Free-form Conversation with Human and Symbolic Avatars in Mixed Reality

Jiarui Zhu*
University of California, Santa Barbara

Radha Kumaran†
University of California, Santa Barbara

Chengyuan Xu‡
University of California, Santa Barbara

Tobias Höllerer§
University of California, Santa Barbara

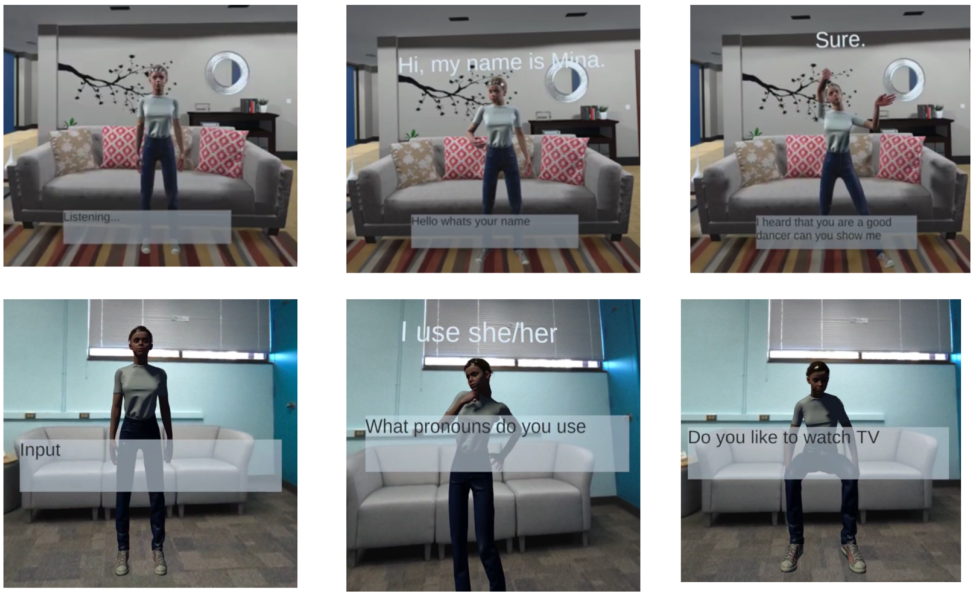

Figure 1: The conversational agent, depicted as a human avatar, is shown in both virtual (top row) and augmented (bottom row) reality environments.

Abstract

The integration of large language models and mixed reality technologies has enabled users to engage in free-form conversations with virtual agents across different "realities". However, if and how the agent's visual representation, especially when combined with mixed reality environments, will affect the conversation content or user experience is not yet fully understood. In this work, we design and conduct a user study involving two types of visual representations (a human avatar and a symbolic avatar) and two mixed reality environments (virtual reality and augmented reality), facilitating a free-form conversation experience with GPT-3 powered agents. We found evidence that the use of virtual or augmented realities can influence conversation content. Users chatting with avatars in virtual reality made significantly more references to the location or the space, suggesting they tended to perceive conversations as occurring in the agent's space, whereas the physical AR environment was perhaps more perceived as the user's space. Conversations with the human avatar improve user recall of the conversation, even though there is no evidence of increased information extracted during the conversation. These observations and our analysis of post-study questionnaires suggest that human avatars can positively impact user memory and experience. We hope our findings and the open-source implementation will help facilitate future research on free-form conversational agents in mixed reality.

Keywords: Artificial intelligence (AI), large language models (LLMs), mixed reality

Index Terms: Human-centered computing—Empirical studies in HCI; Computing methodologies—Mixed / augmented reality Computing methodologies—Virtual reality

1 Introduction

Mixed Reality (MR) is an exciting and rapidly developing technology that seamlessly blends the physical and digital worlds to create immersive experiences. In recent years, the industry and academia have shown increasing interest in this field, as new iterations of research and consumer MR products have brought about and refined the capability of experiencing both augmented reality (AR) and virtual reality (VR) within the same headset, such as Apple Vision Pro, Varjo XR3, HTC Vive XR Elite, and Meta Quest Pro.

At the same time, huge advancements in large language models, particularly in generative models such as GPT-3 [4], ChatGPT [22], Google Bard [1], LLaMA [38], etc., have provided unprecedented conversational capabilities, making it possible for humans to have a free-form continuous conversation with AI agents.

The natural next step is combining these advancements to create interactive conversational agents that can support educational, entertainment, or research applications across virtual and augmented realities, or even freely switch between the two realities as needed. Free-form conversation is the primary method of interaction with voice agents and LLMs today. They are likely to continue to be used

*e-mail: jiarui_zhu@ucsb.edu
†e-mail: rkumaran@ucsb.edu
‡e-mail: cxu@ucsb.edu
§e-mail: holl@cs.ucsb.edu

widely even when MR becomes the primary interaction paradigm, both for general information consumption as well as for more specific applications such as training, healthcare and entertainment. Since MR offers the potential to give virtual agents a human-like appearance with AI-driven avatars, it is important to understand the nature of these interactions both to foster effective communication and to optimize user satisfaction. In terms of the visual representations of the conversational agents, it is worth noting that symbolic avatars or physical objects still hold an important place in many current applications due to their simplicity and lower resource requirements (e.g., Apple's Siri and Amazon's Alexa speakers). On the other hand, numerous attempts have been made to create interactive human avatars in mixed reality [20, 23, 41].

A systematic user study that examines the impacts of both the visual representation of the agents and the mixed reality context, especially in the case of free-form conversation, has yet to be conducted. This is largely due to the fact that true free-form human-AI conversations have only become possible recently. This identifies an opportunity and motivates our investigation into the following research questions:

1. Given the same MR environment and conversation model, what are the benefits or motivations for choosing between a humanoid avatar and a familiar symbolic avatar (e.g., an object akin to current voice assistants) as the visual representation?

2. Given the same visual representation of the agent, how might the nature of the conversation be affected by the virtual or augmented reality environments?

To answer the above questions, we design a 2x2 within-subjects user study shown in Figure 2. The study involves two types of visual representations – a human avatar and a symbolic avatar, and two types of mixed reality environments – virtual reality and augmented reality. Both environments are presented using the same device, the Meta Quest Pro, which mitigates potential confounding factors associated with using different devices to compare VR and AR settings.

The conversational agents in our study are powered by OpenAI's GPT-3 [4], which was the state-of-the-art large language model with publicly accessible APIs at the time of our research[1]. By leveraging the GPT-3 model, both the human avatar and symbolic agent are capable of generating text and speech responses to engage users while maintaining a "memory" of past conversations – a key aspect in providing continuous, contextually relevant responses.

To encourage conversation, we provided the agent with a background story in each of the four conditions and asked participants ($N = 25$) to learn as much information about the agent as they could. Since the user interactions with the AI agent were not limited to a predetermined script, allowing for a more natural conversation flow as opposed to rigid, scripted exchanges, we refer to user interactions with the AI agent as free-form conversation in this paper to highlight the unscripted nature of the interactions. The conversations were automatically transcribed, allowing for the analysis of conversation length, breadth, and sentiments. We reviewed and coded the transcripts as well as user summaries of the conversations after each session, to compare content and recall of information across different avatars and MR settings. Furthermore, subjective ratings were collected to provide additional evidence of user experience.

2 Related Work

2.1 Human Avatars in Mixed Reality

Avatars play a crucial role in mixed reality settings, providing interactive and unique experiences, which has made them a significant research topic. For instance, Kim et al. studied the impact of visual embodiment and social behaviors on the perception of Intelligent Virtual Agents (IVA). They discovered that users trust IVAs more when they exhibit human-like physical and social behaviors instead of just voice feedback [12]. Additionally, when IVAs had a physical form, they aided users in collaborative decision-making by significantly reducing their task load [13]. Chang et al. expanded the research by studying users' cognition and behaviors through EEG when interacting with IVAs in virtual reality [5]. Norouzi et al. conducted a comprehensive review of recent studies on embodied agents in augmented reality using head-mounted displays, highlighting trends, identifying gaps, and offering insights for future research [21].

Slater et al. discussed the potential of immersive virtual reality (IVR) to transform various aspects of life and investigated the factors that contribute to the sense of embodiment in avatars [8,31]. Pakanen et al. evaluated virtual avatars in both AR and VR-based telexistence systems [23], finding that users favored photorealistic full-body human avatars. Mousas et al. studied the impact of virtual characters' appearance and movement on emotional reactivity, determining that these factors significantly influenced emotional response, valence, and reactivity [20]. Reinhardt et al. compared simple and more realistic agent appearances in AR, concluding that users preferred more realistic visualizations due to the additional communication cues they provided. Banos et al. explored the impact of avatar personalization and immersion on the sense of virtual body ownership, presence, and emotional response in virtual environments [41]. Ben Lok's team at the University of Florida researched various avatar characteristics, including skin tone, communication methods, and realism levels. Specifically, You et al. studied how certain avatar traits influence users' willingness to disclose information [44], and Zalake examined how a virtual human's appearance affects users' trust [47]. Stuart et al. analyzed how different virtual human rendering styles affect the perception of visual cues in healthcare [35] and explored how these perceptions impact error detection [34]. Additionally, Zalake et al. evaluated how effectively virtual humans apply persuasion strategies to influence user behavior and intentions [46] [45].

Researchers have been delving into the interactions with virtual characters across various reality settings using head-mounted displays. They've examined factors like perception, plausibility, and

Figure 2: Conversational avatars and MR settings investigated in this study.

[1]OpenAI announced ChatGPT (https://openai.com/blog/chatgpt) during this research, but its APIs did not become publicly available until long after our user study had concluded.

sense of presence, all critical to user experience. For instance, Wolf et al. studied the impact of different XR displays on how users perceive virtual humans [42]. Tang et al. investigated how users' virtual body representations influence their perceptions in both VR and AR [37]. Ghoshal et al. assessed player experiences in a co-located XR game, noting that VR intensifies individual presence, whereas AR boosts player interaction [9].

In recent times, researchers have started to delve deeper into the potential benefits of avatars in mixed reality. Advancements in artificial intelligence and gaming systems have enabled the application of existing models and technologies to real-world challenges. Hassan et al. developed an avatar in VR that simulates an abused child [10], which can aid police and child protection services in interview training. They found that using such an avatar improved users' knowledge and performance. Moreover, Quandt el al. created a system in which signing avatars teach introductory American Sign Language in VR [27], with positive feedback on the potential for learning ASL from an avatar in an immersive virtual reality environment.

2.2 AI Conversation Generation

Retrieval-based and generation-based approaches are two of the most common methods used in AI conversation generation.

In retrieval-based systems, responses are generated by selecting the most appropriate answer from a predefined set of responses [30, 33, 43]. Li et al. addressed the issue of diversity [14] in retrieval-based conversational systems by proposing a new objective function that encourages more diverse and engaging responses. Lowe et al. benchmarked a range of retrieval-based dialogue models, establishing a new standard for this type of approach and introduced the Ubuntu Dialogue Curpus [15]. Bordes et al. demonstrated the use of Memory Networks [3]. Zhou et al. employed a hierarchical attention mechanism to select appropriate responses in multi-turn conversations [48].

Generation-based systems create responses on-the-fly using natural language generation (NLG) techniques, often based on deep learning models such as sequence-to-sequence models [36], recurrent neural networks (RNNs) [40], transformers [39], BERT [7], or GPT-like architectures [4, 22, 28, 29]. OpenAI demonstrated that GPT-3 can perform various natural language processing tasks, including conversational AI, with impressive results.

Our work aims to leverage recent advances in NLG to create interactive mixed reality experiences that enable optimal user performance and experience.

2.3 AI and Conversation Agents in Mixed Reality

AI powered avatars are becoming increasingly popular as they can offer personalized and interactive experiences. Replika is an AI-powered chatbot that creates a personalized avatar to engage users in conversations.

AI-powered NPCs (non-player characters) or avatars in games are becoming more sophisticated, as they can provide more dynamic and engaging experiences for players. Elizabeth - BioShock Infinite: Elizabeth is an AI-driven companion who helps players in the game by providing support and resources. Her AI system allows her to react dynamically to the game environment and player actions.

AI-based conversational avatars have also been used in VR and AR applications for task training [18], studying embodied [32] and non-verbal behavior [24] and teaching social protocols [2]. Most of these works, however, had a specific target group of users and also offered limited conversational abilities since they were based on older conversation generation techniques.

Qu et. al [26] did study participant impressions of a conversational virtual human, but their focus was on participant perception of the virtual human's emotion, and the impact of cultural background on that perception.

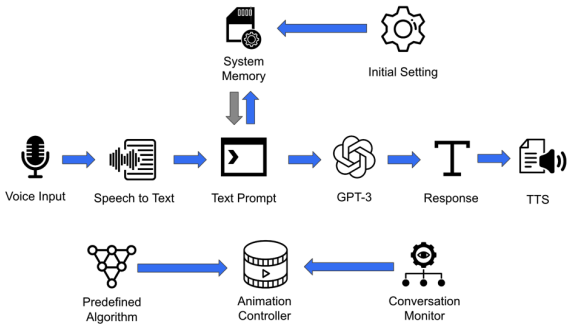

Figure 3: System architecture and key components supporting the interactive conversational agents.

In this work, we focus on studying free-form user conversation with a GPT-3 powered conversational agent, to better understand how users interact with these conversational avatars in an everyday interaction, akin to the use of voice assistants on smart devices.

3 System Design

A high-level overview of the system architecture for our interactive conversational agents is shown in Figure 3. The system incorporates several key components, including: a) the language model responsible for generating human-like conversation responses (OpenAI's GPT-3 [4]); b) the front-end visual and auditory system for interactive avatars (Unity, Oculus Voice SDK, and Meta Oculus Quest Pro); c) a machine learning model that monitors the ongoing conversation to trigger context-aware avatar animations (Wit.ai).

Although our system leverages a suite of existing libraries and tools, the main contribution of this paper lies in the insights learned from the user study. To the best of our knowledge, at the time of our research, our system was the first to synergize a human avatar with a cutting-edge large language model. The unique system facilitates a user study in mixed reality settings, enabling a comparison of free-from conversations with different agent representations.

Our system implementation is publicly accessible [2] to enhance the reproducibility of this work and, more importantly, to aid other researchers in facilitating similar studies.

3.1 Visual and Audio

The Avatars The simple human avatar used in our system is a character from Mixamo[3], a company that provides 3D animation assets and tools for game developers and animators for free. The avatar comes fully textured and rigged, with a pre-built skeleton and controls for animation. While more sophisticated avatars are available (e.g., Omniverse Avatars by NVIDIA), such realistic human depictions could easily fall into the "uncanny valley" [19] and negatively affect the conversation.

A talking symbol can resemble many existing voice assistants, such as Apple's Siri, Amazon's Alexa speakers, or various web-based automated agents. Thus we believe the chosen animation style human avatar and symbolic avatar provided the equally "basic but functional" quality that actually created more equal footing for our user study – the participants understood the nature of their presentation, yet they may have been equally unimpressed with their visual appeal for either design.

The animation component of the human avatar is designed to bring the avatar to life with 15 different animations mimicking human body language and gestures. Some animation examples

[2] https://github.com/gitzhujiarui/avatar_in_mixed_reality

[3] Mixamo https://www.mixamo.com/

include thinking, talking, dancing, sitting down, standing up, and walking around. To make these animations contextually relevant to the conversation, we employ a pre-trained machine learning model that monitors the dialogue in real-time. Every new utterance from the conversation is converted into a pre-configured set of intents with a threshold. Once the threshold is reached, a predefined animation is triggered. The pre-defined ML model is trained with Wit.ai and integrated into the system using Oculus Voice SDK.

Text Display To enhance communication accuracy, the system employs an interactive text display feature. When the user initiates or resumes a conversation with the agent by pressing a designated button on the controller, the message "Listening..." appears in a translucent text box in front of the avatar. At the same time, the user's voice input is transcribed into text using a text-to-speech service provided by the Oculus Voice SDK, replacing the "Listening.." message in the text box. This allows for easy identification and correction of any potential errors or misunderstandings, and ensures that users have a clear understanding of what the avatar "hears".

When the avatar responds, the reply is displayed as a text message floating above the avatar, in addition to the audio response. The combination of text and audio improves information retention and comprehension. Although not the focus of this study, the text input and output feature also makes the conversation accessible to users who may have difficulty hearing. Additionally, users who are not fluent in the language the avatar speaks can benefit from the ability to read and understand the text on the screen.

Auditory The audio response of the avatars is an important part of the system. When the language model generates a response, it is sent to the Oculus Voice SDK's text-to-speech (TTS) feature. The TTS feature uses a cloud-based service to convert the response text into audio output. This audio response is then played through the VR headset's speakers as part of the interactive experience.

3.2 Language Model and Response Generation

The avatars are powered by the GPT-3 (Generative Pre-trained Transformer 3) language model, which is a sophisticated AI language model that is pre-trained on a vast amount of data. The GPT-3 model is used to generate the avatar's response based on user prompts. The system relies on OpenAI's official API to interact with the language model and get responses.

In each of conversational agent and MR setting, the avatar is assigned a unique background setting that includes information such as the avatar's name, age, hobbies, and background story. This setting can be easily customized to give the avatar a unique identity. At the start of each conversation, the system generates a text based on the avatar's background setting, which is then used as part of the seed prompt for the GPT-3 model. This ensures that the avatar's response is centered around its unique identity and background and gives the participants a feeling of talking to different people in different study sessions. Additionally, our system maintains the flow of conversations by saving every interaction, including both the user's input and the agent's response. This information is then used in the prompt for the next interaction with the GPT model. This process enables the AI agents to remember the conversation and generate responses that are aware of the previous exchange.

4 Experiment

This experiment was conducted as a 2x2 within-subjects user study involving 25 participants between the ages of 19 and 28 (M=22.48), 11 of whom identified themselves as female. A full permutation of our four experiment conditions (chosen to maximize statistical power) requires 24 participants, but due to a combination of no-shows and extra recruitments, we ended with 25 participants completing the study. 2 participants had never used VR before, 16 had used it less than 10 times, and 7 had used it more than 10 times. 8 participants had never used AR before, 11 had used it less than 10 times and 6 had used it more than 10 times. Participants were a diverse group of students and affiliates from different departments and graduate levels at a university campus. The study protocol was approved by the university's Human Subjects Committee.

4.1 Design

We studied two independent variables in this experiment, the visual representation of the conversational agents and the mixed reality settings.

Conversational Agents Participants interacted with one of two conversational agents: a *human avatar* and a *symbolic avatar* (an object akin to current voice assistants). Both conversational agents were supported by the same language model. To create the illusion of conversing with distinct individuals, the model was initialized using the same narrative framework but with varying factual details. Due to the task's objective – for participants to understand the conversational agents – each agent had to have a unique backstory. This ensured that participants received different responses to the same questions when posed to different agents.

Mixed Reality Settings Participants experienced each of the conversational agents in two different MR environments: in a completely virtual environment (*VR*) as well as in an augmented reality environment (*AR*). The VR experience was created using a virtual background that resembles the real room where the user study took place, with a couch in front of the participant. The human avatar was positioned in front of the couch to create a seamless integration, allowing for an immersive and realistic VR experience that closely mirrored the real-world environment. In the AR experience, the human avatar was displayed on top of the real-world feed using Meta Quest Pro's color passthrough mode, which captures images of the real world and displays them in the headset. The background was set in the room where the user study took place, and the human avatar was again positioned in front of the couch so as to seamlessly integrate it into the real room. In both VR and AR, the symbolic avatar was also positioned in front of the couch.

Each participant thus experienced four different conditions: interacting with the human avatar in AR, the human avatar in VR, the symbolic avatar in AR, and the symbolic avatar in VR (see Figure 2). In our design, avatars of the same type maintained consistent appearances across both AR and VR. For instance, the human avatars had the same look in both AR and VR settings, as did the symbolic avatars. This design choice was made to avoid introducing additional variables that could affect results. By ensuring a consistent appearance for both human and symbolic avatars, we aimed to focus on key variables influencing the user experience. The sequence in which the conditions were experienced was randomized and counterbalanced across participants.

4.2 Procedure

Participants first filled out a consent form and completed a pre-study questionnaire that collected demographic information. They were then fitted with the Meta Quest Pro headset and taught how to use the hand controllers. Specifically, they were trained to use the controllers for menu navigation and basic control of the headset.

Following this initial training, participants went through a tutorial that helps them get familiar with the application and learn to interact with the voice agents. Specifically, they were taught to press the specific trigger on the controller every time they want to engage in a conversation with the voice agents. The training session also helped participants to get familiar with the application interface and the expected experience with an agent. At the end of the tutorial, participants were given the opportunity to practice interacting with the voice agents exactly as they would in the main experiment, until they felt ready to begin the experiment. The duration of the tutorial was generally about five minutes, subject to the comfort level

of the participants. Participants saw both the virtual and physical environments in this stage, and were thus familiar with it when they began the study.

Each participant completed four experiment trials, one in each of the four conditions described previously. In each trial, the participant was asked to learn as much as they could about the agent within approximately seven minutes, but they are also free to end the conversation earlier if they chose to. The conversation's audio and video were both recorded. After each trial, participants were asked to document everything they had learned about the agent in that trial in a document. To ensure consistency in coding the facts they learned about the agents, participants were advised to use the same writing style (e.g., bullet points or paragraphs depending on their preference) across the four trials when summarizing the conversation. Specifically, they were instructed to take cognizance of their language use, punctuation, and grammar to maintain consistency in their summaries. For instance, if a participant opted to utilize a bullet-point format for summarizing a conversation, the expectation was to consistently adhere to this style for the entirety of the experiment. We can confirm that after careful reading and coding user responses, each participant largely maintained a consistent writing style between the four condition trials. At the end of the four trials, a post-study questionnaire was administered, to understand user experience and preferences among the four conditions.

To combat the issue of demand characteristics, the four experiment conditions were presented in a randomized order. In addition, participants, unaware of the upcoming setting, were asked to recall conversations immediately after each session.

4.3 Analysis

We analyzed the conversation transcripts using quantitative conversation metrics discussed below. We also analyzed user experience based on subjective ratings obtained in the post-study survey.

4.3.1 Quantitative Conversation Metrics

The application automatically saves both a transcript and a video recording of each conversation when a participant completes each trial. These data allow us to robustly measure and compare: 1)conversation content, 2) conversation breadth, 3) conversation recall, 4) conversation length and 5) conversation sentiment.

Conversation content Following a practice similar to open coding and axial coding in the grounded theory method [6], two researchers independently reviewed the transcripts to identify potential topic variations between the experimental conditions for more in-depth analysis. Upon confirming that their disparate criteria converged on similar findings – that under certain conditions, more participants conversed with the agent about the space where the conversation took place or the agent's occupation, they coded the transcripts one more time but using a unified rule.

On the topic of the current space or environment where the conversation takes place, if a human participant initiated one of the following topics: discussing the location (e.g., "Where are we?"), the space itself (e.g., "Is this your apartment?"), or subjects related to the space (e.g., "I like your sofa."), the conversation was assigned one point for each topic mentioned. Therefore, a conversation that touched upon all three topics would receive 3 points, while a conversation that did not discuss space topics would receive 0 points. Conversations that mentioned the agent's occupation or career plans were coded in a binary manner (yes/no). The average of the two researchers' coded scores was subsequently used for statistical analysis.

Conversation recall Immediately after each trial, we asked the participants to write down a summary of the conversation. A third researcher manually coded the transcript into topics to determine the number of unique *facts discovered* in each conversation. The same researcher coded the corresponding user summary with the same rules to count the number of facts that a participant remembers from the conversation (referred to as *facts remembered*). The participants' *fact recall* is computed as the ratio of the facts remembered to the facts discovered:

$$\text{fact recall} = \frac{\text{facts remembered}}{\text{facts discovered}}. \quad (1)$$

This metric of conversation recall was chosen since gist recall is generally higher than verbatim recall [25].

Conversation length and breadth We utilized the transcripts to calculate several quantitative metrics pertaining to user experience. The *total number of words* in each conversation was computed directly from the raw transcript, while the *unique word count* for each conversation was computed after preprocessing the transcript, which included stop word removal and tokenization. To account for differences in the conversation duration among participants, we normalized the *total word count*, *unique word count* and *facts discovered* by the duration of the recorded video (in minutes). Even though the average conversation duration didn't vary significantly across the four conditions, there were noticeable differences in duration among individual participants. We believed that a normalized word count would provide a more accurate metric.

Conversation Sentiment In order to quantify conversation sentiment, each sentence of the conversation was classified as *positive*, *neutral* or *negative* based on the sentiment score returned by the VADER [11] sentiment analysis tool. We then used the proportion of sentences in each conversation that were positive, neutral, and negative respectively as metrics of conversation sentiment.

4.3.2 User Experience and Subjective Ratings

In the post-study survey, participants were requested to provide ratings on a 7-point Likert scale (1=Strongly disagree, 7=Strongly Agree) for their *comfort* level ("I found it comfortable to interact with a {human/symbolic} avatar in {AR/VR} via conversation"), *naturalness* "I found it natural to interact with a {human/symbolic} avatar in {AR/VR} via conversation"), the *enjoyment* derived ("The overall experience of having conversation with a {human/symbolic} avatar in {AR/VR} was enjoyable"), and the amount of *information* they absorbed (1=Very little, 7=A great deal) ("How much information do you feel you got from interacting with a {human/symbolic} avatar in {AR/VR}") when interacting in four distinct settings. Each of the four metrics was analyzed individually to understand user experience in the four experiment conditions.

To assess the impact of the different conditions on the quantitative metrics two-way repeated measures ANOVAs were used. Pairwise comparisons using the Holm correction were used to follow-up significant main effects and interactions. Two-way repeated measures ANOVAs were also used to analyze users' subjective ratings of the different conditions, and followed up by pairwise comparisons using the Holm correction. ANOVAs where the assumption of normality was violated (based on the Shapiro-Wilk test of normality) were re-analyzed using Friedman's test, a non-parametric equivalent of repeated measures ANOVA. The results remained consistent, therefore initial ANOVAs are reported for clarity. Furthermore, ANOVA is known to be robust to violations of normality [17]. These variables included the subjective ratings (comfort [all $W > 0.88$, $p < 0.05$], enjoyment [all $W > 0.8$, $p < 0.05$], information [all $W > 0.86$, $p < 0.05$], naturalness [$W > 0.85$, $p < 0.05$ for avatar in AR, avatar in VR and voice in AR]) and some of the quantitative metrics (unique words [$W = 0.89$, $p < 0.05$ for voice in AR], conversation length [$W > 0.86$, $p < 0.05$ for avatar in AR, voice in VR and voice in AR], space references [all $W > 0.41$, $p < 0.05$], occupation references [all $W > 0.20$, $p < 0.05$] and negative sentiment [$W > 0.73$, $p < 0.05$ for avatar in VR, voice in AR and voice in VR]). Only one variable violated homogeneity of variances (*space references*). This

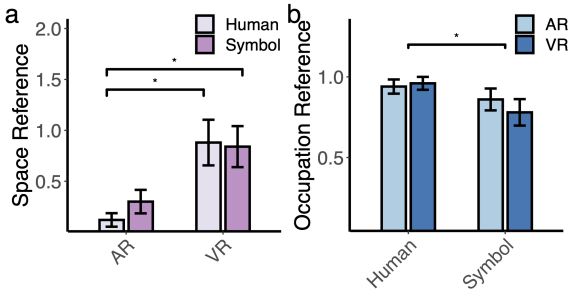

Figure 4: Conversation content. (a) Conversation score for references to current space (0 indicates no references, and 3 indicates references to all of the 3 topics discussed in Section 4.3.1), and (b) Conversation score for references to occupation as the average of two researchers' independent scoring of each conversation (0 indicates no references, 1 indicates mention of the subject), plotted as a function of conversational agent and MR setting. Participants referenced the current space more often in the VR setting than in the AR setting, and inquired about professional occupation more often with the human avatar than the symbolic avatar. For this and all following figures: Error bars = SEM. *Human* refers to the human avatar, and *Symbol* refers to the symbolic avatar.

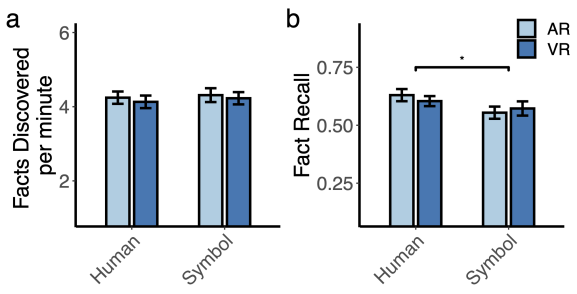

Figure 5: (a) Number of unique facts discovered in the conversation and (b) Recall of discovered facts at the end of the conversation, plotted as a function of conversational agent and MR setting. There was no difference in the number of facts discovered among any of the experiment conditions, which tracks with the unique words in each conversation. Interestingly, participants had a significantly better recall of facts (they remembered more of the conversation) with the human avatar when compared to the symbolic one.

was analyzed using Friedman's test (a non-parametric equivalent of the repeated measures ANOVA), and followed up by pairwise comparisons using the Wilcoxon signed rank test with Holm correction.

5 Results

Participants data was collected from the pre-study survey, post-study survey, conversation transcript, and participants' summary of each conversation. We analyzed the data with a focus on the conversation content, quality, participants' ability to remember in different settings, as well as the overall experience.

5.1 Quantitative Conversation Metrics

We analyzed participant conversations with the conversational agents from five perspectives : conversation content (*space references* and *occupation references*), conversation breadth (*unique words*, *facts discovered*), conversation recall (*fact recall*), conversation length (*total words*) and conversation sentiment.

Conversation Content Figure 4a shows the score for references to the current space in the conversation (on a scale of 0-3) as a function of conversational agent and MR setting. Conversations in VR had a higher score than conversations in AR, as indicated by a Friedman's test ($\chi^2(3) = 9.43$, $p < 0.05$, $effsize = .13$, *medium*). Pairwise comparisons using the Wilcoxon test revealed that the score for Human in VR was higher than Human in AR [$W = 7.5$, $p < 0.05$, $effsize = 0.5$, *large*]. Further, Symbol in VR had higher scores than Human in AR [$W = 6$, $p < 0.05$, $effsize = 0.55$, *large*] and was numerically higher than Symbol in AR, though not significantly so [$W = 8$, $p = 0.06$, $effsize = 0.40$, *large*].

Figure 4b shows the score for references to professional occupation in the conversation, with 0 indicating no references and 1 indicating mentions of professional occupation (as coded by two researchers). There appear to be more such references with the human avatar when compared to the symbolic avatar. Consistent with this pattern, 2x2 [conversational agent; MR setting] repeated measures ANOVA on the *occupation reference* score revealed a main effect of conversational agent ($F(1, 24) = 6.69$, $p < 0.05$, $\eta^2 = .22$, *large*), with more references in conversations with the human avatar than the symbolic avatar [$t(49) = 2.45$, $p < 0.05$, $d = 0.08$, *medium*]. There was no effect of MR setting ($F(1, 24) = 0.59$, $p > 0.05$, $\eta^2 = .02$, *small*) or interaction between the two factors ($F(1, 24) = 0.8$, $p > 0.05$, $\eta^2 = .03$, *small*).

These results indicate that the conversation content was influenced by both the setting and the appearance of the conversational agent, with more references to the current space in VR compared to AR as well as more references to professional occupation with the human avatar compared to the symbolic avatar.

Conversation Breadth Figure 6b shows the total number of *unique words* in the conversation per minute, as a function of conversational agent and MR setting. There appears to be no difference in the number of unique words among conversational agent and MR setting. Consistent with this pattern, a similar repeated measures ANOVA revealed no main effect of conversational agent ($F(1, 24) = 1.44$, $p > 0.05$, $\eta^2 = .05$, *small*), MR setting ($F(1, 24) = 0.61$, $p > 0.05$, $\eta^2 = .02$, *small*) or interaction between the two factors ($F(1, 24) = 0.10$, $p > 0.05$, $\eta^2 = .004$, *small*).

Figure 5a shows the number of *facts discovered* in the conversation per minute, as a function of conversational agent and MR setting. Again, there is no difference in the number of *facts discovered* per minute, among conversational agent and MR setting. Consistent with this pattern, the ANOVA revealed no main effect of conversational agent ($F(1, 24) = 0.77$, $p > 0.05$, $\eta^2 = .03$, *small*), MR setting ($F(1, 24) = 0.77$, $p > 0.05$, $\eta^2 = .03$, *small*) or interaction between the two factors ($F(1, 24) = 0.02$, $p > 0.05$, $\eta^2 = .001$, *small*).

These results suggest that participants didn't necessarily have a wider variety of conversation subjects, or extract more information, from conversations with the human avatar when compared to the symbolic avatar.

Conversation Recall Figure 5b plots participants *recall* of all facts discovered in the conversation, as a function of conversational agent and MR setting. Visual inspection suggests that participants recalled more facts from conversations with the human avatar when compared to the symbolic avatar. Consistent with this, the ANOVA revealed a main effect of conversational agent on fact recall ($F(1, 24) = 5.27$, $p = 0.03$, $\eta^2 = .18$, *large*), such that participants recalled a significantly larger number of facts from conversations with the human avatar when compared to the symbolic avatar [$t(49) = 2.36$, $p = 0.02$, $d = 0.41$, *small*]. There was no main effect of MR setting ($F(1, 24) = 0.05$, $p > 0.05$, $\eta^2 = .002$, *small*) or interaction between the two factors ($F(1, 24) = 0.98$, $p > 0.05$, $\eta^2 = .04$, *small*).

Conversation length A 2x2 [conversational agent; MR setting] repeated measures ANOVA on the *total number of words* in the

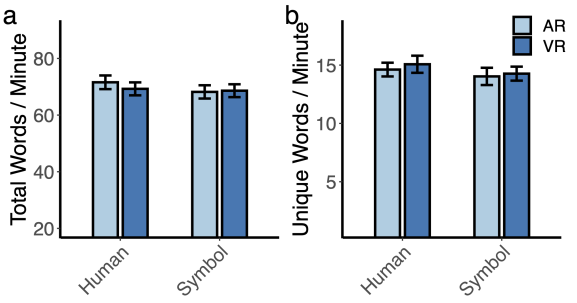

Figure 6: (a) Total number of words in the conversation and (b) Total number of unique words in the conversation, plotted as a function of conversational agent and MR setting. In the AR setting, conversations were significantly longer with the human avatar compared to the symbolic one. However, there was no difference in the number of unique words in conversations among any of the experiment conditions.

conversation per minute (plotted in Figure 6a) revealed no main effect of conversational agent ($F(1, 24) = 1.97$, $p > 0.05$, $\eta^2 = .08$, *medium*), MR setting ($F1, 24) = 0.55$, $p > 0.05$, $\eta^2 = .02$, *small*), or interaction between the two factors($F(1, 24) = 1.12$, $p > 0.05$, $\eta^2 = .04$, *small*). This suggests that there was no impact of conversational agent or MR setting on the length of the conversation.

Conversation Sentiment A 2x2 [conversational agent; MR setting] repeated measures ANOVA on the proportion of positive, neutral and negative sentences in each conversation revealed no main effect of conversational agent [positive ($F(1, 24) = 0.18$, $p > 0.05$, $\eta^2 = .007$, *small*); neutral ($F(1, 24) = 0.15$, $p > 0.05$, $\eta^2 = .006$, *small*), negative ($F(1, 24) = 0.001$, $p > 0.05$, $\eta^2 = 0$, *small*)], MR setting [positive ($F(1, 24) = 0.97$, $p > 0.05$, $\eta^2 = .04$, *small*); neutral ($F(1, 24) = 0.46$, $p > 0.05$, $\eta^2 = .02$, *small*), negative ($F(1, 24) = 0.52$, $p > 0.05$, $\eta^2 = .02$, *small*)] or interaction between the two factors [positive ($F(1, 24) = 0.19$, $p > 0.05$, $\eta^2 = .008$, *small*); neutral ($F(1, 24) = 1.01$, $p > 0.05$, $\eta^2 = .04$, *small*), negative ($F(1, 24) = 1.48$, $p > 0.05$, $\eta^2 = .06$, *small*)].

These results suggest that neither conversational agent nor setting had significant impact on the sentiment of the conversation. Neither the task nor the experiment conditions had any elements that were designed to invoke strong emotions, so the results are unsurprising.

5.2 User Experience and Subjective Ratings

Visual inspection of Figure 7 suggests that participants perceived higher levels of comfort, enjoyment, naturalness and information received with the human avatar when compared to the symbolic avatar. The results of 2x2 repeated measures ANOVAS on these four metrics confirmed this pattern.

For *comfort*, there was a main effect of conversational agent on the level of comfort ($F(1, 24) = 7.11$, $p = 0.01$, $\eta^2 = .23$, *large*), with higher levels of comfort with the human avatar when compared to the symbolic avatar [$t(50) = 3.04$, $p < 0.01$, $d = 0.56$, *medium*]. There was no main effect of MR setting ($F(1, 24) = 3.62$, $p > 0.05$, $\eta^2 = .13$, *large*) and no interaction between the two factors ($F(1, 24) = 0.80$, $p > 0.05$, $\eta^2 = .03$, *small*).

For *enjoyment*, there was a main effect of conversational agent on the level of enjoyment ($F(1, 24) = 10.82$, $p < 0.01$, $\eta^2 = .31$, *large*), with higher levels of enjoyment with the human avatar when compared to the symbolic avatar [$t(50) = 3.68$, $p < 0.001$, $d = 0.54$, *medium*]. There was no main effect of MR setting ($F(1, 24) = 0.85$, $p > 0.05$, $\eta^2 = .03$, *small*) and no interaction between the two factors ($F(1, 24) = 0.04$, $p > 0.05$, $\eta^2 = .00$, *small*).

For *naturalness*, again, there was a main effect of conversational agent on the level of perceived naturalness, ($F(1, 24) = 27.24$, $p < 0.001$, $\eta^2 = .53$, *large*), with higher perceived naturalness with the human avatar when compared to the symbolic avatar [$t(50) = 6.31$, $p < 0.001$, $d = 1.00$, *large*]. There was no main effect of MR setting ($F(1, 24) = 3.61$, $p > 0.05$, $\eta^2 = .13$, *medium*) and no interaction between the two factors ($F(1, 24) = 0.46$, $p > 0.05$, $\eta^2 = .02$, *small*).

Lastly, for *information* received, there was a main effect of conversational agent on the perceived amount of information received in the conversation($F(1, 24) = 19.50$, $p < 0.001$, $\eta^2 = .44$, *large*), with more information perceived to be obtained from the human avatar when compared to the symbolic avatar [$t(50) = 4.95$, $p < 0.001$, $d = 0.80$, *large*]. There was no main effect of MR setting ($F(1, 24) = 2.61$, $p > 0.05$, $\eta^2 = .10$, *medium*) and no interaction between the two factors ($F1, 24) = 0.43$, $p > 0.05$, $\eta^2 = .02$, *small*).

6 Discussion

In this section we analyze the results in the context of our two research questions, and also discuss other factors that could have influenced the results.

6.1 AR vs. VR

Before this study, we expected that the visual and immersion disparity between AR and VR could influence conversation content and quality. For instance, the visually cohesive VR setting might create a relaxed atmosphere, encouraging more casual discussions, while the AR office setup might not stimulate such conversations. On the other hand, the real AR backdrop might have prompted the human participant to root the conversation more in reality rather than fantasy.

The results indicate more references to the current space (environment) in VR conversations when compared to AR conversations, with participants asking more questions about the current location where the conversation takes place and objects visible in the environment in VR when compared to AR. Given that the VR environment was more visually cohesive with the avatars than the AR space, the difference in conversation could have been due to differing perception of the environments - the VR environment being the "avatar's space", and the AR environment being the "participant's space". This would suggest that users might still view virtual avatars as 'out-of-place' in real environments, which could impact their interactions with virtual conversational agents in AR applications. Application designers can consider adding cues that clarify the virtual agents abilities and knowledge of the real environment, to help users interact with both virtual and real elements of the application more seamlessly.

6.2 Human Avatar vs. Symbolic Avatar

The presence of a human avatar as a conversational agent impacted user experience and recall of conversations when compared to the symbolic avatar. The participants remembered more information from conversations with the human avatar than with the symbolic avatar in augmented reality. Interestingly, though, there was no significant difference in the breadth of conversation (and hence, the amount of information discovered) with the human avatar compared to the symbolic avatar even though participants perceived more information received from conversations with the human avatar. This suggests that participants retain more information from interactions with the interactive human avatar, even though there is no evidence that the presence of the human avatar (rather than the symbolic avatar) actually changes the interaction. Recall of information has been shown to be influenced by involvement in the conversation [16], and our results could possibly be explained by participants' involvement in the conversation, as well as their perception of the avatar's involvement in the conversation. We did not, however, collect any data that could confirm this.

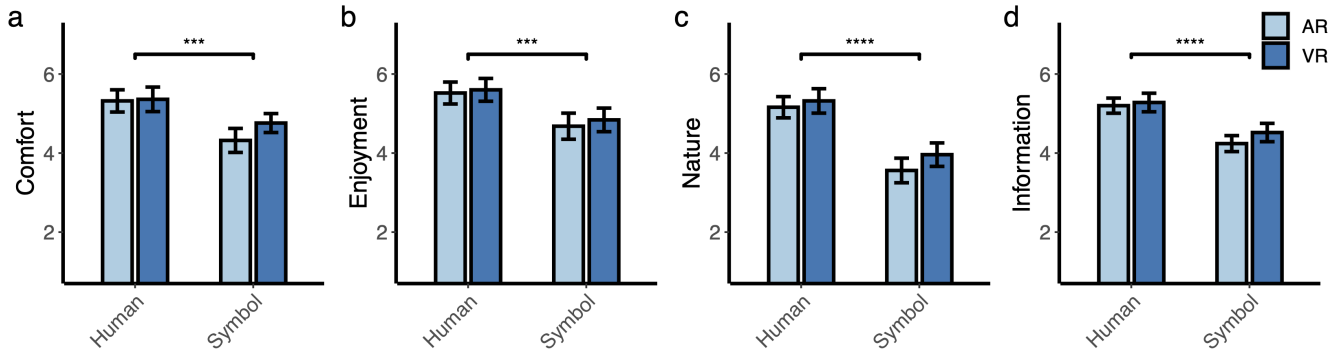

Figure 7: User ratings for subjective responses regarding (a) comfort, (b) enjoyment, (c) naturalness, and (d) information discovered in the conversation, on a scale of 1(low) to 7(high). Participants reported significantly higher levels of comfort, enjoyment and naturalness in the conversations with the human avatar when compared to the symbolic avatar. They also perceived a higher amount of information received from conversations with the human avatar.

Participants also reported higher levels of perceived comfort, enjoyment, and naturalness in their conversations with the interactive human avatar, which could explain the higher recall of information obtained during the dialog. These results highlight the importance of the avatar's appearance in user experience, even when the focus is the interaction with the conversational agent. The tradeoff between the more complex human avatars (which help improve user experience and information retention), and simpler symbolic avatars (which might be better suited to applications that value efficiency of interaction over user experience) are an important factor in the design of future AR systems.

6.3 Limitations

It is important to acknowledge that, while every effort was made to ensure a controlled set of experiment trials, the inherent nature of the free-form conversation task necessitated a degree of freedom in user interactions, in order to study natural human interaction with the avatars. We believe evaluating both AR and VR experiences using the same headset device helps mitigate many potential confounding factors that might affect the results. We observed a difference in avatar size between AR and VR, caused by the different scaling methods in the development environment. Future researchers might investigate how the size and distance of the virtual avatar influence conversation perceptions and overall interaction in mixed reality. The repetitive animations of the avatar during the conversation could also have impacted

We used a virtual couch in the VR setting that is almost identical to the physical couch as an anchor for the virtual agents, in order to make the AR and VR settings appear alike, we acknowledge that the VR environment naturally has a unique appeal and could have appeared more sophisticated. Since the advantage of VR is that it enables users to experience environments that are not constrained by their physical location, we were interested in discovering specific user behaviors that resulted from this difference. A study of how content in the VR environment influences user interaction is also an interesting direction for future research. In the post-processing of the data, we also took extra measures – including normalizing the quantitative measurements by duration and averaging independent coding results from different researchers.

It is also important to note that our use of the term free-form conversation in this paper is meant to emphasize the unscripted nature of the conversation, in contrast to the previous scripted studies we cited. A truly fully free-form conversation is hard to evaluate reliably, thus we designed a task that allows us to quantitatively measure and evaluate the conditions.

Better language models could generate more realistic and human-like conversations. While GPT-3 was a powerful and state-of-the-art language model at the time of our user study, there is room for improvement in the quality of conversation, particularly in the ability to respond to users with questions and expand on and dive deeper into topics when appropriate. Integration with ChatGPT or other more recent language models could improve the quality of the conversations and enhance the interactivity and overall experience of the avatar.

We have made our system implementation publicly available, and have provided the necessary detail to reproduce the user study. We welcome other researchers to conduct free-form conversation studies using our setup, and would love to see our findings tested and refined using even more advanced language models.

7 Conclusion

Recent advances in large language models have enabled the widespread use of AI conversation agents for a variety of applications. In this paper, we examine the potential of conversation agents in MR, specifically as a component of avatars that can converse with and support users in both virtual and augmented reality. We analyzed data collected from a within-subjects experiment that compares free-form conversation with a human avatar and a symbolic avatar, in both VR and AR. To the best of our knowledge, our research is the first to explore free-form conversations with virtual characters in different reality settings and to conduct a user study examining different influencing factors. Our findings suggest that the appearance of the virtual conversation agent impacts user interactions, with better user experience and improved retention of information in conversations with the human avatar. Participants also asked more questions about the space where the conversation took place in VR when compared to AR, suggesting that users perceived the VR environment as the agent's space rather than their own.

Future designs should emphasize crafting realistic avatars to amplify immersion and facilitate effective communication. Avatars that resonate with feelings of comfort and naturalness can further elevate user involvement and recall. On the other hand, symbolic avatars, with their inherent simplicity, are well-suited for applications that demand directness and efficiency, especially when human-like nuances or interactions aren't essential. The overarching challenge for designers is to discerningly choose the avatar type, aligning with the context and specific needs of users.

Acknowledgments

This work was supported in part by ONR grant N00014-23-1-2118.

REFERENCES

[1] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le. Towards a human-like open-domain chatbot, 2020.

[2] S. Babu, E. Suma, T. Barnes, and L. F. Hodges. Can immersive virtual humans teach social conversational protocols? In *2007 IEEE Virtual Reality Conference*, pp. 215–218, 2007. doi: 10.1109/VR.2007.352484

[3] A. Bordes, Y.-L. Boureau, and J. Weston. Learning end-to-end goal-oriented dialog, 2017.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Curran Associates Inc., Red Hook, NY, USA, 2020.

[5] Z. Chang, H. Bai, L. Zhang, K. Gupta, W. He, and M. Billinghurst. Corrigendum: The impact of virtual agents' multimodal communication on brain activity and cognitive load in virtual reality. *Frontiers in Virtual Reality*, 4, 04 2023. doi: 10.3389/frvir.2023.1194313

[6] J. M. Corbin and A. Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1):3–21, Mar. 1990. doi: 10.1007/BF00988593

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[8] R. Fribourg, F. Argelaguet, A. Lécuyer, and L. Hoyet. Avatar and sense of embodiment: Studying the relative preference between appearance, control and point of view. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):2062–2072, 2020. doi: 10.1109/TVCG.2020.2973077

[9] M. Ghoshal, J. Ong, H. Won, D. Koutsonikolas, and C. Yildirim. Co-located immersive gaming: A comparison between augmented and virtual reality. In *2022 IEEE Conference on Games (CoG)*, pp. 594–597, 2022. doi: 10.1109/CoG51982.2022.9893708

[10] S. Z. Hassan, P. Salehi, R. K. Røed, P. Halvorsen, G. A. Baugerud, M. S. Johnson, P. Lison, M. Riegler, M. E. Lamb, C. Griwodz, and S. S. Sabet. Towards an ai-driven talking avatar in virtual reality for investigative interviews of children. In *Proceedings of the 2nd Workshop on Games Systems*, GameSys '22, p. 9–15. Association for Computing Machinery, New York, NY, USA, 2022.

[11] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, vol. 8, pp. 216–225, 2014.

[12] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G. F. Welch. Does a digital assistant need a body? the influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 105–114, 2018. doi: 10.1109/ISMAR.2018.00039

[13] K. Kim, C. M. de Melo, N. Norouzi, G. Bruder, and G. F. Welch. Reducing task load with an embodied intelligent virtual assistant for improved performance in collaborative decision making. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 529–538, 2020. doi: 10.1109/VR46266.2020.00074

[14] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119. Association for Computational Linguistics, San Diego, California, June 2016. doi: 10.18653/v1/N16-1014

[15] R. Lowe, N. Pow, I. Serban, and J. Pineau. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 285–294. Association for Computational Linguistics, Prague, Czech Republic, Sept. 2015. doi: 10.18653/v1/W15-4640

[16] J. B. Miller. Memory for conversation. *Memory*, 4(6):615–632, 1996.

[17] R. G. Miller Jr. *Beyond ANOVA: basics of applied statistics*. CRC press, 1997.

[18] N. Moore, N. Ahmadpour, M. Brown, P. Poronnik, and J. Davids. Designing virtual reality–based conversational agents to train clinicians in verbal de-escalation skills: Exploratory usability study. *JMIR Serious Games*, 10(3):e38669, 2022.

[19] M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012. doi: 10.1109/MRA.2012.2192811

[20] C. Mousas, D. Anastasiou, and O. Spantidi. The effects of appearance and motion of virtual characters on emotional reactivity. *Computers in Human Behavior*, 86:99–108, 2018. doi: 10.1016/j.chb.2018.04.036

[21] N. Norouzi, K. Kim, G. Bruder, A. Erickson, Z. Choudhary, Y. Li, and G. Welch. A systematic literature review of embodied augmented reality agents in head-mounted display environments. In *ICAT-EGVE*, 2020.

[22] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022.

[23] M. Pakanen, P. Alavesa, N. van Berkel, T. Koskela, and T. Ojala. "nice to see you virtually": Thoughtful design and evaluation of virtual avatar of the other user in ar and vr based telexistence systems. *Entertainment Computing*, 40:100457, 2022. doi: 10.1016/j.entcom.2021.100457

[24] T. Pejsa, M. Gleicher, and B. Mutlu. Who, me? how virtual agents can shape conversational footing in virtual reality. In *Intelligent Virtual Agents: 17th International Conference, IVA 2017, Stockholm, Sweden, August 27-30, 2017, Proceedings 17*, pp. 347–359. Springer, 2017.

[25] J. Poppenk, G. Walia, A. McIntosh, M. Joanisse, D. Klein, and S. Köhler. Why is the meaning of a sentence better remembered than its form? an fmri study on the role of novelty-encoding processes. *Hippocampus*, 18(9):909–918, 2008.

[26] C. Qu, W.-P. Brinkman, Y. Ling, P. Wiggers, and I. Heynderickx. Human perception of a conversational virtual human: an empirical study on the effect of emotion and culture. *Virtual Reality*, 17:307–321, 2013.

[27] L. Quandt. Teaching asl signs using signing avatars and immersive learning in virtual reality. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20. Association for Computing Machinery, New York, NY, USA, 2020.

[28] A. Radford and K. Narasimhan. Improving language understanding by generative pre-training. 2018.

[29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019.

[30] A. Severyn and A. Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, p. 373–382. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2766462.2767738

[31] M. Slater and M. Sanchez-Vives. Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3, 12 2016. doi: 10.3389/frobt.2016.00074

[32] H. J. Smith and M. Neff. Communication behavior in embodied virtual reality. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–12, 2018.

[33] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 196–205. Association for Computational Linguistics, Denver, Colorado, May–June 2015. doi: 10.3115/v1/N15-1020

[34] J. Stuart, K. Aul, M. D. Bumbach, A. Stephen, A. G. de Siqueira, and B. Lok. The effect of virtual humans making verbal communication mistakes on learners' perspectives of their credibility, reliability, and trustworthiness. In *2022 IEEE Conference on Virtual Reality and 3D*

User Interfaces (VR), pp. 455–463, 2022. doi: 10.1109/VR51125.2022.00065

[35] J. Stuart, K. Aul, A. Stephen, M. Bumbach, and B. Lok. The effect of virtual human rendering style on user perceptions of visual cues. *Frontiers in Virtual Reality*, 3, 05 2022. doi: 10.3389/frvir.2022.864676

[36] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, p. 3104–3112. MIT Press, Cambridge, MA, USA, 2014.

[37] A. Tang, F. Biocca, and L. Lim. Comparing differences in presence during social interaction in augmented reality versus virtual reality environments: An exploratory study. 2004.

[38] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[40] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator, 2015.

[41] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1643–1652, 2018. doi: 10.1109/TVCG.2018.2794629

[42] E. Wolf, D. Mal, V. Frohnapfel, N. Döllinger, S. Wenninger, M. Botsch, M. E. Latoschik, and C. Wienrich. Plausibility and perception of personalized virtual humans between virtual and augmented reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 489–498, 2022. doi: 10.1109/ISMAR55827.2022.00065

[43] R. Yan, Y. Song, and H. Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, p. 55–64. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2911451.2911542

[44] C. You, R. Ghosh, A. Maxim, J. Stuart, E. Cooks, and B. Lok. How does a virtual human earn your trust? guidelines to improve willingness to self-disclose to intelligent virtual agents. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3514197.3549686

[45] M. Zalake, A. G. de Siqueira, K. Vaddiparti, P. Antonenko, and B. Lok. Towards understanding how virtual human's verbal persuasion strategies influence user intentions to perform health behavior. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, IVA '21, p. 216–223. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3472306.3478345

[46] M. Zalake, A. G. de Siqueira, K. Vaddiparti, and B. Lok. The effects of virtual human's verbal persuasion strategies on user intention and behavior. *International Journal of Human-Computer Studies*, 156:102708, 2021. doi: 10.1016/j.ijhcs.2021.102708

[47] M. Zalake, J. Woodward, A. Kapoor, and B. Lok. Assessing the impact of virtual human's appearance on users' trust levels. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, p. 329–330. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3267851.3267863

[48] X. Zhou, L. Li, D. Dong, Y. Liu, Y. Chen, W. X. Zhao, D. Yu, and H. Wu. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1118–1127. Association for Computational Linguistics, Melbourne, Australia, July 2018. doi: 10.18653/v1/P18-1103