Audiovisual Realism in MR: Investigating the Effects of Room Acoustics on Co-Presence with Photorealistic Avatars

Cyan DeVeaux*
Stanford University, USA

Elizabeth H. Hall†
Meta Reality Labs Research, USA

Andy J. Shaw‡
Meta Reality Labs Research, USA

Andrew Frederick Francl§
Meta Reality Labs Research, USA

Paulus van Horne¶
Meta Reality Labs Research, USA

Frank M. Nieuwenhuizen‖
Meta Reality Labs Research, USA

Sebastià V. Amengual Garí**
Meta Reality Labs Research, USA

Madeline Huberth††
Meta Reality Labs Research, USA

Figure 1: An overview of remote, avatar-mediated conversation in mixed reality.

Abstract

With advances in spatial computing and growth in commercially available head-mounted displays, mixed reality (MR) is an emerging context for social experiences and collaborative interaction. While the role of visual representations in enhancing user experiences has been extensively studied, the contribution of audio has comparatively received little attention. In this exploratory, within-subjects study, we investigate the effects of audio quality on co-presence and associated experiential outcomes during avatar-mediated conversations in MR. In dyads, participants engaged in semi-structured conversation under anechoic and reverberant audio conditions while embodying photorealistic avatars. Results shed a nuanced light on the potential of audio in facilitating co-presence. We conclude by discussing implications for design and future research on audio within audiovisual MR environments.

Index Terms: audio, room acoustics, mixed reality, avatars.

*e-mail: cyanjd@stanford.edu
†e-mail: ehlhall1@gmail.com
‡e-mail: andyjshaw000@gmail.com
§e-mail: francl@meta.com
¶e-mail: paulusvh@meta.com
‖e-mail: franknieu@meta.com
**e-mail: samengual@meta.com
††e-mail: huberth@meta.com

1 Introduction

Over the past decade, mixed reality (MR), an immersive medium that combines virtual content with physical environments, has undergone technological advances and become increasingly accessible to the broader public. As a result, both industry and academia have shown a growing interest in its use for remote communication and collaboration [48]. Through shared spatial references and dynamic social stimuli, MR has the potential to help people in different locations feel as though they are together in the same environment. Accordingly, fostering co-presence, or the feeling of "being there" together [36], is a central goal of collaborative MR experiences.

Audio cues, such as the sound of voices, have the potential to improve the perceptual realism of MR environments [50]. Yet, despite the range of audio rendering techniques that exist, there is little research exploring how this crucial aspect of social MR, and immersive technologies more broadly, influences co-presence [34, 3]. Although prior work has examined the effects of specific aspects of spatial audio [14, 71, 53], comparatively less attention has been paid to the role of room acoustics in this medium. This gap is further complicated by the fact that most research on realism and presence in immersive contexts has focused on fully virtual settings, such as VR [18]. In MR, however, virtual audio is experienced alongside perceptual cues from the physical world, giving audio realism added significance. Mismatches between virtual acoustics and a user's physical environment in MR have been shown to cause a room divergence effect where audio externalization, or the feeling that audio is coming from an outside source, is decreased, causing sounds to be perceived as coming from inside of the listener's head [69]. Accordingly, prior work has emphasized perceptual realism,

perceptual integration with the physical environment, and indistinguishability from its real-world acoustic referent—often framed as plausibility, transfer-plausibility, and authenticity—as important dimensions for evaluating virtual audio in MR [41].

In this paper, we address this gap by investigating how room acoustics shape co-presence during conversations with realistic avatars in MR. In this exploratory, within-subjects study, sixteen dyads engaged in semi-structured conversations under two room acoustic conditions: anechoic and reverberant. Binaural (two-ear), spatialized audio was provided over built-in headset speakers that supported six degrees of freedom (6DoF), allowing participants' head movements to be tracked in three-dimensional space, including position (forward/backward, up/down, left/right) and orientation (pitch, yaw, roll). After engaging in remote MR conversations under each audio condition, participants provided subjective evaluations of co-presence and related experiential outcomes, including co-location, interaction believability, humanness, satisfaction, and perceived turn-taking. They also participated in a group interview, through which we gathered qualitative insights about their experience. Results offer a nuanced understanding of the potential for room acoustics to facilitate co-presence. Specifically, we present preliminary evidence to suggest that who, what, and where could influence when room acoustics best support this experiential outcome. We use these findings to present audio design implications for future collaborative MR interfaces. Moreover, we present our experimental protocol and the insights gained from it as a contribution to guide future audio research in audiovisual MR environments.

2 Related Work

2.1 Defining Co-Presence and Realism

Co-presence is characterized by the sense of being there together in a shared environment [45, 58]. Users experience co-location as a mutual awareness that they are "accessible, available, and subject to one another" [17, 36]. Prior work has frequently conflated co-presence with social presence, the "psychological state that virtual (para-authentic or artificial) actors are experienced as actual social actors" [36, 45]. Though these constructs are interrelated, social presence emphasizes how people perceive the medium's capacity to foster social interaction, whereas co-presence emphasizes co-location and mutual perception of each other [43]. We draw from both social presence and co-presence literature to motivate this work, given their close association and frequent interchangeable use in prior research [45]. However, the present study focuses on co-presence as the primary outcome of interest.

Achieving co-presence is a central goal for computer-mediated communication systems, such as collaborative MR, to improve communication outcomes between remote users [45, 72]. For example, co-presence has been positively linked to trust, enjoyment, and attraction [15, 23]. In its absence, interlocutors can be perceived as artificial [36, 45]. MR remote collaboration systems have aimed to enhance co-presence and related metrics (e.g., social presence and social realism) through rendering shared spatial references and dynamic social stimuli [9, 21, 49].

Realism is another metric of presence defined as whether "a medium can produce seemingly accurate representations of objects, events, and people—representations that look, sound, and/or feel like the "real" thing" [38]. When mediated environments are perceived as real, it can encourage users to behave in realistic ways [64, 74]. Although not all virtual environments strive to create experiences that promote natural behavior [40], realism is relevant in contexts such as the workplace, education, and simulation-based training for real-world scenarios [31]. Furthermore, avatar realism has been highlighted as an important factor for the use of digital representations in the workplace [47].

Most research on realism in immersive virtual environments is drawn from VR [18], but in MR, where elements of the real world are integrated into the scene, realism could take on additional significance. DeMarbre et al. [11] posited that visual coherence of virtual content in MR environments can contribute to presence. We extend this literature by exploring how the interplay of acoustic realism and avatar realism affects co-presence and other aspects of user experience in this emerging medium.

2.2 Spatial Audio and Room Acoustics in Virtual Environments

Audio plays a critical role in the creation of immersive virtual environments. Although presence and MR/VR research has predominantly emphasized visual cues, there is a growing thread of research examining audio's influence in this domain [24, 30, 35]. In particular, spatialization and reverberation of sound are two key auditory elements that can enhance the immersive quality of virtual experiences.

Spatial audio incorporates localization acoustic cues that lead users to perceive sound as coming from specific locations in a three-dimensional environment. In addition to enhancing feelings of presence within a virtual environment [24], spatial audio can also reduce the mental effort needed to process social information [61]. Compared to monaural audio, which has no directional cues, spatial audio can help direct attention to social cues, improving the identification and comprehension of co-located speakers [44, 73]. Additionally, spatial audio has been associated with improved memory, increased perception of turn-taking, and greater social connectedness [44, 46].

Room acoustics is a dimension of audio that allows virtual sound to feel as though it is a part of a listener's environment, thereby enhancing plausibility, or the perceived realism of the sound. This is achieved by modeling the behavior of sound waves as they interact with a physical or virtual space, including reflections, surface absorption, and reverberation, or the persistence of sound after it produced due to reflections [7]. Incorporating room acoustics can improve the perception of externalization, where sound is felt as though it is coming from outside of a listener's head. Specifically, reverberant audio has been associated with greater externalization compared to anechoic audio, or audio lacking reverberation and echoes [2]. Past research also suggests that adding room acoustics to virtual environments can lead to greater presence [35] and enhance the plausibility of auditory augmented reality [42, 2]. Despite the aforementioned value of room acoustics in virtual environments, understanding its effects is an underexplored area of research compared to spatial audio.

Research on audio quality, co-presence, and social presence is mixed. A study by Immohr et al. [27] had participants engage in a free conversation task in-person and in immersive VR under spatialized and non-spatialized audio conditions and found no significant differences in social presence. Similarly, Roßkopf et al. [54] did not observe significant differences in social presence when comparing binaural audio to a loudspeaker in VR. At the same time, there exists evidence to suggest that audio quality may improve co-presence [13]. For example, Shin et al. [60] found that 3D sound in a pre-recorded 360-video of a live concert contributed to higher social presence scores compared to 2D sound. Nowak et al. [44] observed differences in social presence among women between spatial and non-spatial audio during a video conference call.

Altogether, this body of research underscores the potential of audio to enhance social experiences, while also revealing an opportunity for further study in understanding its role in fostering co-presence within collaborative MR environments. Building on this foundation, we explore this phenomenon by comparing the experience of an anechoic audio environment to a reverberant one.

2.3 Crossmodal Audiovisual Interaction

In contrast to audio-only technologies, audiovisual environments exist at the intersection of two primary sensory modalities. While both auditory and visual cues enhance immersion [51] and provide distinct perceptual information, the senses that they engage are connected and can have an influence on each other. For example, visuals can affect the spatial processing of sound (i.e., spatial ventriloquism) and sound can affect the temporal processing of visuals (i.e., temporal ventriloquism) [65]. Hence, the cross-modal influences of audiovisual stimuli can have an impact the quality of user experiences [65].

There is evidence to suggest that audio quality can impact the perceived quality of visual fidelity [63, 29] and vice versa [57]. Storms and Zyda [63] investigated the effects of audiovisual interactions on a computer monitor by manipulating the quality of both the visual and auditory displays. They found that pairing high-quality audio with high-quality visuals improved the perceived visual fidelity compared to visual-only presentations. In contrast, pairing low-quality audio with high-quality video reduced perceptions of audio quality compared to audio-only presentations. Similarly, Schmitt et al. [57] reported that diminishing visual fidelity in a video call also lowered the perceived quality of the audio stream. Joly et al. [29] documented the advantages of high-quality audio in television, showing that lower-quality video paired with non-degraded audio was perceived as higher quality than when audio was absent. Collectively, this work suggests that while low-fidelity visuals can impair perceptions of audiovisual quality, high-fidelity audio can enhance the perception of this combined experience.

It is also possible the visual representation of an avatar could influence expectations for how an avatar sounds. The greater the photorealism of an avatar, the higher the expectation is for behavioral realism, or how naturalistic an avatar behaves [25]. Whereas previous research has primarily investigated behavioral realism through the nonverbal capacities of avatars (e.g., eye gaze, head movements, arm movements, and facial expressions) [1], much less has explored the role of audio-based dimensions of realism.

This body of literature highlights the complementary nature of audio and visual sensory stimuli, however, there is limited research exploring this interaction in collaborative MR environments. Fink et al. [14] investigated the interplay between simple versus rich avatars with monaural versus spatial audio in a remote MR session. While no significant differences in subjective co-presence were reported between these conditions, most participants preferred rich avatars combined with spatial audio over the other combinations. Yang et al. [71] found that using spatial audio for a remote collaborator's voice during a search task helped improve performance compared to when using non-spatial audio. In a non-collaborative context, Weidner et al. [67] observed no significant interaction effects between avatar rendering style and audio spatialization on gaze behavior in an MR storytelling experience. We extend existing work by exploring the impact of room acoustics in audiovisual, social MR.

3 Research Questions and Hypotheses

The primary goal of this exploratory study aims to address the following research question: how does high-fidelity audio influence co-presence and associated experiential outcomes during avatar-mediated conversations in MR? **(RQ1)** In line with this inquiry, we propose the following hypotheses:

- **H1.** MR environments with room acoustics will yield greater co-presence during avatar-mediated conversation compared to anechoic environments.
- **H2.** MR environments with room acoustics will yield greater co-location during avatar-mediated conversation compared to anechoic environments.
- **H3.** MR environments with room acoustics will yield greater interaction believability during avatar-mediated conversation compared to anechoic environments.
- **H4.** MR environments with room acoustics will yield greater satisfaction during avatar-mediated conversation compared to anechoic environments.

We also formed a secondary research question to explore how the relationship between room acoustics and co-presence varies based on participants' immersive tendency and personality trait scores **(RQ2)**. Immersive tendency is one's propensity towards feeling immersed in virtual environments and can effect how much people focus and engage with stimuli [32, 70]. Personality is described as "psychological qualities that contribute to an individual's enduring and distinctive pattern of feeling, thinking, and behaving" [5]. Past research has shown that individual differences in these two traits can influence co-presence in immersive contexts [10, 28, 45], yet less is known how it can moderate the relationship between audio quality and co-presence.

4 Methods

We conducted an Advarra IRB-approved, exploratory experiment to investigate the impact of audio fidelity on co-presence and communication in social MR environments. The methods are detailed in this section.

4.1 Participants

Sixteen dyads, consisting of 32 participants were recruited for this ethics board approved study located in the United States. Four of these participants engaged in a pilot version of the study.

Participants, who self-identified as men (n = 12) and women (n = 20), were between the ages of 18 and 75+ (n_{18-24} = 3, n_{25-34} = 5, n_{35-44} = 12, n_{45-54} = 6, n_{55-64} = 3, n_{65-74} = 1, $n_{75orolder}$ = 1, declined to respond = 1) and identified as American Indian / Alaska Native (n = 2), Asian/Asian American (n = 6), Black/African/African American (n = 6), Latin/Hispanic (n = 1), White/Caucasian (n = 13), multiple racial/ethnic groups (n = 3), or preferred to specify (n = 1; "Mix"). Prior experience with MR/VR, gaming, and avatars in MR/VR are highlighted in Table 1. In regards to prior familiarity with their study partner, 13 participants knew their partner for less than 6 months, 3 for 6 months to a year, 2 for 1-3 years, and 14 for more than three years. 11 never communicated with their partner prior to the study, 1 rarely communicated with their partner, 2 communicated monthly, 6 communicated weekly, and 12 communicated daily. Participants were compensated $75 for each hour of participation, which usually took two to three hours.

4.2 Audio Conditions

Our within-subjects experiment involved two audio conditions under which participant voices were rendered and presented through their avatars:

- **Anechoic Audio Condition:** Served as the control, representing the audio fidelity traditionally available in MR calls at the time of the experiment. Audio was anechoic, lacking any room acoustics.
- **Reverberant Audio Condition:** Represented our higher fidelity audio setting that incorporated room acoustics by simulating reverberation based on the spatial layout of the participant's room. Specifically, the underlying simulation is a bi-directional path tracing (BDPT) informed by the geometry of the space [6].

  The reverberant audio condition computed a ray-traced energy-time curve at a low sample rate (∼50 Hz), then used

Table 1: Participants' prior experience with virtual and mixed reality, video games, and avatars.

Category	Response Option	N
VR/MR Experience	I have never used VR or MR technology.	3
	I have very minimal experience with VR or MR technology (e.g., did a demo).	9
	I have some experience with VR or MR technology (e.g., tried a friend's headset).	15
	I have a lot of experience with VR or MR technology, but do not own a headset.	1
	I have a lot of experience with VR or MR technology and I own and use a VR headset.	4
Video Game Experience	New gamer	2
	Casual gamer	20
	Hardcore or serious gamer	2
	Pro gamer	1
	Non gamer	7
Avatars in VR/MR Experience	I have never used an avatar to represent myself in VR or MR.	12
	I have very minimal experience using an avatar to represent myself in VR or MR (e.g., did a demo).	10
	I have some experience using an avatar to represent myself in VR or MR (e.g., tried multiple VR experiences as an avatar).	10

this to estimate parameters for an ambisonic multiband reverberator. Early reflections were rendered as prioritized delay taps to first-order ambisonics, with up to four significant reflections per source based on intensity. Late reflections were handled by the artificial reverb algorithm, and all audio was processed in frequency bands and rendered in the spherical harmonic domain, enabling efficient application of directional effects and head-related transfer function (HRTF) spatialization for all sources through a single multichannel convolution.

Both audio conditions included voice directivity, were spatialized, and their renderers featured distance dependent direct sound attenuation that enabled 6DoF navigation. Each dyad experienced both conditions in a randomized, counterbalanced order to mitigate ordering effects.

We used geometry-informed BDPT in our reverberant audio condition to prioritize a scalable reverberation approach that could plausibly be used in real-world MR applications. One limitation of this approach is that there may have been some degree of divergence from actual acoustic measurements of the spaces. Transitioning experimental audio algorithms into production requires careful consideration of computational scalability and resource constraints. Ray-traced acoustic rendering becomes particularly computationally demanding when supporting full 6DoF user movement, necessitating trade-offs between simulation fidelity and performance requirements. Scalable implementations must therefore balance computational efficiency with acoustic accuracy, often sacrificing some degree of veridical simulation to meet real-time constraints.

Audio levels were perceptually calibrated by three experimenters through live avatar conversations and iterative adjustments of playback until consensus was reached. At that point, the same level was set for all participants, with data from those who experienced uneven levels prior to calibration excluded from quantitative analysis. In the process, it was considered that both the relative level of the two renderers should be perceptually equivalent, and that the absolute level should be consistent with the vocal effort of the remote caller. This method was chosen due to the large number of variables that could influence instrumental calibration (e.g., differences in microphone distance and renderer-specific attenuation curves that change with 6DoF movement).

4.3 Hardware

Participants used Meta Quest 3 standalone headsets (2064 x 2208 pixel resolution per eye, 110.00° horizontal FOV, 96.00° vertical FOV, 2 RGB cameras with 18 PPD, six degrees of freedom (6DoF) inside-out tracking) and two hand controllers.

Audio was presented to participants over the Meta Quest 3's integrated stereo speakers. We opted for using the headset's built in speakers rather than headphones to reflect the default and more widely used audio configurations of consumer MR.

4.4 Virtual Environment and Avatars

Sessions were hosted in a passthrough MR environment where participants were represented by photorealistic self-avatars. Participants generated these avatars by taking a series of photos of themselves from various angles and facial expressions. The process by which a photorealistic avatar can be created with only small amounts of data includes a conditional representation that can extract person-specific information at multiple scales from a high resolution registered neutral phone scan, and a novel universal avatar prior that has been trained on high resolution multi-view video captures of facial performances of hundreds of human subjects [4].

Avatars were generated using a Universal Gaussian Decoder. This is a neural network-based decoder used in the Codec Avatar pipeline to generate photorealistic, drivable 3D avatars. Unlike person-specific decoders, UGD is trained on a large, diverse dataset of identities, poses, and expressions, enabling it to generalize across many people and body types. Its output is a set of Gaussian splats — compact, view-dependent primitives that encode geometry, color, and other appearance features for rendering avatars in real time [66, 37].

Although the avatars did not reflect participant facial expressions due to hardware constraints, head and hand movements were tracked and rendered to their avatars. Avatar motion was tracked using a specialized diffusion-based synthesizer model designed for driving body motion in avatars using only hand tracking signals. It is an adaptation of the broader cloud-based Avatar Regression Diffusion architecture, optimized for scenarios where full-body tracking is unavailable — such as cloud-based execution or devices lacking body-tracking sensors.

4.5 Physical Environments

We used two different physical environmental stimuli (a bedroom and a conference room; See Figure 1) to represent two different contexts in which remote MR calls might take place with different acoustic properties and to incorporate stimulus sampling into our study design [52]. Participants were randomly assigned to one of the two rooms, within which they experienced both audio conditions. Rather than having each participant experience both rooms, we employed a design intended to reflect real-world use cases, where users typically engage from a singular physical context.

The first room was a bedroom with a 25 m^2 mock-up apartment furnished with a bed and bedside tables. The second room was an office room of approximately 35 m^2 with a large conference table placed in the center. In both rooms, the floors were covered with carpet, the walls were prominently made of plaster, and a dropped ceiling was present. Each room also featured a large window covering most of one wall. In the bedroom, this windowed wall was slanted inward, as shown in Figure 1.

The frequency-dependent reverberation time (RT60) of both rooms was computed using several measurements from different source-receiver locations and is shown in Figure 2. Both rooms

exhibited significant deviations at low frequencies, likely due to strong modes. Additionally, while the bedroom showcases a relatively standard frequency-dependent RT60 profile, with higher RT60 at mid frequencies, the office shows the inverse pattern, with the shortest RT60 at mid frequencies.

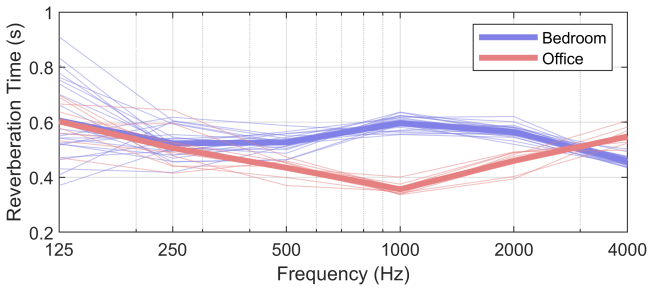

Figure 2: Average reverberation Time (RT60) of the two experiment rooms (thick lines) computed via T30 extracted from several source-receiver measurements (thin lines).

4.6 Tasks

While in MR, participants engaged in two conversation tasks under each audio fidelity condition. First, participants worked together on a survival problem. After reading through a hypothetical survival scenario set either on the moon, in a desert, and/or the ocean [22, 33], participants were instructed to come to a consensus on ranking a list of items from most to least important for survival. This task was selected for its ability to facilitate naturalistic conversation between dyads and its utility in past research on social presence and turn-taking [44]. Separate survival scenarios were used in the different audio fidelity conditions, in a randomly assigned order. Second, participants played a game of charades, where they took turns physically acting out phrases while their partner tried to guess them. This task was selected to showcase location and proximity cues in audio by encouraging physical movement. Altogether, participants spent 10-12 minutes engaging in MR conversation tasks under each audio condition.

Prior to these conversation tasks, participants also completed two audio primer tasks. This methodological choice was guided by pilot testing, which revealed that, through post-task interviews, audio differences went unnoticed or were misattributed to unrelated factors, possibly due to the engaging nature of the task. Informing participants ensured that audio entered their perceptual awareness, enabling more sensitive comparison across conditions.

Participants were given instructions on considerations for when deciding if voices sound realistic and in the same room (e.g., "take into consideration what it feels like to be around and in the same area as them compared to when you're talking on the phone or over a video call"). During the headphone audio primer, which took place prior to going in-headset, participants listened to two practice audio clips (monaural and stereo audio record with in-ear microphones) designed to highlight these distinctions on a pair of headphones. These clips were not part of the main experiment's audio conditions but served as training to help participants better discern audio differences. During the headset audio primer, participants experienced both audio conditions briefly in-headset right before starting the conversation tasks. Under each condition, participants took turns walking around the room while the other read a set of sentences out loud [55]. Participants were not explicitly asked about their perception of audio during the primer activities.

4.7 Procedure

Upon arrival at the study location, one of the experimenters provided participants with basic information about the study and the consent form. After completing their consent forms, participants created their avatars by taking a series of photos. While the avatars were generating (around 45 minutes), participants completed a pre-survey to collect demographic and individual difference information, engaged in their headphone audio primer task, and individually worked on the survival problem that would be discussed in-headset. For any remaining time before avatars finished generating, participants spent time on a filler task (e.g., sudoku).

Once their avatars finished generating, participants were given an overview of their in-headset task and directed into two separate rooms, each with an experimenter. After putting on the MR headset, participants experienced both audio conditions in a headset audio primer task. Participants were not informed what the audio conditions were. Next, in the first audio condition, participants spent 10-12 minutes engaging in two conversation tasks: the survival problem and charades. Afterwards, participants completed the first part of their survey. Participants repeated these conversation tasks in the second audio condition and completed the remainder of their survey. Finally, participants engaged in a group interview about their experience. The entire duration of the study lasted two to three hours.

4.8 Data Collection

During each study session, we collected: (1) survey data and (2) group interview recordings.

4.8.1 Survey Data

All participants completed a pre-survey before their in-MR activities and a post-survey after each of the different conditions. In addition to collecting demographic and prior experience information, our pre-survey included items that measured the following individual differences:

Immersive Tendency. Immersive tendency was measured by eight items adapted from Witmer and Singer [70] rated on 5-point Likert scales. Sample items include *"How good are you at blocking out external distractions when you are involved in something?"* and *"Do you ever become so involved in a daydream that you are unaware of things happening around you?"*

Personality. We measured the Big-5 personality traits (Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism) using the Ten-Item Personality Inventory (TIPI) [19]. This inventory consisted of two items for each personality, each rated on a 7-point Likert scale (1 = Strongly disagree, 7 = Strongly agree).

Our post-survey measured the following constructs for each audio condition:

Co-presence. Co-presence was measured using four items adapted from Herrera et al. [25]'s scale. Participants rated their level of agreement (1 = Strongly disagree; 7 = Strongly agree) with each statement: *"I felt like I was face-to-face with my partner," "I felt like I was in the same room as my partner," "I felt like my partner was aware of my presence,"* and *"I felt like my partner was present."*

Co-location. A more specific measure of co-location was incorporated into the survey. Participants responded to the item, *"Did it seem like you were in the same or different place as your partner's avatar?"* on a 5-point Likert scale (1 = Definitely in different place; 5 = Definitely in the same place)

Interaction Believability. Interaction believability was measured through a 7-point Likert scale response (1 = Very dissimilar; 7 = Very similar) to the item, *"How similar or dissimilar was your interaction to a face-to-face real-life interaction?"*

Humanness. The humanness of how their partner's avatar looked, moved, and sounded was measured by three items on a 5-point Likert scale response (1 = Not at all human-like, 5 = Extremely human-like): *"To what extent does your partner's avatar*

look like a human?" "To what extent did your partner's avatar move like they were a human?" and *"To what extent does your partner's voice sound like a human?"*

Satisfaction. Satisfaction was measured using two items on a 7-point Likert scale (1 = Strongly disagree; 7 = Strongly agree): *"How satisfied or unsatisfied were you with the interaction with your partner's avatar?"* and *"How satisfied or unsatisfied are you with the calling experience?"*

Perceived Turn Taking. Perceived turn-taking was measured using six items on a 7-point Likert scale (1 = Strongly disagree; 7 = Strongly agree), adapted from prior work [59, 44]. Sample items include *"I found it easy to participate in the conversation"* and *"I was able to take control of the conversation when I wanted to."*

At the end of the survey, participants were asked in a free-response question to describe any differences they observed in the audio between the different conditions and explain how these audio factors affected their experience.

4.8.2 Group Interview Recordings

Upon completion of the in-MR activities and post-survey, participants took part in a group interview about their experience in the conference room. Interviews were semi-structured, consisting of pre-written and occasionally improvised questions. Interviews lasted around 10-15 minutes and were audio recorded.

4.9 Data Analysis

Differences in Likert-scale questionnaire responses between the two audio conditions were analyzed using Wilcoxon signed-rank tests with Bonferroni correction applied to control for multiple comparisons. Interaction effects of individual differences were examined using linear mixed-effects models predicting co-presence as a function of audio condition. Due to technical issues (e.g., imbalanced audio volume), 10 participant's data points were dropped from the scope of this analysis. Furthermore, the 4 participants who comprised pilot study participants were excluded from this quantitative analysis.

Interview data and qualitative data of all participants were analyzed using open inductive coding and exploratory analysis. Drawing on the qualitative HCI practice guidelines proposed by McDonald and colleagues [39], we adopted a method focused on identifying recurring topics pertinent to our inquiry and then linking them to form broader themes. Consistent with these guidelines, themes were not required to correspond to the most frequently occurring codes but rather emphasized those most relevant and significant to the present inquiry [39].

5 Results

5.1 Analysis of Quantitative Questionnaire Data

Co-presence scores were relatively high across both the anechoic (M = 5.21, SD = 1.21) and reverberant (M = 5.14, SD = 0.936) audio conditions. However, there were no significant differences in co-presence between these two conditions (V = 49.5, p = 1.00).

To further investigate factors that might influence co-presence, we examined the interaction between audio condition and individual differences. The interaction between audio condition and immersive tendency in predicting co-presence was not significant, b = 0.069, SE = 0.646, t(16) = 0.108, p = .916, with a small effect size (R^2 = 0.000, 95% CI [0.000, 0.136]). In contrast, when considering personality, the interaction between audio condition and extraversion was significant, b = -0.307, SE = 0.140, t(16) = -2.189, p = .044, with a small effect size (R^2 = 0.036, 95% CI [0.000, 0.234]), indicating that the effect of condition on co-presence ratings varies depending on participants' levels of extraversion (See Figure 4). Specifically, participants with lower extraversion seemed to gain more co-presence from room acoustics, while highly extraverted participants experienced less of this benefit. At the same time, predicted scores of co-presence were greater in the anechoic version for more extraverted individuals.

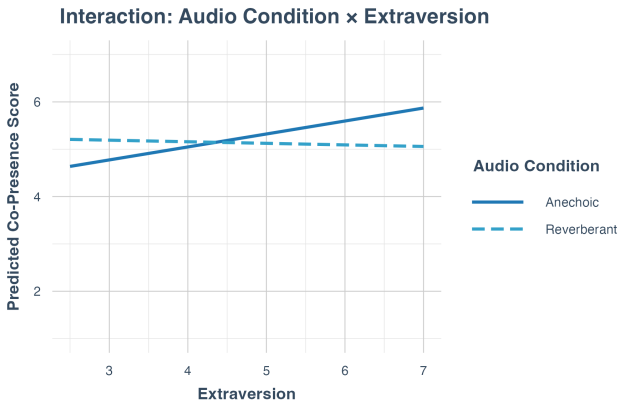

Figure 4: Plot of the interaction between audio condition and extraversion.

Results showed no significant differences in co-location (V = 24, p = 1.00), interaction believability (V = 22.5, p = 1.00), humanness (V = 20.5, p = .930), satisfaction (V = 28, p = 1.00), or perceived turn-taking (V = 93, p = 1.00) between the anechoic and reverberent audio conditions. It is also worth noting that the mean scores for perceived turn-taking ($M_{anechoic}$ = 5.56, $SD_{anechoic}$ = 0.840, $M_{reverberant}$ = 5.60, $SD_{reverberant}$ = 0.632) and satisfaction ($M_{anechoic}$ = 5.11, $SD_{anechoic}$ = 1.29, $M_{reverberant}$ = 5.36, $SD_{reverberant}$ = 1.16) were relatively high. Therefore, anechoic audio did not impair nor improve conversation dynamics and satisfaction with their experience in this context.

Following the primary analyses, we conducted further exploratory examination of the sub-dimensions of humanness and report these results descriptively: humanness–look ($M_{anechoic}$ = 2.83, $SD_{anechoic}$ = 0.92, $M_{reverberant}$ = 2.94, $SD_{reverberant}$ = 1.00), humanness–move ($M_{anechoic}$ = 2.33, $SD_{anechoic}$ = 0.84, $M_{reverberant}$ = 2.72, $SD_{reverberant}$ = 1.02), humanness–sound ($M_{anechoic}$ = 4.33, $SD_{anechoic}$ = 0.69, $M_{reverberant}$ = 4.28, $SD_{reverberant}$ = 0.75).

5.2 Thematic Analysis of Qualitative Data

5.2.1 Audio Facilitates Co-Presence

Feelings of co-presence during participants' MR session were supported by audio quality. Participants noted that the audio quality provided *"the feeling that the other person is in the same room"* (Session 5) and helped them feel *"like there was a physical presence in that room or on that call"* (Session 3). A pilot study participant noted that hearing the room interact with the sound made them feel like they *"were somewhere, somewhere together"* (Session 2). Audio also conveyed social information that facilitated co-presence; participants across two sessions highlighted that they enjoyed being able to hear each other laugh.

5.2.2 Perception of Audio Quality Differences Varied

Overall, participants found the audio quality to be good across both conditions. They described it as *"as very much human like"* (Session 14), *"very live"* (Session 4), and *"better than most of the calls that I take"* (Session 10). However, preferences between the anechoic and reverberant conditions varied. First, there were a number of participants who could not tell a difference between the two audio conditions. They expressed that it *"sounded the same"* (Session 10), *"blended in [their] mind as being equivalent"* (Session 6), and that they *"both [seemed like] they were flawless"* (Session 7). The

Table 2: Comparison of Variables Between Anechoic and Reverberant Audio Conditions

Variable	Anechoic (n = 18)		Reverberant (n = 18)		V-Statistic	p-value	p-value (Bonferroni)
	M	SD	M	SD			
Co-presence	5.21	1.21	5.14	0.936	49.5	0.806	1.00
Co-location	3.11	1.23	3.06	1.21	24	0.903	1.00
Interaction Believability	3.94	1.80	4.56	1.46	22.5	0.194	1.00
Humanness	3.17	0.649	3.31	0.754	20.5	0.155	0.930
Satisfaction	5.11	1.29	5.36	1.16	28	0.405	1.00
Perceived Turntaking	5.56	0.840	5.60	0.632	93	0.760	1.00

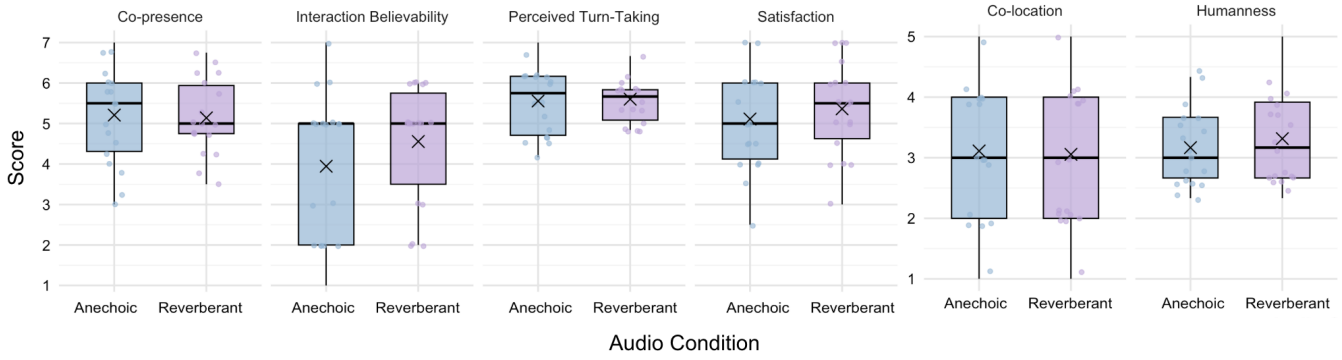

Figure 3: Box plots visualizing the co-presence, interaction believability, perceived turn-taking, satisfaction, co-location, and humanness scores across both audio conditions.

combined effects of the audiovisual environment, conversational nature of the tasks, and overall excitement surrounding the novelty of the experience introduced multiple layers of stimuli, which at times diverted attention away from the audio. For example, participants expressed focusing on *"communication"* (Session 9), *"processing the visual"* (Session 1), *"the task at hand"* (Session 13), *"the motion"* (Session 7), or that they generally were not paying attention to the audio (Session 4 and 12).

Among participants who could discern differences between the audio conditions, while most preferred the reverberant audio, there were still some participants who preferred the anechoic audio. Participants who preferred the reverberant audio felt like it *"captured more [of] the environment"* (Session 3), was *"more clear"* (Session 15), *"more comfortable"* (Session 7), and *"less sterile"* (Session 2). On the topic of co-presence, one participant expressed that under the reverberant audio condition, it *"felt like [their partner's voice] was in the same room"* (Session 13).

Participants who preferred the anechoic audio condition described it as *"more concise"* (Session 14), *"more immersive"* (Session 8), and *"more clear"* (Session 12). When describing differences between the audio conditions, one participant expressed that *"[the reverberant condition] seemed like that one was more realistic than the [anechoic condition]... [The anechoic condition] felt more like virtual reality. Like she stayed in the same place the whole time in my ear."* They preferred the lack of externalization and liked that it made it sound like their partner was closer.

5.2.3 Context Matters for Room Acoustics in MR

Interview responses also reveal that context could play a role in preference for room acoustics. Participants in Session 13 described situations where hearing echoes would make a difference compared to situations where it would not be necessary. One participant said more realistic audio mattered in situations including *"hanging out in something social"* or watching the Olympics and [getting] *"to hear the same echoes as people in like stands or stadiums."* The other said, *"I can see how this could be useful and where spaces matter and where you really need to soak up what's around you."* However, they said for situations where they were just receiving instructions to perform a task or engaging in more of a static activity, such as chess, then audio that sounded like a phone call would be preferred. Virtual performances were also cited as a situation in which room acoustics would matter, with one participant describing *"I would imagine that having some sort of real life audio feedback that sounds like a room of audience members might be an enhanced experience"* (Session 15).

Another participant expressed that neurodiversity could play a role in room acoustic preference as well. They expressed, *"I just don't think that echo was a good idea. Like I said, I'm on the spectrum and noise and colors and stuff. I don't process them the same way other people do"* (Session 9). Therefore, sensitivities to auditory stimuli could impact how participants experience audio quality.

6 Discussion

In this exploratory experimental study, we investigated the role of room acoustics in shaping feelings of co-presence during avatar-mediated conversations in MR. Although our findings did not support our initial hypotheses, we provide preliminary evidence suggesting that personality may moderate the relationship between room acoustics and co-presence. Specifically, our finding reveals the benefit of room acoustics on feeling together with another person in MR may vary based on how extroverted a user is. We explored extraversion as an individual difference based on past research highlighting that individuals with more positive attitudes towards social interaction experience greater levels of social presence, a construct correlated with co-presence [28, 45]. Within the context of our environmental set-up, room acoustics had a greater benefit to co-presence when participants were more introverted. Past research has shown that introverts are more sensitive to noise

and sensory input compared to extroverts [62, 16]. Provided that the room acoustics in our selected environments were not overstimulating, it is possible that the externalization of the audio, enabled by the room acoustics, helped improve listening comfort [20] more than in the anechoic condition and in turn, co-presence. Therefore, even though our results found no significant differences in co-presence between the anechoic and reverberant audio conditions, individual differences between users may shape this dynamic. However, these results should be treated as preliminary, given the small sample size and their dependence on the acoustic characteristics of the rooms selected for this study.

Given that participants experienced each audio condition for long time blocks (around 10 minutes), it is possible that the similar co-presence scores were linked to how they adapted to each acoustic setting over time. Due to the fact there were enough consistent cues between the conditions (i.e, spatial audio, head tracking, distance dependent sound attenuation), it might have made adapting to each audio condition easier, minimized the potential of perceived differences, and led to an overall satisfactory experience. Another explanation for these results and why a number of participants did not notice differences between the audio conditions may be attributed to cross-modal influences [63, 57, 29]. The Colavita visual dominance effect asserts that visual cues can dominate over auditory cues in audiovisual environments [8]; accordingly, avatars and passthrough visuals likely took prominence over audio cues. Prior work has also shown that video resolution can play an important role in shaping audio-visual quality [51], suggesting that the fidelity of the passthrough video, which remained the same across conditions, might have impacted participants' perception of audio. In addition to focusing on visual qualities, their focus on the active tasks might have further diverted attention away from audio. Since participants generally perceived the audio quality as high in both conditions, it is also possible that the communication medium faded into the background, allowing participants to feel immersed in the MR environment [68]. They could focus on their task at hand without consciously noticing the audio.

Among participants who could perceive differences, many preferred the condition with room acoustics. Their explanations revealed that incorporating room acoustics has the potential to make audio feel less sterile and make people in MR calls feel more like they're in the same room together. But while some participants found clarity in this reverberant audio condition, others found the anechoic audio to feel more clear and preferred its lack of externalization. This provides further evidence that audio preferences are not uniform across users. Future collaborative MR interfaces may consider allowing users to toggle between audio settings based on their preferences.

Our findings also highlight that, beyond who the user is, the *what* and *where* of the audio context can play crucial roles in shaping the experience of co-presence in mixed reality (MR). Qualitative data indicate that user preferences for room acoustics vary depending on the environment and activity. Specifically, users expressed a preference for room acoustics in socially dynamic settings (e.g., hanging out in a social setting, performing for an audience, or being in a stadium) where environmental audio cues contribute meaningfully to the sense of co-presence. However, for more static or task-focused contexts, such as receiving instructions, anechoic audio was described as more appropriate, likely due to its clarity. These findings suggest there might not be a universal default for room acoustics in MR environments. Instead, designers should consider the use case and context when considering whether to incorporate room acoustics. Tailoring audio design to the situational needs of users may help enhance co-presence and user experience in MR applications.

6.1 Limitations and Implications for Future Research

This paper presents an experimental protocol exploring how room acoustics affects feelings of co-presence in social MR. While the current protocol can help inform future research on this topic, there are several noteworthy opportunities for subsequent work.

First, in future work we aim to explore how audio with more accurate room acoustics may help improve feelings of co-presence. In this study, we adopted a simulation-based approach to room reverberation to prioritize a design reflective of scalable real-world MR applications. This approach did not include systematic calibration, explicit measurement of the direct-to-reverberant ratio, or validation against measured acoustics in the space. In future work, we aim to use binaural room impulse responses (BRIR) and BRIR extrapolation techniques, together with real-time convolution of the input speech. This will more accurately capture the acoustic characteristics of the environment. It is worth noting that BRIR convolution is only practical in lab environments, due to the need of acoustic measurements and its computation and memory requirements. An alternative approach would involve taking measurements of the space and trying to match the simulated acoustic parameters as closely as possible. This could be done in a first phase of the experiment session, using blind estimation of the acoustic properties with a headset [12] and iterative optimization of the acoustic materials in the acoustic simulation [56].

Second, future studies could incorporate tasks that drive the user's attention more towards audio and include measures that are more sensitive to audio as a modality. For example, this may include shorter and more direct tasks as well as using two-alternative forced choice tests.

Third, we examined two physical contexts where MR-based communication is commonly used: at home and at work. However, future research could explore different social and environmental contexts where room acoustics could make an even greater difference in feelings of co-presence. It is also possible that the configuration of real and virtual sounds can shape the importance of room acoustics. Specifically, blending auditory cues from both real and virtual humans may increase user expectations for how realistic the virtual voices should sound [26]. Future research should further explore this dynamic.

Lastly, we highlight additional limitations of this study. Our results were impacted by the study's small sample size. Future work should explore the relationships presented in this paper with a larger number of participants. Moreover, it is possible that our measure of call satisfaction could have been sensitive to technical issues that occasionally occurred during sessions (e.g., the application crashing and having to re-open it).

Participants experienced both experimental conditions in one of two physical rooms, balanced across participants and within sessions. Acoustic differences between testing environments might have introduced variability in participants' reports that our statistic analyses did not explicitly model. While this environmental setup was selected to reflect realistic usage, future work could more systematically examine how environmental context may influence responses across conditions.

Participants were primed to the audio condition, which may have heightened attention to audio beyond typical incidental listening conditions. This introduces a potential threat to ecological validity: effects observed under directed attention may not generalize to naturalistic MR use where audio is processed incidentally. However, we note three mitigating factors. First, our dependent variables do not directly assess audio, reducing transparent demand cues. Second, pilot testing demonstrated that without priming, audio differences did not register on experiential measures, necessitating the attentional guidance. Third, the predominantly null results across conditions are inconsistent with attention-driven response inflation, suggesting participants reported genuine experi-

ences rather than perceived experimenter expectations. Nonetheless, future work should examine whether spatial audio effects on social presence emerge under incidental listening conditions.

This study did not explicitly measure audio plausibility or authenticity. We instead prioritized outcome measures, such as co-presence, that were more directly aligned with the study's research questions. Because plausibility and authenticity represent high perceptual thresholds [41] that are difficult to achieve under simulated reverberation, these measures were beyond the scope of the present work. Future studies, especially those employing measurement-based room reverberation, may consider incorporating such perceptual evaluations.

7 Conclusion

Despite the critical role that audio plays in enhancing the immersive and social experiences in MR, little work has explored how this modality shapes feelings of co-presence in this medium. To address this critical gap, we presented a study exploring how room acoustics impacts co-presence during avatar-mediated conversations in MR. Participants created photorealistic avatars of themselves and engaged in semi-structured conversation tasks under anechoic and reverberant audio conditions. Results from this exploratory study of audiovisual realism in MR shed a nuanced light on how room acoustics can facilitate co-presence. There were no significant differences between the two conditions in co-presence. However, we presented preliminary evidence to suggest that who, what, and where can influence when room acoustics best supports this experiential outcome. We discussed implications for the design of future social MR interfaces and shared insights to guide future audio research in immersive, audiovisual environments.

Acknowledgments

We thank Alexey Bezugly, David Airapetyan, Dilara Semerci Frey, Giancarlo Di Biase, Kevin Blot, and Peifeng Jing for their technical contributions, valuable feedback, and discussions that informed this work.

References

[1] J. N. Bailenson, N. Yee, D. Merget, and R. Schroeder. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and co-presence in dyadic interaction. *Presence: Teleoperators and Virtual Environments*, 15(4):359–372, 2006. 3

[2] J. Bhattacharyya, L. Picinali, A. Vinciarelli, and S. Brewster. Investigating the influence of environmental acoustics and playback device for audio augmented reality applications. In *Audio Engineering Society Conference: AES 2024 International Audio for Games Conference*. Audio Engineering Society, 2024. 2

[3] I. d. V. Bosman, O. uruk, K. Jørgensen, and J. Hamari. The effect of audio on the experience in virtual reality: a scoping review. *Behaviour & Information Technology*, 43(1):165–199, 2024. 1

[4] C. Cao, T. Simon, J. K. Kim, G. Schwartz, M. Zollhoefer, S. Saito, S. Lombardi, S.-E. Wei, D. Belko, S.-I. Yu, et al. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 4

[5] D. Cervone and L. A. Pervin. *Personality: Theory and research*. John Wiley & Sons, 2022. 3

[6] C. Chen, C. Schissler, S. Garg, P. Kobernik, A. Clegg, P. Calamia, D. Batra, P. Robinson, and K. Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems*, 35:8896–8911, 2022. 3

[7] J.-T. Chien. *Source separation and machine learning*. Academic Press, 2018. 2

[8] F. B. Colavita. Human sensory dominance. *Perception & Psychophysics*, 16(2):409–412, 1974. 8

[9] T. Combe, R. Fribourg, L. Detto, and J.-M. Normand. Exploring the influence of virtual avatar heads in mixed reality on social presence, performance and user experience in collaborative tasks. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2206–2216, 2024. 2

[10] J. Cortese and M. Seo. The role of social presence in opinion expression during ftf and cmc discussions. *Communication Research Reports*, 29(1):44–53, 2012. 3

[11] E. DeMarbre, J. Henderson, and R. J. Teather. Investigating presence across rendering style and ratio of virtual to real content in mixed reality. In *Proceedings of the 2024 ACM Symposium on Spatial User Interaction*, pp. 1–8, 2024. 2

[12] T. Deppisch, N. Meyer-Kahlen, and S. V. A. Garí. Blind identification of binaural room impulse responses from smart glasses. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4052–4065, 2024. doi: 10.1109/TASLP.2024.3454964 8

[13] C. Dicke, V. Aaltonen, A. Rämö, and M. Vilermo. Talk to me: The influence of audio quality on the perception of social presence. In *Proceedings of HCI 2010*. BCS Learning & Development, 2010. 2

[14] D. I. Fink, M. Skowronski, J. Zagermann, A. V. Reinschluessel, H. Reiterer, and T. Feuchtner. There is more to avatars than visuals: Investigating combinations of visual and auditory user representations for remote collaboration in augmented reality. *Proceedings of the ACM on Human-Computer Interaction*, 8(ISS):540–568, 2024. 1, 3

[15] B. J. Fogg and H. Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 80–87, 1999. 2

[16] R. G. Geen. Preferred stimulation levels in introverts and extroverts: Effects on arousal and performance. *Journal of Personality and Social Psychology*, 46(6):1303, 1984. 8

[17] E. Goffman. *Behavior in public places*. Simon and Schuster, 2008. 2

[18] G. Goncalves, P. Monteiro, H. Coelho, M. Melo, and M. Bessa. Systematic review on realism research methodologies on immersive virtual, augmented and mixed realities. *IEEE Access*, 9:89150–89161, 2021. 1, 2

[19] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003. 5

[20] V. Grimaldi, G. Courtois, L. S. Simon, and H. Lissek. Externalization of virtual sounds using low computational cost spatialization algorithms for hearables. In *Forum Acusticum*, pp. 917–921, 2020. 8

[21] J. E. S. Grønbæk, K. Pfeuffer, E. Velloso, M. Astrup, M. I. S. Pedersen, M. Kjær, G. Leiva, and H. Gellersen. Partially blended realities: Aligning dissimilar spaces for distributed mixed reality meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2023. 2

[22] J. Hall and W. H. Watson. The effects of a normative intervention on group decision-making performance. *Human relations*, 23(4):299–317, 1970. 5

[23] K. Hassanein and M. Head. Manipulating perceived social presence through the web interface and its impact on attitude towards online shopping. *International journal of human-computer studies*, 65(8):689–708, 2007. 2

[24] C. Hendrix and W. Barfield. The sense of presence within auditory virtual environments. *Presence: Teleoperators & Virtual Environments*, 5(3):290–301, 1996. 2

[25] F. Herrera, S. Y. Oh, and J. N. Bailenson. Effect of behavioral realism on social interactions inside collaborative virtual environments. *Presence*, 27(2):163–182, 2020. 3, 5

[26] A. HOFMANN, N. MEYER-KAHLEN, S. J. SCHLECHT, and T. LOKKI. Audiovisual congruence and localization performance in virtual reality. 2024. 8

[27] F. Immohr, G. Rendle, A. Neidhardt, S. Göring, R. R. Ramachandra Rao, S. Arevalo Arboleda, B. Froehlich, and A. Raake. 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*, pp. 209–215, 2023. 2

[28] S.-A. A. Jin. Parasocial interaction with an avatar in second life: A typology of the self and an empirical test of the mediating role of social presence. *Presence*, 19(4):331–340, 2010. 3, 7

[29] A. Joly, N. Montard, and M. Buttin. Audio-visual quality and interac-

tions between television audio and video. In *Proceedings of the Sixth International Symposium on Signal Processing and its Applications (Cat. No. 01EX467)*, vol. 2, pp. 438–441. IEEE, 2001. 3, 8

[30] D. Kao, R. Ratan, C. Mousas, A. Joshi, and E. F. Melcer. Audio matters too: How audial avatar customization enhances visual avatar customization. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–27, 2022. 2

[31] A. D. Kaplan, J. Cruit, M. Endsley, S. M. Beers, B. D. Sawyer, and P. A. Hancock. The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: A meta-analysis. *Human factors*, 63(4):706–726, 2021. 2

[32] K. J. Kim, E. Park, and S. S. Sundar. Caregiving role in human–robot interaction: A study of the mediating effects of perceived benefit and social presence. *Computers in Human Behavior*, 29(4):1799–1806, 2013. 3

[33] J. C. Lafferty, P. M. Eady, and J. Elmers. The desert survival problem. *Experimental learning methods*, 1974. 5

[34] P. Larsson, A. Väjamäe, D. Västfjäll, and M. Kleiner. Auditory-induced presence in mediated environments and related technology. *Threshold*, 1(e2):e3, 2005. 1

[35] P. Larsson, D. Västfjäll, and M. Kleiner. Effects of auditory information consistency and room acoustic cues on presence in virtual environments. *Acoustical Science and Technology*, 29(2):191–194, 2008. 2

[36] K. M. Lee. Presence, explicated. *Communication theory*, 14(1):27–50, 2004. 1, 2

[37] J. Li, C. Cao, G. Schwartz, R. Khirodkar, C. Richardt, T. Simon, Y. Sheikh, and S. Saito. Uravatar: Universal relightable gaussian codec avatars. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024. 4

[38] M. Lombard and T. Ditton. At the heart of it all: The concept of presence. *Journal of computer-mediated communication*, 3(2):JCMC321, 1997. 2

[39] N. McDonald, S. Schoenebeck, and A. Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–23, 2019. 6

[40] J. McVeigh-Schultz and K. Isbister. The case for "weird social" in vr/xr: a vision of social superpowers beyond meatspace. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pp. 1–10, 2021. 2

[41] N. Meyer-Kahlen, S. Schlecht, S. V. A. Garí, and T. Lokki. Testing auditory illusions in augmented reality: Plausibility, transfer-plausibility, and authenticity. *Journal of the Audio Engineering Society*, 72(11):797–812, 2024. 2, 9

[42] A. Neidhardt, C. Schneiderwind, and F. Klein. Perceptual matching of room acoustics for auditory augmented reality in small rooms-literature review and theoretical framework. *Trends in Hearing*, 26:23312165221092919, 2022. 2

[43] K. Nowak. Defining and differentiating copresence, social presence and presence as transportation. In *presence 2001 conference, Philadelphia, PA*, vol. 2, pp. 686–710, 2001. 2

[44] K. Nowak, L. Tankelevitch, J. Tang, and S. Rintel. 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*, pp. 1–10, 2023. 2, 5, 6

[45] C. S. Oh, J. N. Bailenson, and G. F. Welch. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5:114, 2018. 2, 3, 7

[46] V. Y. Oviedo, K. A. Johnson, M. Huberth, and W. O. Brimijoin. Social connectedness in spatial audio calling contexts. *Computers in Human Behavior Reports*, 15:100451, 2024. 2

[47] V. Phadnis, K. Moore, and M. Gonzalez-Franco. The work avatar face-off: Knowledge worker preferences for realism in meetings. In *2023 IEEE international symposium on mixed and augmented reality (ISMAR)*, pp. 960–969. IEEE, 2023. 2

[48] T. Piumsomboon, A. Day, B. Ens, Y. Lee, G. Lee, and M. Billinghurst. Exploring enhancements for remote mixed reality collaboration. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*, pp. 1–5. 2017. 1

[49] T. Piumsomboon, G. A. Lee, J. D. Hart, B. Ens, R. W. Lindeman, B. H. Thomas, and M. Billinghurst. Mini-me: An adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–13, 2018. 2

[50] C. Pörschmann. One's own voice in auditory virtual environments. *Acta Acustica united with Acustica*, 87(3):378–388, 2001. 1

[51] T. Potter, Z. Cvetković, and E. De Sena. On the relative importance of visual and spatial audio rendering on vr immersion. *Frontiers in Signal Processing*, 2:904866, 2022. 3, 8

[52] B. Reeves, L. Yeykelis, and J. J. Cummings. The use of media in media psychology. *Media psychology*, 19(1):49–71, 2016. 4

[53] G. Rendle, F. Immohr, C. Kehling, A. Lammert, A. Kreskowski, K. Brandenburg, A. Raake, and B. Froehlich. Influence of audiovisual realism on communication behaviour in group-to-group telepresence. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 569–579. IEEE, 2025. 1

[54] S. Roßkopf, L. O. Kroczek, F. Stärz, M. Blau, S. Van de Par, and A. Mühlberger. The impact of binaural auralizations on sound source localization and social presence in audiovisual virtual reality: converging evidence from placement and eye-tracking paradigms. *Acta Acustica*, 8:72, 2024. 2

[55] E. H. Rothauser. Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246, 1969. 5

[56] C. Schissler, C. Loftin, and D. Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics*, 24(3):1246–1259, 2018. doi: 10.1109/TVCG.2017.2666150 8

[57] M. Schmitt, J. Redi, P. Cesar, and D. Bulterman. 1mbps is enough: Video quality and individual idiosyncrasies in multiparty hd videoconferencing. In *2016 Eighth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6. IEEE, 2016. 3, 8

[58] R. Schroeder. Copresence and interaction in virtual environments: An overview of the range of issues. In *Presence 2002: Fifth international workshop*, pp. 274–295, 2002. 2

[59] A. J. Sellen. Speech patterns in video-mediated conversations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 49–59, 1992. 6

[60] M. Shin, S. W. Song, S. J. Kim, and F. Biocca. The effects of 3d sound in a 360-degree live concert video on social presence, parasocial interaction, enjoyment, and intent of financial supportive action. *International Journal of Human-Computer Studies*, 126:81–93, 2019. 2

[61] J. Skowronek and A. Raake. Assessment of cognitive load, speech communication quality and quality of experience for spatial and non-spatial audio conferencing calls. *Speech Communication*, 66:154–175, 2015. 2

[62] L. Standing, D. Lynn, and K. Moxness. Effects of noise upon introverts and extroverts. *Bulletin of the Psychonomic Society*, 28(2):138–140, 1990. 8

[63] R. L. Storms and M. J. Zyda. Interactions in perceived quality of auditory-visual displays. *Presence: Teleoperators & Virtual Environments*, 9(6):557–580, 2000. 3, 8

[64] P. M. Strojny, N. Dużmańska-Misiarczyk, N. Lipp, and A. Strojny. Moderators of social facilitation effect in virtual reality: Co-presence and realism of virtual agents. *Frontiers in psychology*, 11:1252, 2020. 2

[65] J. Vroomen and B. de Gelder. Temporal ventriloquism: sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance*, 30(3):513, 2004. 3

[66] S. Wang, T. Simon, I. Santesteban, T. Bagautdinov, J. Li, V. Agrawal, F. Prada, S.-I. Yu, P. Nalbone, M. Gramlich, et al. Relightable full-body gaussian codec avatars. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025. 4

[67] F. Weidner, J. Hartbrich, S. Arevalo Arboleda, C. Kunert, C. Schneiderwind, C. Diao, C. Gerhardt, T. Surdu, W. Broll, S. Werner, et al. Eyes on the narrative: Exploring the impact of visual realism and audio presentation on gaze behavior in ar storytelling. In *Proceedings of*

the 2024 Symposium on Eye Tracking Research and Applications, pp. 1–7, 2024. 3

[68] M. Weiser. Creating the invisible interface: (invited talk). In *Proceedings of the 7th annual ACM symposium on User interface software and technology*, p. 1, 1994. 8

[69] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg. A summary on acoustic room divergence and its effect on externalization of auditory events. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6. IEEE, 2016. 1

[70] B. G. Witmer and M. J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7(3):225–240, 1998. 3, 5

[71] J. Yang, P. Sasikumar, H. Bai, A. Barde, G. Sörös, and M. Billinghurst. The effects of spatial auditory and visual cues on mixed reality remote collaboration. *Journal on Multimodal User Interfaces*, 14(4):337–352, 2020. 1, 3

[72] J. Yin, W. Zheng, Y. Wang, X. Tong, and Y. Yan. A comparison study understanding the impact of mixed reality collaboration on sense of co-presence. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 580–590. IEEE, 2025. 2

[73] S. Zhong, L. Rosset, M. Papinutto, D. Lalanne, and H. S. Alavi. Binaural audio in hybrid meetings: Effects on speaker identification, comprehension, and user experience. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24, 2022. 2

[74] K. Zibrek, S. Martin, and R. McDonnell. Is photorealism important for perception of expressive virtual humans in virtual reality? *ACM Transactions on Applied Perception (TAP)*, 16(3):1–19, 2019. 2