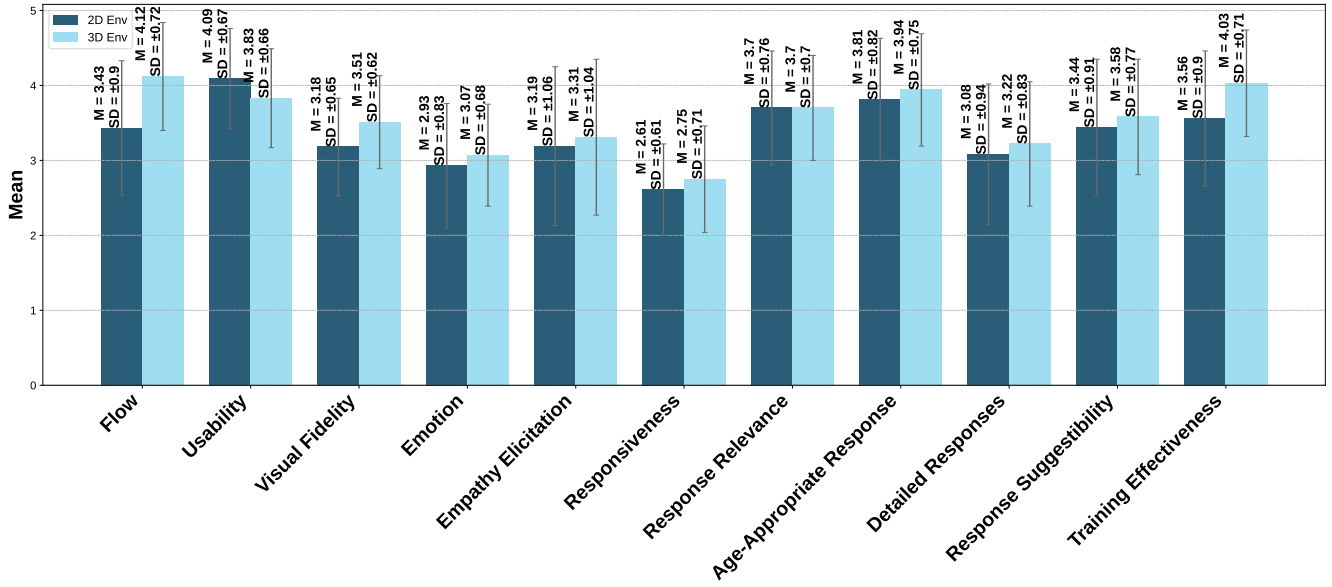


FIGURE 3. Bar-plot (95% confidence interval) of mean scores across seven evaluative factors, along with four independent questions of *Empathy Elicitation*, *Age-Appropriate Response*, *Detailed Responses*, and *Response Suggestibility*, in 2D and 3D environments.

(p-value ≤ 0.05), meaning that the probability that the observed results are based on experimental conditions rather than random variation. 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 important 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:

Virtual 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 *Virtual 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.

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 fundamental 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

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

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 important 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 essential, especially where accurate and detailed perceptions of movement are important for user immersion and realism [56]. 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* ($p = 0.037$). 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 [57], where non-experts could reliably evaluate general aspects of interaction, but domain experts identified subtle and key 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 important 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 feel 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. In Figure 4 and Table 5, 'Env' refers to Environment and 'Exp' refers to Expertise. The other variables are based on the tags listed in Table 2.

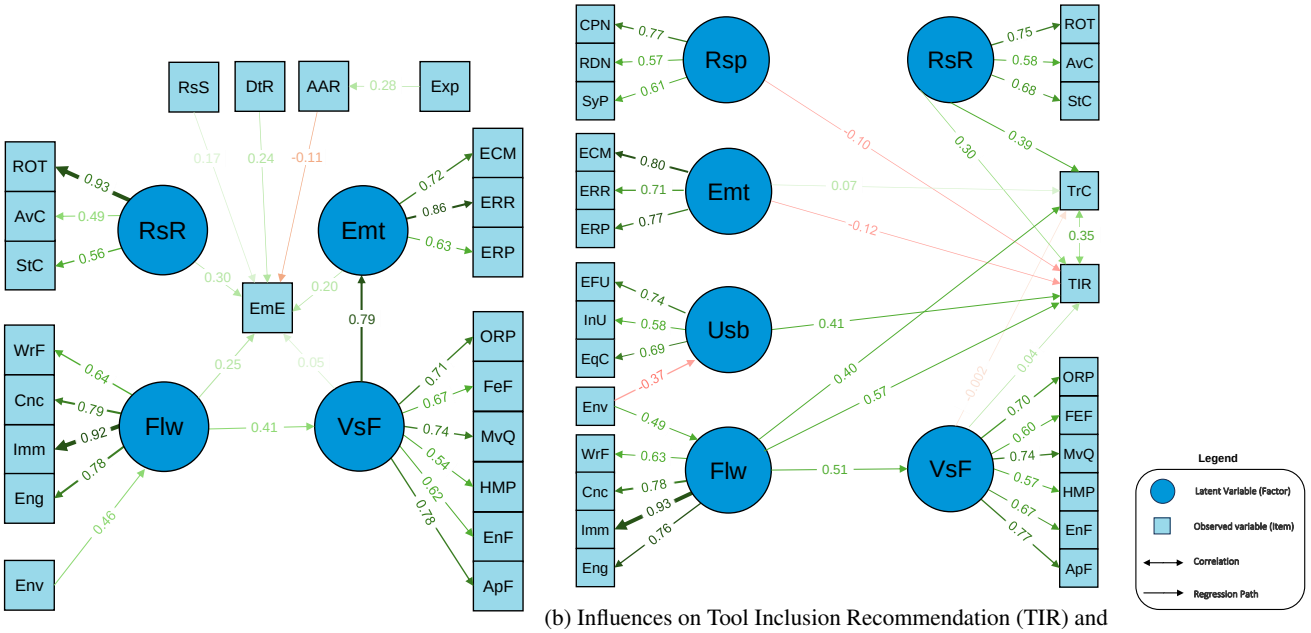
1) Impact on Empathy Elicitation

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 *Virtual 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 consistent and on topic but 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



(a) Influences on Empathy Elicitation (EmE)

(b) Influences on Tool Inclusion Recommendation (TIR) and Training Comparability (TrC)

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.

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-value ≤ 0.05 are indicated. For a graphical representation of these relationships, refer to Figure 4.

SEM	Predictor	Outcome	Coefficient	Std. Error	t-value	p-value	Standardized Coeff. (β)
Influence on EmE	Env	Flow	1.047	0.361	2.898	0.004	0.464
	Flow	Visual Fidelity	0.399	0.181	2.210	0.027	0.411
	Visual Fidelity	Emotion	1.187	0.450	2.641	0.008	0.793
	Exp	Age-Appropriate Response	0.268	0.094	2.846	0.004	0.284
	Flow	Empathy Elicitation	0.228	0.101	2.252	0.024	0.252
	Visual Fidelity	Empathy Elicitation	0.046	0.313	0.146	0.884	0.049
	Response Relevance	Empathy Elicitation	0.308	0.123	2.500	0.012	0.302
	Response Suggestibility	Empathy Elicitation	0.214	0.127	1.686	0.092	0.175
	Age-Appropriate Response	Empathy Elicitation	-0.147	0.136	-1.084	0.278	-0.113
	Emotion	Empathy Elicitation	0.127	0.201	0.633	0.527	0.205
	Detailed Responses	Empathy Elicitation	0.274	0.112	2.444	0.015	0.235
Influence on TIR and TrC	Env	Flow	1.123	0.341	3.289	0.001	0.490
	Env	Usability	-0.806	0.333	-2.418	0.016	-0.374
	Flow	Visual Fidelity	0.512	0.183	2.795	0.005	0.506
	Flow	Tool Inclusion Recommendation	0.359	0.099	3.613	0.0003	0.573
	Usability	Tool Inclusion Recommendation	0.271	0.071	3.807	0.0001	0.406
	Visual Fidelity	Tool Inclusion Recommendation	0.023	0.083	0.278	0.781	0.037
	Emotion	Tool Inclusion Recommendation	-0.087	0.098	-0.888	0.375	-0.121
	Responsiveness	Tool Inclusion Recommendation	-0.070	0.102	-0.685	0.493	-0.097
	Response Relevance	Tool Inclusion Recommendation	0.215	0.156	1.377	0.169	0.299
	Flow	Training Comparability	0.380	0.130	2.913	0.004	0.400
	Visual Fidelity	Training Comparability	-0.002	0.176	-0.013	0.990	-0.002
	Emotion	Training Comparability	0.081	0.220	0.369	0.712	0.075
	Response Relevance	Training Comparability	0.429	0.141	3.053	0.002	0.394

relatable and emotionally resonant, prompting a stronger empathetic response from the user.

The result also identified several areas for improvement. Despite our expectation, the research revealed that emotional expressions and *Virtual Fidelity* were not significant factors in enhancing empathy, often due to their lack of realism and depth. Additionally, *Age-Appropriate Response*, while designed to simulate child-like responses, tended to have a negative impact.

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* ($\beta = 0.573$, $p = 0.0003$) and *Usability* ($\beta = 0.406$, $p = 0.0001$). As the significance of *Flow* and *Usability* were anticipated based on mixed-effect results, this outcome underscores the importance of both factors in the adoption of this tool for training applications. The 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.

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* ($\beta = -0.121$) 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 ($\beta = -0.097$). 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 [58].

The coefficient for *Response Relevance* ($\beta = 0.215$, $p = 0.169$) shows trends but does not reach statistical significance. This underscores the importance of accurate, contextually relevant, and on-topic responses in enhancing user trust and reliance on the tool.

However, the coefficient for *Virtual Fidelity* ($\beta = 0.037$) indicates that it has no significant effect. This suggests that this factor may not have met users' expectations. Improving the realism of avatars' appearance and movement could enhance their effectiveness.

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 *Flow* and *Response Relevance* in this aspect. The significant impact of *Flow* ($\beta = 0.400$, $p = 0.004$) shows that the engagement and immersion in the virtual environment are vital for making the interaction feel comparable to real-life training.

Furthermore, *Response Relevance* ($\beta = 0.394$, $p = 0.002$) is a strong predictor of training effectiveness. Accurate and relevant avatar responses make the training comparable to human-facilitated sessions, providing meaningful feedback for improving interviewing skills. This highlights its essential role in creating realistic and effective training scenarios that can rival interactions with human trainers.

However, the insignificant results concerning *Virtual Fidelity* and *Emotion* may reflect that the current system falls short of simulating the human-like nuances necessary for truly immersive and empathetic interactions. This consistency with the findings suggests that both the visual details and emotional expression of the avatars need significant enhancement to meet user expectations.

VI. INTERVIEW ANALYSIS

Quantitative data alone cannot capture detailed in-depth insights into the user experience. This section enriches the quantitative findings by providing user perspectives and contextual details through additional detailed interview, thereby strengthening the overall conclusions of the study. We have therefore examined user feedback from the interviews to identify recurring themes and patterns in participant responses. The users' responses to the prompts (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.

A. VISUAL FIDELITY

In virtual environments, realism plays a significant role in influencing the quality of the user experience and interactions. 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

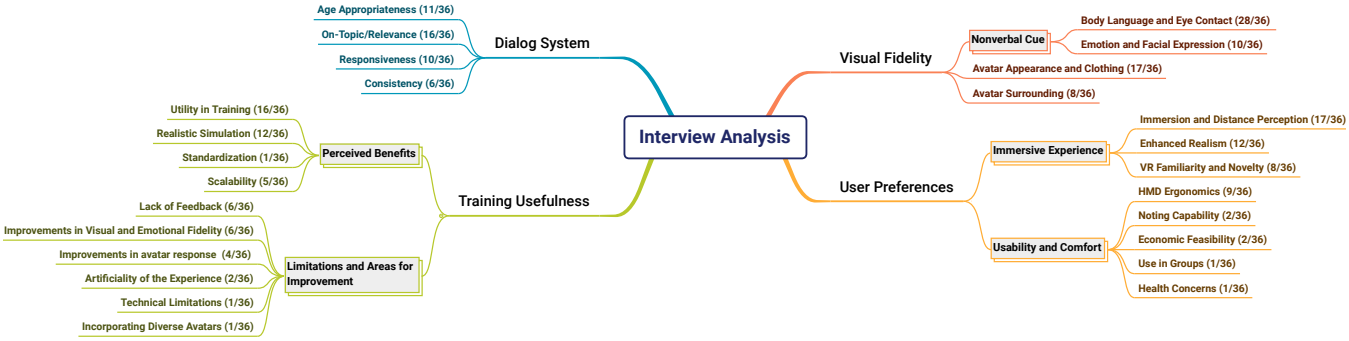


FIGURE 5. Overview of Interview Analysis: Breakdown of key themes with participant counts (x/36), where 'x' represents the number of participants who emphasized each theme out of a total of 36.

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 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 while 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 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 her, she is looking over that side. That's normal for kids."*. This observation suggests that even imperfect body language can convey significant information in this case, 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 so important 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—she has 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 relevant 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 somewhat*

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. [59] 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, an 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 inappropriate. 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, and I wasn't sure whether I spoke too softly. Or sometimes no response came for quite a while. So not sure whether it's what the child would know or not."* Another participant said: *"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 an essential step 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. 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 participant observed that the child avatar initially denied having siblings but later mentioned living with a little brother. Participants associated this inconsistency with the avatar's inability 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? She said, No. I then asked, Who do you live with? She replied, Our mum and his little brother."*

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, addressing them is necessary 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

Participants overwhelmingly preferred the 3D virtual environments for their ability to create a more engaging and realistic training experience. This heightened sense of presence allows users to interact with virtual avatars in a way that closely mimics real-life interactions, thereby enhancing the overall effectiveness of the training sessions.

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. One individual 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, despite 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 [60]. 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 important 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, taking notes was impossible."* Participants suggested that the ability to easily take notes as they would in traditional learning environments remains an essential 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 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. 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 practice 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 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, 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 training package that can be delivered."*. Avatar 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 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 [61]. 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 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 necessary 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 for the variety."* 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 [62]. 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

This study aimed to evaluate the effectiveness of an AI-driven training platform that integrates 2D and 3D virtual environments. The following sections provide an in-depth discussion of findings, limitations, and future enhancements.

A. RESULTS AND FINDINGS

We incorporated qualitative data from semi-structured interviews alongside our quantitative analyses. The insights from these interviews 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. Despite many users remarking on the impressive sense of immersion, the discomfort and complexity of the VR setup highlighted the necessity of user comfort and simple interfaces in the adoption of new training technologies.

Additionally, participants with previous VR experience reported higher *Avatar Interaction Comfort* levels, suggesting that familiarity with VR technology can alleviate some of the initial usability challenges and discomfort linked to the use of 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 relevant responses of the avatars played a key role in *Empathy Elicitation*, in the context of our study outweighing the importance of *Virtual Fidelity*. This finding underscores the importance of continuous improvements in the AI-driven dialogue system to ensure realistic interactions. While the non-significance of *Emotion* and *Virtual Fidelity* in *Empathy Elicitation* can be considered as a reason for its lack of realism according to user expectations, this issue was identified in various analyses within this study.

B. LIMITATIONS

Qualitative and quantitative results highlighted significant areas for improvement in the avatars' emotional reactions and *Virtual 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 potentially impact the training. Even if many children do not express emotions during investigative interviews, emotional expressiveness has been linked to increased informativeness when interviewers show emotional support and follow best practice guidelines [63]. Quantitative analysis supported these observations, indicating that emotion in facial expressions and *Virtual 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.

However, some user have indicated a preference for human trainers over avatars at present. We believe that once the identified issues with the avatar are addressed, a more thorough comparison with human trainers would be justified.

C. FUTURE DIRECTIONS

Recognizing the importance of realism and the lack of this vital element in our research, future iterations of the system will prioritize the development of more lifelike avatars and environments. These features are very important when it comes to training scenarios, because the more realistic they appear, the better we will be able to relate and empathise with them and subsequently have a much greater effect of our training.

Incorporating Metahuman in Unreal Engine technology with NVIDIA Omniverse Audio2Face[5] will improve realistic avatars with emotional reactions to closely mimic real human interactions. These advancements are expected to make the virtual interactions more authentic and emotionally resonant, fostering deeper empathetic connections between users and avatars. This focus on realism will ultimately provide a more effective and immersive training tool for professionals in child welfare and law enforcement, ensuring that they are better prepared for real-world scenarios.

Furthermore, enhancing realism can also make interactions in the 2D environment sufficiently engaging without the need for a VR headset, accommodating users who find use of VR headsets uncomfortable. Additionally, we aim to improve the system by automating response processing through advanced speech recognition technology, eliminating the need for users to send their speech manually by pressing a certain keyword. These improvements will reduce delays and provide a more seamless interaction experience, addressing user concerns about processing time.

Notably, the system developed and tested in this study serves as a proof of concept for enhancing training in child welfare and law enforcement. Initial testing has included a variety of groups such as police officers and CPS personnel in Norway and Australia. The system has also been showcased at conferences in the United States. As part of our ongoing efforts to advance this work, we are currently in the process of transitioning the system into a robust professional training tool. This endeavor is underway in Norway, with the aim of incorporating this technology into the training curricula for child welfare and law enforcement professionals.

[5]https://www.nvidia.com/en-us/omniverse/apps/audio2face/

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 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 *Virtual 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 key role in *Empathy Elicitation*, in this case outweighing *Virtual 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-Appropriate Response*. Future work should focus on improving realism, including different aspects such as facial expressions and nonverbal cues, and the responsiveness of the system to better leverage AI-driven training tools for professional usage.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support and assistance from the Queensland Police Service in undertaking this research. The views expressed in this publication are not necessarily those of the Queensland Police Service and any errors of omission or commission are the responsibility of the authors. We also would like to thank M. Cayetana López Cano for her great assistance in collecting data in Australia. We acknowledge the use of AI-assisted technologies to improve the language and readability of this manuscript.

REFERENCES

[1] M. Stoltenborgh, M. J. Bakermans-Kranenburg, M. H. van Ijzendoorn, and L. R. A. Alink, "Cultural–geographical differences in the occurrence of child physical abuse? A meta-analysis of global prevalence," International Journal of Psychology, vol. 48, no. 2, pp. 81–94, 2013, doi: 10.1080/00207594.2012.697165.

[2] W. A. Walsh, L. M. Jones, T. P. Cross, and T. Lippert, "Prosecuting child sexual abuse: The importance of evidence type," Crime & Delinquency, vol. 56, no. 3, pp. 436–454, 2010, doi: 10.1177/0011128708320484.

[3] M. D. Everson and J. M. Sandoval, "Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements," Child Abuse & Neglect, vol. 35, no. 4, pp. 287–298, 2011, doi: 10.1016/j.chiabu.2011.01.001.

[4] C. J. Brainerd and V. F. Reyna, "Reliability of children's testimony in the era of developmental reversals," Developmental Review, vol. 32, no. 3, pp. 224–267, 2012, doi: 10.1016/j.dr.2012.06.008.

[5] S. P. Brubacher, M. S. Benson, M. B. Powell, J. Goodman-Delahunty, and N. J. Westera, "An overview of best practice investigative interviewing of child witnesses of sexual assault," Child Sexual Abuse, pp. 445–466, 2020, doi: 10.1016/b978-0-12-819434-8.00022-2.

[6] M. Johnson, S. Magnussen, C. Thoresen, K. Lønnum, L. V. Burrell, and A. 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, vol. 29, no. 5, pp. 661–668, 2015, doi: 10.1002/acp.3147.

[7] G. Baugerud, M. S. Johnson, H. B. G. Hansen, S. Magnussen, and M. E. Lamb, "Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015–2017)," Applied Cognitive Psychology, vol. 34, no. 3, pp. 654–663, 2020, doi: 10.1002/acp.3647.

[8] G. A. Baugerud, R. K. Røed, H. B. G. Hansen, J. S. Poulsen, and M. S. Johnson, "Evaluating Child Interviews Conducted by Child Protective Services Workers and Police Investigators," The British Journal of Social Work, vol. 53, no. 5, pp. 2784–2803, 2023, doi: 10.1093/bjsw/bcac245.

[9] A.-C. Cederborg, Y. Orbach, K. J. Sternberg, and M. E. Lamb, "Investigative interviews of child witnesses in Sweden," Child Abuse & Neglect, vol. 24, no. 10, pp. 1355–1361, 2000, doi: 10.1016/s0145-2134(00)00183-6.

[10] M. Yi, E. Jo, and M. E. Lamb, "Effects of the NICHD Protocol Training on Child Investigative Interview Quality in Korean Police Officers," Journal of police and criminal psychology, vol. 31, no. 2, pp. 155–163, 2016, doi: 10.1007/s11896-015-9170-9.

[11] K. Faller, "Forty Years of Forensic Interviewing of Children Suspected of Sexual Abuse, 1974–2014: Historical Benchmarks," Social Sciences, vol. 4, no. 1, pp. 34–65, 2014, doi: 10.3390/socsci4010034.

[12] M. B. Powell, R. P. Fisher, and C. H. Hughes-Scholes, "The effect of intra-versus post-interview feedback during simulated practice interviews about child abuse," Child Abuse & Neglect, vol. 32, no. 2, pp. 213–227, 2008, doi: 10.1016/j.chiabu.2007.08.002.

[13] P. Salehi et al., "Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children," Big Data and Cognitive Computing, vol. 6, no. 2, p. 62, 2022, doi: 10.3390/bdcc6020062.

[14] S. Z. Hassan et al., "Enhancing investigative interview training using a child avatar system: a comparative study of interactive environments," Scientific Reports, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-47368-2.

[15] P. Salehi et al., "Immersive Virtual Reality in Child Interview Skills Training: A Comparison of 2D and 3D Environments," Proceedings of the International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE), pp. 1–7, 2024, doi: 10.1145/3652212.3652219.

[16] K. C. Dalli, "Technological Acceptance of an Avatar Based Interview Training Application: The Development and technological acceptance study of the AvBIT application.," M.S. thesis, Dept. of CS and Media Tech. (CM), Linnaeus University, Småland, Sweden, 2021. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1596930/FULLTEXT01.pdf.

[17] B. Guadagno and M. Powell, "E-Simulations for the Purpose of Training Forensic (Investigative) Interviewers," In D. Holt, S. Segrave, & J. L. Cybulski (Eds.), Professional education using e-simulations: Benefits of blended learning design, pp. 71–86, IGI Global, doi: 10.4018/978-1-61350-189-4.ch005.

[18] R. K. Røed, M. B. Powell, M. A. Riegler, and G. A. Baugerud, "A field assessment of child abuse investigators' engagement with a child-avatar to develop interviewing skills," Child Abuse & Neglect, vol. 143, p. 106324, 2023, doi: 10.1016/j.chiabu.2023.106324.

[19] M. S. Benson and M. B. Powell, "Evaluation of a comprehensive interactive training system for investigative interviewers of children," Psychology, Public Policy, and Law, vol. 21, no. 3, pp. 309–322, 2015, doi: 10.1037/law0000052.

[20] M. B. Powell, B. Guadagno, and M. Benson, "Improving child investigative interviewer performance through computer-based learning activities," Policing and Society, vol. 26, no. 4, pp. 365–374, 2016, doi: 10.1080/10439463.2014.942850.

[21] F. Pompedda, "Training in Investigative Interviews of Children: Serious Gaming Paired with Feedback Improves Interview Quality," Ph.D. thesis, Åbo Akademi University, Åbo, Finland, 2018. [Online]. Available: https://www.doria.fi/handle/10024/152565.

[22] N. Krause, F. Pompedda, J. Antfolk, A. Zappalá, and P. Santtila, "The Effects of Feedback and Reflection on the Questioning Style of Untrained Interviewers in Simulated Child Sexual Abuse Interviews," Applied Cognitive Psychology, vol. 31, no. 2, pp. 187–198, 2017, doi: 10.1002/acp.3316.

[23] F. Pompedda, A. Zappalà, and P. Santtila, "Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality," Psychology, Crime & Law, vol. 21, no. 1, pp. 28–52, 2015, doi: 10.1080/1068316x.2014.915323.

[24] S. Haginoya, S. Yamamoto, F. Pompedda, M. Naka, J. Antfolk, and P. Santtila, "Online Simulation Training of Child Sexual Abuse Interviews With Feedback Improves Interview Quality in Japanese

University Students," Frontiers in psychology, vol. 11, p.998, 2020, doi: 10.3389/fpsyg.2020.00998.

[25] N. Krause 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 , vol. 149, p. 106677, 2024, doi: 10.1016/j.chiabu.2024.106677.

[26] T. B. Brown et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[27] K. A. Mills and A. Brown, "Immersive virtual reality (VR) for digital media making: transmediation is key," Learning, Media and Technology, vol. 47, no. 2, pp. 179–200, 2022, doi: 10.1080/17439884.2021.1952428.

[28] F. Rasheed, P. Onkar, and M. Narula, "Immersive virtual reality to enhance the spatial awareness of students," Proceedings of the Indian Conference on Human-Computer Interaction (IndiaHCI), 2015, doi: 10.1145/2835966.2836288.

[29] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," Computers & Education, vol. 147, p. 103778, 2020, doi: 10.1016/j.compedu.2019.103778.

[30] T. K. Metzinger, "Why Is Virtual Reality Interesting for Philosophers?," Frontiers in Robotics and AI, vol. 5, 2018, doi: 10.3389/frobt.2018.00101.

[31] L. Jensen and F. Konradsen, "A review of the use of virtual reality head-mounted displays in education and training," Education and Information Technologies, vol. 23, no. 4, pp. 1515–1529, 2018, doi: 10.1007/s10639-017-9676-0.

[32] C. Fowler, "Virtual reality and learning: Where is the pedagogy?," British Journal of Educational Technology, vol. 46, no. 2, pp. 412–422, 2015, doi: 10.1111/bjet.12135.

[33] S. G. Fussell and D. Truong, "Accepting virtual reality for dynamic learning: an extension of the technology acceptance model," Interactive Learning Environments, vol. 31, no. 9, pp. 5442–5459, 2023, doi: 10.1080/10494820.2021.2009880.

[34] P. Üstel et al., "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, vol. 39, no. 3, pp. 219–231, 2021, doi: 10.1080/15228835.2021.1915924.

[35] M. Kandaurova and S. H. (Mark) Lee, "The effects of virtual reality (VR) on charitable giving: the role of empathy, guilt, responsibility, and social exclusion," Journal of Business Research, vol. 100, pp. 571–580, 2019, doi: 10.1016/j.jbusres.2018.10.027.

[36] D. Shin, "Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience?," Computers in Human Behavior, vol. 78, pp. 64–73, 2018, doi: 10.1016/j.chb.2017.09.012.

[37] K.-W. K. Lai and H.-J. H. Chen, "A comparative study on the effects of a VR and PC visual novel game on vocabulary learning," Computer Assisted Language Learning, vol. 36, no. 3, pp. 312–345, 2023, doi: 10.1080/09588221.2021.1928226.

[38] E. Krokos, C. Plaisant, and A. Varshney, "Virtual memory palaces: immersion aids recall," Virtual Reality, vol. 23, no. 1, pp. 1–15, 2019, doi: 10.1007/s10055-018-0346-3.

[39] J. Madden, S. Pandita, J. P. Schuldt, B. Kim, A. S. Won, and N. G. Holmes, "Ready student one: Exploring the predictors of student learning in virtual reality," PLoS ONE, vol. 15, no. 3, p. e0229788, 2020, doi: 10.1371/journal.pone.0229788.

[40] S. M. Slobounov, W. Ray, B. Johnson, E. Slobounov, and K. M. Newell, "Modulation of cortical activity in 2D versus 3D virtual reality environments: An EEG study," International Journal of Psychophysiology, vol. 95, no. 3, pp. 254–260, 2015, doi: 10.1016/j.ijpsycho.2014.11.003.

[41] A. Dan and M. Reiner, "EEG-based cognitive load of processing events in 3D virtual worlds is lower than processing events in 2D displays," International Journal of Psychophysiology, vol. 122, pp. 75–84, 2017, doi: 10.1016/j.ijpsycho.2016.08.013.

[42] F. Tian, X. Wang, W. Cheng, M. Lee, and Y. Jin, "A Comparative Study on the Temporal Effects of 2D and VR Emotional Arousal," Sensors, vol. 22, no. 21, p. 8491, 2022, doi: 10.3390/s22218491.

[43] H. Wang, V. Gaddy, J. R. Beveridge, and F. R. Ortega, "Building an Emotionally Responsive Avatar with Dynamic Facial Expressions in Human—Computer Interactions," Multimodal Technologies and Interaction, vol. 5, no. 3, p. 13, 2021, doi: 10.3390/mti5030013.

[44] A. Visconti, D. Calandra, and F. Lamberti, "Comparing technologies for conveying emotions through realistic avatars in virtual reality-based metaverse experiences," Computer Animation and Virtual Worlds, vol. 34, no. 3–4, 2023, doi: 10.1002/cav.2188.

[45] E. Dzardanova, V. Nikolakopoulou, V. Kasapakis, S. Vosinakis, I. Xenakis, and D. Gavalas, "Exploring the impact of non-verbal cues on user experience in immersive virtual reality," Computer Animation and Virtual Worlds, Virtual, vol. 35, no. 1, 2024, doi: 10.1002/cav.2224.

[46] L. Luo, D. Weng, N. Ding, J. Hao, and Z. Tu, "The effect of avatar facial expressions on trust building in social virtual reality," The Visual Computer, vol. 39, no. 11, pp. 5869–5882, 2023, doi: 10.1007/s00371-022-02700-1.

[47] D. Y. Kim, H. K. Lee, and K. Chung, "Avatar-mediated experience in the metaverse: The impact of avatar realism on user-avatar relationship," Journal of Retailing and Consumer Services, vol. 73, p. 103382, 2023, doi: 10.1016/j.jretconser.2023.103382.

[48] P. Salehi et al., "Is More Realistic Better? A Comparison of Game Engine and GAN-based Avatars for Investigative Interviews of Children," Proceedings of the ACM Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR), pp. 41-49, 2022, doi: 10.1145/3512731.3534209.

[49] R. K. Røed et al., "Enhancing questioning skills through child avatar chatbot training with feedback," Frontiers in Psychology, vol. 14, 2023, doi: 10.3389/fpsyg.2023.1198235.

[50] J. Brooke, "SUS: A Quick and Dirty Usability Scale," Usability Evaluation in Industry, vol. 189, no. 194, pp. 4–7, 1996, doi: 10.1201/9781498710411-35.

[51] F. Biocca and M.R. Levy,"Communication in the Age of Virtual Reality," Routledge, 2013.

[52] E. Wilson, D. G. Hewett, B. C. Jolly, S. Janssens, and M. 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, vol. 3, no. 1, 2018, doi: 10.1186/s41077-018-0080-7.

[53] S. Schmidt, "Assessing the Quality of Experience of Cloud Gaming Services," M.S. thesis, Springer T-Labs Series in Telecommunication Services, Springer Nature, Berlin, Germany, 2022. [Online]. Available: https://link.springer.com/content/pdf/10.1007/978-3-031-06011-3.pdf

[54] Q. Zhu, L. Cui, W.-N. Zhang, F. Wei, and T. Liu, "Retrieval-Enhanced Adversarial Training for Neural Response Generation," arXiv preprint, 2019, doi: 10.18653/v1/p19-1366.

[55] S. Z. Hassan et al., "Towards an AI-driven talking avatar in virtual reality for investigative interviews of children," In Proceedings of the Workshop on Games Systems (GameSys), pp. 9-15, 2022, doi: 10.1145/3534085.3534340.

[56] R. Skarbez, F. P. Brooks, Jr., and M. C. Whitton, "A Survey of Presence and Related Concepts," ACM computing surveys (CSUR), vol. 50, no. 6, pp. 1–39, 2018, doi: 10.1145/3134301.

[57] S. S. Sabet et al., "Comparison of Crowdsourced and Remote Subjective User Studies: A Case Study of Investigative Child Interviews," Proceedings of the International Conference on Quality of Multimedia Experience (QoMEX), pp. 1-6, 2022, doi: 10.1109/qomex55416.2022.9900900.

[58] S. S. Sabet, "The influence of delay on cloud gaming quality of experience," Springer, Berlin/Heidelberg, Germany, pp. 1-127, 2023, doi: 10.1007/978-3-030-99869-1_3.

[59] M. Lammerse, S. Z. Hassan, S. S. Sabet, M. A. Riegler, and P. Halvorsen, "Human vs. GPT-3: The challenges of extracting emotions from child responses," Proceedings of the International conference on quality of multimedia experience (QoMEX), pp. 1-4, 2022, doi: 10.1109/qomex55416.2022.9900885.

[60] A. Mehrfard, J. Fotouhi, G. Taylor, T. Forster, N. Navab, and B. Fuerst, "A comparative analysis of virtual reality head-mounted display systems," arXiv preprint, Dec. 2019. doi: 10.48550/arXiv.1912.02913

[61] M. E. Lamb, K. J. Sternberg, Y. Orbach, I. Hershkowitz, D. Horowitz, and P. W. Esplin, "The Effects of Intensive Training and Ongoing Supervision on the Quality of Investigative Interviews With Alleged Sex Abuse Victims," Applied Developmental Science, vol. 6, no. 3, pp. 114–125, 2002, doi: 10.1207/s1532480xads0603_2.

[62] N. Ellemers, "Gender Stereotypes," Annual review of psychology, vol. 69, no. 1, pp. 275–298, 2018, doi: 10.1146/annurev-psych-122216-011719.

[63] Y. Karni-Visel, I. Hershkowitz, M. E. Lamb, and U. Blasbalg, "Facilitating the Expression of Emotions by Alleged Victims of Child Abuse During Investigative Interviews Using the Revised NICHD Protocol," Child Maltreat, vol. 24, no. 3, pp. 310–318, 2019, doi: 10.1177/1077559519831382.

PEGAH SALEHI is currently a PhD candidate at SimulaMet, 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 is currently a PhD candidate at SimulaMet, Norway, and Oslo Metropolitan University (OsloMet). His research work is focused on natural language processing and virtual reality systems. He has worked on designing lifelike avatars integrated within a VR framework, leveraging LLM and vision technologies to simulate realistic interactions, and integrating a LLM to provide real-time feedback on user performance.

GUNN ASTRID BAUGERUD is an associate professor at Oslo Metropolitan University (OsloMet), and she 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 child protective services (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 University 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 former PostDoc at SimulaMet and currently a developer and machine learning engineer at Forzasys AS, Norway. 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 (QoE) in multimedia applications.

MICHAEL A. RIEGLER received the Ph.D. degree from the Department of Informatics, University of Oslo, Norway, in 2015. He is currently working as a Chief Research Scientist at SimulaMet, 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, Norway, a Full Professor at Department of Computer Science, Oslo Metropolitan University (OsloMet), and an Adjunct Professor at Department of Informatics, University of Oslo, Norway. His research interest includes multimedia systems operating systems, processing, storage, retrieval, communication, distribution, and data analysis.