Mitigating Response Delays in Free-Form Conversations with LLM-powered Intelligent Virtual Agents

Mykola Maslych
University of Central Florida
Orlando, Florida, USA
mykola.maslych@ucf.edu

Mohammadreza Katebi
NeuroVulpis
Orlando, Florida, USA
m.r.katebii@gmail.com

Christopher Lee
Virginia Tech
Blacksburg, Virginia, USA
chrislee24@vt.edu

Yahya Hmaiti
University of Central Florida
Orlando, Florida, USA
Yohan.Hmaiti@ucf.edu

Amirpouya Ghasemaghaei
University of Central Florida
Orlando, Florida, USA
aghaei.ap@gmail.com

Christian Pumarada
University of Central Florida
Orlando, Florida, USA
cpuma@ucf.edu

Janneese Palmer
University of Central Florida
Orlando, Florida, USA
ja448612@ucf.edu

Esteban Segarra Martinez
University of Central Florida
Orlando, Florida, USA
esteban.segarra@ucf.edu

Marco Emporio
University of Verona
Verona, Italy
marco.emporio@univr.it

Warren Snipes
University of Central Florida
Orlando, Florida, USA
warren.snipes@ucf.edu

Ryan P. McMahan
Virginia Tech
Blacksburg, Virginia, USA
rpm@vt.edu

Joseph J. LaViola Jr.
University of Central Florida
Orlando, Florida, USA
jlaviola@ucf.edu

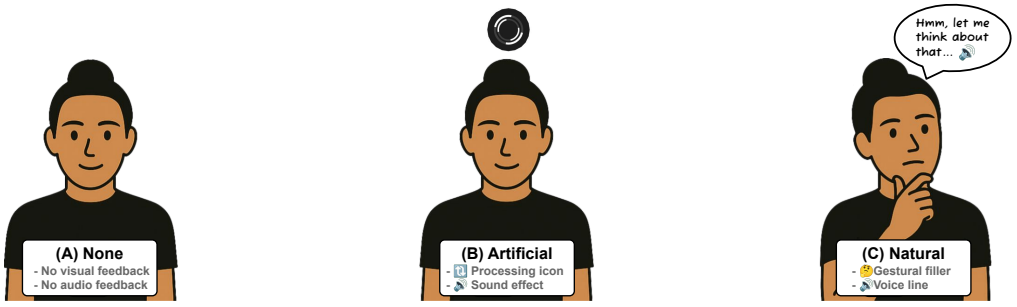

Figure 1: In this paper, we investigated response delays with intelligent virtual agents and the effects of conversational fillers: (A) *None*: The user receives no feedback while waiting on the agent; (B) *Artificial*: The user receives a processing icon and sound effect; and (C) *Natural*: The user receives feedback in the form of social interaction cues (gesture and voice).

ABSTRACT

We investigated the challenges of mitigating response delays in free-form conversations with virtual agents powered by Large Language Models (LLMs) within Virtual Reality (VR). For this, we used conversational fillers, such as gestures and verbal cues, to bridge delays between user input and system responses and evaluate their effectiveness across various latency levels and interaction scenarios. We found that latency above 4 seconds degrades quality of experience, while natural conversational fillers improve perceived response time, especially in high-delay conditions. Our findings provide insights for practitioners and researchers to optimize user engagement whenever conversational systems' responses are delayed by network limitations or slow hardware. We also contribute an open-source pipeline that streamlines deploying conversational agents in virtual environments.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; **Natural language interfaces**; **Empirical studies in HCI**.

KEYWORDS

Conversational interfaces, LLM, VR, user study, response latency.

CUI '25, July 8–10, 2025, Waterloo, ON, Canada

ACM Reference Format:
Mykola Maslych, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, Warren Snipes, Ryan P. McMahan, and Joseph J. LaViola Jr.. 2025. Mitigating Response Delays in Free-Form Conversations with LLM-powered Intelligent Virtual Agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25), July 8–10, 2025, Waterloo, ON, Canada.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3719160.3736636

1 INTRODUCTION

Advancements in artificial intelligence (AI), particularly in large language models (LLMs), automatic speech recognition (ASR) and text-to-speech (TTS), catalyzed the integration of intelligent virtual agents (IVAs) into mainstream systems [84, 97, 98]. These advancements make IVAs more accessible and capable than ever before, evolving beyond simple scripted responses. They offer personalized interactions that adapt in real-time, recall past conversations, and provide a truly realistic and engaging user experience (UX). The usefulness of these intelligent agents extends beyond entertainment [15, 76, 97] to encompass medical [1, 18, 62, 80, 87], educational [24, 37, 77, 93, 98], personal [3, 104, 108] and social [79, 84] applications.

More recent applications of IVAs emphasize the support of free-from conversation. To achieve this, most modern architectures for IVAs rely on TTS to vocalize LLM-generated responses to speech transcribed with ASR [7, 15, 31, 37, 62, 66, 84, 86, 97, 98, 104, 108]. These components are computationally demanding, and when running locally, they compete for system resources with graphics and physics simulations [72], leading to system response time (SRT) that exceeds the natural duration of silence between interlocutors' turns. To alleviate this, most prior architectures offload processing to cloud-based models. However, cloud solutions are inherently unpredictable, prone to network congestion, connection instability, and system downtime, all of which exacerbate latency. Whether due to hardware limitations, cloud infrastructure, or local system contention, conversational systems' response latency remains a significant challenge, degrading Quality of Experience (QoE) [57, 65, 81].

Long response times from virtual humans, robots, and voice assistants cause impatience [73, 105], frustration [53, 69, 105], and general dissatisfaction [12, 30, 41]. However, findings on response latency and its mitigation in conversations with virtual agents relied on scripted questions and answers [12, 49, 53], used WoZ protocol [29, 83, 101], or rated pre-recorded conversations [64, 74], limiting their generalization to real-time interactions. Addressing this, our work is the first to examine the effects of response latency with *truly interactive* conversational agents powered by LLMs to support free-form conversations, and we aim to answer the following research questions: **(1)** *Do prior findings on perception of latency remain consistent when using free-form conversational IVAs?* **(2)** *Do natural conversational fillers mitigate the negative effects of latency when interacting with free-form IVAs?* **(3)** *Do artificial wait indicators mitigate the negative effects of latency when interacting with free-form IVAs?*

We investigated the effects of response latency and conversational fillers on perceived response time and broader dimensions of user experience in free-form task-guided voice conversations with LLM-based IVAs in Virtual Reality (VR). Participants experienced three virtual worlds, conversing with a total of 9 virtual agents under three filler types: *None*, *Artificial*, and *Natural* (see Figure 1), along with three response latency levels: *Low (1.5s)*, *Medium (4.0s)*, and *High (6.5s)* (see subsubsection 3.1.1). We used an LLM and TTS to generate responses to users' speech transcribed with ASR in real-time. Latency was simulated by delaying the agents' response onset to study the effect of different conversational fillers and response latencies on users' perception of the agents.

We found that delay significantly worsened participants' perceived response time and broader perception metrics, and was less bearable beyond 4 seconds. *Natural* conversational fillers mitigated some of these effects, however *Artificial* wait indicators did not significantly affect user experience. To streamline voice-based conversational agent studies, we provide an open-source library[1], which includes a system that uses natural language processing, capable of real-time analysis of conversation history to identify and manage task transitions seamlessly. Overall, our contributions inform the design of IVAs and embodied conversational agents (ECAs) applicable to inherently virtual, as well as physical agents, digital twins of which can be rendered in VR.

2 RELATED WORK

Latency in conversations (response delay) refers to the delay between one speaker's utterance and the other's response. In human communication, brief delays and silences are natural [20, 48], yet research suggests that response times beyond two seconds begin to feel unnatural, and silences exceeding four seconds can disrupt conversational flow, signaling a breakdown in communication [70]. Studies on turn-taking in vocal conversations report sub-second latency on average [38, 50, 88], and show that faster response times signal greater social connection [92]. They also show that in natural dialogues, some responses begin even before the previous speaker has finished [38, 58, 88]. In conversational user interfaces, fast, seamless interactions improve engagement, while excessive delays can cause frustration, disrupt the conversational flow, and reduce trust in the system. This underscores the importance of minimizing response latency in human-computer interactions.

When designing virtual agents that rely on ASR, LLMs, and TTS systems, latency poses an even greater challenge due to computational constraints. While techniques like incremental response generation [85, 94] exist, they are not always viable — a useful response often depends on the information contained at the end of the user's turn, which has not been fully processed before the agent starts responding. Considering delays caused by external factors unrelated to the process of response generation, we focus on latency mitigation through UI-based strategies, drawing from research in domains where delays are an inherent constraint: web and mobile user interfaces, text-based chat interfaces, and embodied human-agent (HAI) and human-robot interactions (HRI). In the following subsections, we outline the strategies used to reduce the perceived delay across various interface types.

[1] github.com/ISUE/iva-cui

2.1 Latency Mitigation in 2D Interfaces

Overestimation of wait times by users is a persistent UI design concern [42], so traditional research on delays in human-computer interfaces recommends always displaying a visual progress indicator to show the system is processing a task [71].

In web browser applications, user satisfaction significantly decreases when response times exceed 12 seconds, often leading users to abandon the application [43]. To address this, some research focused on progress bars behaviors, which significantly impact users' perceived wait times on websites [13, 35], although results regarding the optimal progress indicator speed are mixed. Some work shows that indicators that start fast and end slow induce encouragement, reducing abandonment rates [21, 51, 96] and increasing satisfaction rate [60]. Faster initial countdown intervals [52] and animated ribbing moving backward while decelerating [36] can reduce perceived wait duration. Conversely, other work found that users are more tolerant of initial negative progress with an observed preference for linear progress bars, which lead to shorter perceived wait times [2, 13, 35]. Furthermore, interactive animations substantially reduce perceived wait times by distracting user attention, making wait times feel shorter, and increasing user satisfaction in contrast with passive animations and standard progress bars [39].

In mobile applications, interactive and color-changing loading screens improve user satisfaction by shrinking perceived wait times compared to passive ones [17]. A study comparing bar and cartoon-bar indicators at constant, accelerating, and decelerating speeds found that decelerating progress bars reduced perceived wait time, and cartoon-bars increased user acceptance and satisfaction [61].

2.2 Latency in Text-based Chat Interfaces

Much research on human-to-human text chats and instant messaging (IM) has explored response latency. Users prefer seeing a typing indication (speech bubble or live typing) over nothing while waiting for a reply [47]. However, under time pressure, users rated their partners as less involved, reporting frustration even in the presence of a typing indicator [46]. Interestingly, third-party observers were more forgiving of delayed responses in support chats, as long as the conversation felt contingent [59]. While typos are common in human messages, chatbots were rated as less human when their replies contained typos in a WoZ study [100].

In text chatbot interfaces, users prefer short delay over long delay or zero delay, because they perceive it as more realistic [4, 32, 40, 82], and presence of a typing indicator during response generation increased the feeling of social presence for novice users [33]. Inspired by chatbot interfaces familiar to users, LLM-powered text chat interfaces adopted similar approaches to mitigate response delays, accompanied by the live appearance of response text as it is being generated. Systems like ChatGPT[2], Claude[3], Gemini[4], MetaAI[5], and Perplexity AI[6] employ distinct loading feedback mechanisms to enhance user experience: ChatGPT uses a pulsating black dot; Claude employs a pulsing star-like object; Gemini utilizes progress indicators including a rotating star and sequential progress bars; Perplexity AI uses a rotating ellipsis resembling pages of a book; MetaAI integrates a spinning wheel. These techniques inform the user about the ongoing process of response generation, potentially reducing the perceived wait time.

[2] chatgpt.com (Accessed May 6, 2025)
[3] anthropic.com/claude (Accessed May 6, 2025)
[4] gemini.google.com (Accessed May 6, 2025)
[5] meta.ai (Accessed May 6, 2025)
[6] perplexity.ai (Accessed May 6, 2025)

2.3 Latency Mitigation for Embodied Agents

An embodied agent interacts with its physical/virtual environment through physical/virtual body and, unlike text-based chat interfaces, must manage both verbal and visual cues to maintain engagement despite response delays. Excessively long latencies negatively impact UX and users' perception of virtual agents and robots, often leading to user discomfort [8, 12, 30, 41] and assumptions that an error has occurred [30]. Wait times of 4.5 to 5 seconds can also be thought of as negative responses [78]. During task-solving with robots, delays lead to increased frustration and anger, and decreased satisfaction and future use intention [105]. Additionally, quicker rather than delayed responses from a robot receptionist lead to increased tolerance and reported interaction quality in users [73].

Research on mitigating response delays has identified conversational fillers [49, 63, 64, 83] common in everyday speech as a promising direction. These fillers (e.g. 'uh...' and 'uhm...') serve several para-linguistic functions [9], including turn-taking management [89], social approval [19], and co-creation of pragmatic and discourse [90], making them essential for smooth and successful natural conversations [11, 91]. While simple fillers can reduce a robot's perceived intelligence and likability [49], more complex ones mitigate delay effects without harming perceptions of intelligence [101] or virtual agent's competence [53]. These complex fillers include *pensive fillers* (e.g., 'let me think' paired with gestures like chin-scratching) and *acknowledgment fillers* (e.g., 'aha' with rapid head-nodding). Systems using such fillers — whether generic ("uh") or context-aware ("I've found a flight for you") – were rated as more appropriate than silent systems, even with equal delay [64].

While the literature is clear on the negative effects of latency on UX, studies in perception of conversational agents relied on predetermined sets of questions and responses [12, 49, 53], followed WoZ protocol [29, 83, 101], or rated pre-recorded conversations [64, 74]. This limits their generalization to cases where virtual agents are truly interactive, afforded by recent advancements in the speed and quality of speech and text processing models. Addressing this, we explored how latency affects the perceived virtual agent responsiveness and whether its negative effects can be mitigated by conversation fillers through a study where agents responded to any free-form queries in real-time across multiple scenarios.

3 METHODOLOGY

This study involved immersive VR scenarios where participants interacted with virtual agents using their speech, under varied response delays and multiple delay mitigation strategies. We selected VR as a medium for our study to make it easily reproducible [75] and to reduce external factors and distractions [54], focusing on variables of interest [22]. We expect research on embodied conversational agents to continue expanding, given the rise of VR social applications where users interact with embodied virtual avatars [67], and indications that immersive VR induces greater

social influence from virtual characters compared to standard desktop applications [5, 55]. The remainder of this section outlines key components of the experiment design, conditions, system implementation, apparatus, participants, and the collected data.

3.1 Experiment Design

The experiment included three delay levels: *Low (1.5s)*, *Medium (4.0s)*, *High (6.5s)*, and three latency mitigation types: *None*, *Artificial Wait Indicator*, *Natural Conversational Filler*, leading to a total of 9 conditions. This matched the number of IVAs present in our study. The order of IVAs was the same across all participants, and the order of conditions was counter-balanced using Balanced Latin Square (18 orders). This way, the conditions applied to IVAs varied across participants.

3.1.1 Delay Levels. Our design included three delay levels: (1) *Fast* at 1.5 seconds, (2) *Medium* at 4.0 seconds, (3) *Slow* at 6.5 seconds. Under *Low (1.5s)* latency, responses were played as soon as they were generated (SRT average). The longer delays increased in 2.5-second steps: *Medium (4.0s)* matched the comfortable silence threshold reported in prior work, while *High (6.5s)* exceeded it (see section 2).

3.1.2 Filler Types. For latency mitigation, we used three filler types: (1) *None* with no filler; (2) *Artificial Wait Indicator (WI)* with a loading icon and sound; and (3) *Natural Conversational Filler (CF)* with a thinking gesture and a voice line (see Figure 1). Both *Artificial* and *Natural* fillers combined visual and auditory cues. We treated these as unified conditions, as prior work showed that multimodal fillers work outperform unimodal ones (see subsection 2.3). Fillers appeared during the wait period (varying by delay level) between the end of participant speech and the agent's response.

In the *None* condition (baseline), agents provided no visual or auditory cues during the delay. They remained in the 'attentive idle' animation, facing the participant, and began speaking once the delay elapsed.

In the *Artificial* condition, a rotating visual indicator (concentric quarter-circles) appeared above the agent's head, accompanied by a continuous processing sound (similar to the ChatGPT mobile app). The agents stayed in the 'attentive idle' state throughout. Both the icon and sound ended when the delay expired.

In the *Natural* condition, each agent randomly selected from three 'thinking' gestures and six filler voice lines at runtime. To simulate deliberation, agents turned their head, touched their chin or the back of their head, and said a conversational filler. They held this pose with subtle breathing motion until the delay ended, then returned to 'attentive idle' before responding. Fillers were inspired by prior work [53, 101] and included: "Hmm, let's see...", "Okay, hmm...", "Uhmmm...", "Ah...", "Hmm, one moment...", "Hmmm...".

3.1.3 Hypotheses. We derived six hypotheses from prior work: (**H1a**, **H1b**) on the negative effects of high latency on user experience (section 2); (**H2a**, **H2b**) on the benefits of conversational fillers in mitigating perceived delay (subsection 2.3); and (**H3a**, **H3b**) on the role of loading indicators in improving perceived latency in classical UIs (subsection 2.1). These hypotheses guided our investigation into how latency and conversational fillers affect the perception of embodied IVAs responding to free-form queries in immersive VR:

- **H1a:** Latency degrades perceived response time of conversational agents.
- **H1b:** Latency degrades broader perception dimensions of conversational agents.
- **H2a:** Under latency, *Natural* conversational fillers improve perceived response time of conversational agents.
- **H2b:** Under latency, *Natural* conversational fillers improve broader perception dimensions of conversational agents.
- **H3a:** Under latency, *Artificial* wait indicators improve perceived response time of conversational agents.
- **H3b:** Under latency, *Artificial* wait indicators improve broader perception dimensions of conversational agents.

3.2 Study Questionnaires

To test our hypotheses and gather additional insights into user perception, we created two custom questionnaires informed by prior literature, as no standardized survey exists for interface latency and its mitigation. The first was administered after each condition, and the second after completing all experimental conditions.

3.2.1 Post-condition Questions. After each agent interaction, participants completed an in-VR survey by selecting from five labeled response options: *Agree*, *Somewhat Agree*, *Neutral*, *Somewhat Disagree*, and *Disagree*. Q1 assessed perceived response latency and included the word "meaningfully" to avoid bias toward *Natural* fillers that included speech. Q2–Q6 measured other perception dimensions, drawn from prior literature [14, 49, 53, 64] and our study objectives. Q4 and Q5 were adapted from the Robotic Social Attributes Scale (RoSAS) [14], corresponding to discomfort and competence dimensions.

- **(Q1)** **Response Time**: From the moment I stopped talking, the agent was quick to start responding meaningfully.
- **(Q2)** **Engagement**: I felt absorbed during my interaction with this virtual agent.
- **(Q3)** **Good Impression**: The virtual agent left a good impression on me.
- **(Q4)** **Discomfort**: I felt awkward, scared, and strange when talking to this agent.
- **(Q5)** **Competence**: This agent was reliable, competent, and interactive.
- **(Q6)** **Willingness to Interact Again**: I would be willing to interact and spend time with this virtual agent again.

3.2.2 Post-study Questions. After completing the VR experience, participants filled out a post-study survey (Table 1). The first six questions assessed their overall impressions of the agents and whether they noticed the study conditions (i.e., filler types). Participants then watched a 40-second video showing a single agent performing each of the three fillers at *Medium (4.0s)* to clarify differences in case they had gone unnoticed. All participants viewed the same video to ensure a consistent comparison baseline. This was shown *after* the initial questions to avoid bias. The final four questions focused on perceptions of the filler types and depended on participants' awareness of them. While PSQ10 resembled PSQ4 and PSQ5 (future use intent), it was placed after the video to reflect informed responses.

3.3 Interaction Scenarios

Participants interacted with nine virtual agents across three environments (Figure 2) corresponding to distinct scenarios: **Store**, **Hotel**, **Museum**. Each condition targeted at least five conversational

Table 1: Questions ordered as they appeared in the post-VR survey. Rank-order question order was randomly initialized. ⋆: text entry answer justification required; ⋆⋆: optional.

Label	Question	Response anchors
PSQ1	Overall, which agent did you like the most? ⋆	Each of 9 agents
PSQ2	Overall, which agent did you like the least? ⋆	Each of 9 agents
PSQ3	I felt like the agents understood me. ⋆	Agree → Disagree
PSQ4	I would use a system where agents replied fastest. ⋆	Agree → Disagree
PSQ5	I would use a system where agents replied slowest. ⋆	Agree → Disagree
PSQ6	Did you notice any conversational fillers that different agents had? ⋆⋆	Yes, Maybe, No
	Video explanation of the three filler types.	
PSQ7	Rank the conversational fillers that you saw in terms of your preference for future applications.	Natural, Artificial, None
PSQ8	For **Natural** fillers, were *gestures* or *voice lines* more helpful to fill conversations? ⋆⋆	Gestures, Voice lines, Same
PSQ9	For **Artificial** fillers, were *wait indicators* or *sound effects* more helpful to fill conversations? ⋆⋆	Indicators, Sounds, Same
PSQ10	I would use a system where agents replied slowest with conversational fillers. ⋆	Agree → Disagree

Figure 2: Intelligent Embodied Virtual Agents in their corresponding virtual environments.

turns to ensure the study conditions were noticeable [12], influencing participants' survey responses. To maintain engagement, the scenarios were designed with a gamified structure (similar to video game quests) and clear objectives, encouraging free-form conversations. Agents' responses were generated at runtime using LLM prompts (see project repository) and user queries, and supported small-talk (e.g., greetings, environment awareness). When queried with off-topic or role-breaking input, agents naturally redirected the conversation back to the scenario context. Participants were free to explore each environment and speak with agents, guided by an on-hand UI that updated after the transition-check system (Figure 3) detected relevant dialogue context in the message history.

In the **Store** scenario, participants retrieved a shirt from the *Friend* agent and returned it to a store across the street. There, the *Clerk* directed them to the *Manager* for approval, as he was still in training. In the **Hotel** scenario, participants checked in with the *Receptionist*, then visited their room where *Maintenance* worker informed them it wasn't ready. The worker apologized and redirected them back to the receptionist to receive a complimentary dinner voucher. At the restaurant, the *Waiter* asked about food preferences, dietary restrictions, and the voucher. In the **Museum** scenario, participants played a student working on a school assignment about human rights. They obtained a ticket from the *Host* agent, then

spoke with *Volunteer 1* about the Cyrus Cylinder[7] and *Volunteer 2* about the U.S. civil rights movement.

3.3.1 Intelligent Agents' Avatars. Agents were represented using avatars that matched the visual fidelity of our virtual environments (VEs). Figure 2 shows the agents in context, with race and gender distributions reflecting our university's student demographics. While we initially considered Rocketbox avatars [34], we selected avatars from the VALID library [25] due to their validation through user perception studies, ensuring accurate and representative character design. Agent outfits aligned with scenario roles and the visual style of each VE. Avatar folder names and corresponding Edge-TTS voices are listed in Table 2.

Table 2: Scenarios, roles, avatars, and abbreviated TTS voices used for avatars in our experiment. Full Edge-TTS voice identifiers are available in the source code repository.

Scenario	Role	VALID Avatar	Voice
Store	Friend	`Black_F_2_Casual`	Ava
Store	Clerk	`White_M_3_Casual`	Andrew
Store	Manager	`Hispanic_F_1_Busi`	Aria
Hotel	Receptionist	`Hispanic_F_2_Casual`	Michelle
Hotel	Maintenance	`Hispanic_M_2_Util`	Guy
Hotel	Waiter	`MENA_M_2_Casual`	Brian
Museum	Host	`White_F_2_Busi`	Emma
Museum	Volunteer 1	`White_M_1_Casual`	Florian
Museum	Volunteer 2	`Asian_F_1_Casual`	Yan

3.3.2 Agent Animations. Each agent performed a scenario-specific 'busy' animation until first addressed by the user. For example, *Waiter* and *Host* interacted with small displays, *Maintenance* worker manipulated wires on an electrical panel, and the *Receptionist* and *Manager* alternated between typing and looking at a monitor. Upon user interaction, agents transitioned to an 'attentive idle' state (passive listening pose with subtle breathing). While users spoke, agents turned their heads toward them. Interactions were only possible within a defined proximity; agents would look away and return to their 'busy' animation once the user exited this area.

3.4 Conversational System Implementation

Our system (see Figure 3) was implemented as a Unity application connected via HTTP requests to a local server that served the Automatic Speech Recognition (ASR) and LLM engines. When a user spoke within the agent's defined area, the user's voice was transcribed with the FasterWhisper[8] medium model, then added to the message history of that agent, and then sent to the locally-hosted Llama3.1-8b-Q5 LLM[9] to generate a response from the agent's perspective (ran on an RTX4090 GPU on Pop!_OS22[10]). This text was then passed to the Edge-TTS API[11], which generated an audio file, storing it on our local server, and returning a static download link to Unity. The audio was then downloaded by the application and played from an audio source on the agent, with OVR Lipsync[12] used to animate the agent's mouth movements. The average system response time (SRT) was approximately 1.5 seconds ($\mu = 1.47$, $\sigma = .23$), constrained by the processing time of the locally hosted ASR, LLM, and TTS components (see Figure 4 for a breakdown). This value defined the *Low (1.5s)* condition in our experiment.

3.5 Apparatus

We used a Meta Quest Pro HMD, with a resolution of 1800×1920 pixels per eye and a FOV of $106^\circ \times 96^\circ$ connected to a PC running the Unity 2022 LTS application. We developed the VR interactions using the XR Interaction Toolkit package[13]. Our application ran at constant 70 frames per second (FPS). Participants wore sanitized earphones, and navigated the VE using the Quest Touch Pro controllers. The thumbsticks controlled movement and turning, the back trigger selected in-VR survey responses, the grip button selected objects within the VE, and the 'A' button on the right controller activated the microphone.

3.6 Participants

We used G*Power [28] to estimate a minimum required sample size of 33 participants, assuming a medium-to-large effect size (0.3), a within-subjects design, and repeated measures ANOVA with 9 measurements. To increase statistical power and accommodate counterbalancing, our final participant pool included 54 university participants (three full rotations of the 18-order Latin square), comprising 23 females (43%) and 31 males (57%), aged 18-56 ($\bar{x} = 22.07$, $s = 6.55$). Participants self-reported varying experiences:

- **VR use:** Daily: 2, Weekly: 6, Monthly: 11, Yearly: 18, Never: 17
- **Gaming:** Daily: 16, Weekly: 16, Monthly: 12, Yearly: 7, Never: 3
- **Social VR[14]:** Daily: 0, Weekly: 2, Monthly: 13, Yearly: 8, Never: 31

All participants could read and speak English, wore an HMD for 35 minutes while seated, used both hands for controllers, and had normal or corrected vision. Those with glasses or contact lenses kept them on during the study. No participants reported color blindness, neurological conditions, or physical disabilities.

3.7 Procedure

Participants arrived at the study location and were screened for eligibility. After confirming eligibility, the consent process was administered, including answering any participant questions. Participants were then asked to complete a demographics survey electronically. Following this, participants were assisted in wearing the VR headset while seated. We adjusted the participants' seated in-VR height to their real-world standing height to make their conversations with the agents feel more natural. Before starting each scenario, we presented participants with a brief introduction, outlining the scenario's theme without revealing the outcome. In each scenario, participants had to navigate the environment and engage in conversations with three agents, totaling nine virtual agents across all

[7] en.wikipedia.org/wiki/Cyrus_Cylinder (Accessed May 6, 2025)
[8] github.com/SYSTRAN/faster-whisper (Accessed May 6, 2025)
[9] huggingface.co/bullerwins/Meta-Llama-3.1-8B-Instruct-GGUF Llama 3.1 achieves state of the art performance on a number of benchmarks at the time of our research (Accessed May 6, 2025)
[10] pop.system76.com (Accessed May 6, 2025)
[11] github.com/rany2/edge-tts (Accessed May 6, 2025)
[12] developers.meta.com/horizon/documentation/unity/audio-ovrlipsync-unity/ (Accessed May 6, 2025)
[13] Unity Docs | XR Interaction Toolkit (Accessed May 6, 2025)
[14] VRChat (vrchat.com), Meta Horizon (horizon.meta.com) (Accessed May 6, 2025)

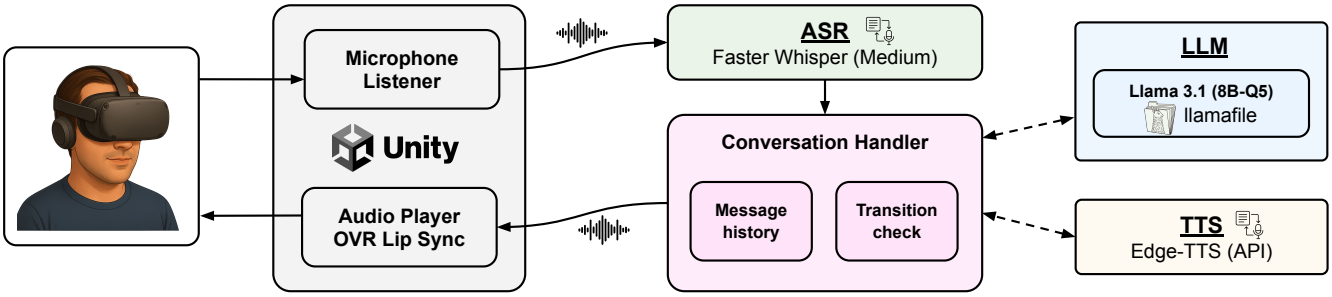

Figure 3: System architecture: speech is recorded in Unity, passed to the ASR model, combined with the message history, passed to an LLM, the generated response from which updates message history and checks for transitions. Generated text is then passed to TTS, and the generated voice is played from an audio source that controls OVR lip-sync in Unity.

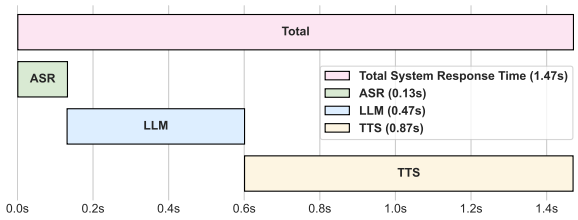

Figure 4: Breakdown of latencies achieved in our system. Latency is the time between the user finishing their microphone input (by pressing the 'A' button) and when the agent starts responding with the LLM-generated text in voice.

scenarios (see subsection 3.1). To facilitate the exploring of the environment, participants had access to an on-hand task list attached to their left hand. This list only gave participants an overview of their active task without providing hints about the conversation flow. Once completed, a new task appeared, while the previous task remained as strike-through. Upon completing the 3 in-VR scenarios, participants removed the HMD and filled out a web-based post-study survey. We paid participants $10 and thanked them for their time.

4 RESULTS

4.1 Post-condition Responses

Since our data consists of Likert-scale responses, the Aligned Rank Transform (ART) [102] was used for a 3×3 full-factorial repeated-measures ANOVA. This analysis examined main and interaction effects across three delay levels (*Low (1.5s)*, *Medium (4.0s)*, and *High (6.5s)*) and three filler types (*None*, *Artificial Wait Indicator*, and *Natural Conversational Filler*). Post-hoc ART-C tests [27] were conducted to test the hypotheses, with p-values adjusted for 36 comparisons using the Holm-Bonferroni correction. Responses were coded from −2 (Disagree) to 2 (Agree), and error bars in Figure 5 and Figure 6 show ±1 SEM based on this scale.

Delay had a significant main effect on (Q1) Response Time, (Q2) Engagement, (Q3) Good Impression, (Q5) Competence, and (Q6) Willingness to Interact Again (all $p < 0.0001$). It also significantly affected (Q4) Discomfort ($p < 0.05$). The main effect of fillers was

Table 3: Repeated-Measures ANOVA on ART-Transformed Data. η_p^2 is the effect size of a specific independent variable on the dependent variable (metric measured by the question), while controlling for effects of other independent variables.

Question	Factor	*Df*	*F*	*p*	*Sig.*	η_p^2
(Q1) Response Time	Delay	(2, 424)	154.55	$p < 0.0001$	****	0.422
	Filler	(2, 424)	18.61	$p < 0.0001$	****	0.081
	Delay x Filler	(4, 424)	4.66	$p < 0.01$	**	0.042
(Q2) Engagement	Delay	(2, 424)	25.33	$p < 0.0001$	****	0.107
	Filler	(2, 424)	4.68	$p < 0.01$	**	0.022
	Delay x Filler	(4, 424)	1.08	$p = 0.365$		0.010
(Q3) Good Impression	Delay	(2, 424)	30.92	$p < 0.0001$	****	0.127
	Filler	(2, 424)	5.03	$p < 0.01$	**	0.023
	Delay x Filler	(4, 424)	0.36	$p = 0.834$		0.003
(Q4) Discomfort	Delay	(2, 424)	4.65	$p < 0.05$	*	0.021
	Filler	(2, 424)	1.28	$p = 0.280$		0.006
	Delay x Filler	(4, 424)	1.55	$p = 0.187$		0.014
(Q5) Competence	Delay	(2, 424)	14.74	$p < 0.0001$	****	0.065
	Filler	(2, 424)	3.74	$p < 0.05$	*	0.017
	Delay x Filler	(4, 424)	1.18	$p = 0.320$		0.011
(Q6) Willingness to Interact Again	Delay	(2, 424)	16.93	$p < 0.0001$	****	0.074
	Filler	(2, 424)	2.26	$p = 0.106$		0.011
	Delay x Filler	(4, 424)	0.42	$p = 0.791$		0.004

significant for (Q1) Response Time ($p < 0.0001$), (Q2) Engagement ($p < 0.01$), (Q3) Good Impression ($p < 0.01$), and (Q5) Competence ($p < 0.05$). Interactions between latency levels and Fillers were significant only for (Q1) Response Time ($p < 0.01$).

Post-hoc pairwise comparisons between delay levels in the absence of fillers (Figure 5) revealed significant differences in (Q1) Response Time across all conditions ($p < 0.0001$), supporting **H1a**. For other metrics, participants' responses significantly differed between *Low (1.5s)* and *High (6.5s)* delay levels on (Q2) Engagement ($p < 0.0001$), (Q3) Good Impression ($p < 0.01$), (Q4) Discomfort ($p < 0.01$), (Q5) Competence ($p < 0.01$), and (Q6) Willingness to Interact Again ($p < 0.001$), supporting **H1b**. Additionally, (Q2) Engagement differed significantly between *Medium (4.0s)* and *High (6.5s)* delays ($p < 0.05$).

Pairwise comparisons between filler types at fixed delay levels (Figure 6) showed significant differences between *Natural* and *None* fillers on (Q1) Response Time at *Medium (4.0s)* ($p < 0.01$) and *High (6.5s)* ($p < 0.0001$) delays, supporting **H2a**. *Natural* and *Artificial* fillers also significantly differed at *High (6.5s)* delay on (Q1) Response Time ($p < 0.05$). However, no support was found for **H2b**, **H3a**, or **H3b**. We discuss implications of these results

in subsection 5.1, and a full report of all main effects and pairwise comparisons is available in the supplementary material.

4.2 Post-study Responses

4.2.1 Preference for agents depending on response speed. The first two post-study questions assessed whether participants preferred agents that responded quickly. The frequency of each speed condition was recorded for agent selections. Among the agents participants liked the most (PSQ1), *Low (1.5s)* responses occurred 31 times (57.41%), *Medium (4.0s)* 12 times (22.22%), and *High (6.5s)* 11 times (20.37%), ($\chi^2_2(N = 54) = 14.12, p < 0.0001$). Among the agents participants liked the least (PSQ2), *Low (1.5s)* responses occurred 9 times (16.67%), *Medium (4.0s)* 12 times , and *High (6.5s)* 33 times (61.11%), ($\chi^2_2(N = 54) = 19.0, p < 0.0001$).

4.2.2 Agents' understanding and noticeability of conversational fillers. Overall, 49 participants (90.74%) indicated that the agents understood them through answers to PSQ3 (*Strongly Agree*: 22, *Agree*: 27, *Neither agree nor disagree*: 3, *Disagree*: 2, *Strongly disagree*: 0). A chi-square test revealed a significant difference from a uniform

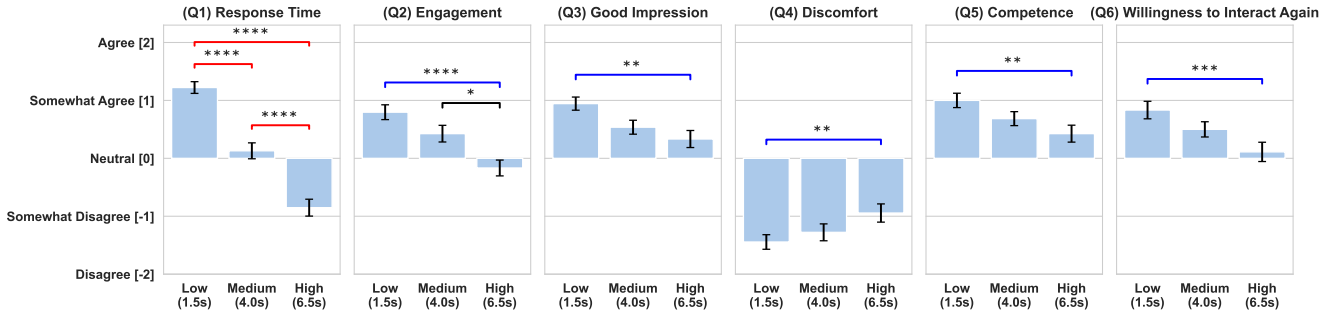

Figure 5: Effect of response delay on user perception of intelligent embodied virtual agents in the *None* filler condition, collected through post-condition survey responses. H1a – Significant effect of delay on (Q1) Response Time; H1b – Significant effect of delay on other perception dimensions. Error bars show 1 SEM (standard error of the mean). Significance annotations [16] from ART-C pairwise comparisons adjusted using Holm-Bonferroni ($^* = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$, $^{****} = p < 0.0001$).

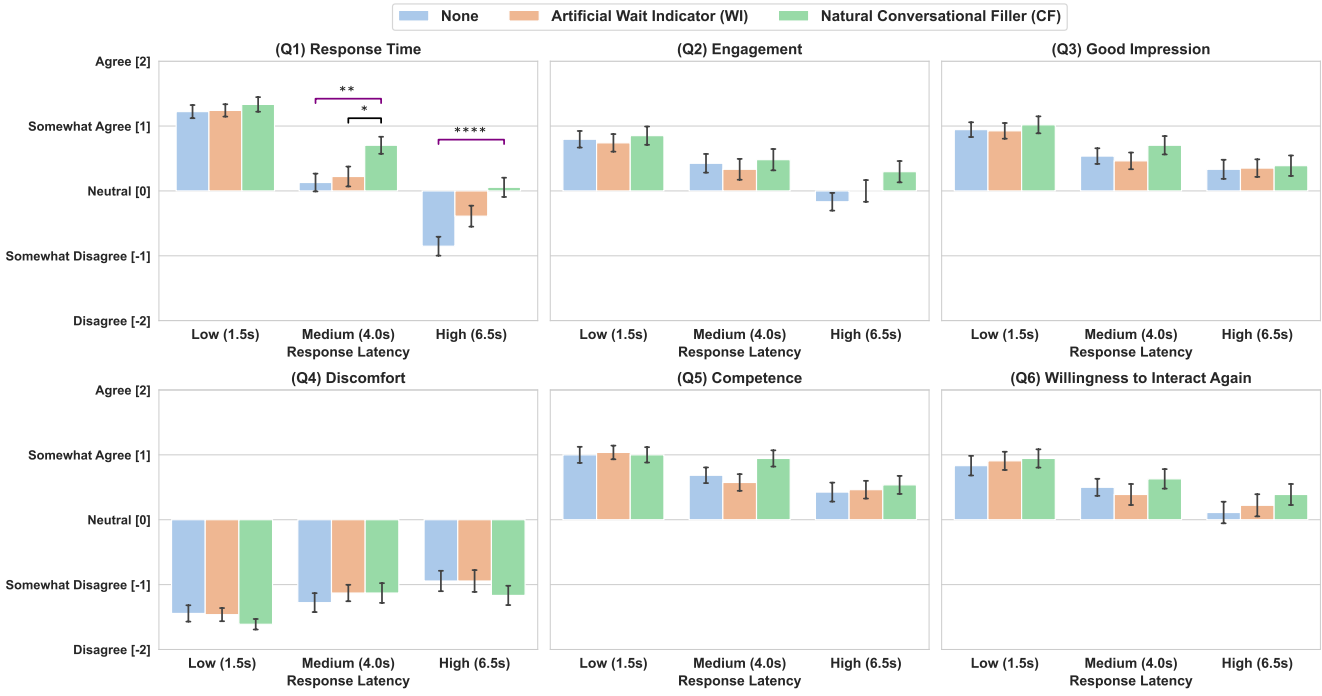

Figure 6: Effect of delay with all filler types on user perception of intelligent embodied virtual agents. H2a – Significant effect of *Natural Conversational Fillers* on (Q1) Response Time; Error bars show 1 SEM (standard error of the mean). Significance annotations [16] from ART-C pairwise comparisons adjusted using Holm-Bonferroni ($^* = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$).

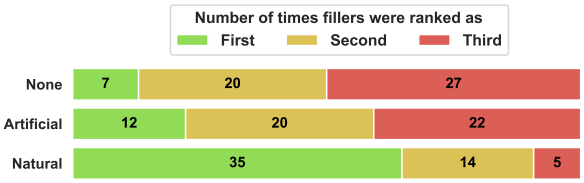

Figure 7: Number of times each filler was ranked in terms of preference of seeing it in future applications.

choice distribution ($\chi^2_4(N = 54) = 36.82, p < 0.0001$). Before being introduced to the filler types through a video, most participants reported noticing them on PSQ6 (*Yes*: 33, *Maybe*: 11, *No*: 10), with non-uniform responses distribution ($\chi^2_2(N = 54) = 18.78, p < 0.0001$).

4.2.3 Preference for filler types in future applications. In PSQ7, Participants ranked the three types of fillers based on their preference for their use in future applications. Figure 7 shows that the majority of participants (35/54, 64.81%) selected *Natural* filler as the most preferred, while 12 participants (22.22%) favored the *Artificial* fillers instead. Seven (12.96%) participants chose *None* (no filler) as their top preference.

4.2.4 Helpfulness of visual and auditory modalities of fillers. For *Natural* fillers (PSQ8), 16 participants (29.70%) found gestures more helpful, 14 (25.93%) chose voice lines, and 24 (44.44%) rated both equally ($\chi^2_2(N = 54) = 3.12, p = 0.21$). For *Artificial* fillers (PSQ9), 19 participants (35.18%) preferred the visual wait indicator, 10 (18.52%) preferred sound effects, and 25 (46.30%) reported that both contributed equally ($\chi^2_2(N = 54) = 6.34, p < 0.05$).

4.2.5 Statements about using the system depending on the speed of agent's responses. In PSQ4, PSQ5 and PSQ10, participants responded to whether they would use the system with agents at different response latencies and with or without conversational fillers (see Figure 8). A chi-square test revealed that the distribution of choices for the use of the fastest agents ($\chi^2_4(N = 54) = 65.81, p < 0.001$), slowest agents ($\chi^2_4(N = 54) = 104.33, p < 0.001$), and slowest agents with conversational fillers ($\chi^2_4(N = 54) = 9.89, p < 0.05$) were not uniform. A comparison of participants' willingness to use the slowest system with and without conversational fillers revealed greater acceptance when fillers were present ($\chi^2_4(N = 54) = 61.56, p < 0.001$).

4.2.6 Analysis of text justifications. To analyze participants' text justifications of their post-study questionnaire answers, we conducted a preliminary inductive analysis using in vivo excerpts and descriptive coding to generate general thematic insights [10], and counted the frequencies of similar statements across participants.

5 DISCUSSION

To our knowledge, our study is the first to vary both response latency and conversational fillers in an experiment where conversational virtual agents were *truly interactive*, leveraging ASR, LLM and TTS to support free-form conversations. Previous research on latency mitigation has largely studied it in isolation, rather than including it as an explicit factor in studies with fully interactive embodied conversational agents. We found that response latency significantly degrades user perceptions of the agents, and that conversational fillers improve perceived response latency in *Medium (4.0s)* and *High (6.5s)* delay levels. These findings align with prior work that investigated conversational fillers in WoZ studies [12, 29, 49, 53, 83, 101] and those that used pre-recorded human-agent conversations [64, 74]. This section addresses our research questions by interpreting participants' post-condition responses to evaluate hypotheses (subsection 5.1, subsection 5.2). We then interpret the remaining data to derive insights about acceptable response latency in subsection 5.3, effect of human-likeness of agents in subsection 5.4, visual and auditory modalities in subsection 5.5, and implications on future work in subsection 5.6, informing future studies and the design of human-agent conversational user interfaces.

5.1 Effects of Response Latency

Response latency had a significant negative impact on perceived (Q1) Response Time (Figure 5), with all pairwise comparisons between delay conditions showing significant differences ($p < 0.0001$), confirming **H1a**. This aligns with prior research on response delays in human-computer interaction [12, 29, 49, 53, 64, 74, 83, 101], reinforcing that conversational latency directly impacts perceived system efficiency. Additionally, our results support **H1b** — higher response latency was associated with lower (Q2) Engagement, (Q3) Good Impression, (Q5) Competence, (Q6) Willingness to Interact Again, as well as increased (Q4) Discomfort. These findings suggest that beyond perceived response time, delays in responses affect user trust and willingness to continue interacting with virtual agents, which is consistent with earlier work on conversational interruptions and delay tolerance [53].

Post-study responses to PSQ1 and PSQ2 revealed that most participants favored agents that corresponded to the *Low (1.5s)* latency (31/54, 57.41%), and disliked agents that corresponded to *High (6.5s)* latency (33/54, 61.11%). Given that Delay-Filler conditions applied to agents were counterbalanced between participants (implying an equal distribution (18/54, 33.33%) per delay level if chosen randomly), these preferences suggest that latency shaped participants' impressions of agents, even though they were not explicitly informed about response speeds. In text justifications for these two questions, 25/54 (46.29%) participants directly mentioned fast response times as the main reason behind choosing their favorite agent, and 22/54 (40.74%) participants mentioned slow response times as the reason for choosing an agent as their least favorite. Results from PSQ4 and PSQ5 (Figure 8) suggest that response speed was a key determinant in participants' system preference. Most participants indicated that they would use a system with the fastest response times, but rejected systems that responded slowest, which aligns with literature indicating that increased delay reduces future use intent [43]. These findings should be further validated through future studies involving conversational systems, through questionnaires containing concrete latency reference points and use case scenarios.

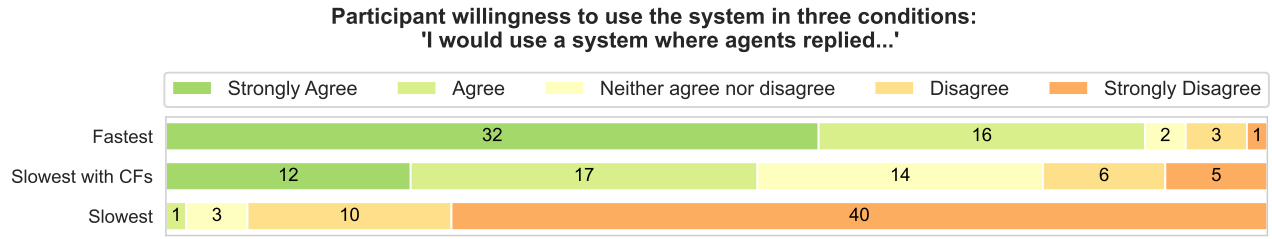

Figure 8: Participants' responses to question on whether they would use the system depending on the speed of agents' responses (PSQ4, PSQ5) and presence of conversational fillers (PSQ10).

5.2 Effects of Natural and Artificial Fillers

Natural conversational fillers significantly improved participants' ratings on (Q1) Response Time at *Medium (4.0s)* and *High (6.5s)* latency levels ($p < 0.01, p < 0.0001$, respectively), supporting **H2a**. This finding is consistent with prior studies indicating that conversational fillers facilitate smoother interactions and reduce the cognitive burden of waiting [12, 53]. Although *Natural* fillers improved average ratings on broader user experience dimensions over no fillers (*None* filler condition), enhancements in Engagement, Good Impression, Discomfort, Competence, or Willingness to Interact Again were not significant. Therefore, **H2b** was not supported, suggesting that while *Natural* conversational fillers improve perceived response latency, they do not mitigate its broader effects on user experience.

Although average Response Time ratings were slightly higher with *Artificial* fillers at *Medium (4.0s)* and *High (6.5s)* latency levels, they were not significantly different from the *None* condition (see Figure 6). This result does not support **H3a**. Similarly, participants' ratings on broader perception dimensions were not significantly different from *None* fillers, so **H3b** was also not supported. These findings suggest that passive visual and auditory wait indicators (e.g., visual loading icons and sounds) are not engaging enough to mitigate response latency or improve user experience on broader dimensions.

Responses to PSQ7 showed that *Natural* fillers ranked highest among the participants (35/54, 64.81%), as compared to *Artificial* fillers and no fillers present (see Figure 7). Most participants also rejected the idea of using a system where agents responded slowest (PSQ5). However, they were more open to such system when conversational fillers were present (PSQ10), with a statistically significant shift in responses ($\chi^2_4(N = 54) = 61.56, p < 0.001$). While this suggests that fillers can improve tolerance for delayed responses, participants still showed an overall preference for faster responses.

5.3 Insights on Acceptable Response Latency

Through our study, we empirically demonstrated that **response latency above 4 seconds significantly degrades user experience** in conversations with intelligent embodied virtual agents: (1) perceived latency degraded at *Medium (4.0s)* and *High (6.5s)* latencies (Figure 5); (2) future use intent was lowest under high-delay conditions (Figure 8); (3) agents with slowest response times were disproportionally ranked lowest in the post-study survey (subsection 5.1). These findings are especially relevant amid the growing adoption of conversational user interfaces powered by ASR, LLM, and TTS pipelines. Recent studies using similar pipelines for free-form interaction frequently omit explicit reporting of system response speed, despite descriptions suggesting relatively high latencies [31, 106, 108]. Based on our results, we recommend that future studies minimize and mitigate response delays, aiming for latencies under 4 seconds. Failing to do so risks negatively biasing participants' perceptions of conversational agents and degrading overall system usability.

While incremental response generation techniques [85, 94] can improve perceived response time, they are not always viable — critical information at the end of a user's turn may require a complete restructuring of the generated response. LLM-to-TTS transfer, where voice generation can begin as soon as the first sentence is produced, is the most parallelizable component in recent pipelines. Advancements, such as OpenAI's GPT-4o[15], demonstrate the potential for responses with negligible latency by processing input and output directly in speech form, bypassing sequential steps that involve text. However, these advancements do not address latency introduced by network instability or hardware limitations, which persistently cause latency in real-world deployments.

Despite ongoing efforts to accelerate response generation through model optimization and hardware improvements, the demand for more advanced reasoning capabilities inherently increases computational overhead. Techniques such as retrieval-augmented generation (RAG) and web search integration improve factual accuracy by dynamically incorporating external knowledge, but at the cost of additional processing time. Similarly, chain-of-thought (CoT) reasoning enhances logical inference by explicitly generating multi-step responses, while test-time optimization (TTO) adapts model outputs based on recent interactions — both substantially increasing SRT. These reasoning techniques are crucial in areas where accuracy outweighs speed, such as medical, legal, financial, and educational domains. However, their computational costs underscore a repeating pattern: as systems become faster, expectations rise, and increasingly complex compute causes latency to reemerge. Thus, mitigating IVA response latency at the user interface level — through adaptive UI design and conversational fillers — is just as critical as optimizing SRT. Rather than treating response delay as a purely technical limitation, future work should treat latency as

[15] openai.com/index/hello-gpt-4o (Accessed May 6, 2025)

an inevitable factor in high-quality conversational AI and design interactions that minimize its negative perceptual impact.

5.4 Participants' Preference for Human-likeness of Avatars

When designing our study, we focused on making the agents appear, behave, and sound human-like (natural). For this reason, we used VALID avatars (see subsubsection 3.3.1), animated all agents with 'busy', 'attentive idle', and 'thinking' states, as well as used a TTS engine that generated realistic voices for responses. In addition, conditions with *Natural* fillers included a thinking gesture and a voice line while agents generated their responses. In justifications for choosing their favorite agent, a majority of participants (38/54, 70.37%) included naturalness and relatability of conversations, in addition to quick responses. Some participant quotes provide more insight: (1) "*She **responded surprisingly quickly** and her **answers felt lifelike** compared to most of the other AIs I interacted with*", (2) "*The agent was very **personable** and acted like an actual friend would. The most **natural** of all of the agents*". Conversely, participants who selected specific agents as their least favorite often cited issues related to the lack of human-likeness: unnaturalness (20/54, 37.03%), awkwardness (11/54, 20.37%), and lack of motion during wait time (20/54, 37.03%). For example: (1) "*Felt the most **robot like** of the agents, with **delayed responses** and the dialogue did not feel smooth*", (2) "*It was fast but **creepy***".

While participants favored human-like avatars, presence of *Natural* fillers at *Low (1.5s)* latency did not significantly change perception on neither (Q1) Response Time, nor the broader dimensions of perception of the agents. This suggests that when agents are quick to respond, designers should prioritize delay being filled with animations consistent with agent's behavior, as execution of *Natural* fillers in a short time frame could be perceived as rushed and exaggerated. Moreover, in medical and other high-stakes scenarios, human tendency to trust and be more forgiving of anthropomorphized AI [44, 45, 99] may discourage users from scrutinizing accuracy. Future work should further investigate the relationship between filler anthropomorphism and risk profile: richer, natural cues for social or entertainment settings; more neutral or artificial signals where critical judgment matters.

5.5 Relative Importance of Visual and Audio Modalities

Responses on the different modalities of the *Natural* and *Artificial* fillers revealed that many participants felt they contributed similarly (see subsubsection 4.2.4). However, quotes from participants with opposing opinions provide interesting insights. For *Natural* fillers, quotes in favor of gestures include: (1) "*Gestures are a normal motion people do when prompted with questions or problems. So seeing the agents do it made it more of a real world experience.*", (2) "*It made me subconsciously realize they were thinking. But I preferred when they made a short hmm and a gesture*". On the other hand, some participants favored voice fillers over gestures, reasoning about higher naturalness, giving agents a buffer time to think, and confirming that the agents heard them; sample quotes include: (1) "*It seemed unrealistic and almost cartoony that the AI would go into a thinking pose every time I said something to them. The filler voice line, however, was surprising at first, though it felt like a natural buffer to give them time to think that you would see in a regular person.*", and (2) "*It let me know that the agent heard me. If they just used gestures, I would assume it's an idle animation.*".

For *Artificial* filler conditions, participants favored visual cues because of familiarity, for example: "*I'm used to the loading UI because it appears on websites, but the sound effect I'm not used to. I like the spinning UI because it appears exactly when I finish and disappears when they respond, so it's easy to tell that the world is generating message*". As for participants who chose auditory cues, they believed it was better as it confirmed the agent heard them, was less distracting, and helped reduce silence awkwardness when the agent was thinking. Some example quotes: (1) "*the spinning ui feels off, although it lets the user know the agent is thinking, it feels a bit inhuman*", (2) "*Because it let me know the response was loading through audio*". Two participants also mentioned thinking that a system error has occurred when they saw the visual wait indicator element of the *Artificial* filler for the first time.

The split in participants' opinions on the relative importance of auditory and visual filler modalities (subsubsection 4.2.4) suggests that no one specific set of features will work for every user, despite prior work indicating that the combination of gestures and voice utterances is best on average [53]. We recommend that researchers allow participants to choose their preferred filler modality as long as it is not a factor in their study, and that practitioners allow users to personalize conversational fillers for IVAs. While *Artificial* fillers effectively signaled that the system was processing input, future studies should explore more communicative and socially expressive indicators, such as familiar icons that users associate with thinking, or sequences that represent distinct stages of the response generation process.

5.6 Implications for Research and Design

Based on our findings, we distilled recommendations applicable to future research and design of embodied conversational agents.

Minimize response latency. Minimizing the turn-taking delay in agents' responses is crucial for sustainably-high QoE. Response delays of over 4 seconds worsen participants' perception on broader metrics about embodied agents, skewing collected data and limiting its generalizability.

Choose fillers consistent with the system's purpose. While *Natural* conversational fillers improve perceived system response time, their appropriateness depends on the system's purpose. If accuracy is more important than response speed or being liked by users, *Artificial* fillers or no fillers could fit better.

Allow response filler personalization. Presenting different conversational filler options and allowing users to select among them will account for preferences in visual and auditory modalities. This extends to the granularity of individual animations and phrases, as cultural backgrounds influence the interpretation of gestures.

Maintain participant engagement. Keeping participants engaged for the entire duration of a user study is important for collecting quality data [107]. Designing quest-like scenarios and minimizing response latency are possible ways to maintain high participant engagement in experiments.

Select a medium that minimizes distractions. Immersive VR environments can reduce external visual and auditory distractions [54], making them well-suited for studies involving embodied conversational agents. In our experiment, this helped ensure that participant responses were influenced by agent behavior rather than uncontrolled environmental factors.

6 LIMITATIONS AND FUTURE WORK

While the questions that we used in our experiment served us to get insight into user perception of IVAs, we acknowledge the need for a more standardized survey to collect perception metrics in this context. Our results serve as an initial comparison point with future work on IVAs. This is apparent in post-condition questions, where all aspects of discomfort and competence metrics from RoSAS [14] were aggregated into Q4 and Q5, respectively. This may have reduced the probability of detecting significant differences between conditions. Post-study questions PSQ4, PSQ5 and PSQ10 gauged future use intent; however, the order in which they were presented and the absence of clear application scenarios or reference points may have inadvertently influenced participants' responses. We recommend future work to consider such nuances when designing questionnaires.

The *Natural* and *Artificial* fillers used in our study were context-neutral and appropriate for the chosen scenarios. However, in serious scenarios (e.g., medical), a "laugh" filler would be less appropriate than "let me check your record", and the opposite applies in a playful scenario where human-likeness is more valued. In conversations between humans, prosodic features [38], conversational fillers, body gestures, and facial expressions, appear before an interlocutor finishes responding [6, 26, 68]. The challenge of adding these features to IVAs is rooted in the same reason that IVA response delays exists — it takes time to predict the appropriate voice line, gesture, and facial expression. Prior work integrated contextual animations for virtual agents, however, they were only triggered after multiple responses of the same type were generated, thus being played too late to mitigate response delay [108]. Future work should address this by designing fast NLP-based models that would process conversation history within a time shorter than 300ms [88], or potentially even before the user finishes speaking.

Applying conversational fillers to other applications of embodied conversational agents could yield valuable insights. For instance, conversational IVAs could *interrupt* humans, mimicking the natural flow of human-human dialogue [38, 58, 88]. As LLMs gain stronger multilingual capabilities [95], fillers could be adapted for culturally-sensitive timing in language learning scenarios [24, 93], potentially increasing IVA acceptance. Time-sensitive settings may also amplify the effects of latency: urgency can compress users' tolerance for delay [73, 105] and increase cognitive load [103]. While our study avoided time pressure by allowing participants to proceed at their own pace, future work should evaluate fillers in high-stakes or time-constrained contexts, such as a desert survival task [56], where embodied assistants have been shown to ease cognitive load [23]. Lastly, we find that research on conversational agents is increasingly conducted in immersive VR, however, there is a need for a more scrutinized evaluation of virtual mediums that can be used for studying ECAs.

7 CONCLUSION

We explored conversational fillers' impact on mitigating response delays in free-form conversations with LLM-powered embodied conversational agents in VR. We found that *Natural* fillers enhance VR user experience by significantly improving participants' perceived response time and reducing latency's negative repercussions. Our findings also indicate that *Artificial* fillers, namely wait indicators and processing sound effects, were not effective at reducing perceived response time. Our results contribute to the nascent work in optimizing user experiences with LLM-powered virtual agents, where network latency or hardware constraints inflate conversational system response time, especially in VR simulations of HRI and HAI scenarios. We outline design recommendations based on our findings and contribute an open-source pipeline as a solution to deploy LLM-based intelligent virtual agents in VR, advancing research efforts in developing more immersive and human-like interactions in VR.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful feedback and the ISUE lab members for their support. This work was in part supported by NSF Award CNS-2326134, DEVCOM Award W912CG2320004, and the Florida High Tech Corridor Council Industry Matching Research Program.

REFERENCES

[1] Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K. Schubert, and Ehsan Hoque. 2020. A Virtual Conversational Agent for Teens with Autism Spectrum Disorder: Experimental Results and Design Lessons. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3383652.3423900

[2] T. S. Amer and Todd L. Johnson. 2016. Information Technology Progress Indicators: Temporal Expectancy, User Preference, and the Perception of Process Duration. *International Journal of Technology and Human Interaction (IJTHI)* 12, 4 (Oct. 2016), 1—-14. https://doi.org/10.4018/IJTHI.2016100101

[3] Deepali Aneja, Rens Hoegen, Daniel McDuff, and Mary Czerwinski. 2021. Understanding Conversational and Expressive Style in a Multimodal Embodied Conversational Agent. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3411764.3445708

[4] Jana Appel, Astrid von der Pütten, Nicole C. Krämer, and Jonathan Gratch. 2012. Does Humanity Matter? Analyzing the Importance of Social Cues and Perceived Agency of a Computer System for the Emergence of Social Reactions during Human-Computer Interaction. *Advances in Human-Computer Interaction* 2012, 1 (2012), 324694. https://doi.org/10.1155/2012/324694

[5] Jakki O. Bailey, Jeremy N. Bailenson, Jelena Obradović, and Naomi R. Aguiar. 2019. Virtual Reality's Effect on Children's Inhibitory Control, Social Compliance, and Sharing. *Journal of Applied Developmental Psychology* 64 (July 2019), 101052. https://doi.org/10.1016/j.appdev.2019.101052

[6] Janet Beavin Bavelas and Nicole Chovil. 1997. Faces in Dialogue. In *The Psychology of Facial Expression*, James A. Russell and José Miguel Fernández-Dols (Eds.). Cambridge University Press, Cambridge, 334–346. https://doi.org/10.1017/CBO9780511659911.017

[7] Rojin Bayat, Elios De Maio, Jacopo Fiorenza, Massimo Migliorini, and Fabrizio Lamberti. 2024. Exploring Methodologies to Create a Unified VR User-Experience in the Field of Virtual Museum Experiences. In *2024 IEEE Gaming, Entertainment, and Media Conference (GEM)*. IEEE, Turin, Italy, 1–4. https://doi.org/10.1109/GEM61861.2024.10585452 ISSN: 2766-6530.

[8] Amy L. Baylor and Rinat B. Rosenberg-Kima. 2006. Interface agents to alleviate online frustration, In Proceedings of the 7th International Conference on Learning Sciences (Bloomington, Indiana). *Proceedings of ICLS 2006* 1, 30–36. https://repository.isls.org//handle/1/3514

[9] Štefan Beňuš and Marián Trnka. 2014. Prosody, Voice Assimilation, and Conversational Fillers. In *Speech Prosody 2014*. ISCA, Online, 75–79. https://doi.org/10.21437/SpeechProsody.2014-3

[10] Andrea J. Bingham. 2023. From Data Management to Actionable Findings: A Five-Phase Process of Qualitative Data Analysis. *International Journal of Qualitative Methods* 22 (Jan. 2023), 1–11. https://doi.org/10.1177/16094069231183620 Publisher: SAGE Publications Inc.

[11] Heather Bortfeld, Silvia D Leon, Jonathan E Bloom, Michael F Schober, and Susan E Brennan. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech* 44, 2 (2001), 123–147. https://doi.org/10.1177/00238309010440020101 PMID: 11575901.

[12] Halim-Antoine Boukaram, Micheline Ziadee, and Majd F Sakr. 2021. Mitigating the Effects of Delayed Virtual Agent Response Time Using Conversational Fillers. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21)*. Association for Computing Machinery, New York, NY, USA, 130–138. https://doi.org/10.1145/3472307.3484181

[13] Russell J Branaghan and Christopher A Sanchez. 2009. Feedback Preferences and Impressions of Waiting. *Human Factors* 51, 4 (2009), 528–538. https://doi.org/10.1177/0018720809345684 PMID: 19899362.

[14] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The Robotic Social Attributes Scale (RoSAS): Development and Validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. Association for Computing Machinery, New York, NY, USA, 254–262. https://doi.org/10.1145/2909824.3020208

[15] Llogari Casas, Samantha Hannah, and Kenny Mitchell. 2024. MoodFlow: Orchestrating Conversations with Emotionally Intelligent Avatars in Mixed Reality. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, Orlando, FL, USA, 86–89. https://doi.org/10.1109/VRW62533.2024.00021

[16] Florian Charlier, Marc Weber, Dariusz Izak, Emerson Harkin, Marcin Magnus, Joseph Lalli, Louison Fresnais, Matt Chan, Nikolay Markov, Oren Amsalem, Sebastian Proost, Agamemnon Krasoulis, getzze, and Stefan Repplinger. 2022. *Statannotations*. Statannotations Contributors. https://doi.org/10.5281/zenodo.7213391

[17] Anping Cheng, Dongming Ma, Hao Qian, and Younghwan Pan. 2024. The Effects of Mobile Applications' Passive and Interactive Loading Screen Types on Waiting Experience. *Behaviour & Information Technology* 43, 8 (2024), 1652–1663. https://doi.org/10.1080/0144929X.2023.2224901

[18] Vuthea Chheang, Shayla Sharmin, Rommy Márquez-Hernández, Megha Patel, Danush Rajasekaran, Gavin Caulfield, Behdokht Kiafar, Jicheng Li, Pinar Kullu, and Roghayeh Leila Barmaki. 2024. Towards Anatomy Education with Generative AI-based Virtual Assistants in Immersive Virtual Reality Environments. In *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. IEEE, Piscataway, NJ, USA, 21–30. https://doi.org/10.1109/AIxVR59861.2024.00011 ISSN: 2771-7453.

[19] Nicholas Christenfeld. 1995. Does It Hurt to Say Um? *Journal of Nonverbal Behavior* 19 (1995), 171–186. https://doi.org/10.1007/BF02175503

[20] Herbert H Clark. 1996. *Using language*. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511620539

[21] Frederick G Conrad, Mick P Couper, Roger Tourangeau, and Andy Peytchev. 2010. The Impact of Progress Indicators on Task Completion. *Interacting with computers* 22, 5 (2010), 417–427. https://doi.org/10.1016/j.intcom.2010.03.001

[22] Sarah H. Creem-Regehr, Jeanine K. Stefanucci, and Bobby Bodenheimer. 2022. Perceiving Distance in Virtual Reality: Theoretical Insights from Contemporary Technologies. *Philosophical Transactions of the Royal Society B: Biological Sciences* 378, 1869 (Dec. 2022), 1–12. https://doi.org/10.1098/rstb.2021.0456

[23] Celso M. de Melo, Kangsoo Kim, Nahal Norouzi, Gerd Bruder, and Gregory Welch. 2020. Reducing Cognitive Load and Improving Warfighter Problem Solving With Intelligent Virtual Assistants. *Frontiers in Psychology* 11 (2020), 12 pages. https://doi.org/10.3389/fpsyg.2020.554706

[24] Rahul R. Divekar*, Jaimie Drozdal*, Samuel Chabot*, Yalun Zhou, Hui Su, Yue Chen, Houming Zhu, James A. Hendler, and Jonas Braasch. 2022. Foreign Language Acquisition via Artificial Intelligence and Extended Reality: Design and Evaluation. *Computer Assisted Language Learning* 35, 9 (Dec 2022), 2332–2360. https://doi.org/10.1080/09588221.2021.1879162

[25] Tiffany D. Do, Steve Zelenty, Mar Gonzalez-Franco, and Ryan P. McMahan. 2023. VALID: A Perceptually Validated Virtual Avatar Library for Inclusion and Diversity. *Frontiers in Virtual Reality* 4 (Nov. 2023), 15 pages. https://doi.org/10.3389/frvir.2023.1248915

[26] Filippo Domaneschi, Marcello Passarelli, and Carlo Chiorri. 2017. Facial Expressions and Speech Acts: Experimental Evidences on the Role of the Upper Face as an Illocutionary Force Indicating Device in Language Comprehension. *Cognitive Processing* 18, 3 (Aug. 2017), 285–306. https://doi.org/10.1007/s10339-017-0809-6

[27] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 754–768. https://doi.org/10.1145/3472749.3474784

[28] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behavior Research Methods* 41, 4 (Nov 2009), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

[29] Kotaro Funakoshi, Kazuki Kobayashi, Mikio Nakano, Seiji Yamada, Yasuhiko Kitamura, and Hiroshi Tsujino. 2008. Smoothing Human-Robot Speech Interactions by Using a Blinking-light as Subtle Expression. In *Proceedings of the 10th international conference on Multimodal interfaces (ICMI '08)*. Association for Computing Machinery, New York, NY, USA, 293–296. https://doi.org/10.1145/1452392.1452452

[30] Markus Funk, Carie Cunningham, Duygu Kanver, Christopher Saikalis, and Rohan Pansare. 2020. Usable and Acceptable Response Delays of Conversational Agents in Automotive User Interfaces. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 262–269. https://doi.org/10.1145/3409120.3410651

[31] Irene Lopez Garcia, Ephraim Schott, Marcel Gohsen, Volker Bernhard, Benno Stein, and Bernd Froehlich. 2024. Speaking with Objects: Conversational Agents' Embodiment in Virtual Museums. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, Piscataway, NJ, USA, 279–288. https://doi.org/10.1109/ISMAR62088.2024.00042 ISSN: 2473-0726.

[32] Ulrich Gnewuch, Stefan Morana, Marc Adam, and Alexander Maedche. 2018. Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In *European Conference on Information Systems*. AIS Electronic Library (AISeL), Portsmouth, UK, 1–17. https://aisel.aisnet.org/ecis2018_rp/113

[33] Ulrich Gnewuch, Stefan Morana, Marc Adam, and Alexander Maedche. 2018. "The Chatbot is typing ..." – The Role of Typing Indicators in Human-Chatbot Interaction. In *SIGHCI 2018 Proceedings*. AIS Electronic Library (AISeL), San Francisco, CA, 1–5. https://aisel.aisnet.org/sighci2018/14

[34] Mar Gonzalez-Franco, Eyal Ofek, Ye Pan, Angus Antley, Anthony Steed, Bernhard Spanlang, Antonella Maselli, Domna Banakou, Nuria Pelechano, Sergio Orts-Escolano, Veronica Orvalho, Laura Trutoiu, Markus Wojcik, Maria V. Sanchez-Vives, Jeremy Bailenson, Mel Slater, and Jaron Lanier. 2020. The Rocketbox Library and the Utility of Freely Available Rigged Avatars. *Frontiers in Virtual Reality* 1 (2020), 20 pages. https://doi.org/10.3389/frvir.2020.561558

[35] Chris Harrison, Brian Amento, Stacey Kuznetsov, and Robert Bell. 2007. Rethinking the Progress Bar. In *Proceedings of the 20th annual ACM symposium on User interface software and technology (UIST '07)*. Association for Computing Machinery, New York, NY, USA, 115–118. https://doi.org/10.1145/1294211.1294231

[36] Chris Harrison, Zhiquan Yeo, and Scott E. Hudson. 2010. Faster Progress Bars: Manipulating Perceived Duration with Visual Augmentations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1545–1548. https://doi.org/10.1145/1753326.1753556

[37] Masum Hasan, Cengiz Ozel, Sammy Potter, and Ehsan Hoque. 2023. SAPIEN: Affective Virtual Agents Powered by Large Language Models*. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, Cambridge, MA, USA, 1–3. https://doi.org/10.1109/ACIIW59127.2023.10388188

[38] Mattias Heldner and Jens Edlund. 2010. Pauses, Gaps and Overlaps in Conversations. *Journal of Phonetics* 38, 4 (Oct. 2010), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

[39] Jess Hohenstein, Hani Khan, Kramer Canfield, Samuel Tung, and Rocio Perez Cano. 2016. Shorter Wait Times: The Effects of Various Loading Screens on Perceived Performance. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 3084–3090. https://doi.org/10.1145/2851581.2892308

[40] T. M. Holtgraves, S. J. Ross, C. R. Weywadt, and T. L. Han. 2007. Perceiving Artificial Social Agents. *Computers in Human Behavior* 23, 5 (Sept. 2007), 2163–2174. https://doi.org/10.1016/j.chb.2006.02.017

[41] Kate Hone. 2006. Empathic Agents to Reduce User Frustration: The Effects of Varying Agent Characteristics. *Interacting with computers* 18, 2 (2006), 227–245. https://doi.org/10.1016/j.intcom.2005.05.003

[42] Jacob Hornik. 1984. Subjective vs. Objective Time Measures: A Note on the Perception of Time in Consumer Behavior. *Journal of consumer research* 11, 1 (1984), 615–618. https://doi.org/10.1086/208998

[43] John Hoxmeier and Chris DiCesare. 2000. System Response Time and User Satisfaction: An Experimental Study of Browser-based Applications. , 6 pages. https://aisel.aisnet.org/amcis2000/347

[44] Qian Hu and Zhao Pan. 2024. Is Cute AI More Forgivable? The Impact of Informal Language Styles and Relationship Norms of Conversational Agents on Service Recovery. *Electronic Commerce Research and Applications* 65 (May 2024), 1–14. https://doi.org/10.1016/j.elerap.2024.101398

[45] Zuwen Huang and Ada Lo. 2025. Human vs. Robot Service Provider Agents in Service Failures: Comparing Customer Dissatisfaction and The Mediating Role of Forgiveness and Service Recovery Expectation. *Information Technology & Tourism* 27 (Feb. 2025), 1–32. https://doi.org/10.1007/s40558-025-00314-6

[46] Sun Young Hwang, Negar Khojasteh, and Susan R. Fussell. 2019. When Delayed in a Hurry: Interpretations of Response Delays in Time-Sensitive Instant Messaging. *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (Dec. 2019), 1–20. https://doi.org/10.1145/3361115

[47] Zainab Iftikhar, Yumeng Ma, and Jeff Huang. 2023. "Together But Not Together": Evaluating Typing Indicators for Interaction-Rich Communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 724, 12 pages. https://doi.org/10.1145/3544548.3581248

[48] Joseph Jaffe and Stanley Feldstein. 1970. *Rhythms of Dialogue*. Academic Press, New York, NY, USA. https://cir.nii.ac.jp/crid/1130282269381241984

[49] Yuin Jeong, Juho Lee, and Younah Kang. 2019. Exploring Effects of Conversational Fillers on User Perception of Conversational Agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 6 pages. https://doi.org/10.1145/3290607.3312913

[50] Takayuki Kanda, Masayuki Kamasima, Michita Imai, Tetsuo Ono, Daisuke Sakamoto, Hiroshi Ishiguro, and Yuichiro Anzai. 2007. A Humanoid Robot that Pretends to Listen to Route Guidance from a Human. *Auton. Robots* 22, 1 (Jan. 2007), 87–100. https://doi.org/10.1007/s10514-006-9007-6

[51] Woojoo Kim, Shuping Xiong, and Zhuoqian Liang. 2017. Effect of Loading Symbol of Online Video on Perception of Waiting Time. *International Journal of Human–Computer Interaction* 33, 12 (2017), 1001–1009. https://doi.org/10.1080/10447318.2017.1305051

[52] Takanori Komatsu, Chenxi Xie, and Seiji Yamada. 2024. Waiting Time Perceptions for Faster Count-downs/ups Are More Sensitive Than Slower Ones: Experimental Investigation and Its Application. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3613904.3641942

[53] Junyeong Kum and Myungho Lee. 2022. Can Gestural Filler Reduce User-Perceived Latency in Conversation with Digital Humans? *Applied Sciences* 12, 21 (Jan. 2022), 10972. https://doi.org/10.3390/app122110972

[54] Vishal Kiran Kuvar, Jeremy N. Bailenson, and Caitlin Mills. 2024. A novel quantitative assessment of engagement in virtual reality: Task-unrelated thought is reduced compared to 2D videos. *Computers & Education* 209 (Feb. 2024), 104959. https://doi.org/10.1016/j.compedu.2023.104959

[55] Christos Kyrlitsias and Despina Michael-Grigoriou. 2022. Social Interaction With Agents and Avatars in Immersive Virtual Environments: A Survey. *Frontiers in Virtual Reality* 2 (Jan. 2022), 13 pages. https://doi.org/10.3389/frvir.2021.786665 Publisher: Frontiers.

[56] J.C. Lafferty, P.M. Eady, A.W. Pond, and Human Synergistics. 1974. *The Desert Survival Problem: A Group Decision Making Experience for Examining and Increasing Individual and Team Effectiveness: Manual*. Experimental Learning Methods, Plymouth, Michigan: Experimental Learning Methods. https://books.google.com/books?id=X37-GwAACAAJ

[57] Asif Ali Laghari, Hui He, Muhammad Shafiq, and Asiya Khan. 2019. Application of Quality of Experience in Networked Services: Review, Trend & Perspectives. *Systemic Practice and Action Research* 32, 5 (Oct. 2019), 501–519. https://doi.org/10.1007/s11213-018-9471-x

[58] Stephen C. Levinson and Francisco Torreira. 2015. Timing in Turn-taking and its Implications for Processing Models of Language. *Frontiers in Psychology* 6 (2015), 731. https://doi.org/10.3389/fpsyg.2015.00731

[59] Zijian Lew, Joseph B Walther, Augustine Pang, and Wonsun Shin. 2018. Interactivity in Online Chat: Conversational Contingency and Response Latency in Computer-mediated Communication. *Journal of Computer-Mediated Communication* 23, 4 (July 2018), 201–221. https://doi.org/10.1093/jcmc/zmy009

[60] Shasha Li and Chien-Hsiung Chen. 2019. The Effect of Progress Indicator Speeds on Users' Time Perceptions and Experience of a Smartphone User Interface. In *Human-Computer Interaction. Recognition and Interaction Technologies: Thematic Area, HCI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II*. Springer-Verlag, Berlin, Heidelberg, 28–36. https://doi.org/10.1007/978-3-030-22643-5_3

[61] Shasha Li and Chien-Hsiung Chen. 2019. The Effects of Visual Feedback Designs on Long Wait Time of Mobile Application User Interface. *Interacting with Computers* 31, 1 (Jan. 2019), 1–12. https://doi.org/10.1093/iwc/iwz001

[62] Jose Llanes-Jurado, Lucía Gómez-Zaragozá, Maria Eleonora Minissi, Mariano Alcañiz, and Javier Marín-Morales. 2024. Developing Conversational Virtual Humans for Social Emotion Elicitation Based on Large Language Models. *Expert Systems with Applications* 246 (Jul 2024), 123261. https://doi.org/10.1016/j.eswa.2024.123261

[63] Soledad López Gambino, Sina Zarrieß, and David Schlangen. 2017. Beyond On-hold Messages: Conversational Time-buying in Task-oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis (Eds.). Association for Computational Linguistics, Saarbrücken, Germany, 241–246. https://doi.org/10.18653/v1/W17-5529

[64] Soledad López Gambino, Sina Zarrieß, and David Schlangen. 2019. Testing Strategies For Bridging Time-To-Content In Spoken Dialogue Systems. In *9th International Workshop on Spoken Dialogue System Technology*, Luis Fernando D'Haro, Rafael E. Banchs, and Haizhou Li (Eds.). Springer, Springer, Singapore, 103–109. https://doi.org/10.1007/978-981-13-9443-0_9

[65] Redowan Mahmud, Satish Narayana Srirama, Kotagiri Ramamohanarao, and Rajkumar Buyya. 2019. Quality of Experience (QoE)-aware Placement of Applications in Fog Computing Environments. *J. Parallel and Distrib. Comput.* 132 (Oct. 2019), 190–203. https://doi.org/10.1016/j.jpdc.2018.03.004

[66] Mykola Maslych, Christian Pumarada, Amirpouya Ghasemaghaei, and Joseph J. LaViola Jr. 2024. Takeaways from Applying LLM Capabilities to Multiple Conversational Avatars in a VR Pilot Study. arXiv:2501.00168 [cs.HC]

[67] Mykola Maslych, Difeng Yu, Amirpouya Ghasemaghaei, Yahya Hmaiti, Esteban Segarra Martinez, Dominic Simon, Eugene Matthew Taranta, Joanna Bergström, and Joseph J. LaViola Jr. 2024. From Research to Practice: Survey and Taxonomy of Object Selection in Consumer VR Applications. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (Seattle, WA, USA) *(ISMAR '24)*. IEEE, Piscataway, NJ, USA, 990–999. https://doi.org/10.1109/ISMAR62088.2024.00115

[68] David Matsumoto and Hyisung C. Hwang. 2018. Microexpressions Differentiate Truths From Lies About Future Malicious Intent. *Frontiers in Psychology* 9 (Dec. 2018), 2545. https://doi.org/10.3389/fpsyg.2018.02545

[69] Thomas McWilliams, Bryan Reimer, Bruce Mehler, Jonathan Dobres, and Hale McAnulty. 2015. A Secondary Assessment of the Impact of Voice Interface Turn Delays on Driver Attention and Arousal in Field Conditions. *Driving Assessment Conference* 8, 2015 (June 2015), 408–414. https://pubs.lib.uiowa.edu/driving/article/id/28561/

[70] Robert B. Miller. 1968. Response Time in Man-computer Conversational Transactions. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I on - AFIPS '68 (Fall, part I)*. ACM Press, San Francisco, California, 267. https://doi.org/10.1145/1476589.1476628

[71] Brad A Myers. 1985. The Importance of Percent-Done Progress Indicators for Computer-Human Interfaces. *ACM SIGCHI Bulletin* 16, 4 (1985), 11–17. https://doi.org/10.1145/1165385.317459

[72] Matthias Müller-Brockhausen, Giulio Barbero, and Mike Preuss. 2023. Chatter Generation through Language Models. In *2023 IEEE Conference on Games (CoG)*. IEEE, Boston, MA, USA, 6. https://doi.org/10.1109/CoG57401.2023.10333244 ISSN: 2325-4289.

[73] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, Dilip Kumar Limbu, Swee Lan See, and Alvin Hong Yee Wong. 2010. Socializing with Olivia, the Youngest Robot Receptionist Outside the Lab. In *Social Robotics*, Shuzhi Sam Ge, Haizhou Li, John-John Cabibihan, and Yeow Kee Tan (Eds.). Springer, Springer, Berlin, Heidelberg, 50–62. https://doi.org/10.1007/978-3-642-17248-9_6

[74] Naoki Ohshima, Keita Kimijima, Junji Yamato, and Naoki Mukawa. 2015. A Conversational Robot with Vocal and Bodily Fillers for Recovering from Awkward Silence at Turn-takings. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Kobe, Japan, 325–330. https://doi.org/10.1109/ROMAN.2015.7333677

[75] Xueni Pan and Antonia F. de C. Hamilton. 2018. Why and How to Use Virtual Reality to Study Human Social Interaction: The Challenges of Exploring a New Research Landscape. *British Journal of Psychology* 109, 3 (2018), 395–417. https://doi.org/10.1111/bjop.12290

[76] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC]

[77] Gustav Bøg Petersen, Aske Mottelson, and Guido Makransky. 2021. Pedagogical Agents in Educational VR: An in the Wild Study. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3411764.3445760

[78] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3174214

[79] Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation with LLM-Powered Chatbot Avatars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3613904.3642105

[80] Amir Bani Saeed, Zahra Moussavi, and Bruce Hardy. 2024. Developing an Avatar in Virtual Reality for Mental Health Treatment. *CMBES Proceedings* 46 (June 2024), 1–1. https://proceedings.cmbes.ca/index.php/proceedings/article/view/1107

[81] Junaid Shaikh, Markus Fiedler, and Denis Collange. 2010. Quality of Experience from User and Network Perspectives. *Annals of Telecommunications - Annales Des Télécommunications* 65, 1 (Feb. 2010), 47–57. https://doi.org/10.1007/s12243-009-0142-x

[82] Nicole Shechtman and Leonard M. Horowitz. 2003. Media Inequality in Conversation: How People Behave Differently When Interacting with Computers and

People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. Association for Computing Machinery, New York, NY, USA, 281–288. https://doi.org/10.1145/642611.642661

[83] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2009. How Quickly Should a Communication Robot Respond? Delaying Strategies and Habituation Effects. *International Journal of Social Robotics* 1, 2 (April 2009), 141–155. https://doi.org/10.1007/s12369-009-0012-8

[84] Alon Shoa, Ramon Oliva, Mel Slater, and Doron Friedman. 2023. Sushi with Einstein: Enhancing Hybrid Live Events with LLM-Based Virtual Humans. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents (IVA '23)*. Association for Computing Machinery, New York, NY, USA, 6 pages. https://doi.org/10.1145/3570945.3607317

[85] Gabriel Skantze and Anna Hjalmarsson. 2013. Towards Incremental Speech Generation in Conversational Systems. *Computer Speech & Language* 27, 1 (2013), 243–262. https://doi.org/10.1016/j.csl.2012.05.004

[86] Sruti Srinidhi, Edward Lu, and Anthony Rowe. 2024. XaiR: An XR Platform that Integrates Large Language Models with the Physical World. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, Bellevue, WA, USA, 759–767. https://doi.org/10.1109/ISMAR62088.2024.00091

[87] Ian Steenstra, Farnaz Nouraei, Mehdi Arjmand, and Timothy Bickmore. 2024. Virtual Agents for Alcohol Use Counseling: Exploring LLM-Powered Motivational Interviewing. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*. ACM, GLASGOW United Kingdom, 1–10. https://doi.org/10.1145/3652988.3673932

[88] Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and Cultural Variation in Turn-taking in Conversation. *Proceedings of the National Academy of Sciences* 106, 26 (June 2009), 10587–10592. https://doi.org/10.1073/pnas.0903616106

[89] Jan Svartvik (Ed.). 1990. *The London–Lund Corpus of Spoken English: Description and Research*. Lund Studies in English, Vol. 82. Lund University Press, Lund, Sweden. https://lup.lub.lu.se/record/c9ccd3ca-4a6e-4885-9ca5-1f939baa977f

[90] Marc Swerts. 1998. Filled Pauses as Markers of Discourse Structure. *Journal of pragmatics* 30, 4 (1998), 485–496. https://doi.org/10.1016/S0378-2166(98)00014-9

[91] Maite Taboada. 2006. Spontaneous and Non-spontaneous Turn-taking. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)* 16, 2-3 (2006), 329–360. https://doi.org/10.1075/prag.16.2-3.04tab

[92] Emma M. Templeton, Luke J. Chang, Elizabeth A. Reynolds, Marie D. Cone LeBeaumont, and Thalia Wheatley. 2022. Fast Response Times Signal Social Connection in Conversation. *Proceedings of the National Academy of Sciences of the United States of America* 119, 4 (Jan. 2022), e2116915119. https://doi.org/10.1073/pnas.2116915119

[93] Oguzhan Topsakal and Elif Topsakal. 2022. Framework for A Foreign Language Teaching Software for Children Utilizing AR, Voicebots and ChatGPT (Large Language Models). *The Journal of Cognitive Systems* 7, 2 (Dec 2022), 33–38. https://doi.org/10.52876/jcs.1227392

[94] Vivian Tsai, Timo Baumann, Florian Pecune, and Justine Cassell. 2019. Faster Responses Are Better Responses: Introducing Incrementality into Sociable Virtual Personal Assistants. In *9th International Workshop on Spoken Dialogue System Technology*, Luis Fernando D'Haro, Rafael E. Banchs, and Haizhou Li (Eds.). Springer, Singapore, 111–118. https://doi.org/10.1007/978-981-13-9443-0_10

[95] Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, and Kartik Kuckreja et al. 2024. All Languages Matter: Evaluating LMMs on Culturally Diverse 100 Languages. arXiv:2411.16508 [cs.CV]

[96] Ana Villar, Mario Callegaro, and Yongwei Yang. 2013. Where am I? A Meta-analysis of Experiments on the Effects of Progress Indicators for Web Surveys. *Social Science Computer Review* 31, 6 (2013), 744–762. https://doi.org/10.1177/0894439313497468

[97] Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. 2024. Building LLM-based AI Agents in Social Virtual Reality. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3613905.3651026

[98] Zhan Wang, Lin-Ping Yuan, Liangwei Wang, Bingchuan Jiang, and Wei Zeng. 2024. VirtuWander: Enhancing Multi-modal Interaction for Virtual Tour Guidance through Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3613904.3642235

[99] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (Jan. 1966), 36–45. https://doi.org/10.1145/365153.365168

[100] David Westerman, Aaron C. Cross, and Peter G. Lindmark. 2019. I Believe in a Thing Called Bot: Perceptions of the Humanness of "Chatbots". *Communication Studies* 70, 3 (May 2019), 295–312. https://doi.org/10.1080/10510974.2018.1557233

[101] Noel Wigdor, Joachim de Greeff, Rosemarijn Looije, and Mark A Neerincx. 2016. How to Improve Human-robot Interaction with Conversational Fillers. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, IEEE, New York, NY, USA, 219–224. https://doi.org/10.1109/ROMAN.2016.7745134

[102] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[103] Charley Wu, Eric Schulz, Timothy Pleskac, and Maarten Speekenbrink. 2022. Time pressure changes how people explore and respond to uncertainty. *Scientific Reports* 12 (March 2022), 4122. https://doi.org/10.1038/s41598-022-07901-1

[104] Takato Yamazaki, Tomoya Mizumoto, Katsumasa Yoshikawa, Masaya Ohagi, Toshiki Kawamoto, and Toshinori Sato. 2023. An Open-Domain Avatar Chatbot by Exploiting a Large Language Model. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani (Eds.). Association for Computational Linguistics, Prague, Czechia, 428–432. https://doi.org/10.18653/v1/2023.sigdial-1.40

[105] Euijung Yang and Michael C. Dorneich. 2015. The Effect of Time Delay on Emotion, Arousal, and Satisfaction in Human-Robot Interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59, 1 (Sept. 2015), 443–447. https://doi.org/10.1177/1541931215591094 Publisher: SAGE Publications Inc.

[106] Fu-Chia Yang, Kevin Duque, and Christos Mousas. 2024. The Effects of Depth of Knowledge of a Virtual Agent. *IEEE Transactions on Visualization and Computer Graphics* 30, 11 (Nov. 2024), 7140–7151. https://doi.org/10.1109/TVCG.2024.3456148

[107] Difeng Yu, Qiushi Zhou, Benjamin Tag, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2020. Engaging Participants during Selection Studies in Virtual Reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Atlanta, GA, USA, 500–509. https://doi.org/10.1109/VR46266.2020.00071 ISSN: 2642-5254.

[108] Jiarui Zhu, Radha Kumaran, Chengyuan Xu, and Tobias Höllerer. 2023. Free-form Conversation with Human and Symbolic Avatars in Mixed Reality. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, Sydney, Australia, 751–760. https://doi.org/10.1109/ISMAR59233.2023.00090 ISSN: 2473-0726.