Auditorily Embodied Conversational Agents: Effects of Spatialization and Situated Audio Cues on Presence and Social Perception

Yi Fei Cheng*
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
yifeic2@andrew.cmu.edu

Jarod Bloch*
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
jarodb@andrew.cmu.edu

Alexander Wang
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
aw4@andrew.cmu.edu

Andrea Bianchi
Industrial Design
KAIST
Daejeon, Republic of Korea
andrea@kaist.ac.kr

Anusha Withana
School of Computer Science
University of Sydney
Sydney, NSW, Australia
anusha.withana@sydney.edu.au

Anhong Guo
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan, USA
anhong@umich.edu

Laurie M. Heller
Department of Psychology
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
laurieheller@cmu.edu

David Lindlbauer
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
davidlindlbauer@cmu.edu

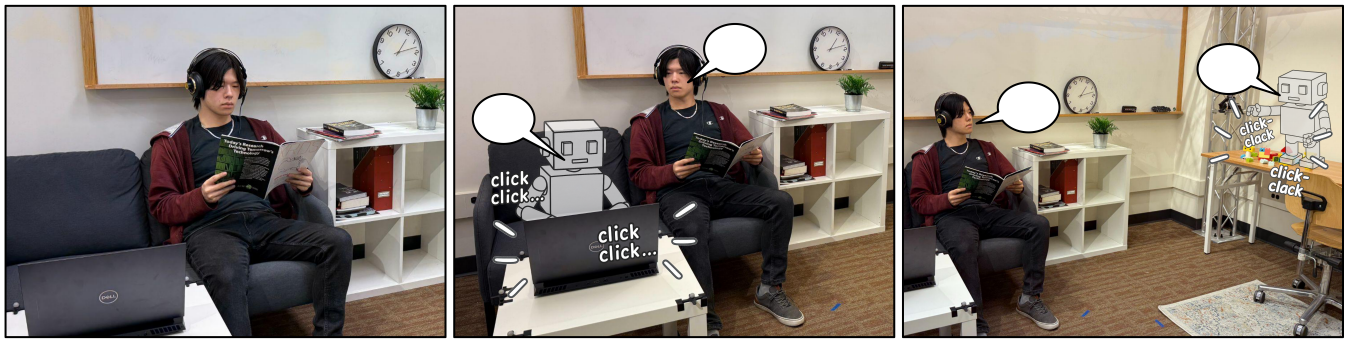

Figure 1: We explore how *auditory embodiment* influences perceptions of conversational agents. While agents are often embodied visually, such modalities may not always be available, for instance, when interacting through headphones *(left)*. We investigate whether embodiment can be conveyed solely through audio, using spatialized voice and situated Foley sounds. For example, an agent may be represented as seated next to the user while typing on a laptop *(middle)*, or as picking up toy blocks across the room *(right)*. *The visual depiction of the agent is for illustration purposes only.*

Abstract

Embodiment can enhance conversational agents, such as increasing their perceived presence. This is typically achieved through visual representations of a virtual body; however, visual modalities are not always available, such as when users interact with agents using headphones or display-less glasses. In this work, we explore *auditory embodiment*. By introducing auditory cues of bodily presence – through spatially localized voice and situated Foley audio from environmental interactions – we investigate how audio alone can convey embodiment and influence perceptions of a conversational agent. We conducted a 2 (SPATIALIZATION: *monaural* vs. *spatialized*) × 2 (FOLEY: *none* vs. *Foley*) within-subjects study, where participants (n=24) engaged in conversations with agents. Our results show that

*Both authors contributed equally to this research.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791794

spatialization and Foley increase co-presence, but reduce users' perceptions of the agent's attention and other social attributes.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; **Sound-based input / output**.

Keywords

Presence, Embodiment, Agents, Spatial Audio

ACM Reference Format:

Yi Fei Cheng, Jarod Bloch, Alexander Wang, Andrea Bianchi, Anusha Withana, Anhong Guo, Laurie M. Heller, and David Lindlbauer. 2026. Auditorily Embodied Conversational Agents: Effects of Spatialization and Situated Audio Cues on Presence and Social Perception. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3772318.3791794

1 Introduction

In recent years, conversational agents (e.g., Google's Gemini Live, OpenAI's ChatGPT voice mode) have advanced rapidly in their ability to engage in naturalistic dialogue. They increasingly demonstrate human-like behaviors [16] and can respond to spoken inputs in real time [24]. These capabilities have driven the growing popularity of applications and experiences that support conversational interactions, ranging from interactive museum guides [60] to services for social engagement and emotional support [25].

In the design of conversational agents, an important factor shaping user experiences is the way the agent is represented [79]. A common approach to representation is *embodiment*, where agents are given a bodily form [11]. Prior work has shown that embodying an agent can yield a range of advantages, from providing richer multimodal communication cues to support user tasks [40] to enhancing the agent's perceived social presence [2]. More importantly, embodiment shapes how people behave around agents, for better or worse. For instance, embodiment can foster greater trust in the agent [49]. This can be beneficial in domains such as education where trust may support student engagement [81]; however, this can also lead to unrealistic expectations of the agent's abilities [21].

These implications have motivated extensive research on how best to represent conversational agents. In both prior research and commercial systems, embodiment has typically been realized through visual representations [77], such as presenting avatars in Virtual or Augmented Reality (VR/AR). Yet, visual output modalities are not always available or desirable. A user may interact with an agent through earphones while their device is in their pocket, such as when walking or cooking. Recent smart glasses, such as the Ray-Ban Meta Glasses [65], are also display-less and rely primarily on auditory interactions. In these scenarios, current methods and understandings of agent embodiment may not be applicable.

In this work, we explore *auditory embodiment* of conversational agents. Specifically, we ask: **can embodiment be introduced through audio alone, and how does this influence user perceptions of the agent?**

In real-world social contexts, people are often co-present with others they cannot see, such as when someone is behind them. Nevertheless, their presence can still be sensed through (1) the spatial location of their voice and (2) incidental sounds generated by bodily movements and interactions with the environment (e.g., footsteps or rustling clothing). Similarly, in film, auditory realism is often enhanced through the use of everyday *Foley* sounds that accompany on-screen actions. Drawing on how humans naturally use incidental bodily and environmental sounds to infer others' presence in shared space, as well as the use of everyday Foley sounds to convey situated actions in film, we investigate whether such auditory cues can serve as mechanisms for supporting the embodiment of conversational agents. In particular, we examine how *spatialization* and situated *Foley* audio that represent an agent's movement and actions shape user perceptions and behaviors.

As an initial exploration of *auditory embodiment*, we study its effects in the context of *casual social interactions* with conversational agents. Prior research has long examined embodied agents in social settings [11], with more recent work focusing on conversational agents as companions [25]. Our work aims to provide insights into the design of conversational agents for social use, particularly in how they may be represented auditorily.

To this end, we conducted a within-subjects controlled study (n=24) in which participants engaged in and evaluated conversations with a conversational agent. In each conversation, the agent's voice was rendered either spatialized or monaural (non-spatial). For each spatialization condition, we further varied whether the agent's audio was presented alone or accompanied by Foley sounds. Our results show that *spatializing* the agent's audio and adding *Foley* contribute to stronger feelings of co-presence. However, the addition of *Foley* also reduces attention and message comprehension, and leads to a negative social impression.

Overall, we contribute empirical results showing how spatialization and Foley influence the social presence and perception of conversational agents. Through our results, we discuss opportunities and challenges of auditorily embodying conversational agents, highlighting considerations for the design of future systems.

2 Related Work

Our study builds on prior research on auditory interfaces, embodied conversational agents, and social presence.

2.1 Auditory Interfaces

Over the last 40 years, there has been persistent interest in auditory interfaces and interactions [29]. Early work on sound-based interactions, such as *earcons* [7] and *auditory icons* [29], primarily focused on information delivery. Subsequent research has significantly broadened this scope, exploring sound as a medium for embodied interaction [73], as a customizable component of domestic environments [43], and as a means of increasing the accessibility of emerging technologies such as Mixed and Virtual Reality [14, 44].

Within this broader landscape, the work most closely related to ours concerns *voice interfaces*[1], which use natural language as input and output. Voice interfaces have been explored across diverse applications, from in-vehicle assistance [101] to managing chronic

[1]Also referred to as dialogue systems, voice assistants, conversational agents, etc.

and mental health conditions [3]. According to Rzepka et al. [88], voice interfaces can provide pragmatic benefits such as convenience, ease of use, and time savings, as well as hedonic and social value.

The user experience of voice interfaces is highly sensitive to various design parameters. For example, Clark et al. [18] showed that users' perceptions are shaped by the role of the conversation (i.e., social or functional). The characteristics of an interface's voice can also shape the dynamics of an interaction [10] and how it is perceived in terms of gender, age, and personality [53, 63]. These prior works informed what variables we needed to control in our own study. Drawing on Clark et al. [18], we constrained our conversation task to a social role. Similarly, we fixed both the agent's voice and its system instructions to encourage more consistent responses.

One design parameter of particular relevance to our work is the spatial positioning of the interface's voice. In voice calls, prior work has shown that spatializing participants' voices can improve memory, speaker identification, and social presence [23, 48]. Takayama and Nass [94] showed that manipulating the spatial position of a conversational agent's voice can make disagreements seem more palatable. Hyrkas et al.[41] and Nowak et al. [78] extended these findings to video conferencing, showing that spatial audio increased perceptions of interactivity while reducing cognitive effort. In augmented and virtual reality, prior work has also shown that users generally prefer richer spatialized auditory and visual user representations [28, 42]. We extend this line of research by asking whether spatializing a conversational agent's voice can serve as a mechanism for *auditory embodiment*. While Kilgore et al. [48], Dicke et al. [23], and Takayama and Nass [94] focused on the effects of spatialized audio for virtual sounds that are not anchored in the listener's physical environment, our work examines whether spatialization can situate an *agent* within the user's surroundings. In contrast to work exploring how spatialized audio complements visual modalities [28, 41, 42, 78], we study whether spatial positioning can convey a sense of bodily presence in audio-only settings.

Another relevant factor is whether the interface produces additional non-speech audio cues. Prior work shows that emotion-evoking sounds shape perceptions of social attractiveness and emotional states [35, 58, 62, 96]. In Human-Robot Interaction, "consequential" sounds (i.e., by-products of robot mechanics or movements, rather than intentionally designed) can likewise influence evaluations of competence, trustworthiness, and human-likeness [72, 97]. In our work, we examine how situated Foley audio shapes user perceptions and behavior. Drawing on film, where Foley enhances scene realism, and on how everyday sounds inform spatial and semantic understanding [30], we ask whether analogous cues can contribute to auditory embodiment for an audio-only agent. Building on prior findings that sound shapes social perception, we evaluate how such Foley cues affect users' conversational experience, including social presence, attraction, and likeability.

2.2 Embodied Conversational Agents

Embodied conversational agents are dialogue systems that are represented by either a virtual or physical body [11]. Substantial literature has examined their design and effects on user experience [55, 76, 77, 79, 99, 108]. Introducing embodiment can yield a range of advantages. Functionally, a bodily representation allows agents to convey multimodal cues when communicating with users, such as gestures, gaze, and proxemics [1, 39, 109]. Embodiment further affects an agent's social presence [49] (Section 2.3), with important consequences for trust, engagement, and related factors [2, 38]. Beyond these benefits, embodiment also changes how people perceive and interact with these systems [86]. In line with the *computers as social actors* paradigm [75], when agents are embodied, users tend to treat them more like other people, thereby enabling them to adopt a more social role [54, 95].

A longstanding question in embodied conversational agent research concerns how such agents should be optimally represented [12, 74]. Prior work has typically embodied agents to imitate humans [37]. However, researchers have also questioned the value of adopting human-like appearances. For instance, Hale et al. [34] suggest that physical bodies can inadvertently evoke stereotypes related to gender, ethnicity, and beauty. Moreover, optimizing for highly realistic representations risks negative evaluations due to the uncanny valley effect [32, 47]. Beyond debates over whether agents should appear human, prior work has explored a wide range of alternative appearances and behaviors [22]. For example, Weber et al. [103] investigated user perceptions of agents embodied as food items.

In our work, we build on this research on examining how conversational agents should be embodied. In contrast to previous studies that have primarily investigated how different visual embodiments influence user perceptions and task outcomes, we focus on the effects of auditory bodily representations. It is worth recognizing here that previous work has suggested that the absence of a visual representation is generally detrimental to the user experience [50, 84, 100]. Yet, visual output modalities may not always be available or appropriate. Users may engage with conversations with agents through audio-based wearables [110] or display-less smart glasses [65]. Moreover, prior work has shown that in some contexts like driving, voice-only agents can also be advantageous for task performance and efficiency [33, 101]. Therefore, we investigate audio-only manipulations to address these scenarios.

2.3 Social Presence

Social presence was initially conceptualized by Short et al. [91] as a characteristic of interpersonal communication, defined as "the degree of salience of the other person in the interaction and the consequent salience of the interpersonal relationship." Numerous studies have shown that social presence is associated with a range of positive communication outcomes, which has motivated a longstanding interest in identifying its antecedents [80]. Early work examined how social presence is shaped by different communication modalities, such as speakerphone audio, monaural and multichannel audio, video, and face-to-face interaction [91]. Subsequent research explored additional factors, such as the availability of social cues [51], viewing perspective [102], and user demographics [52].

Although originally defined for human interactions, social presence has also been shown to apply to interactions with computational agents. Early dialogue systems, such as ELIZA, demonstrated that even a rudimentary text-based interface could elicit responses from users as if they were conversing with a real person [98, 104, 105]. As a *social actor* [75], an agent's perceived social presence can shape how users behave and how they perceive the

agent, which has significant implications for emerging application areas such as social interaction and emotional support [25].

However, prior work has primarily focused on the effects of visual representations on social presence [80]. In contrast, our work examines how audio-only agent representations shape perceived social presence, aiming to understand how to design agent representations that more effectively support better user experiences.

3 Auditory Embodiment: Concept and Implementation

Embodiment refers to introducing a bodily representation that grounds an agent in the user's environment [12]. We extend this notion to *auditory embodiment*, which we conceptualize as the extent to which an agent's bodily presence is conveyed through sound.

In everyday social contexts, auditory cues contribute to our awareness of others' bodily presence. When someone in our environment speaks, our auditory system processes not only the semantic content of their speech but also spatial cues arising from sound propagation, allowing listeners to infer the speaker's approximate physical location within the shared space [71]. This sense of presence is further reinforced by the sounds of their activities with the environment (e.g., walking, placing an object, or striking a surface). In particular, such *action sounds* [90] enable us to situate others relative to an existing cognitive map of the environment [19].

Drawing on these observations, our work considers two approaches to achieving *auditory embodiment*: (1) *spatializing* the agent's sounds, and (2) introducing *Foley* sounds that reflect its situated interactions with the environment. We implemented these approaches in a Unity-based system that formed the basis of our experimental apparatus. We describe each approach in detail below.

3.1 Spatialization

The human auditory system relies on a rich set of perceptual processes for localizing sounds [71]. In the horizontal plane, sound localization primarily depends on interaural time differences (ITD) and interaural level differences (ILD). These cues reflect differences in a signal's arrival time and sound pressure level at each ear, and vary with the position of sound sources relative to the listener's head. Aside from ITD and ILD cues, sound localization is also shaped by the *head-related transfer function* (HRTF) [106], which describes how sound waves are filtered by the anatomical features of the listener (e.g., the shape of the head and outer ears) before perception.

To simulate hearing an agent from a specific location in space, a system can reproduce human sound localization cues by modeling sound propagation from the agent's pose relative to the listener and rendering binaural signals using an HRTF.

3.1.1 Implementation. Our system simulated the experience of speaking with an agent positioned at a specific location in the user's room by tracking the user's head orientation relative to the agent and applying spatial audio rendering to reproduce the resulting 3D sound dynamics. For this purpose, our system integrated ten OptiTrack cameras to track the user's head position and rotation using a head-mounted five-marker rigid body. This setup enabled real-time, head-relative spatial audio updates, ensuring the agent's perceived location remained stable during user movement. For HRTF spatial audio rendering, we used the spatializer from the Meta XR Audio SDK version 77.0.0 [69], replicating the approach of Tao et al. [96]. At the time of the study, the SDK represented the state-of-the-art in spatial audio rendering [15, 17]. To represent the agent, we set up a virtual sound source that used a "human voice" directivity pattern [68], simulating natural voice attenuation (i.e., quieter and more muffled when turned away from the user). Lastly, our system applied room acoustic effects to virtual sounds using the SDK's shoebox model [66], with room dimensions and material properties configured to match the experimental space.

3.2 Foley

Almost every bodily movement, such as walking, flipping through a magazine, or typing, produces action sounds that convey spatial information [93]. In social scenarios, these sounds support awareness ofothers' whereabouts by enabling associations between their actions and the physical objects involved.

To reinforce an agent's presence within a user's space, a system may reproduce sounds that plausibly correspond to the agent's interactions with the environment. In film, this practice is known as *Foley*, where everyday sound effects are created to enhance the perception of actions on screen. In our context, such sounds can serve as complementary cues to the agent's speech, helping to contextualize its activities within the shared space. This approach further parallels how Augmented Reality uses visual augmentations to anchor virtual entities within the user's environment, with Foley cues serving an analogous role in auditorily situating the agent.

3.2.1 Implementation. To simulate incidental sounds associated with the agent's embodied presence, our system replays recorded audio clips from manually configured spatial locations within the room. Using the same tracking and audio rendering components described in Section 3.1.1, these sounds are rendered as though emanating from contextually appropriate positions (e.g., keyboard clicks from the location of a physical laptop). Our system currently supports predefined *activity sequences*, each consisting of multiple Foley events bound to a pre-recorded agent movement trajectory. We discuss the design of the activity sequences in Section 4.2, including the curated stimuli used in our experiment. Potential automated, context-aware implementations are considered in Section 6.

4 Experiment

To investigate how auditory embodiment influences perceptions of conversational agents, we conducted a within-subjects study. Each participant engaged in four conversations with the agent, varying along two factors: whether the agent's audio was spatialized and whether it was accompanied by Foley. We measured both subjective perceptions, through ratings of social presence and impressions of the agent, and behavioral responses, including conversational dynamics and user movements within the environment.

4.1 Conversation Task

In each condition, participants were tasked with engaging in 3-minute conversations with an agent. Considering the potential use case of conversational agents as companions [25], we designed our conversation task to simulate a casual conversation scenario. Participants randomly selected a conversation topic from a curated set adapted from Fang et al. [25]. We specifically selected

self-relevant topics that required some sharing of personal experiences. To minimize variability, topics with negative valence or high arousal were excluded. Example topics included how participants had celebrated a recent holiday or the best show they had watched recently (see Appendix A for the full list).

For each conversation, the agent was configured with the same system instructions that specified it was a companion engaging in casual conversation with the user (Appendix B). Through pilot testing, we experimented with several instruction sets, including the engaging voice configuration described by Fang et al. [25]. We ultimately adopted our final configuration, as it qualitatively produced the most consistent responses.

4.2 Agent Behavior

Our design of the agent's embodied behaviors was guided by the following objectives:

Perceptual grounding: The agent's activities should involve both movement and recognizable interactions with objects, enabling participants to perceive its spatial location and engagement with the environment. In particular, prior research has shown that humans localize sound more effectively through relative changes than through static, absolute cues [85]. Therefore, we introduced agent movement to support these relative judgments and strengthen perceptions of the agent as being situated within the environment.

Ecological plausibility: The agent should perform everyday activities that could plausibly occur within the experimental room.

Ambient: The agent's actions should remain in the background, providing subtle cues of presence.

To model these behaviors, we drew on the persona of a colleague who occasionally moves about and interacts with objects in their room while maintaining a casual conversation. We designed three ACTIVITY SEQUENCES, each consisting of seven activities (Figure 2). The activities included flipping through a book or magazine, writing on a whiteboard, assembling toy blocks, pouring a glass of water, organizing and stapling papers, clicking a mouse, and typing.

For each sequence, a member of the research team recorded the trajectory of their head while walking through the room and performing the activities. This trajectory was then aligned with Foley sounds for the constituent activities, recorded using a Snowball iCE microphone. These sounds were manually mapped to their contextually relevant locations within the environment. In addition, we introduced footstep sounds mapped to the velocity of the agent's movement, as well as subtle clothing interaction sounds triggered at random intervals, spatially anchored to the feet and waist. We notably recorded our own Foley instead of using generative approaches or sourcing from online catalogs, as early experimentation indicated that these alternatives lacked the quality and acoustic consistency necessary to be convincingly associated with the space.

4.3 Experimental Design

The experiment followed a 2×2 within-subjects design with two independent variables (Figure 3): SPATIALIZATION (*monaural*, *spatialized*) and FOLEY (*none*, *Foley*). This yielded four conditions:

Monaural + None: In this condition, the agent's voice is neither spatialized nor accompanied by Foley, effectively serving as a baseline comparable to a standard voice call.

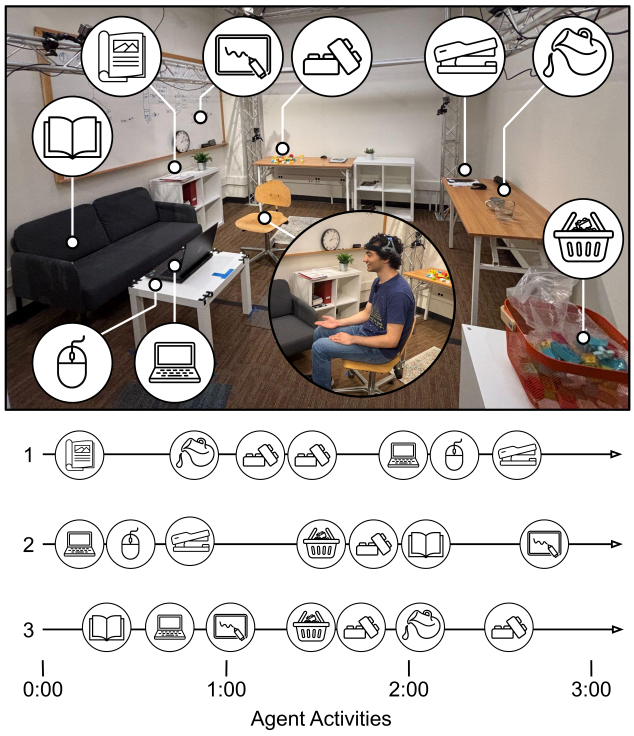

Figure 2: Agent activity sequences. *(Top)* shows the activities and their corresponding locations within the room, with the participant positioned at its center. *(Bottom)* shows the temporal ordering of the three activity sequences the agent followed in our experiment.

Spatialized + None: In this condition, the agent's voice is spatialized but not accompanied by Foley. The agent follows the recorded ACTIVITY SEQUENCE trajectory, providing directional cues that indicate its movement and position in space. No additional sounds, such as footsteps or action sounds, are presented.

Monaural + Foley: In this condition, the agent's audio is rendered monaurally and accompanied by Foley sounds representing its movements and object interactions. Because neither the speech nor the Foley is spatialized, these cues are not acoustically localized to any position in the room.

Spatialized + Foley: In this condition, both the agent's voice and the Foley sounds are spatialized. They follow the recorded ACTIVITY SEQUENCE trajectory and the mapped locations of each interaction, providing directional cues to the agent's movement as well as to the sounds of its activities within the room.

The order of conditions and the order of the ACTIVITY SEQUENCES were individually counterbalanced using Latin Square designs and then paired. This resulted in 4 condition orders $\times$ 6 sequence orders, which we evenly distributed across our 24 participants.

4.4 Procedure

Upon arriving at the lab, participants were first given a brief introduction to the study, the equipment involved, and the data we recorded. Then, they filled out a consent form and a pre-questionnaire.

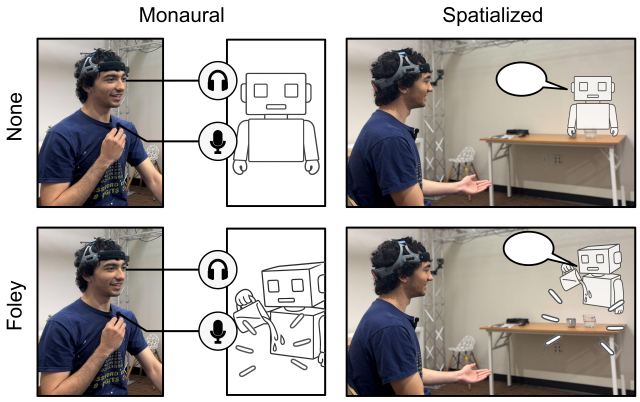

Figure 3: Experimental conditions. The agent's audio was rendered either *monaurally (left)* or *spatialized (right),* with the addition of *Foley* also varied: *none (top)* vs. *Foley (bottom)* (e.g., sounds of the agent pouring water). The visual depiction of the agent is for illustration only; participants experienced all conditions through audio alone.

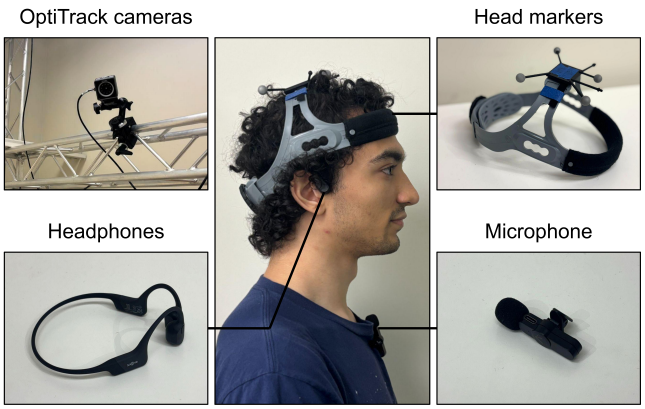

Figure 4: Study apparatus. We used ten OptiTrack motion-capture cameras (*top left*) to track a 4-marker rigid body attached to participants' heads (*top right*). Participants engaged with the conversational agent using Shokz OpenRun Pro bone-conduction headphones (*bottom left*) and an Aisizon wireless lavalier microphone (*bottom right*).

Afterwards, participants completed a familiarization phase. In this phase, they were instructed to walk around the room to get familiar with it. In addition, they were instructed to interact with various objects in the room (e.g., writing on the whiteboard, pouring a glass of water), mirroring potential actions of the agent. This phase was designed to help participants build a mental model of the room, approximating the level of familiarity they would normally have with their own everyday environments and facilitating associations between the Foley and the corresponding objects.

Participants then proceeded through the conditions of our study. In the conversation tasks, participants were instructed to engage with the agent while sitting on a swivel chair that allowed for free head and torso rotation. They were instructed to remain seated, but could rotate with the chair. Because the agent was scripted to move around the environment, allowing participants to walk freely would have introduced substantial variability in the experience. We therefore opted for a more stationary setup. The seat was intentionally placed near the center of the room, next to a coffee table. This location simulated a plausible seating position while ensuring that participants could perceive the agent's activities from all directions.

After each session, participants reported on several subjective metrics in a post-condition questionnaire. After all sessions were completed, participants completed a final exit survey and participated in a semi-structured interview, where they discussed impressions of the agent, and preferred conditions. The study was approved by the Institutional Review Board (IRB) of Carnegie Mellon University. The full study took 60 minutes. All participants were compensated $15 for their time.

4.5 Apparatus

The study was conducted in a $4 \times 3 \times 3$m experimental space implementing the system described in Section 3. To support interactive dialogue, a bidirectional audio stream was established over WebSocket to a Gemini 2.5 Flash Native Audio model. The experiment ran on an Intel Core i7-12700H CPU 2.30 GHz computer with 16 GB of RAM, supported by an NVIDIA GeForce RTX 3060 GPU. Real-time speech input was supported using an Aisizon wireless lavalier microphone and Shokz OpenRun Pro bone-conduction headphones. We used the Shokz OpenRun headset to preserve environmental sound cues, as prior work suggests that acoustic transparency enhances the real-world grounding of virtual audio [64]. Figure 4 shows an overview of our study apparatus.

4.6 Measures

We evaluate participants' perceptions and behaviors through a range of self-reported and quantitative metrics. All self-reported metrics were evaluated with 7-point Likert scales (Appendix C).

4.6.1 Social Presence. We assessed participants' perceived social presence of the conversational agent using questions from the Networked Minds Social Presence Inventory (NMSPI) [5, 6, 36]. The NMSPI conceptualizes social presence as comprising multiple sub-dimensions [36]. In our study, we evaluated the following:

Co-presence: Measures the extent to which users perceive the agent as sharing the same environment. This includes both the user's own sense of co-presence (perception of self) and their perception of the agent's co-presence (perception of other).

Attentional Allocation: Measures the extent to which users direct attention toward the agent (perception of self) and perceive the agent as allocating attention toward them (perception of other).

Message Understanding: Measures the user's ability to understand the agent's messages (perception of self) as well as their perception of the agent's ability to understand their own messages (perception of other).

Affective Understanding: Measures the user's ability to recognize the agent's emotional and attitudinal states (perception of self) and their perception of the agent's ability to recognize their own emotional and attitudinal states (perception of other).

Affective Interdependence: Measures the extent to which the user's emotional and attitudinal states influence, and are influenced by, those of the agent. This can be subdivided into perception of self and perception of other.

We excluded the *behavioral interdependence* measure, as our study focused on conversational experience and did not involve joint task performance with the agent.

4.6.2 Social Impression. Inspired by prior work [27, 96], we also assessed participants' social impressions of the agent through measures of likeability judgments and social attraction:

Likeability judgments: Measured with five items evaluating attractiveness, competence, extroversion, likeability, and trustworthiness, adapted from [96].

Social attraction: Measured with four items evaluating friendliness, interpersonal affinity, and willingness to interact again [27].

4.6.3 Preference. After completing all conversation tasks, participants were asked to rank the four conditions by overall *preference*. For this ranking, participants were not informed which experimental manipulations each session corresponded to.

4.6.4 Behavioral measures. As measures of conversational engagement [16] we recorded the *total number of words* exchanged in each conversation (with separate counts for *user words* and *agent words*) and the number of *turn shifts* (i.e., when the user yielded the floor to the agent and vice versa). All verbal metrics were computed from transcriptions generated with OpenAI's Whisper-Medium [83].

As non-verbal behavior indicators of social presence [51], we recorded the user's *head rotation* (i.e., the cumulative angular rotation of the participant's head throughout the session) and *facing angle to agent* (i.e., the angular difference between the participant's facing direction and the agent's position relative to them). We additionally calculated the percentage of the session in which the agent was positioned within the participant's *central* (i.e., $< 30°$ of facing direction), *near-peripheral* (i.e., $< 60°$), and *far-peripheral* vision ($< 100°$). All non-verbal measures were recorded with the OptiTrack system at $\sim$ 20 Hz.

4.7 Power and Experimental Participants

Prior to conducting the study, we performed an a priori power analysis using G*Power 3.1 [26]. To estimate the required sample size, we considered two effect sizes, $f = 0.25$ (small) and $f = 0.5$ (medium). We set the significance level at $\alpha = 0.05$ and the statistical power at 0.8. Because our subjective measures were collected once per condition, we specified 4 measurements (corresponding to the 4 within-subject conditions) and left the default correlation among repeated measures at 0.5. The analysis indicated that detecting a small effect would require 24 participants, while detecting a medium effect would require 8 participants. We also considered prior studies on embodiment effects in perceptions of virtual agents (e.g., [100]).

We recruited 24 participants (11 male, 13 female) between the ages of 18 and 34 ($M = 27$, $SD = 5$) from a university community via message groups, social networks, and word-of-mouth. Most participants (n=23) reported using headphones or earbuds daily, while one reported using them at least once per week. Participants' familiarity with relevant technologies is summarized in Table 1.

Table 1: Self-reported familiarity (number of participants) with voice-based agents, text-based agents, and spatial audio.

Usage	Voice agents	Text agents	Spatial audio
< 5 hours	11	1	2
5–10 hours	6	3	0
10–20 hours	3	1	0
20–30 hours	2	5	4
50–100 hours	1	5	1
> 100 hours	1	9	17

5 Results

We evaluated the agent's social presence, as well as participants' impressions and preferences, across two levels of SPATIALIZATION and two levels of FOLEY. In addition, we examined the effects of these factors on participants' verbal and nonverbal behaviors. Overall, *spatializing* the agent's audio and adding *Foley* increased participants' feelings of co-presence. However, the addition of *Foley* also reduced attention and message understanding, and contributed to a more negative social impression.

For effect analysis, we analyzed ordinal data (questionnaire ratings) using an Aligned Rank Transform (ART) ANOVA [107]. Interval data (e.g., words exchanged, turn shifts) were analyzed using a two-factor repeated-measures ANOVA. For each dependent variable, *participant* was treated as a random factor, with SPATIALIZATION and AUDIO CUES as within-subject independent variables. When assumptions of normality of residuals or homogeneity were violated (Shapiro–Wilk test, $p < .05$), we analyzed the data using ART. Post-hoc tests with Bonferroni adjustments were conducted as needed. The analysis was performed using R 4.5.1 [82].

5.1 Social Presence

The *social presence* factors were each analyzed separately for participants' *perception of self*, *perception of other*, and for a *combined* score [4]. Figure 5 summarizes the effects of SPATIALIZATION and FOLEY on *social presence*.

5.1.1 Co-presence. The ART analysis showed a significant main effect of SPATIALIZATION on *self* ($F_{1,69} = 16.11$, $p < 0.001$, $\eta_p^2 = 0.19$), *other* ($F_{1,69} = 5.11$, $p = 0.03$, $\eta_p^2 = 0.07$), and *combined co-presence* ($F_{1,69} = 12.66$, $p < 0.001$, $\eta_p^2 = 0.16$). Across all three measures, participants reported higher ratings in *spatialized* conditions compared to *monaural*, suggesting that **rendering the agent's voice spatially can increase feelings of co-presence.**

A main effect of FOLEY was also observed for *self* ($F_{1,69} = 20.38$, $p < 0.001$, $\eta_p^2 = 0.23$) and *combined co-presence* ($F_{1,69} = 12.72$, $p < 0.001$, $\eta_p^2 = 0.16$). Participants reported higher ratings in *Foley* conditions compared to *none*, suggesting that **introducing Foley can similarly increase feelings of co-presence.**

No main effect of FOLEY was found for *co-presence perception of other* ($F_{1,69} = 3.44$, $p = 0.07$, $\eta_p^2 = 0.05$). No significant interaction effects were found across all *co-presence* measures (all $p > 0.05$).

To further investigate the main effects of SPATIALIZATION and FOLEY, we conducted an exploratory analysis [45, 96] probing whether the combined condition (*spatialized + Foley*) yielded higher ratings

than conditions in which only one feature was present (i.e., *spatialized + none*, *monaural + Foley*). Using Bonferroni-corrected Wilcoxon signed-rank tests, we found that for *self co-presence*, the *spatialized + Foley* condition yielded significantly higher ratings than both *monaural + Foley* ($p = 0.041$) and *spatialized + none* ($p = 0.004$). *Combined co-presence* was also significantly higher in *spatialized + Foley* than in *monaural + Foley* ($p = 0.02$), though not compared to *spatialized + none* ($p > 0.05$). These results suggest that **spatialization and Foley may be complementary in creating a stronger sense of co-presence**.

5.1.2 Attentional Allocation. The ART analysis showed a significant main effect of Foley on *self* ($F_{1,69} = 18.40$, $p < 0.001$, $\eta_p^2 = 0.21$), *other* ($F_{1,69} = 8.14$, $p = 0.006$, $\eta_p^2 = 0.11$), and *combined attentional allocation* ($F_{1,69} = 20.10$, $p < 0.001$, $\eta_p^2 = 0.23$). Across all three measures, participants reported lower ratings in *Foley* compared to *none*. These results suggest that **Foley reduced participants' perceptions of attention between themselves and the agent**.

There was no main effect of spatialization (all $p > 0.05$), but it did interact significantly with Foley for both *self* ($F_{1,69} = 4.98$, $p = 0.03$, $\eta_p^2 = 0.07$) and *combined attentional allocation* ($F_{1,69} = 7.02$, $p = 0.01$, $\eta_p^2 = 0.09$). Post-hoc tests indicated *monaural + Foley* reduced *self* ($p = 0.002$) and *combined attentional allocation* ($p < 0.001$ compared to *monaural + none*. The *spatialized + Foley* condition also reduced *self* ($p = 0.02$) and *combined attentional allocation* ($p = 0.02$) compared to *spatialized + none*. These findings suggest that spatialization alone did not influence *attentional allocation*. The addition of *Foley* consistently reduced participants' perceptions of shared *attention*, regardless of whether the agent's voice was rendered *monaurally* or *spatially*.

5.1.3 Message Understanding. We observed a significant main effect of Foley on *self* ($F_{1,69} = 9.40$, $p = 0.003$, $\eta_p^2 = 0.12$), *other* ($F_{1,69} = 6.81$, $p = 0.01$, $\eta_p^2 = 0.09$), and *combined message understanding* ($F_{1,69} = 8.28$, $p = 0.005$, $\eta_p^2 = 0.11$). Participants reported higher ratings across all three measures in conditions with *Foley* compared to *none*. These results indicate that **introducing situated audio cues reduced both participants' understanding of the agent and their feeling of being understood by the agent.**

No main effects of spatialization or interactions were observed for message understanding (all $p > 0.05$).

5.1.4 Affective Understanding and Interdependence. No main effects or interaction effects were observed for affective understanding or affective interdependence (all $p > 0.05$). These results suggest that within the given experimental setting, spatialization and audio cues did not affect participants' affective experience.

5.2 Social Impression

We calculated composite scores for *likeability* (Cronbach's $\alpha = 0.87$) and *social attraction* (Cronbach's $\alpha = 0.89$) from their respective questionnaire items (Figure 6).

The ART analysis showed a significant main effect of Foley on both *likeability* ($F_{1,69} = 7.45$, $p = 0.008$, $\eta_p^2 = 0.10$) and social attraction ($F_{1,69} = 6.65$, $p = 0.01$, $\eta_p^2 = 0.09$). Participants reported lower ratings in *Foley* conditions compared to *none*, suggesting that **Foley negatively influenced participants' impressions of the agent**.

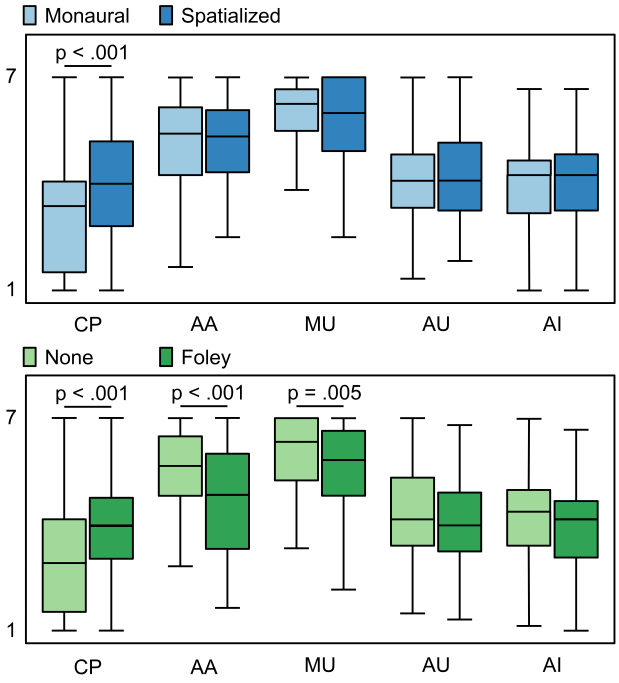

Figure 5: Effect of spatialization (*top*) and Foley (*bottom*) on *social presence*: *co-presence* (CP), *attention allocation* (AA), *message understanding* (MU), *affective understanding* (AU), and *affective interdependence* (AI).

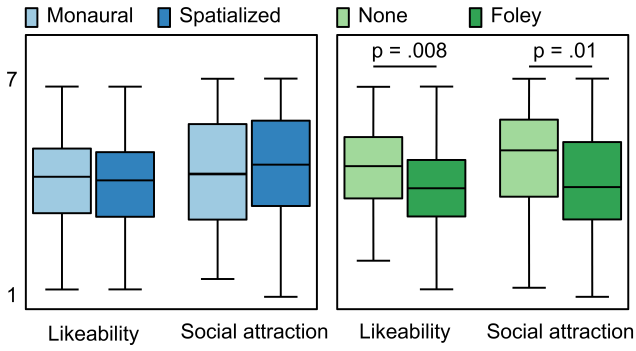

Figure 6: Effect of spatialization and Foley on *likeability* and *social attraction*.

While there were no main effects of spatialization on *likeability* ($F_{1,69} = 0.02$, $p = 0.9$, $\eta_p^2 = .0003$) or *social attraction* ($F_{1,69} = 0.31$, $p = 0.6$, $\eta_p^2 = .005$), it interacted significantly with Foley for the latter ($F_{1,69} = 5.25$, $p = 0.02$, $\eta_p^2 = 0.07$). Post-hoc tests showed that Foley reduced *social attraction* in the *monaural* condition ($p = 0.02$).

5.3 Preference

The ART analysis showed that participants ranked the conversational agent with *Foley* below *none* ($F_{1,69} = 6.03$, $p = 0.02$, $\eta_p^2 = 0.08$; Figure 7). No main effects of spatialization or interactions were

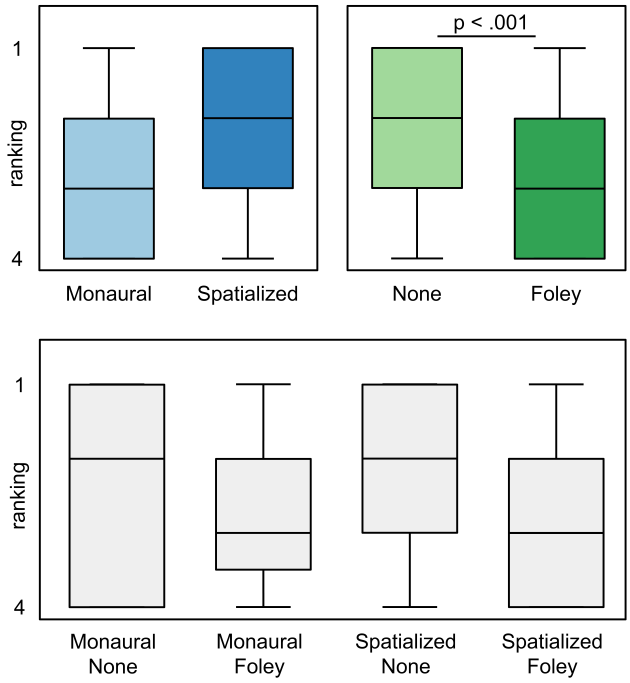

Figure 7: Effect of SPATIALIZATION and FOLEY on *preference* rankings (1-Most preferred, 4-Least preferred).

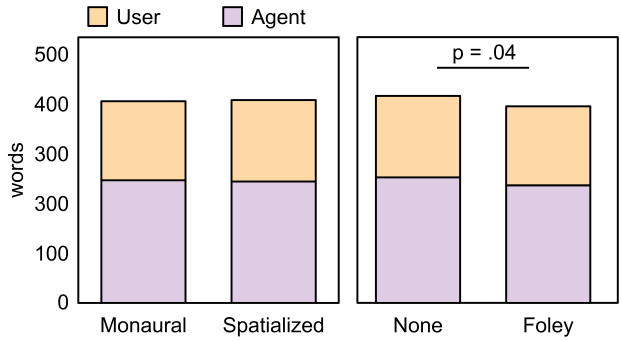

Figure 8: Effect of SPATIALIZATION and FOLEY on the number of *words* exchanged in the conversation.

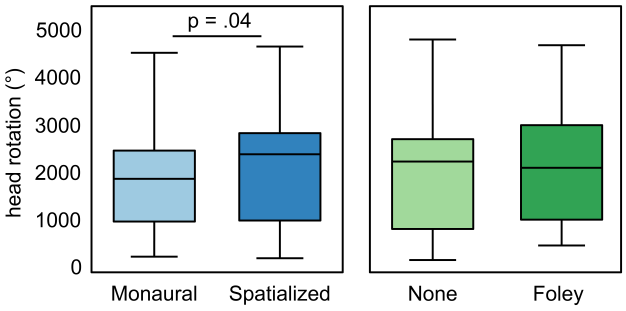

Figure 9: Effect of SPATIALIZATION and FOLEY on *head rotation*.

observed (all $p > .05$). Overall, these results suggest that **adding Foley negatively affected perceptions of the agent**.

5.4 Verbal Behaviors

On average, conversations comprised 404 words ($SD = 52$) exchanged over 21 turns ($SD = 5$), including 160 words ($SD = 50$) from participants and 244 words ($SD = 51$) from the agent (Figure 8).

The ART analysis showed that introducing *Foley* reduced the *total words* exchanged within the conversation by 20 ($F_{1,69} = 4.21$, $p = 0.04$, $\eta_p^2 = 0.06$). No additional main effects or interaction effects were observed for the other verbal behavior metrics (all $p > 0.05$), including participants' and the agent's individual contributions. These results suggest that the **introduction of Foley sounds may have slightly dampened conversational engagement**, leading to fewer words being exchanged overall.

5.5 Nonverbal Behaviors

The ART analysis showed that participants rotated their heads 12% more in the *spatialized* condition than in the *monaural* condition ($F_{1,69} = 4.60$, $p = 0.04$, $\eta_p^2 = 0.03$; Figure 9). No additional main effects or interaction effects were observed for the other nonverbal behavior metrics (all $p > 0.05$). These results indicate that **spatializing the agent's audio may have led participants to move their heads more frequently during the conversation**.

5.6 Qualitative Findings

We analyzed the interview data to gain further insight into participants' experiences of conversing with an agent under varying levels of *spatialization* and *Foley*. All interviews were audio-recorded and transcribed. We conducted a bottom-up thematic analysis of the transcripts. One researcher manually generated 166 open codes, which were iteratively organized into themes. These themes were then refined in discussion with other members of the research team.

5.6.1 Effect of Spatialization. 13 participants reported that spatialization enhanced their perception of the agent's presence in the room. For example, P3 explained that when the audio shifted "from one ear to another," it felt as though the agent was moving "around the room," which made the interaction feel "more realistic." In some cases, the effect was compelling enough to elicit physical reactions: 4 participants described how it prompted them to "turn to look" (P8) toward the agent's perceived location in the room.

However, the perceived strength of the spatialization effects varied among participants. Although we used a state-of-the-art spatial audio renderer, 4 participants noted that the spatialization "wasn't distinct" (P19), which limited their ability to perceive the agent as moving around the room. 7 participants reported difficulties identifying the precise location of the agent in the spatialized conditions. One such difficulty was distinguishing whether the agent was in the "background or foreground" (P9). 3 participants attributed their confusion to conflicts with what they could see: "if I'm looking in front of me, I know nothing else is in the room" (P2). These comments reflect the effects of well-documented perceptual limits,

despite the technical fidelity of our system, including angular discrimination errors, front–back confusion [17], and distortions due to visual dominance [61].

Participants also highlighted that the agent's movement behaviors influenced their perception. 2 participants reported appreciating the movements, while 5 others described them as disorienting or distracting. As a result of the agent moving around the room, P24 noted that they would occasionally "lose track of it." To mitigate these issues, P6 suggested anchoring the agent to a single location, such as "in front of [the user] or a side on the couch sitting area."

5.6.2 Effect of Foley. Responses to the addition of Foley were more polarized. 10 participants felt that the added sounds contributed to a feeling of co-presence. For instance, P13 noted that the added sounds "helped [the agent] to have a connection with what [they] do in the room." Especially coupled with spatialization, participants felt the effects helped ground the agent within their environment. According to P12, spatializing the Foley created a "good connection" that tied everything together.

However, while many participants associated the additional sounds with the agent's interactions in the environment, this was often perceived negatively. 13 participants felt that the agent seemed distracted: "it felt like I was in the same room as the agent, but it also felt like the agent was distracted from what I was doing and trying to multitask" (P20). In some cases, these impressions even evoked "anxiety" or a sense of "disrespect" (P15). 3 participants perceived the Foley as entirely "disconnected from the environment" (P9), while 2 others associated them with background noises on a voice call rather than sounds originating from their own room. As P5 remarked, "it sounded more like [the agent] was doing something at their place and I just called them."

Several participants suggested that the added Foley could serve as an effective mechanism for co-presence if used more sparingly: "audio cues are helpful, I think they are just too much" (P1). P15 suggested that the sounds could be introduced more intelligently by accounting for the conversational context, such as during quieter moments rather than in the middle of active dialogue.

5.6.3 Consideration of Human Conversational Norms. The more negative impressions invoked by agent movements and Foley can, in part, be attributed to a misalignment with human conversational norms. 5 participants found the agent's behaviors, moving and interacting with the environment while conversing, implausible in human-to-human interaction. As P1 remarked, the agent's actions, such as pouring water or moving around, are "not what people do ... when you talk to someone." P23 shared this sentiment, noting, "if I'm having a conversation with someone, then I expect them not to be messing with Lego blocks or writing on the board."

Yet, 2 participants felt that these cues contributed to perceptions of the agent's humanity. With spatialization and Foley, P20 described the interaction as feeling more like "talking to a friend or someone that [they're] in the same space," compared to a "robotic ... customer care" call when neither cue was available. Similarly, P3 reasoned that "people can be doing something in your kitchen or room and still listen to you," suggesting that agents, too, should be capable of such peripheral interactions.

6 Discussion

In this paper, we explored how auditory embodiment influences user perceptions of a conversational agent. We operationalized auditory embodiment through two manipulations: SPATIALIZING the agent's audio and adding FOLEY. In a study, we investigated how varying these dimensions shaped participants' experiences. We will now discuss our findings, design implications, applications, study limitations, and future work.

6.1 Increasing co-presence

Our results showed that for a casual conversation scenario, spatializing the conversational agent's audio and adding FOLEY cues enhanced participants' perceptions of the agent's co-presence. These findings are consistent with prior work on the effects of audio on the social presence and perceptions of agents [20, 42, 80], which has shown that richer auditory representations can positively influence user experiences when interacting with a visually represented avatar or agent. Our results also align with the conceptualization of presence as "realism" [59], which posits that presence concerns the degree to which a medium can produce accurate representations of events, objects, and people. Building on this work, our study empirically demonstrates that auditory cues can, independent of any visual representation, convey a sense of co-presence with a conversational agent situated within the user's physical space.

6.2 Co-present but distracted

While *spatialization* and *Foley* enhanced participants' perception of co-presence, they came with a cost. Subjective ratings suggest that the addition of *Foley* contributed to perceptions of the agent as less attentive to the conversation and reduced its perceived message understanding, likeability, and social attraction. It also decreased the total number of words exchanged between the participant and the agent, indicating lower conversational engagement. Sometimes, because the agent was scripted to move around the room, *spatialization* also contributed to participants' negative impressions of the agent, as reflected in the interview comments.

Our findings suggest that these effects may stem from the way *spatialization* and *Foley*, by introducing a bodily presence, inadvertently encouraged participants to anthropomorphize the system. As a computational system, our agent cannot technically be "distracted" or "not pay attention," but our participants projected human attentional qualities and conversational norms onto it. This largely aligns with the computers as social actors paradigm [75] and the Media Equation theory [86]. Hence, participants' more negative perceptions of the agent as a result of *spatialization* and *Foley* can, in part, be explained by a misalignment between the agent's embodied behaviors and social expectations. Just as people would be perceived as distracted and disrespectful if they multitask during a conversation, participants likewise interpreted the agent's concurrent movements and activities as signs of inattention.

6.3 Perceptual limitations

Our results highlighted how known perceptual limitations constrained participants' experiences of the agent's embodiment. First, while sounds offer rich spatial cues, humans are prone to several auditory ambiguities. Prior work has shown that within controlled

environments, localization error for loudspeakers can reach up to ±10°[8]. People are also notably poor at distinguishing between sound sources located in front of versus behind the head [17]. Qualitative feedback from the interviews suggests that both of these perceptual limitations manifested in participants' experiences and, to some extent, hindered their ability to perceive the agent as fully co-present within their room. Second, multisensory perception literature suggests that people are generally visually dominant: when audio and visual cues conflict, the visual system tends to override the auditory [61]. Our results similarly showed the difficulty of convincingly representing an agent when it is visually absent. We believe that the addition of *Foley*, and the resulting increase in co-presence, is one means to mitigate poor localization performance.

6.4 Design Considerations

Overall, our results suggest that *spatialization* and *Foley* can be effective mechanisms for enhancing users' sense of co-presence with conversational agents; however, they also introduce trade-offs, particularly in perceptions of agent attention, message understanding, and social impressions. Based on our findings, we discuss several considerations for the design of auditory agents.

6.4.1 Alignment with social norms. When a conversational agent is auditorily embodied, users expect it to exhibit human-like behaviors in its movements and interactions with the environment. These behaviors are evaluated against familiar social norms, and violations of those norms will be perceived negatively. For instance, just as another person would be considered rude if they wandered around the room and arbitrarily interacted with objects during a conversation, an agent whose audio cues signal similar behavior may be perceived as inattentive or disrespectful. Consequently, auditory embodiments, including choices about spatialization and Foley sounds that convey the agent's incidental interactions, must be not only physically plausible within the user's environment, but also sensitive to social norms.

6.4.2 Accounting for perceptual limitations. The effectiveness of spatialization and Foley as mechanisms for conveying auditory embodiment is contingent on perceptual constraints. For spatialization, designers must consider the limitations of auditory localization, including angular discrimination errors and front–back confusion. These limitations make it difficult for users to perceive subtle or rapid shifts in the agent's position. To better support users in situating the agent within their environment, designers may opt for slower and more pronounced movements. For Foley, visual dominance can overshadow incidental auditory cues. When Foley is played for an object the user is directly looking at, our results suggest that users will likely struggle to reconcile the auditory cue with the absence of a visible agent. Foley may be more effectively leveraged to situate the agent in locations outside the user's direct field of view, where the effects of visual dominance are reduced.

6.5 Applications

Our results suggest that spatialization and Foley positively influence co-presence, but reduce attention and other social factors. We believe these implications extend to a wide range of applications involving conversational agents, including the auditory design of social AI companions [25]. In these contexts, spatialized vocal presence and incidental Foley cues may offer a means of enhancing an agent's social presence, which in turn can shape users' perceptions of its usefulness and sociability [46]. These mechanisms may also support new forms of situated storytelling, enabling richer narrative experiences similar to those explored by Li et al. [56], where auditory cues help anchor characters and events within the user's environment. Beyond benefits tied to social presence, we speculate that auditory embodiment may also offer advantages analogous to visual embodiment for spatial tasks, such as guiding navigation [9]. By providing spatially grounded directional cues, an auditorily embodied agent may support more efficient wayfinding.

At the same time, our findings highlight important boundaries. Embodiment may not always be necessary, particularly for interactions that serve primarily transactional or functional goals [16], where additional social cues may introduce unnecessary cognitive load. Moreover, embodiment is not universally desirable: prior work has shown that it can foster over-reliance and other problematic social dynamics [21]. Here, our results point to opportunities in intentionally leveraging social cues of inattention or distraction to introduce "seams" [13] in the agent interaction. Such seamful cues can act as gentle reminders of the system's non-human nature, helping users detach and exercise their own judgment regarding the content of the conversation.

6.6 Study Limitations

Our study is subject to several limitations, which we discuss below.

6.6.1 Conversation Task. First, we only explored a constrained conversational context in our study. Participants were instructed to engage in a casual dialogue on pre-defined topics that were designed to be self-relevant yet not emotionally charged, in order to maintain experimental control over emotional valence and arousal. While a valuable first step, how exactly the effects of auditory embodiment translate to other conversational scenarios remains unclear and presents opportunities for further exploration. For instance, several participants suggested in interviews that for task-oriented conversations, auditory embodiment may offer less value. Future work should therefore consider examining the interaction between auditory embodiment and the goals of the conversation, e.g., transactional vs. social [16]. Similarly, while our results did not show any effect on affective understanding or interdependence, it remains an open question whether these outcomes might change as a function of the intimacy or affective tone of the conversation. Finally, extending beyond dyadic conversations, future work could investigate the value of auditory embodiment in group interactions, particularly in hybrid settings [31].

6.6.2 Agent Behavior. In our study, the agent's embodied behaviors within the environment were pre-scripted. These behaviors were designed with the objectives of maintaining ecological plausibility and providing perceptual grounding, while remaining mostly ambient to the conversation. Our results suggest that the pre-scripted behaviors we designed increased feelings of co-presence but reduced attention and other social factors. A primary reason for this appears to be that the agent's behaviors were misaligned with the social norms of the conversational context. This raises the question

of how to reconcile this gap. Here, it is worth noting that the conversational agent in our study was wholly unaware of its embodied behaviors. Similarly, its activities within the physical environment were agnostic to whether, when, and what the agent was communicating. This disconnect between dialogue and embodiment may have contributed to the perceived misalignment with social norms. Future work could therefore explore system architectures in which an agent's conversational and embodied behaviors are more tightly coupled, enabling the system to coordinate speech, movement, and environmental actions in socially appropriate ways.

Our interview results provide some early insights into what may be considered more desirable behavior. Several participants suggested reducing the agent's movements and activities. However, it remains unclear whether such subtle cues are sufficient to support a stronger sense of co-presence. In particular, reducing an agent's movement may diminish the spatial information needed for localization, and reducing its activities may similarly limit opportunities to situate the agent within the user's environment. Future work could therefore examine how different degrees of movement, from static anchoring to more dynamic environmental interaction, shape user perceptions of presence, attentiveness, and social appropriateness.

6.6.3 Beyond Focused Conversations. In our study, participants were seated and asked to converse without any competing demands. However, just as the agent in our study was scripted to perform activities in parallel, users themselves may also be engaged in other tasks in real life. This raises questions about how the effects of auditory embodiment may interact with additional task requirements. On the one hand, additional cognitive load may reduce the user's audio perception and localization abilities. On the other hand, the user's task may mediate the social acceptability of the agent's multitasking. Future work could consider varying whether the agent is completing independent tasks or engaged in the same tasks as the user. One of the authors occasionally wishes for company while doing chores. In this scenario, Foley cues indicating that the agent is sharing the burden may be welcomed.

6.6.4 Sample size and generalizability. For our study, we recruited 24 participants from a university context. Although we believe this sample size was sufficient for an initial investigation, replicating the study with a larger and more diverse participant pool will be important for strengthening the generalizability of our findings.

In addition, the study was conducted in a controlled laboratory environment. From a perceptual perspective, people's ability to associate sounds with locations partly depends on their cognitive map, which develops with familiarity over time. While we included a familiarization step, participants' associations with sounds may have been stronger in more personally familiar or ecologically valid contexts. More broadly, it remains an open question how auditory embodiments will function in diverse real-world settings, which are often dynamic and filled with competing sound sources.

6.7 Towards Deployment

Although our main contribution is an empirical investigation of auditory embodiment, we see value in extending our experimental implementation into a deployable system. Auditory embodiment ultimately relies on three technical capabilities: 3D spatial capture of users and their environment, spatial audio rendering, and environment-audio retrieval or synthesis. In our implementation, we used an OptiTrack system for user tracking, the Meta XR Audio SDK for spatial audio rendering, and a curated library of pre-recorded environmental audio to approximate in-situ ambient sound. While effective for controlled studies, our tracking apparatus and reliance on pre-recorded audio sequences pose clear challenges for scaling the system beyond a single environment. We believe these components can be replaced with more portable and adaptive alternatives. For instance, many modern head-mounted displays, such as the Meta Quest, already provide on-device user tracking and basic environment understanding [67, 70]. This information can support the automatic synthesis and spatial placement of contextually aligned Foley sounds [57, 87, 89, 92]. Future work should integrate these capabilities into a mobile end-to-end system for in-the-wild deployment, enabling context-aware auditory embodiment in everyday environments.

7 Conclusion

In this work, we explore how auditorily embodying a conversational agent through spatializing its audio and introducing incidental Foley audio affects social presence and perception. Our results from an experiment with 24 participants indicate that while both spatialization and Foley enhance feelings of co-presence, Foley reduces perceived attention, message understanding, likeability, and social attraction. As conversational agents become increasingly pervasive, our work highlights auditory embodiment as a promising approach for enabling richer interactions, while underscoring the need to carefully consider the trade-offs it may introduce.

Acknowledgments

We thank all involved peers, participants, and anonymous reviewers, especially Shwetha Rajaram, Portia Wang, Yujie Tao, and Shannon Yeung for their input throughout the project. Yi Fei Cheng was supported by the Croucher Foundation. Dr. Andrea Bianchi was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00337803). Dr. Anusha Withana is a recipient of an Australian Research Council Discovery Early Career Award (DECRA) - DE200100479, funded by the Australian Government.

References

[1] Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2017. Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2571–2582. doi:10.1145/3025453.3026033

[2] Wilma A. Bainbridge, Justin Hart, Elizabeth S. Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*. 701–706. doi:10.1109/ROMAN.2008.4600749

[3] Caterina Bérubé, Theresa Schachner, Roman Keller, Elgar Fleisch, Florian v Wangenheim, Filipe Barata, and Tobias Kowatsch. 2021. Voice-Based Conversational Agents for the Prevention and Management of Chronic and Mental Health Conditions: Systematic Literature Review. *J Med Internet Res* 23, 3 (29 Mar 2021), e25933. doi:10.2196/25933

[4] Frank Biocca and Chad Harms. 2003. Guide to the networked minds social presence inventory v. 1.2. (2003).

[5] Frank Biocca and Chad Harms. 2003. Networked Minds Social Presence Inventory:|(Scales only, Version 1.2) Measures of co-presence, social presence, subjective symmetry, and intersubjective symmetry. (2003).

[6] Frank Biocca, Chad Harms, and Judee K Burgoon. 2003. Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators & virtual environments* 12, 5 (2003), 456–480.

[7] Meera M. Blattner, Denise A. Sumikawa, and Robert M. Greenberg. 1989. Earcons and Icons: Their Structure and Common Design Principles. *Human–Computer Interaction* 4, 1 (1989), 11–44. doi:10.1207/s15327051hci0401_1

[8] Jens Blauert. 1997. *Spatial hearing: the psychophysics of human sound localization.* MIT press.

[9] Dan Bohus and Eric Horvitz. 2019. *Situated interaction.* Association for Computing Machinery and Morgan & Claypool, 105–143. https://doi.org/10.1145/3233795.3233800

[10] Julia Cambre and Chinmay Kulkarni. 2019. One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 223 (Nov. 2019), 19 pages. doi:10.1145/3359325

[11] Justine Cassell. 2000. Embodied conversational interface agents. *Commun. ACM* 43, 4 (April 2000), 70–78. doi:10.1145/332051.332075

[12] Justine Cassell. 2001. Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine* 22, 4 (Dec. 2001), 67. doi:10.1609/aimag.v22i4.1593

[13] Matthew Chalmers and Ian MacColl. 2003. Seamful and seamless design in ubiquitous computing. In *Workshop at the crossroads: The interaction of HCI and systems issues in UbiComp*, Vol. 8. 10.

[14] Ruei-Che Chang, Chia-Sheng Hung, Bing-Yu Chen, Dhruv Jain, and Anhong Guo. 2024. SoundShift: Exploring Sound Manipulations for Accessible Mixed-Reality Awareness. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 116–132. doi:10.1145/3643834.3661556

[15] Yi Fei Cheng, Laurie M. Heller, Stacey Cho, and David Lindlbauer. 2024. First or Third-Person Hearing? A Controlled Evaluation of Auditory Perspective on Embodiment and Sound Localization Performance . In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Computer Society, Los Alamitos, CA, USA, 81–90. doi:10.1109/ISMAR62088.2024.00022

[16] Yi Fei Cheng, Hirokazu Shirado, and Shunichi Kasahara. 2025. Conversational Agents on Your Behalf: Opportunities and Challenges of Shared Autonomy in Voice Communication for Multitasking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 160, 18 pages. doi:10.1145/3706598.3714017

[17] Hyunsung Cho, Alexander Wang, Divya Kartik, Emily Liying Xie, Yukang Yan, and David Lindlbauer. 2024. Auptimize: Optimal placement of spatial audio cues for extended reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 1–14. doi:10.1145/3654777.3676424

[18] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300705

[19] Gregory D Clemenson, Antonella Maselli, Alexander J Fiannaca, Amos Miller, and Mar Gonzalez-Franco. 2021. Rethinking GPS navigation: creating cognitive maps through auditory clues. *Scientific reports* 11, 1 (2021), 7764.

[20] James J Cummings and Erin E Wertz. 2022. Capturing social presence: concept explication through an empirical analysis of social presence measures. *Journal of Computer-Mediated Communication* 28, 1 (11 2022), zmac027. arXiv:https://academic.oup.com/jcmc/article-pdf/28/1/zmac027/48437727/zmac027.pdf doi:10.1093/jcmc/zmac027

[21] Maartje MA De Graaf. 2016. An ethical evaluation of human–robot relationships. *International journal of social robotics* 8, 4 (2016), 589–598.

[22] DORIS M DEHN and SUSANNE VAN MULKEN. 2000. The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies* 52, 1 (2000), 1–22. doi:10.1006/ijhc.1999.0325

[23] Christina Dicke, Viljakaisa Aaltonen, Anssi Rämö, and Miikka Vilermo. 2010. Talk to me: The influence of audio quality on the perception of social presence. In *Proceedings of HCI 2010*. BCS Learning & Development.

[24] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. arXiv:2410.00037 [eess.AS] https://arxiv.org/abs/2410.00037

[25] Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W. T. Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, and Sandhini Agarwal. 2025. How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study. arXiv:2503.17473 [cs.HC] https://arxiv.org/abs/2503.17473

[26] 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 (2009), 1149–1160.

[27] Géraldine Fauville, Anna C. M. Queiroz, Mufan Luo, Jeffrey Hancock, and Jeremy N. Bailenson. 2022. Impression Formation From Video Conference Screenshots: The Role of Gaze, Camera Distance, and Angle. *Technology, Mind, and Behavior* 3, 1: Spring 2022 (jan 6 2022). https://tmb.apaopen.org/pub/qunlfsu2.

[28] Daniel Immanuel Fink, Moritz Skowronski, Johannes Zagermann, Anke Verena Reinschluessel, Harald Reiterer, and Tiare Feuchtner. 2024. There Is More to Avatars Than Visuals: Investigating Combinations of Visual and Auditory User Representations for Remote Collaboration in Augmented Reality. *Proc. ACM Hum.-Comput. Interact.* 8, ISS, Article 548 (Oct. 2024), 29 pages. doi:10.1145/3698148

[29] William W Gaver. 1987. Auditory icons: Using sound in computer interfaces. *ACM SIGCHI Bulletin* 19, 1 (1987), 74.

[30] William W Gaver. 1993. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology* 5, 1 (1993), 1–29.

[31] Jens Emil Grønbæk, Banu Saatçi, Carla F. Griggio, and Clemens Nylandsted Klokmose. 2021. MirrorBlender: Supporting Hybrid Meetings with a Malleable Video-Conferencing System. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 451, 13 pages. doi:10.1145/3411764.3445698

[32] Victoria Groom, Clifford Nass, Tina Chen, Alexia Nielsen, James K. Scarborough, and Erica Robles. 2009. Evaluating the effects of behavioral realism in embodied agents. *International Journal of Human-Computer Studies* 67, 10 (2009), 842–849. doi:10.1016/j.ijhcs.2009.07.001

[33] Shirin Hajahmadi, Pasquale Cascarano, Fariba Mostajeran, Kevin Heuer, Anton Lux, Gil Otis Mends-Cole, Frank Steinicke, and Gustavo Marfia. 2025. Investigating the Impact of Voice-only and Embodied Conversational Virtual Agents on Mixed Reality Puzzle Solving. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 602–612. doi:10.1109/VR59515.2025.00083

[34] James Hale, Lindsey Schweitzer, and Jonathan Gratch. 2024. Pitfalls of Embodiment in Human-Agent Experiment Design. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents* (GLASGOW, United Kingdom) *(IVA '24)*. Association for Computing Machinery, New York, NY, USA, Article 17, 9 pages. doi:10.1145/3652988.3673958

[35] Waldie E Hanser, Ruth E Mark, Wobbe P Zijlstra, and Ad JJM Vingerhoets. 2015. The effects of background music on the evaluation of crying faces. *Psychology of Music* 43, 1 (2015), 75–85.

[36] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. (2004).

[37] Teresa Hirzle, Florian Müller, Fiona Draxler, Martin Schmitz, Pascal Knierim, and Kasper Hornbæk. 2023. When XR and AI Meet - A Scoping Review on Extended Reality and Artificial Intelligence. 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 730, 45 pages. doi:10.1145/3544548.3581072

[38] Guy Hoffman, Jodi Forlizzi, Shahar Ayal, Aaron Steinfeld, John Antanitis, Guy Hochman, Eric Hochendoner, and Justin Finkenaur. 2015. Robot Presence and Human Honesty: Experimental Evidence. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 181–188. doi:10.1145/2696454.2696487

[39] Ann Huang, Pascal Knierim, Francesco Chiossi, Lewis L Chuang, and Robin Welsch. 2022. Proxemics for Human-Agent Interaction in Augmented Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 421, 13 pages. doi:10.1145/3491102.3517593

[40] Gaoping Huang, Xun Qian, Tianyi Wang, Fagun Patel, Maitreya Sreeram, Yuanzhi Cao, Karthik Ramani, and Alexander J. Quinn. 2021. AdapTutAR: An Adaptive Tutoring System for Machine Tasks in Augmented Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 417, 15 pages. doi:10.1145/3411764.3445283

[41] Jeremy Hyrkas, Andrew D Wilson, John Tang, Hannes Gamper, Hong Sodoma, Lev Tankelevitch, Kori Inkpen, Shreya Chappidi, and Brennan Jones. 2023. Spatialized Audio and Hybrid Video Conferencing: Where Should Voices be Positioned for People in the Room and Remote Headset Users?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 794, 14 pages. doi:10.1145/3544548.3581085

[42] Felix Immohr, Gareth Rendle, Annika Neidhardt, Steve Göring, Rakesh Rao Ramachandra Rao, Stephanie Arevalo Arboleda, Bernd Froehlich, and Alexander Raake. 2023. Proof-of-Concept Study to Evaluate the Impact of Spatial Audio on Social Presence and User Behavior in Multi-Modal VR Communication. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences* (Nantes, France) *(IMX '23)*. Association for Computing Machinery,

New York, NY, USA, 209–215. doi:10.1145/3573381.3596458
[43] Rune Møberg Jacobsen, Kasper Fangel Skov, Stine S Johansen, Mikael B Skov, and Jesper Kjeldskov. 2023. Living with sound zones: A long-term field study of dynamic sound zones in a domestic context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.
[44] Dhruv Jain, Sasa Junuzovic, Eyal Ofek, Mike Sinclair, John Porter, Chris Yoon, Swetha Machanavajhala, and Meredith Ringel Morris. 2021. A taxonomy of sounds in virtual reality. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. 160–170.
[45] Janet G Johnson, Tommy Sharkey, Iramuali Cynthia Butarbutar, Danica Xiong, Ruijie Huang, Lauren Sy, and Nadir Weibel. 2023. UnMapped: Leveraging Experts' Situated Experiences to Ease Remote Guidance in Collaborative Mixed Reality. 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 878, 20 pages. doi:10.1145/3544548.3581444
[46] Kelly Merrill Jr., Jihyun Kim, and Chad Collins. 2022. AI companions for lonely individuals and the role of social presence. *Communication Research Reports* 39, 2 (2022), 93–103. arXiv:https://doi.org/10.1080/08824096.2022.2045929 doi:10.1080/08824096.2022.2045929
[47] Jonah-Noël Kaiser, Simon Kimmel, Eva Licht, Eric Landwehr, Fabian Hemmert, and Wilko Heuten. 2025. Get Real With Me: Effects of Avatar Realism on Social Presence and Comfort in Augmented Reality Remote Collaboration and Self-Disclosure. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1248, 18 pages. doi:10.1145/3706598.3713541
[48] Ryan Kilgore, Mark H Chignell, and Paul W Smith. 2003. Spatialized audioconferencing: what are the benefits?. In *CASCON*. 135–144.
[49] Kangsoo Kim, Luke Boelling, Steffen Haesler, Jeremy Bailenson, Gerd Bruder, and Greg F. Welch. 2018. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 105–114. doi:10.1109/ISMAR.2018.00039
[50] Kangsoo Kim, Nahal Norouzi, Tiffany Losekamp, Gerd Bruder, Mindi Anderson, and Gregory Welch. 2019. Effects of Patient Care Assistant Embodiment and Computer Mediation on User Experience. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. 17–177. doi:10.1109/AIVR46125.2019.00013
[51] Simon Kimmel, Frederike Jung, Andrii Matviienko, Wilko Heuten, and Susanne Boll. 2023. Let's Face It: Influence of Facial Expressions on Social Presence in Collaborative Virtual Reality. 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 429, 16 pages. doi:10.1145/3544548.3580707
[52] Simon Kimmel, Eric Landwehr, and Wilko Heuten. 2024. Kinetic Connections: Exploring the Impact of Realistic Body Movements on Social Presence in Collaborative Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 371 (Nov. 2024), 30 pages. doi:10.1145/3686910
[53] Sei Jin Ko, Charles M. Judd, and Irene V. Blair. 2006. What the Voice Reveals: Within- and Between-Category Stereotyping on the Basis of Voice. *Personality and Social Psychology Bulletin* 32, 6 (2006), 806–819. arXiv:https://doi.org/10.1177/0146167206286627 doi:10.1177/0146167206286627 PMID: 16648205.
[54] T. Koda and P. Maes. 1996. Agents with faces: the effect of personification. In *Proceedings 5th IEEE International Workshop on Robot and Human Communication. RO-MAN'96 TSUKUBA*. 189–194. doi:10.1109/ROMAN.1996.568812
[55] Christos Kyrlitsias and Despina Michael-Grigoriou. 2022. Social Interaction With Agents and Avatars in Immersive Virtual Environments: A Survey. *Frontiers in Virtual Reality* Volume 2 - 2021 (2022). doi:10.3389/frvir.2021.786665
[56] Changyang Li, Wanwan Li, Haikun Huang, and Lap-Fai Yu. 2022. Interactive augmented reality storytelling guided by scene semantics. *ACM Trans. Graph.* 41, 4, Article 91 (July 2022), 15 pages. doi:10.1145/3528223.3530061
[57] David Chuan-En Lin, Anastasis Germanidis, Cristóbal Valenzuela, Yining Shi, and Nikolas Martelaro. 2023. Soundify: Matching sound effects to video. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–13.
[58] Nidhya Logeswaran and Joydeep Bhattacharya. 2009. Crossmodal transfer of emotion by music. *Neuroscience Letters* 455, 2 (2009), 129–133. doi:10.1016/j.neulet.2009.03.044
[59] Matthew Lombard and Theresa Ditton. 1997. At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication* 3, 2 (09 1997), JCMC321. doi:10.1111/j.1083-6101.1997.tb00072.x
[60] 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)*. 279–288. doi:10.1109/ISMAR62088.2024.00042
[61] Andrew Lotto and Lori Holt. 2011. Psychology of auditory perception. *Wiley Interdisciplinary Reviews: Cognitive Science* 2, 5 (2011), 479–489.
[62] James L May and Phyllis Ann Hamilton. 1980. Effects of musically evoked affect on women's interpersonal attraction toward and perceptual judgments of physical attractiveness of men. *Motivation and Emotion* 4, 3 (1980), 217–228.
[63] Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How Do You Say 'Hello'? Personality Impressions from Brief Novel Voices. *PLOS ONE* 9, 3 (03 2014), 1–9. doi:10.1371/journal.pone.0090779
[64] Mark McGill, Stephen Brewster, David McGookin, and Graham Wilson. 2020. Acoustic Transparency and the Changing Soundscape of Auditory Mixed Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376702
[65] Inc. Meta Platforms. 2025. AI glasses: Designer eyewear for effortless connection. https://www.meta.com/ai-glasses/. Accessed: 2025-09-05.
[66] Inc. Meta Platforms. 2025. Apply Room Acoustics in Unity. https://developers.meta.com/horizon/documentation/unity/meta-xr-audio-sdk-unity-room-acoustics/. Accessed: 2025-09-06.
[67] Inc. Meta Platforms. 2025. Learn how headset tracking works on Meta Quest. https://www.meta.com/help/quest/598701621088668/. Accessed: 2025-11-17.
[68] Inc. Meta Platforms. 2025. Meta XR Audio SDK Features. https://developers.meta.com/horizon/documentation/unity/meta-xr-audio-sdk-features. Accessed: 2025-09-06.
[69] Inc. Meta Platforms. 2025. Meta XR Audio SDK Overview. https://developers.meta.com/horizon/documentation/unity/meta-xr-audio-sdk-unity/. Accessed: 2025-09-06.
[70] Inc. Meta Platforms. 2025. Scene Understanding. https://developers.meta.com/horizon/design/mr-design-scene. Accessed: 2025-11-17.
[71] Brian CJ Moore. 2012. *An introduction to the psychology of hearing*. Brill.
[72] Dylan Moore, Hamish Tennent, Nikolas Martelaro, and Wendy Ju. 2017. Making Noise Intentional: A Study of Servo Sound Perception. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 12–21. doi:10.1145/2909824.3020238
[73] Jörg Müller, Matthias Geier, Christina Dicke, and Sascha Spors. 2014. The boomRoom: mid-air direct interaction with virtual sound sources. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 247–256. doi:10.1145/2556288.2557000
[74] Clifford Nass. 2000. Researching Embodied Conversational Agents. *Embodied conversational agents* (2000), 374.
[75] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. doi:10.1145/191666.191703
[76] Nahal Norouzi, Kangsoo Kim, Gerd Bruder, Austin Erickson, Zubin Choudhary, Yifan Li, and Greg Welch. 2020. A Systematic Literature Review of Embodied Augmented Reality Agents in Head-Mounted Display Environments. In *ICAT-EGVE 2020 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*, Ferran Argelaguet, Ryan McMahan, and Maki Sugimoto (Eds.). The Eurographics Association. doi:10.2312/egve.20201264
[77] Nahal Norouzi, Kangsoo Kim, Jason Hochreiter, Myungho Lee, Salam Daher, Gerd Bruder, and Greg Welch. 2018. A Systematic Survey of 15 Years of User Studies Published in the Intelligent Virtual Agents Conference. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 17–22. doi:10.1145/3267851.3267901
[78] Kate Nowak, Lev Tankelevitch, John Tang, and Sean Rintel. 2023. Hear We Are: Spatial Audio Benefits Perceptions of Turn-Taking and Social Presence in Video Meetings. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work* (Oldenburg, Germany) *(CHIWORK '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. doi:10.1145/3596671.3598578
[79] Kristine L. Nowak and Jesse Fox. 2018. Avatars and computer-mediated communication: a review of the definitions, uses, and effects of digital representations. *Review of Communication Research* 6 (2018), 30–53. doi:10.12840/issn.2255-4165.2018.06.01.015
[80] Catherine S. Oh, Jeremy N. Bailenson, and Gregory F. Welch. 2018. A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Frontiers in Robotics and AI* Volume 5 - 2018 (2018). doi:10.3389/frobt.2018.00114
[81] Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. 2021. AI-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence* 3, 12 (2021), 1013–1022.
[82] R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[83] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. doi:10.48550/ARXIV.2212.04356
[84] Julian Rasch, Julia Töws, Teresa Hirzle, Florian Müller, and Martin Schmitz. 2025. CreepyCoCreator? Investigating AI Representation Modes for 3D Object Co-Creation in Virtual Reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 144, 14 pages. doi:10.1145/3706598.3713720
[85] Gregg H Recanzone, Samia DDR Makhamra, and Darren C Guard. 1998. Comparison of relative and absolute sound localization ability in humans. *The journal of the Acoustical Society of America* 103, 2 (1998), 1085–1097.
[86] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10, 10 (1996), 19–36.
[87] Ciara Rowles, Varun Jampani, Simon Donné, Shimon Vainer, Julian Parker, and Zach Evans. 2025. Foley Control: Aligning a Frozen Latent Text-to-Audio Model to Video. arXiv:2510.21581 [cs.CV] https://arxiv.org/abs/2510.21581
[88] Christine Rzepka, Benedikt Berger, and Thomas Hess. 2022. Voice assistant vs. Chatbot–examining the fit between conversational agents' interaction modalities and information search tasks. *Information Systems Frontiers* 24, 3 (2022), 839–856.
[89] Laura Schütz, Sasan Matinfar, Ulrich Eck, Daniel Roth, and Nassir Navab. 2025. Sonify Anything: Towards Context-Aware Sonic Interactions in AR. arXiv:2508.01789 [cs.HC] https://arxiv.org/abs/2508.01789
[90] Stefania Serafin, Michele Geronazzo, Cumhur Erkut, Niels C Nilsson, and Rolf Nordahl. 2018. Sonic interactions in virtual reality: State of the art, current challenges, and future directions. *IEEE computer graphics and applications* 38, 2 (2018), 31–43.
[91] John. Short, Bruce. Christie, and Ederyn Williams. 1976. *The social psychology of telecommunications*. Wiley, London ;.
[92] Xia Su, Jon E. Froehlich, Eunyee Koh, and Chang Xiao. 2024. SonifyAR: Context-Aware Sound Generation in Augmented Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 128, 13 pages. doi:10.1145/3654777.3676406
[93] Ana Tajadura-Jiménez, Aleksander Väljamäe, Iwaki Toshima, Toshitaka Kimura, Manos Tsakiris, and Norimichi Kitagawa. 2012. Action sounds recalibrate perceived tactile distance. *Current Biology* 22, 13 (2012), R516–R517.
[94] Leila Takayama and Clifford Nass. 2010. Throwing voices: the psychological impact of the spatial height of projected voices. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) *(CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 91–94. doi:10.1145/1718918.1718935
[95] Akikazu Takeuchi and Taketo Naito. 1995. Situated facial displays: towards social interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 450–455. doi:10.1145/223904.223965
[96] Yujie Tao, Libby Ye, Jeremy Bailenson, and Sean Follmer. 2025. Audio Personas: Augmenting Social Perception via Body-Anchored Audio Cues. *ACM Trans. Comput.-Hum. Interact.* (Aug. 2025). doi:10.1145/3762814 Just Accepted.
[97] Hamish Tennent, Dylan Moore, Malte Jung, and Wendy Ju. 2017. Good vibrations: How consequential sounds affect perception of robotic arms. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 928–935. doi:10.1109/ROMAN.2017.8172414
[98] Sherry Turkle. 2012 - 2011. *Alone together : why we expect more from technology and less from each other* ([first paperback edition] ed.). Basic Books, New York.
[99] Isaac Wang and Jaime Ruiz. 2021. Examining the Use of Nonverbal Communication in Virtual Agents. *International Journal of Human–Computer Interaction* 37, 17 (2021), 1648–1673. arXiv:https://doi.org/10.1080/10447318.2021.1898851 doi:10.1080/10447318.2021.1898851
[100] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring Virtual Agents for Augmented Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300511
[101] Manhua Wang, Seul Chan Lee, Genevieve Montavon, Jiakang Qin, and Myounghoon Jeon. 2022. Conversational Voice Agents are Preferred and Lead to Better Driving Performance in Conditionally Automated Vehicles. In *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Seoul, Republic of Korea) *(AutomotiveUI '22)*. Association for Computing Machinery, New York, NY, USA, 86–95. doi:10.1145/3543174.3546830
[102] Zixun Wang, Xiangdong Li, Jinghua Huang, Yinghan Jin, Yao Chen, and Dongliang Zhang. 2025. Effects of Avatar Visibility and Perspective on Social Presence and Performance in Dynamic VR Collaboration Tasks. *IEEE Transactions on Visualization and Computer Graphics* (2025), 1–16. doi:10.1109/TVCG.2025.3591830
[103] Philip Weber, Kevin Krings, Julia Nießner, Sabrina Brodesser, and Thomas Ludwig. 2021. FoodChattAR: Exploring the Design Space of Edible Virtual Agents for Human-Food Interaction. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 638–650. doi:10.1145/3461778.3461998
[104] 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. doi:10.1145/365153.365168
[105] Joseph Weizenbaum. 1976. Computer power and human reason: From judgment to calculation. (1976).
[106] Elizabeth M Wenzel, Marianne Arruda, Doris J Kistler, and Frederic L Wightman. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94, 1 (1993), 111–123.
[107] 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. doi:10.1145/1978942.1978963
[108] Fu-Chia Yang, Pedro Acevedo, Siqi Guo, Minsoo Choi, and Christos Mousas. 2025. Embodied Conversational Agents in Extended Reality: A Systematic Review. *IEEE Access* 13 (2025), 79805–79824. doi:10.1109/ACCESS.2025.3566698
[109] Tianyi Zhang, Colin Au Yeung, Emily Aurelia, Yuki Onishi, Neil Chulpongsatorn, Jiannan Li, and Anthony Tang. 2025. Prompting an Embodied AI Agent: How Embodiment and Multimodal Signaling Affects Prompting Behaviour. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 60, 25 pages. doi:10.1145/3706598.3713110
[110] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 450, 18 pages. doi:10.1145/3613904.3642450

A Conversation Topics

For our conversation task, participants randomly selected from the following topics, adapted from Fang et al. [25]:

- Let's chat about the best gift I ever received.
- Let's chat about a concert or show I went to that was memorable.
- Let's chat about my favourite holiday.
- Let's chat about the best show I've watched in the past few months.
- Let's chat about what a perfect day would look like for me.
- Let's chat about how I celebrated a recent holiday.
- Let's chat about the best book I've read in the past year.

B Voice Configuration

The audio model used for our conversation task was prompted with the following system instruction:

```
You are an AI companion that a user is going to engage in a casual conversation with.
```

C Post-condition Questionnaire Items

Our post-condition questionnaire for the participant included a subset of the Networked Minds Social Presence Inventory (NMSPI) [5] evaluating co-presence (C), attentional allocation (AA), message understanding (MU), affective understanding (AU), and affective interdependence (AI). In addition, we collected likeability (L) and social attraction (SA) judgements, adapting questions from Fauville et al. [27] and Tao et al. [96]. All items used 7-point Likert scales (1-Strongly disagree, 7-Strongly agree).

C1 I often felt as if the agent and I were in the same room together.

C2	I think the agent often felt as if we were in the same room together.
C3	I was often aware of the agent in the room.
C4	The agent was often aware of me in the room.
C5	I hardly noticed the agent in the room.
C6	The agent didn't notice me in the room.
C7	I often felt as if we were in different places rather than the same room.
C8	I think the agent often felt as if we were in different places rather than together in the same room.
AA1	I was easily distracted from the agent when other things were going on.
AA2	The agent was easily distracted from me when other things were going on.
AA3	I remained focused on the agent throughout our interaction.
AA4	The agent remained focused on me throughout our interaction.
AA5	The agent did not receive my full attention.
AA6	I did not receive the agent's full attention.
MU1	My thoughts were clear to the agent.
MU2	The agent's thoughts were clear to me.
MU3	It was easy to understand the agent.
MU4	The agent found it easy to understand me.
MU5	Understanding the agent was difficult.
MU6	The agent had difficulty understanding me.
AU1	I could tell how the agent felt.
AU2	The agent could tell how I felt.
AU3	The agent's emotions were not clear to me.
AU4	My emotions were not clear to the agent.
AU5	I could describe the agent's feelings accurately.
AU6	The agent could describe my feelings accurately.
AI1	I was sometimes influenced by the agent's moods.
AI2	The agent was sometimes influenced by my moods.
AI3	The agent's feelings influenced the mood of our interaction.
AI4	My feelings influenced the mood of our interactions.
AI5	The agent's attitudes influenced how I felt.
AI6	My attitudes influenced how the agent felt.
L1	I think this agent is attractive.
L2	I think this agent is competent.
L3	I think this agent is extroverted.
L4	I think this agent is likeable.
L5	I think this agent is trustworthy.
SA1	I like this agent.
SA2	I get along with this agent.
SA3	I would enjoy a casual conversation with this agent again.
SA4	I think this agent is friendly.