There Is More to Avatars Than Visuals: Investigating Combinations of Visual and Auditory User Representations for Remote Collaboration in Augmented Reality

DANIEL IMMANUEL FINK, University of Konstanz, Germany
MORITZ SKOWRONSKI, University of Konstanz, Germany
JOHANNES ZAGERMANN, University of Konstanz, Germany
ANKE VERENA REINSCHLUESSEL, University of Konstanz, Germany
HARALD REITERER, University of Konstanz, Germany
TIARE FEUCHTNER, University of Konstanz, Germany and Aarhus University, Denmark

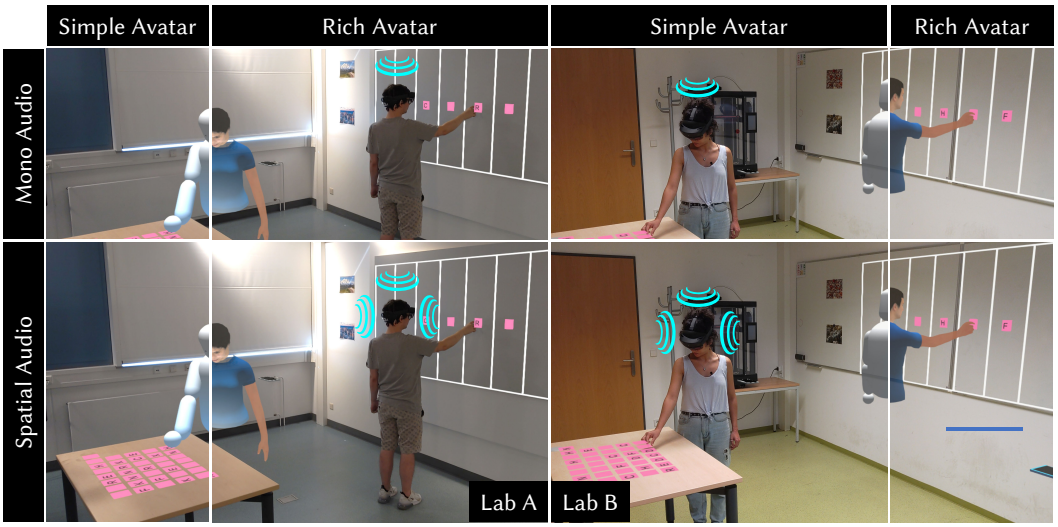

Fig. 1. We investigate the interplay of visual and auditory user representation characteristics for remote collaboration in augmented reality. In our study, two remote collaborators perceive each other as virtual avatars, whereby we compared two different visual user representations, *Simple Avatar* and *Rich Avatar*, and two different auditory conditions, *Mono Audio* (illustrated as single sound wave) and *Spatial Audio* (illustrated as three soundwaves).

Authors' Contact Information: Daniel Immanuel Fink, University of Konstanz, Konstanz, Germany, daniel.immanuel.fink@uni-konstanz.de; Moritz Skowronski, University of Konstanz, Konstanz, Germany, moritz@moritz-skowronski.com; Johannes Zagermann, University of Konstanz, Konstanz, Germany, johannes.zagermann@uni-konstanz.de; Anke Verena Reinschluessel, University of Konstanz, Konstanz, Germany, anke.reinschluessel@uni-konstanz.de; Harald Reiterer, University of Konstanz, Konstanz, Germany, harald.reiterer@uni-konstanz.de; Tiare Feuchtner, University of Konstanz, Konstanz, Germany and Aarhus University, Aarhus, Denmark, tiare.feuchtner@uni-konstanz.de.

ACM 2573-0142/2024/12-ART548
https://doi.org/10.1145/3698148

Supporting remote collaboration through augmented reality facilitates the experience of co-presence by presenting the collaborator's avatar in the user's physical environment. While visual user representations are continuously researched and advanced, the audio configuration – especially in combination with different visualizations – is rarely considered. In a user study (n = 48, 24 dyads), we evaluate the combination of two visual (*Simple vs. Rich Avatar*) with two auditory (*Mono vs. Spatial Audio*) user representations, to investigate their impact on user's overall experience, performance, and perceived social presence during collaboration. Our results show a preference for rich auditory and visual user representations, as *Spatial Audio* supports completion of parallel tasks and the *Rich Avatar* positively influence user experience. However, the *Simple Avatar* draws less attention, which potentially benefits task efficiency, advocating for simpler visualizations in performance-oriented settings. Our findings contribute to a deeper understanding of how visual and auditory user representation combinations impact remote collaboration in augmented reality.

CCS Concepts: • **Human-centered computing → Mixed / augmented reality**; **User studies**; **Collaborative interaction**.

Additional Key Words and Phrases: user representation, augmented reality, remote collaboration, avatar, spatial audio, mono audio, thematic analysis

ACM Reference Format:
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 (December 2024), 29 pages. https://doi.org/10.1145/3698148

1 Introduction

Recent and current crises, such as the COVID-19 pandemic or the climate crisis, drive the need for solutions that facilitate remote collaboration. Consumer solutions currently support creative combinations of video conferencing (e.g., Zoom), shared content editors (e.g., Google Docs), and online whiteboards (e.g., Miro), but typically lack key qualities of in-person interaction, such as workspace awareness, social presence, and non-verbal communication [8, 13]. This may be addressed through advancements in immersive technologies that allow for highly detailed user representations, i.e., avatars that inhabit the same physical space through augmented reality (AR), creating the illusion that remote users are co-located. Beyond verbal communication, modern tracking solutions support faithful reconstruction of body and hand movements and even facial expressions. Thereby the mentioned key qualities of in-person interaction become achievable, allowing users to communicate, jointly manipulate, and share virtual content [12]. In contrast to virtual reality (VR), AR brings the benefit that users can still interact with people and objects in their physical environment.

Recent research has generally investigated ways to facilitate remote collaboration in AR (e.g., [8, 12]) and studied the impact of the visual representation of the remote person – ranging from rather abstract user representations (e.g., virtual viewing frustums [35] with additional gaze ray visualization [40]), to mannequin-style avatars [28], to highly realistic virtual avatars [39], while aiming to avoid the uncanny valley effect [32]. Additionally, researchers have investigated the role of auditory information in conveying the remote person's location or activity for workspace awareness [8, 42]. Studying the interplay of visual and auditory user representation characteristics for remote collaboration is promising, because the complementarity of audio and vision can benefit the processing of available information and attention [9], and serves as confirmation of ambiguous stimuli [3, 31]. Prior research investigated the interplay of visual and auditory user representation characteristics for a remote navigation scenario involving an asymmetry of devices, user roles, user representations, tasks, and capabilities [58]. However, it remains unclear to what extent these results can be transferred to other scenarios.

Therefore, we set out to explore combinations of visual and auditory user representations for *symmetric synchronous* remote collaboration in AR, i.e., where users with equal roles collaborate through identical interfaces and devices. We represent users with full upper body avatars that are either abstract plain white mannequin-style avatars (*Simple Avatar*) or detailed avatars with animated fingers and simulated facial expressions (based on lip-sync and not mirroring the real facial movements), which are personalized to the user's appearance (*Rich Avatar*). In a user study with 24 dyads working in remote physical spaces, we investigate the impact on user's *experiences* (RQ1), *subjective perception of social presence* (RQ2), and *performance* (RQ3). Our findings indicate an overall preference for *Spatial Audio*, regardless of the visual user representations. The rich visual user representation (*Rich Avatar*) significantly improved user experience in terms of hedonic aspects and led to significantly more visual attention paid to the avatar's head compared to the *Simple Avatar* conditions. Interestingly, the visual representation did not affect task performance, but perceived performance was reported to be better with *Simple Avatar* due to it being less distracting. Further, as expected, *Spatial Audio* led to faster completion times for parallel tasks.

In summary, our paper contributes (1) an investigation of meaningful combinations of visual and auditory user representations in a user study, emphasizing the complementarity of these information channels. The results of our user study lead to (2) design perspectives for collaborative systems. For example, in scenarios prioritizing social interaction and increased user experience (e.g., playful or creative activities), a *Rich x Spatial* user representation is recommended. However, as some *Rich Avatar* features were perceived as task-distracting, a reduced visual user representation is preferable for "performance"-oriented activities, requiring high individual attention on the task (e.g., sorting/clustering tasks). Further, we propose (3) directions for future research towards adaptive user representations and refinement of the volume roll-off curve for *Spatial Audio* to balance between speech intelligibility and realistic sound simulation.

2 Related Work

In the following, we review prior research relating to the representation of users in mixed reality (MR) with focus on remote collaboration scenarios, and review the role of avatar appearance and auditory cues.

2.1 Remote Collaboration in Mixed Reality Environments

In comparison to video communication platforms (e.g., [11]), MR technology offers the potential to bring multiple people into the same (virtual) space, where they can feel co-present, communicate, and collaborate in a shared workspace – simulating key qualities of co-located collaboration. Recent research has explored a wide range of technologies to support such activities, e.g., using handheld devices [35, 36, 38] or AR/VR head-mounted displays (HMDs) [39, 41, 57, 58], thereby spanning the MR continuum [49]. VR has emerged as go-to technology for commercial solutions, such as Mozilla Hubs[1], MeetinVR[2], Spatial[3], Mesh for Microsoft Teams[4], and more recently Meta Horizon Workrooms[5], all of which offer collaboration tools and customization of the user's (comic style) avatar. For AR, only few commercial solutions exist (e.g., Microsoft Remote Assist[6] or Spatial[3]).

While recent research highlights the potential of remote collaboration in AR [8], it also reveals a number of challenges, especially with regard to dissimilar physical environments and spatial

[1]Mozilla Hubs: https://hubsfoundation.org/ – last accessed: July 1st, 2024
[2]MeetinVR: https://www.meetinvr.com – last accessed: July 1st, 2024
[3]Spatial: https://youtu.be/PG3tQYlZ6JQ? – last accessed: August 30th, 2024
[4]Mesh: https://www.microsoft.com/en-us/mesh – last accessed: July 1st, 2024
[5]Workrooms: https://www.meta.com/work/workrooms – last accessed: July 1st, 2024
[6]Microsoft Remote Assist: https://dynamics.microsoft.com/de-de/mixed-reality/remote-assist/ – last accessed: July 1st, 2024

layouts [12]. For example, different physical spaces can lead to the Jamais Vu effect, a sensation when "familiar real-world interactions suddenly feel[ing] unfamiliar" [56]. However, it is unclear whether the design of user representations might also contribute to this effect. User representations for AR are an active field of research [46, 55] resulting in a wide variety of approaches for representing a remote person [8, 35, 36, 39–41, 59]. However, most of these focus mainly on visual attributes of user representations, whereby the relevance and impact of realistic appearance, visual detail and fidelity on user experience and task performance remain underexplored. Recent research has shown the benefits of using AR to support remote collaboration, but likewise illustrated its challenges – emphasizing the need for a deeper understanding of what makes remote collaboration in AR successful. We aim to contribute to this understanding by studying the influences of different visual and auditory user representations on remote collaboration.

2.2 User Representations in Immersive Mixed Reality Environments

Most collaborative interactive systems facilitate a sense of co-presence and awareness of others' actions through some form of visual user representation [52]. Importantly, the user representation can convey the users' capabilities and actions, which may, for example, be limited due to the users' interaction context and the used device [38]. For interactions in immersive MR environments, these user representations range from abstract and low-cost (e.g., abstract visualization of a user's virtual view frustum [35, 58] with gaze ray [40]) to highly realistic yet hardware-heavy representations (e.g., Holoportation [39]). A popular user representation that combines attributes from both abstract and realistic techniques are 3D avatars, again ranging from (parts of) an abstract generic body to almost photo-realistic reconstructions of individual persons [45]. While visual realism has been found to positively impact the sense of presence and body ownership [45, 62], it may also be prone to the uncanny valley effect [32].

Exploring the impact of avatar appearance during social interactions and collaboration in immersive multi-user systems, prior work has shown that even abstract avatars can elicit a strong sense of social presence [59] and communication behavior that is comparable to co-located collaboration [38, 51]. Although a full-body avatar is desirable [7], representing the upper body seems most critical while the lower extremities are less relevant [30].

We designed our avatars (see section 3.2.1) based on related work, opting for an anthropomorphic visualization instead of an abstract hand and head position [59] but limiting this to the upper body [30]. Further, we chose to compare an abstract avatar, which has been found to elicit social presence [58] and support behaviors comparable to physical co-location [51], to a more realistic avatar that can positively influence avatar perception [45] and represents the state-of-the-art in consumer applications. With this comparison, we contribute to a better understanding of the influence of different levels of visual and motion fidelity of user representations on effects such as presence and users' experience, as suggested by recent research [45].

2.3 Spatial Audio for Awareness and Presence

While "sound is an underused modality in VR" [48], it undeniably brings opportunities for designing effective interactions in immersive environments: Our auditory perception is "always on", inherently temporal (i.e., occurs briefly as the effect of an event) and omnidirectional, and can be processed in parallel to visual information [9]. This allows the reduction of visual clutter and makes it well-suited for presenting information that might otherwise be overlooked or occluded.

In a virtual space, we can differentiate between *action sounds* caused by the user (e.g., walking, hitting something) and *environmental sounds* (e.g., sound caused by objects in the environment) [48]. For example, in a remote collaboration scenario [42], we may support spatial awareness by communicating the collaborators' location in a room through *action sounds* and when they are speaking [43].

To achieve this, various physical aspects must be simulated, such as sound propagation (i.e., movement of sound from its source to the user's ear) and binaural rendering (i.e., how sounds reach each of the user's ears) [48]. Head-tracked speakers or headphones allow us to simulate locations of sound sources around the user through spatial audio (e.g., placing a sound source in mid-air [34]). Using spatial audio for speech has been shown to aid speaker identification, improve intelligibility, reduce cognitive load, and increase social presence [2, 37, 43, 50] and spatialized environmental sounds can lead to increased spatial awareness of objects in the space [58].

While our ability for echolocation is surprisingly good [34], we constantly refer to further sensory information for confirmation and disambiguation of uncertain stimuli. Primarily, we rely on vision, which dominates to such a degree that it can even lead to erroneous sound source localization (e.g., ventriloquist effect [3]) or sound interpretation, e.g., when heard utterances depend on observed lip movements (McGurk effect [31]). However, in most cases, vision aids hearing, e.g., through lip-reading – how much we rely on this remains greatly unnoticed until this ability is hampered, as was experienced by many when wearing face masks during the COVID-19 pandemic. In this context, recent research on video conferencing reports improved conversation comprehension and confidence when audio was spatialized, in particular when wearing face masks [44].

As previous research highlights the relevance of spatial audio for awareness, co-presence and communication, but is technically more challenging to realize [48], we compare a *Spatial Audio* representation, which is state-of-the-art in MR, to a *Mono Audio* representation, as state-of-the-art for video conferencing (see section 3.2.2). With this comparison, in combination with different levels of visual user representations, we contribute to a better understanding of the complementarity of both modalities, to draw conclusions for user representation design targeted to specific scenarios.

3 System Design and Implementation

We developed a study prototype that supports a synchronous collaborative task in AR, whereby the visual and auditory richness of the remote user representation can be varied. The design and implementation of the task and user representations are described in the following sections.

3.1 Designing for Collaboration in Augmented Reality

A benefit of AR settings is that users can still interact with the local physical environment. However, when collaborating remotely, this can lead to limited spatial awareness of each others' spaces as two physical environments are rarely the same. While various approaches have been proposed to address this [8, 12, 60], we avoid this issue by assuming two identical room setups. In our collaborative workspace remote users perceive each other as virtually co-located with the same type of visual-auditory user representations, have the same interaction capabilities, and obtain equal roles in the task (symmetric collaboration). Aiming for a task that takes advantage of AR, i.e., combining physical and virtual space, we chose the scenario of collaboratively working with sticky notes used in creativity methods, e.g., affinity diagramming [15]. While sticky notes can be attached anywhere in the real world, we adapted this task for AR, as described below.

3.1.1 Collaborative Sticky Note Interaction. In our AR application, virtual sticky notes (8×8 cm) can be picked up using a pinch gesture of the dominant hand (see Figure 2b) and attached to two physical surfaces – a table and a whiteboard (see Figure 2a) – or to each other (see Figure 2d), by releasing the pinch. In addition, sticky notes can be attached to the palm of the non-dominant hand and thereby stacked for easier transport (see Figure 2c). Sticky notes snap to dedicated surfaces when released near them, whereby a white rectangle shows a preview of the snapping position (see Figure 2b). If a sticky note is not released near a surface, it will fall down, similar to a physical sticky note. However, to avoid users getting frustrated with having to pick sticky notes up from the

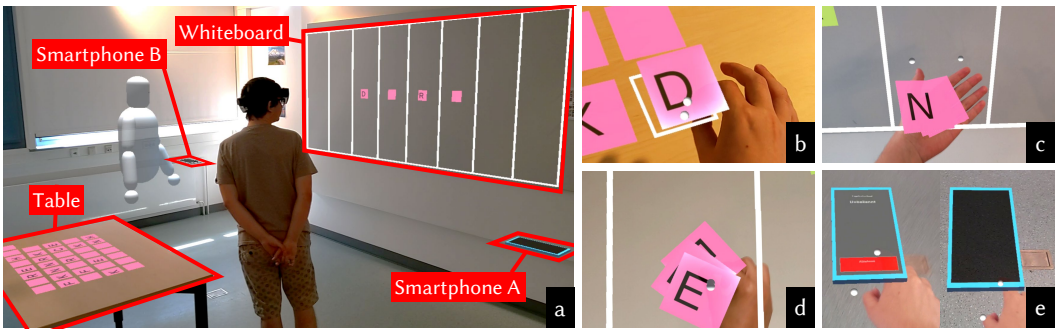

Fig. 2. The collaboration space provides a table and a whiteboard where virtual sticky notes can be attached. In addition, the space includes two virtual smartphones (one for each collaborator) that can interrupt the collaboration with an incoming call (a). The sticky notes can be moved with a pinch gesture (b), collected in the palm of the non-dominant hand (c), and attached to each other (d). When a virtual smartphone rings, its screen lights up (e left) and the call can be declined with the red button (e right).

floor, they are returned to their previous attachment positions. Auditory feedback (action sound) is given when sticky notes are picked up, dropped, and stuck to a surface (snapping). The position and rotation of all sticky notes is synchronized between remote spaces in real time. However, when a sticky note is held by one user, it cannot be pinched by the other to prevent picking up sticky notes from the collaborator's hands.

3.1.2 Individual Smartphone Interaction. The space also includes two virtual smartphones (one blue and one white, one for each collaborator) floating at table height on opposite sides of the room (see Figure 2a). The ringing smartphones represent environmental sounds in the scene, and their sole function is to interrupt the primary task with an incoming call (see Figure 2e left), which the 'owner' needs to decline (see Figure 2e right).

3.1.3 Remote Communication. Our prototype supports verbal and non-verbal communication. The first is achieved through live voice transmission using the microphone of the HMD. For the latter, the avatars are animated by tracking the user's head, hand and finger poses in real-time (by the HMD), revealing the user's position and movement trajectory, gaze direction, arm movements etc., which constitute relevant information for effectively collaborating in MR [51, 58].

3.2 User Representation Design

To study visual and auditory user representations, we designed variants for each with two richness levels. In the following, we describe the user representations in more detail.

3.2.1 Visual User Representations. We explore avatars with an upper body consisting of a head with eyes and a torso with connected arms and hands, based on our analysis of previous work (cf. section 2.2) and state-of-the-art consumer applications (cf. section 2.1). We designed two avatars with distinct visual appearances: *Simple Avatar* and *Rich Avatar*, which reflect two opposite manifestations on the range from 'stickman' to 'realistic,' as categorized by Weidner et al. [55].

Simple Avatar. The avatar's body consists of basic geometric shapes (see Figure 3a). The head is a white capsule with two black spheres as eyes that serve to indicate head gaze direction. The upper body has a large white capsule as torso, two connected capsule segments form each arm, and a white sphere represents each hand. This allows only limited hand gestures, but provides precise enough pointing for task execution, since no task objects are much smaller than the hand spheres.

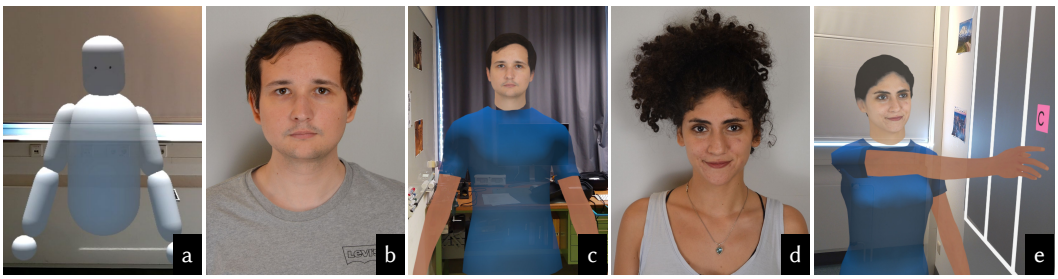

Fig. 3. We study two different visual user representations. The *Simple Avatar* (a) is independent of the appearance of the real person. The *Rich Avatar* is generated from a photo (b,d) and the manual selection of two different torso variations (c,e).

Rich Avatar. The avatar is personalized based on a portrait photo of the real person (see Figure 3b and Figure 3d). The generated head reflects the person's skin color, facial features and hair within certain limits (e.g., glasses are not recognized - hence, users were asked to remove their glasses for the photo). The avatar's face is animated with periodical eye blinking with pseudo-random variation and simulated lip synchronization based on the audio input of the user's voice (cf. section 3.3). The lip synchronization also triggers subtle movements of the eyebrows and other parts of the face, to simulate facial expressions (i.e., no direct face-tracking was used). Between the head and humanoid torso a small gap serves to reduce potential uncanny valley effects that might emerge from differently deforming meshes and merging of two different texture styles (i.e., the photo-generated head texture vs. uniform torso skin texture), which is a common approach for research prototypes [20]. The humanoid torso of the avatar is available in two variations, one with a subtly feminine (see Figure 3e) and one with a rather masculine body shape (see Figure 3c). The study conductor manually selected the torso based on the body shape of the participant. In both variations, the torso wears a blue shirt. The hands with fingers are animated in real-time using the transmitted hand tracking data and inverse kinematics. The rich visualization of the hands allows for precise pointing gestures and can reveal specific interactions (e.g., pinching) with virtual objects.

Both avatars have a comparable size and circumference and are animated using the tracking data from the HMD (see section 3.1.3) and inverse kinematics. When no hand-tracking data is transmitted (e.g., because the hands are outside the tracking range), the arms and hands hover next to the torso. While the non-verbal communication cues conveyed through both avatars may differ in richness, we consider comparability given, as accurate pointing and detailed hand gestures are not task relevant, and research showed no significant effect on collaboration [46].

3.2.2 Auditory User Representations. We explore two auditory user representations – *Mono* and *Spatial Audio* – for verbal communication and auditory feedback when interacting with virtual objects (action sounds). Sound was played from the speakers of the HMD, instead of opting for noise canceling headphones, as this allows simultaneously hearing sounds from the physical environment as well (e.g., the own voice). The auditory user representations differ in the audio playback quality as described below.

Mono Audio. All sounds are played on all speakers at the same volume, without considering the position in the scene. The perception of the remote user's voice is comparable to the audio of video conferencing software.

Spatial Audio. We spatialized all sound sources in the environment, following approaches from existing research (e.g., [58]), whereby the loudness and direction of any audio source in the scene is adjusted based on a head-related transfer function (HRTF). The function uses a linear volume roll-off curve (attenuation of sound based on distance), mapping distance to loudness. The sound of any audio source in the virtual space (the remote collaborator speaking, the smartphone ringing, or sticky note manipulation sounds) is played with the correct temporal and spectral cues to create the illusion of coming from its specific direction and distance.

3.3 Implementation

The study prototype was implemented using Unity[7] (2019.4), the Mixed Reality Toolkit (MRTK)[8] (v2), and the Microsoft HoloLens 2[9] AR HMD. For a consistent mapping of virtual content to the physical space, we use Vuforia Image Targets[10] (8.5.9). The Image Target is placed at a fixed position in each room and needs to be viewed once (for several seconds) to align the virtual coordinate system with the physical space.

The head of the *Rich Avatar* is generated using the Avatar SDK[11] (HEAD 2.0). The generated head provides blend shapes for the animation of eyes and mouth. We generate the facial movements using SALSA Lip-sync[12] (2.4). The eye blinking animation is realized with the "Eyes" module. The interval between eye blinks is randomized with values between 0.5 and 5 s. One blink lasts 0.3 s. Due to the limited memory of the HoloLens 2, we are only able to use three blend shapes of the avatar's head (visemes "DD", "CH", and "aa") for mouth movements. The viseme "DD" is triggered for normalized amplitude values (of the voice data) < 0.11, "CH" for values between 0.11 and 0.43, and "aa" for values > 0.43.

The torso variations of the *Rich Avatar* were created using the Unity Multipurpose Avatar[13] (2.1) plugin, which generates fully rigged avatars including fingers. The body of the *Simple Avatar* was generated using Unity's default "Capsule" and "Sphere" objects.

The animation of the body parts is realized with the inverse kinematics solution VRIK of FinalIK[14] (1.9). Additionally, a custom finger mapping is applied whereby the 25 hand joints of the MRTK hand tracking are mapped to the 20 hand joints of the avatar by disregarding the five fingertips, which are determined based on the distal joints.

For audio spatialization, we use the Microsoft HRTF Spatializer, which is part of the Windows Mixed Reality[15] package for Unity. For *Spatial Audio*, we set "spatialize" to "true", "spatialBlend" to 1 (3D), and apply a linear volume roll-off curve for all audio sources. For *Mono Audio*, we set "spatialize" to "false", "spatialBlend" to 0 (2D), and apply the default logarithmic curve. The audio sources are positioned at the center of the corresponding virtual objects, e.g., the mouth in case of the *Rich Avatar*.

The interaction with sticky notes is realized with the manipulation handler of MRKT, with far manipulation disabled. The sticky note snapping uses multiple raycasts to detect snappable surfaces

[7]Unity: https://unity.com/ – last accessed: July 1st 2024
[8]Mixed Reality Toolkit: https://github.com/microsoft/MixedRealityToolkit-Unity – last accessed: July 1st 2024
[9]HoloLens 2: https://www.microsoft.com/hololens/hardware – last accessed: July 1st 2024
[10]Vuforia Image Target: https://library.vuforia.com/objects/image-targets – last accessed: July 1st 2024
[11]Avatar SDK: https://avatarsdk.com/– last accessed: July 1st 2024
[12]SALSA Lip-Sync: https://crazyminnowstudio.com/unity-3d/lip-sync-salsa/– last accessed: July 1st 2024
[13]Unity Multipurpose Avatar: https://assetstore.unity.com/packages/3d/characters/uma-2-unity-multipurpose-avatar-35611– last accessed: July 1st 2024
[14]FinalIK: https://assetstore.unity.com/packages/tools/animation/final-ik-14290 – last accessed: July 1st 2024
[15]Windows Mixed Reality for Unity: https://docs.unity3d.com/Packages/com.unity.xr.windowsmr.metro@4.2/manual/index.html – last accessed: July 1st, 2024

at a max distance of 20 cm. When a sticky note is released and no snappable surface is detected, gravity is applied (by disabling "isKinematic" on the rigid body) and the sticky note falls down.

For data transmission, we used Colibri [21]. Position and rotation data of the head, hands, finger joints, and sticky notes are transmitted in real-time using TCP. For the processing of the microphone data, we use the Dissonance Voice Chat[16] (6.4.6) plugin. The voice data is transmitted using UDP. The generated *Rich Avatars* are uploaded to and downloaded from Colibri's data store using its REST API.

4 User Study

We conducted a user study aimed to better understand the role of combinations of visual and auditory user representations for remote collaboration in AR. We address this overall goal by investigating the following research questions:

RQ1: How does the visual and auditory representation of remote collaborators influence users' *experiences*?

RQ2: How does the visual and auditory representation of remote collaborators influence users' *subjective perception of social presence*?

RQ3: How does the visual and auditory representation of remote collaborators influence users' *performance*?

We chose a within-subjects design with the visual and auditory user representation as independent variables. Combinations of these were presented in our prototype (see section 3), resulting in four conditions: *Simple x Mono*, *Simple x Spatial*, *Rich x Mono*, and *Rich x Spatial*. To account for potential carry-over effects, the order of conditions was determined using a balanced Latin square.

While we expect that all conditions effectively support the collaborative task, prior research suggests that increased visual realism of the avatar can enhance embodiment and social presence [45, 59]. Further, spatial audio can contribute to intelligibility, reduced cognitive load, and social presence [2, 37, 42, 43, 50, 58]. Hence, we expect conditions with *Rich Avatar* and *Spatial Audio* to be superior. However, the relative importance and interplay of the visual and auditory modalities remains unclear, and is subject of this investigation. In the following, we describe how we operationalize our dependent variables using quantitative and qualitative data collection methods.

Users' Experiences. We assessed participants' experience with the User Experience Questionnaire (UEQ-S) [47], measured their task load using the NASA Task Load Index (NASA TLX) [17, 18], and asked participants to rank conditions by preference in the concluding semi-structured interview. The interview also focused on how participants perceived the four conditions and how these facilitated collaborative activities.

Subjective Perception of Social Presence. Following prior research (e.g., [45, 62]), we assessed social presence using the Networked Minds Social Presence Inventory [16]. To additionally investigate the visual attention and proxemics, we logged participants' eye gaze behavior (e.g., viewing duration on the avatar) and the distance between the local participant and the remote participant's avatar. These aspects were also covered in the concluding interview to gain qualitative insights.

Performance. We studied the dyads' completion time for the primary task (see section 4.2) and investigated the individual call rejection time in the interruption task as an indicator for the effectiveness of spatial audio. The concluding semi-structured interview further allowed us to investigate perceived performance.

[16]Dissonance Voice Chat: https://placeholder-software.co.uk/dissonance/ – last accessed: July 1st 2024

4.1 Setup

Two labs (*Lab A* and *Lab B*) at our university served as distributed work environments (see Figure 1). Each was equipped with one standard-size table, a whiteboard, and a printed marker for room calibration. The virtual smartphones were positioned on the right and left end of the room, halfway between the whiteboard and table (see Figure 4 for exact placements). Virtual whiteboards of identical size were overlayed on the physical whiteboards in both labs, to ensure symmetric setups. Both participants were equipped with a Microsoft HoloLens 2, using the same 5GHz WiFi in both labs for synchronization. An experimenter was present in each lab, to control and monitor the study. Participants' portraits were captured in *Lab A* using a Nikon D5100 camera and studio lights.

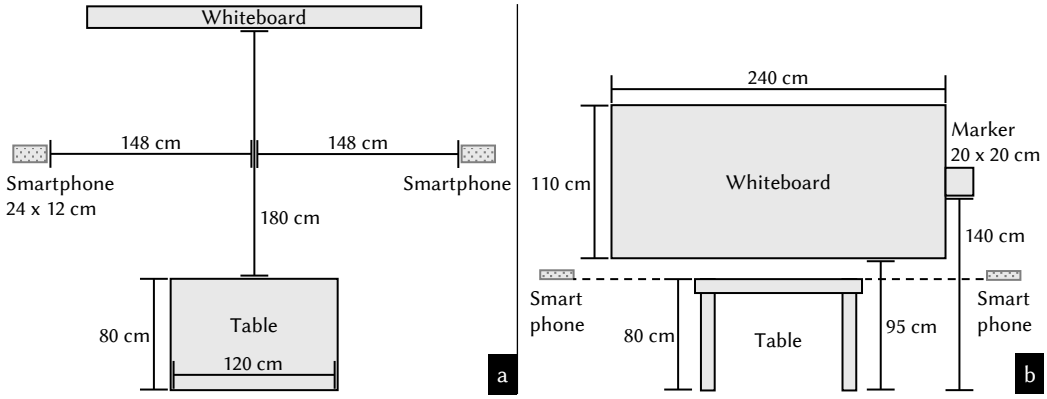

Fig. 4. The study setup of both labs as a top-down view (a) and a front view (b), including the measurements of the physical and virtual objects and their distances. Virtual objects (i.e., smartphones) are marked with dots.

4.2 Task

To reflect a common remote collaboration scenario, like brainstorming or doing thematic analyses with virtual sticky notes on online whiteboards, we chose the abstract classification task by Liu et al. [27], where participants have to sort items depending on their content (i.e., a letter of the alphabet). This task has previously been applied for individual use of wall-sized displays [25, 27], multi-device environments [61], as well as co-located and remote collaboration [1, 26]. It fulfills our requirements of being demanding enough to engage participants with the task itself and each other, being adaptable to our study setup and number of conditions, and not requiring specific prior knowledge. The complexity and task duration can be controlled by the number of letters and items, and we defined item sets as follows: Participants had to sort 40 sticky notes, each showing one of eight different letters, into eight white rectangles on the whiteboard, so that each only contains sticky notes with the same letter ($n = 5$). As suggested by previous work [61], we used the following letters: C, D, E, F, H, K, N, and R as they guarantee equal legibility [24]. While both participants could see and manipulate all 40 sticky notes, each participant could only read the letters on half of the sticky notes, while their partner saw the other half (see also Figure 1). This served to simulate different levels of knowledge among real-world collaborators and encourage communication (e.g., "what's the letter on this sticky note?"). Further, we ensured that a single participant could read a maximum of 4 (out of 5) sticky notes with the same letter, to ensure that no rectangle could be completely filled individually. Four distinct item sets of sticky notes with letters were precomputed based on the aforementioned rules - one per study trial.

At the beginning, 36 sticky notes were placed on the table in a 8 × 5 grid with a gap of 2 cm, allocating an area of 78 cm × 38 cm, while four sticky notes were pre-sorted on the whiteboard inside four rectangles (see Figure 2a). For each participant, 18 letters were visible on the table and two on the whiteboard. Once a rectangle was correctly filled with five sticky notes of the same letter, these turned green. The task was completed when all sticky notes were correctly sorted.

In our variation of a dual-task paradigm, we interrupted the primary task with a simulated incoming phone call on a virtual smartphone. Each participant was assigned their respective smartphone and was asked to decline incoming calls as fast as possible by pressing the virtual red "Decline" button. Importantly, participants were only able to decline calls on their own smartphone. This served multiple purposes: The call increased the realism of the task and forced the participants to reorient themselves in the shared environment, allowing us to study the effectiveness and relevance of spatial audio, as well as aspects of awareness of each other's activities. In each study session, there were a total of 20 interruptions: Five interruptions per condition, with each phone ringing at least once per trial and 10 times over all trials. The timing of the incoming calls was matched to the progress of the primary task: Every time six additional sticky notes were placed on the whiteboard one of the smartphones rang with a delay of two seconds, to prevent participants from perceiving a causal relationship between actions.

4.3 Procedure

The participant dyads were welcomed by two examiners in *Lab A*, they provided informed consent and demographic data, and were photographed to generate their *Rich Avatars*. Afterward, one of the examiners guided *Participant B* to *Lab B*, while *Participant A* remained in *Lab A*. Participants were individually introduced to the HMD and the application, following a protocol. After performing the calibration of eye-tracking and the room, participants learned the possible interactions (with sticky notes and the virtual smartphone) in an interactive tutorial with text, audio, and video instructions.

After the tutorial, the examiner in *Lab A* started the data recording (video, audio, data) and triggered the training phase for both participants. They were given a sheet describing their tasks and performed a complete test run with a training item set. Here, participants were represented in a minimal way: three white spheres visually indicated the head and hand position, and the auditory representation was *Mono Audio*.

Then followed the task phase, which included four trials, one per condition in counterbalanced order (*Simple x Mono*, *Simple x Spatial*, *Rich x Mono*, or *Rich x Spatial*). In each trial, participants were first asked to (1) describe each other's user representation to make them focus on the current visual and auditory combination, (2) perform a short hearing test to ensure that the spatial audio was perceived, by locating a ringing sound from each direction (i.e., the locations of both smartphones), (3) discuss the strategy for performing the task to separate discussion time from the task completion time, (4) collaboratively complete the task, and (5) individually fill out the questionnaires (UEQ-S, NASA TLX, and Networked Minds Social Presence Inventory). Each trial was initiated by the examiner in *Lab A*.

After finishing all four trials, the concluding interview was jointly performed with both participants in *Lab A*. Each study session lasted about 90 minutes, and participants were compensated for their time with the local minimum wage. We followed all ethical and sanitary guidelines of our university.

4.4 Participants

We recruited 48 participants (24 female, 24 male, 0 non-binary; age 18-37 years, $M = 22.31$, $SD = 3.55$) in dyads that previously knew each other. Of these 24 dyads, 8 were female-female, 8 male-male, and 8 mixed-gender dyads. All had normal or corrected vision (e.g., glasses) and

indicated no color vision deficiency and normal hearing. Participants were recruited at the local university and consisted of 43 students from diverse backgrounds (e.g., biology, economics, or sociology), three PhD students, and two employees. Rating their prior experience on a Likert scale (1 inexperienced – 5 very experienced), participants indicated low familiarity with AR applications ($M = 1.5$, $SD = 0.8$; 15 had experienced AR before) and similarly low VR experience ($M = 1.75$, $SD = 0.96$; 24 had experienced VR before).

4.5 Data Analysis

Our statistical analysis is based on testing against a null hypothesis of no effect, which assumes that there is no significant difference between the conditions studied. All quantitative data were analyzed using IBM SPSS 29 [23], assuming an alpha level of .05. We tested all data, except the questionnaires and preference scores, with the Shapiro-Wilk test to determine whether a normal distribution can be assumed. If the data were normally distributed, we applied a one-way repeated measures ANOVA, whereby we treat all four combinations as distinct conditions. For questionnaire ratings or non-normally distributed data, we used Friedman's ANOVA to compare the four conditions and a Wilcoxon test to compare aggregated avatar or audio conditions. Post-hoc pairwise comparisons were done using Dunn's test with Bonferroni correction (but alpha level was adjusted back to .05). Two independent coders performed an inductive thematic analysis [5] of the transcribed interview data using MAXQDA 2022 [54], with substantial intercoder reliability (Cohen's Kappa = 0.7, based on segmental overlap of 4 interviews and adjusted P_c value [6]).

5 Results

In this section, we report the results of our data analysis from the three questionnaires and the subjective ranking, the user position (proxemic zones) and eye gaze data (visual attention), the task performance, and the interview.

5.1 Questionnaires and Subjective Ranking

In the following, we report on responses to the UEQ-S and NASA TLX questionnaires as well as preference rankings, as measures for the users' experience, followed by the Networked Minds Social Presence Inventory.

5.1.1 UEQ-S. The results visualized in Figure 5 (top left) show no statistically significant differences in terms of Pragmatic Quality ($\chi^2(3) = 2.553$, $p = .466$). However, we found significant differences in Hedonic Quality ($\chi^2(3) = 47.285$, $p < .001$). Post-hoc pairwise comparisons reveal significant differences between all conditions with different *visual* user representation (*Rich x Spatial* vs. *Simple x Mono* ($z = -4.19$, $p < .001$); *Rich x Mono* vs. *Simple x Mono* ($z = -4.822$, $p < .001$); *Rich x Mono* vs. *Simple x Spatial* ($z = -4.19$, $p < .001$)), while the audio appears to have had no influence.

5.1.2 NASA TLX. Participants' overall task load was low across all conditions: *Simple x Mono* ($M = 26.61$, $SD = 18.17$), *Simple x Spatial* ($M = 26.67$, $SD = 16.69$), *Rich x Mono* ($M = 25.47$, $SD = 17.41$), and *Rich x Spatial* ($M = 25.69$, $SD = 16.25$) (see Figure 5 bottom). Results show no significant differences between conditions in the overall score ($\chi^2(3) = .853$, $p = .837$). For the subscales, significant differences were only found for Effort ($\chi^2(3) = 7.938$, $p = .047$), but post-hoc pairwise comparisons revealed no significant differences.

5.1.3 Networked Minds Social Presence Questionnaire. We assessed participants' social presence based on three subscales (see Figure 5 top right). Co-presence scores were high across all conditions: *Simple x Mono* ($M = 5.51$, $SD = 1.04$), *Simple x Spatial* ($M = 5.42$, $SD = 1.28$), *Rich x Mono*

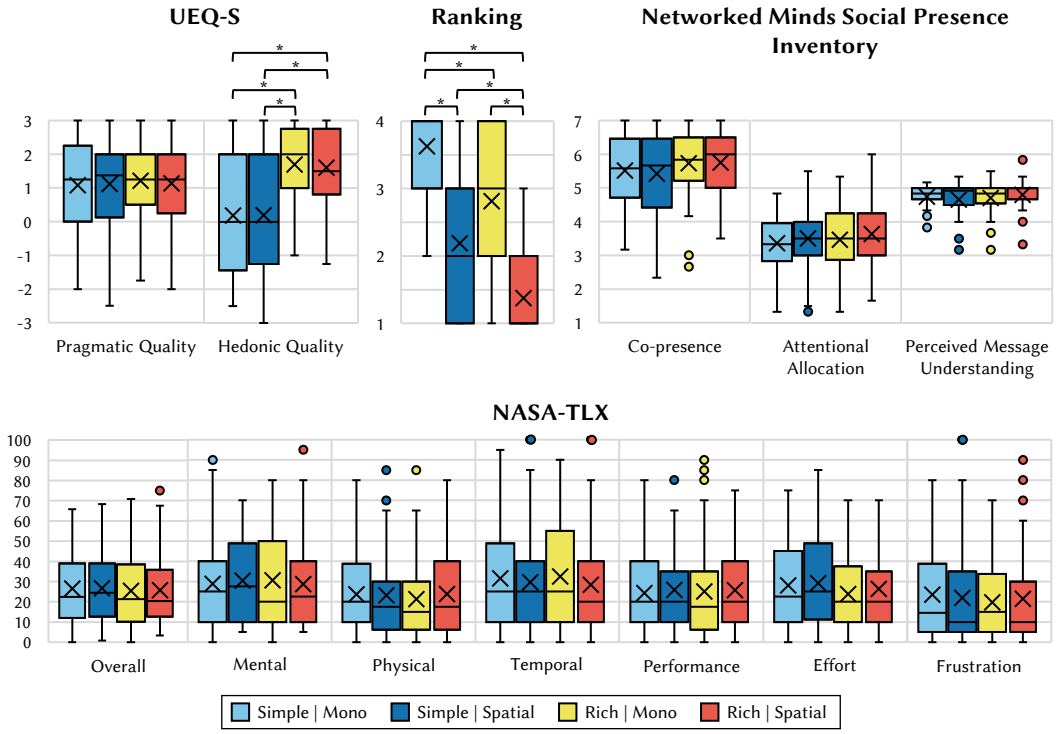

Fig. 5. Box plots visualizing the UEQ-S scores (ranging from -3 to 3), the subjective ranking asked in the interview (1 most preferred - 4 least preferred), the scores of the Networked Minds Social Presence Inventory subscales "Co-presence", "Attentional Allocation", and "Perceived Message Understanding" (ranging from 1 to 7), and the overall score and all subscales of the NASA TLX (ranging from 0 to 100).

($M = 5.73$, $SD = 0.99$), and *Rich x Spatial* ($M = 5.75$, $SD = 0.94$). Attentional Allocation reveals medium mutual attention during the task: *Simple x Mono* ($M = 3.36$, $SD = 0.86$), *Simple x Spatial* ($M = 3.5$, $SD = 0.92$), *Rich x Mono* ($M = 3.47$, $SD = 0.88$), and *Rich x Spatial* ($M = 3.64$, $SD = 0.86$). Perceived Message Understanding was predominantly good: *Simple x Mono* ($M = 4.74$, $SD = 0.33$), *Simple x Spatial* ($M = 4.65$, $SD = 0.52$), *Rich x Mono* ($M = 4.71$, $SD = 0.41$), and *Rich x Spatial* ($M = 4.8$, $SD = 0.4$). Results show no significant differences between conditions: Co-presence ($\chi^2(3) = 3.325$, $p = .344$), Attentional Allocation ($\chi^2(3) = 4.434$, $p = .218$), and Perceived Message Understanding ($\chi^2(3) = 2.768$, $p = .429$).

5.1.4 Subjective Ranking. In the concluding interview, we asked participants to rank all four conditions by their subjective preference from 1 (most preferred) to 4 (least preferred). On average, participants preferred *Rich x Spatial* ($M = 1.38$, $SD = 0.64$, $n = 34$ most preferred) followed by *Simple x Spatial* ($M = 2.19$, $SD = 0.87$, $n = 13$ most preferred), *Rich x Mono* ($M = 2.81$, $SD = 0.87$, $n = 1$ most preferred), and *Simple x Mono* ($M = 3.63$, $SD = 0.64$, $n = 0$ most preferred). Rankings were significantly different between conditions ($\chi^2(3) = 78.525$, $p < .001$). Post-hoc pairwise comparisons revealed significant differences between all conditions except *Simple x Spatial* and *Rich x Mono*, as can be see in Figure 5 (top middle): *Rich x Spatial* vs. *Simple x Mono* ($z = 8.538$, $p < .001$), *Rich x Spatial* vs. *Simple x Spatial* ($z = 3.083$, $p = .012$), *Rich x Spatial* vs.*Rich x Mono* ($z = 5.455$, $p < .001$), *Rich x Mono* vs. *Simple x Mono* ($z = 3.083$, $p = .012$), *Simple x Spatial* vs.

Simple x Mono ($z = 5.455$, $p < .001$). In the interviews, half of the participants ($n = 24$) indicated that their decision was based on the visual representation, while the other half based it on the auditory representation.

5.2 Proxemic Zones

To compute participants' interpersonal distance, we assume round torsos with a diameter of 30 cm. We categorize the distance between avatars based on the proxemic zones by Hall [14]: Intimate (< 18 in ≈ 46 cm), Personal (< 4 ft ≈ 122 cm), Social (< 12 ft ≈ 366 cm), and Public ($>= 12$ ft ≈ 366 cm). Additionally, we measured the time that virtual torsos overlapped. We used the relative duration (percentages) to the total task duration, so that it is comparable across sessions (see Figure 6). Results show no significant differences for avatar overlap ($\chi^2(3) = .356$, $p = .94$) between conditions. While none of the dyads ever went into the Public zone, the durations spent in each of the other proxemic zones was similar across conditions (i.e., no significant effects): Intimate ($F(3, 69) = .292$, $p = .831$), Personal ($\chi^2(3) = 2.95$, $p = .399$), and Social ($\chi^2(3) = .85$, $p = .837$). On average, participants spent most time within Personal space ($M = 44.38$ %, $SD = 1.25$ %), followed by Intimate ($M = 27.11$ %, $SD = 0.85$ %), and Social ($M = 25.28$ %, $SD = 0.29$ %), and rarely let avatars overlap ($M = 3.24$ %, $SD = 0.37$ %).

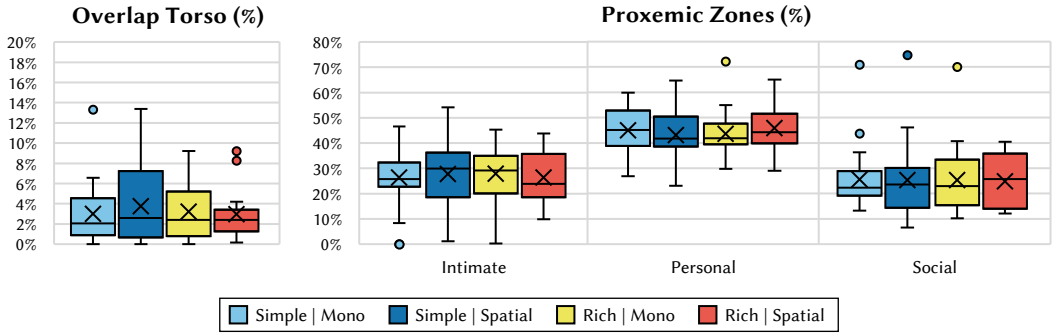

Fig. 6. Box plots visualizing the percentage of the total duration that the torsos of the avatars overlapped (left), and the participants' avatars were within the proxemic zones "Intimate", "Personal", and "Social" (right). None of the dyads were in the "Public" zone.

5.3 Visual Attention

We evaluated participants' gaze based on eye tracking data recorded on the HoloLens 2. For comparability across sessions, we compute the relative durations of looking at the other's avatar and specifically their head (see Figure 7). In addition to the analysis per condition, we also aggregated conditions with the same visual or auditory user representation. Results show no significant differences between conditions for viewing the avatar overall ($\chi^2(3) = 1.75$, $p = .626$), and neither when the conditions are aggregated by avatar ($z = -1.487$, $p = .137$) or by audio ($z = -.072$, $p = .943$). However, viewing the avatar's head was significantly different between conditions ($\chi^2(3) = 14.26$, $p = .003$). Post-hoc pairwise comparisons revealed that participants viewed the avatar's head significantly longer during *Rich x Spatial* compared to *Simple x Spatial* ($z = -3.162$, $p = .009$), as well as during *Rich x Mono* compared to *Simple x Spatial* ($z = -3.202$, $p = .008$). Here, aggregating the avatar conditions reveals that participants viewed the head of the *Rich Avatar* significantly longer than the head of the *Simple Avatar* ($z = -2.79$, $p = .005$). No significant difference was found when aggregating audio conditions ($z = -1.026$, $p = .305$).

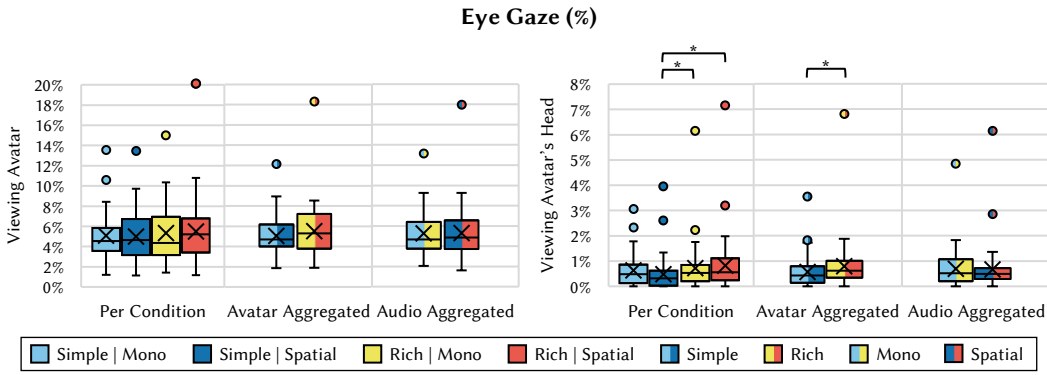

Fig. 7. Box plots visualizing the percentage of time that participants viewed the avatar and the avatar's head during the task.

5.4 Task Performance

Task completion time was measured from the start of the sorting task until all notes were correctly placed on the whiteboard (see Figure 8, left). Here statistical analysis reveals no significant differences between conditions ($\chi^2(3) = 1.15$, $p = .765$). Average task completion times were 240.07 s ($SD = 111.43$) for *Simple x Mono*, 238.21 s ($SD = 60.41$) for *Simple x Spatial*, 235.09 s ($SD = 81.45$) for *Rich x Mono*, and 240.21 s ($SD = 77.29$) for *Rich x Spatial*.

In addition, we measured the call response time for the individual call interruption from the start of the ringing until the "Decline" button was pressed (see Figure 8, right). Results show no differences in call response time per condition ($\chi^2(3) = 7.37$, $p = .061$). When aggregating conditions with the same auditory user representation, to investigate whether audio had an effect on performance in this predominantly auditory task, significant difference do become apparent: Participants responded significantly faster to a call with *Spatial Audio* than with *Mono Audio* ($z = -1.998$, $p = .047$).

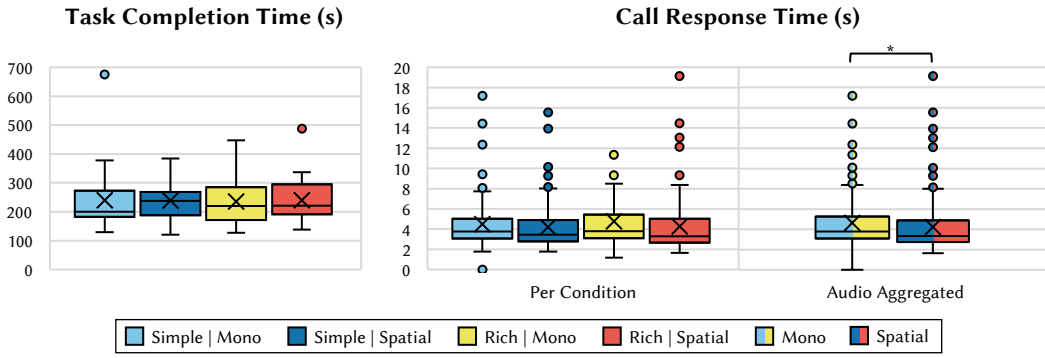

Fig. 8. Box plots visualizing measured task completion time (s) and call response time (s) by condition and when aggregating conditions with the same auditory user representation.

5.5 Thematic Analysis of Interview Data

The analysis of the interview transcripts resulted in 137 codes, that were combined into five themes, including 24 sub-themes. These address topics related to all three research questions. In the following theme descriptions, we indicate the number of individual participants whose interviews were coded (and not the number of coded segments) by '($n = x$)' after each mentioned code. The complete list of themes and sub-themes, including the associated RQ and codes, and the number of interviews coded with the respective codes can be found in Appendix A.

5.5.1 Rich avatars are perceived as stimulating. Participants reported that avatars – in general, but specifically the *Rich Avatar*– are 'fun', 'fascinating', or even 'entertaining' ($n = 12$).

5.5.2 Visual avatar richness affects the sense of being together with another person. Both visual avatar representations resulted in a feeling of being together with someone by communicating their activity and position ($n = 16$), with a more prominent effect for the *Rich Avatar*, even though few participants ($n = 3$) mentioned that it was hard to spot. For the *Rich Avatar* participants also mentioned that it was more like 'face-to-face' interaction ($n = 18$) and more intimate ($n = 8$), while the *Simple Avatar* was described as 'anonymous' ($n = 7$). Participants also felt hesitation regarding touching the *Rich Avatar* ($n = 5$).

5.5.3 Visual and auditory richness impact perceived realism. For both aspects – auditory and visual – of the user representations, participants mentioned that the richer version was more realistic (visual aspect: $n = 21$; auditory aspect: $n = 13$) because participants recognized the real person the avatar was based on ($n = 32$). Especially in comparison to the *Simple Avatar*, 'consisting of simple shapes' ($n = 36$), they described the *Rich Avatar* as more human-like ($n = 19$). They attributed this to the display of body language with gestures ($n = 14$), hands with details such as fingers ($n = 25$) and facial expressions ($n = 17$). Yet, on the other hand, some participants felt that *Rich Avatar* was not yet real (enough) ($n = 14$), as they experienced that the visualization did not truly match the real person as they would have expected ($n = 11$).

5.5.4 Spatial audio supports orientation and task completion but can impact speech intelligibility. Many participants reported that *Spatial Audio* was 'helpful' ($n = 27$) as solving the task was 'more efficient' and 'easier' ($n = 26$) with it. They could identify where the sound originated ($n = 44$) with high precision ($n = 28$). While some participants attributed good or better intelligibility ($n = 10$) with a low perceived latency ($n = 15$) to *Spatial Audio*, others perceived the audio as 'faint' ($n = 3$), and 'not clear' ($n = 3$) resulting in lower intelligibility ($n = 9$), describing *Spatial Audio* as 'demanding' ($n = 6$). In line with this, the same number of participants ($n = 6$) reported that *Mono Audio* is more efficient. Yet, the majority mentioned that *Mono Audio* increased the need to rely on visual cues ($n = 28$), making the task more challenging ($n = 12$).

5.5.5 Avatars can support specific tasks, while reduced visual features can lower distractions. Having an avatar during the collaboration task at hand was perceived as helpful by 25 participants. Yet, 20 participants felt it was unnecessary for the task. Many described the *Simple Avatar* as (more) efficient ($n = 20$), because its visualization provided all necessary information ($n = 23$), such as where the user is and what they are doing ($n = 13$). The *Rich Avatar* was described as 'distracting' ($n = 14$).

6 Discussion

In the following sections, we discuss our results, present design perspectives, and provide directions for future research. Statements are attributed to individual participants according to the following scheme: D1A and D1B are the participants A and B of dyad 1.

6.1 Users' Experiences

RQ1 relates to how the visual and auditory user representations influence the users' experiences. Based on the preference rankings, the conditions with *Spatial Audio* were perceived as superior, regardless of the visual user representation (see Figure 5). However, in the interview, only half of the participants stated that the auditory user representation was decisive for their ranking, suggesting a more balanced importance of visual and auditory characteristics, possibly also reflecting individual differences. Participant D11B, for example, stated that "[...] for this task I would prefer to have a good sound over a nice avatar"[17], while another participant saw a greater benefit of the visual aspects: "when you see the person, you collaborate more with the person" [D3A]. This ambiguity of whether the individual preference is based on auditory or visual user representation characteristics highlights the importance of investigating these aspects in combination instead of individually.

A similar pattern is reflected in the results of the UEQ-S, where conditions with a *Rich Avatar* had significantly higher ratings for hedonic qualities, regardless of the auditory user representation. This is also supported by the theme 'Rich avatars are perceived as stimulating' from our thematic analysis. During the interviews, participants referred to the *Rich Avatar* as "easy on the eyes" [D17B], or "more fun" [D18B]. This may suggest that while the *Rich Avatar* had more hedonic qualities, the *Spatial Audio* had other advantages, e.g., in terms of performance. For example, participants reported that *Spatial Audio* was helpful in solving the task efficiently because they were more aware of their surroundings, which is in line with previous research [9]. The different qualities of the *Rich Avatar* and *Spatial Audio* possibly complemented each other, which may have contributed to the vast majority ($n = 34$) preferring *Rich x Spatial*, as expected.

> **Design Perspective:** As the rich user representations were generally preferred, we recommend the use of rich visual and auditory representations in application scenarios where hedonic qualities are important for collaboration, e.g., social activities, like playing games or creative tasks.

6.2 Perceived Realism of Visual Aspects

Several participants stated that they perceived the *Rich Avatar* as more realistic (e.g., "it looked real" [D9A]), they even recognized the real person (e.g., "[directed at the other person] it looked like you" [D7B]), and positively noted the high level of detail. Yet, some participants also mentioned that it did not meet their expectations for realism. D2B said "[t]he *Rich Avatar* aimed to look human, but was not detailed enough", and added "it was wishy-washy, it was supposed to look like them but didn't." One possible reason is that the torso did not match the real person and was "comic like" [D19A], there is a small gap between the head and the torso, and it has no legs, as mentioned by eight participants. The latter were conscious design decisions since the HMD cannot track the user's legs, and the merging of the head and torso textures may look even more unrealistic (cf. section 3.2). Our findings suggest that the automatic generation of the avatar from a photo was perceived positively. However, users should be given the opportunity to customize the generated avatar further to ensure that other body parts, such as the torso or hair, can be adjusted to better match the real person (cf. RQ1). We envision a hybrid approach that combines the advantages of fast, automatic avatar generation and high customizability of manual avatar generators, while offering opportunities to resolve perceived mismatches and associated feelings of eeriness. For example, this could be incorporated in a virtual mirror, as participants also mentioned that they would have loved to see themselves.

[17] Quotes were translated from the participants' mother tongue to English.

In the qualitative results, we can clearly see that participants appreciated hand gestures, facial expressions, blinking eyes, and moving lips when speaking: "you could see emotions [...] or how the avatar moves the lips" [D6A] or "one could recognize facial expressions [...] one could see a smile" [D10A]. This confirms the importance of facial expressions for social interaction in virtual settings, as was already expressed in prior research (e.g., [4, 33, 53]), where participants reportedly missed it, or the authors criticized the limited support thereof in the investigated virtual meeting platforms. The fully articulated hands of the *Rich Avatar* were also mentioned frequently ($n = 25$, e.g., "fingers were a plus" [D20A]). In contrast, participants unsurprisingly described the face of the *Simple Avatar* as "cold" [D13A], "lifeless" [D2B], or "indifferent" [D6A]. In combination with the missing ability to perform hand gestures, this strongly hampered non-verbal communication with the *Simple Avatar*.

Design Perspective: The photo-generated avatar was generally well received, yet some participants demanded more visual detail that was not accurately captured or well-reflected. Therefore, we suggest a hybrid method – combining photo-generated user representations with manual customization – allowing the creation of pleasing, rich user presentations. An avatar creation interface could support this by allowing users to upload a portrait photo that is used to pre-select matching options from the supported avatar feature set.

6.3 Perceived Realism of Auditory Aspects

Participants mentioned in the interview that *Spatial Audio* increased the perceived realism of the collaborator's speech and environmental sounds (cf. RQ1). For example, they described *Spatial Audio* as "closer to reality and more fitting" [D17B], or noted that speech seemed "more human-like – not so robotic" [D23B]. However, most statements regarding the benefit of the *Spatial Audio* focused on that it was "more pleasant to listen to" [D1B], "more fun to know the direction of the sound" [D9B]. As expected, with regards to the task, it was "[...] more comfortable as one knew where the sound, for example, from the phone originated" [D18A]. Arguably, in a scenario with more than one remote collaborator, this benefit of *Spatial Audio* would become evident also with respect to identifying the current speaker [2, 22, 44, 50].

Interestingly, participants' opinions of speech intelligibility with *Spatial Audio* were mixed, with some mentioning good or even better intelligibility and low perceived latency, while a small subset perceived audio as faint and unclear, as also reported in some related work [58]. Notably, some participants remarked for *Mono Audio* that it made "communication [...] easier" [D20A]. Two possible explanations for this difference in perception of *Spatial Audio* could be (1) the used HRTF, which matches individual sound perception to a different extent (though our hearing tests ensured that all participants could correctly identify the direction of the ringing), or (2) the applied volume roll-off curve to simulate the interaural level difference (ILD). While we already used a linear (instead of logarithmic, cf. section 3) volume roll-off curve to compensate for previously reported problems (e.g., [58]), this may still have reduced speech intelligibility when participants stood further apart or back to back. Therefore, we suggest further refinement of the *Spatial Audio* parameters. For example, a different volume roll-off curve could be used specifically for speech playback, to further reduce its impact within shorter distances. However, this needs further investigation.

Notably, the qualities discussed above are mainly pragmatic and *Spatial Audio* did not evidently contribute to hedonic quality ratings in the UEQ-S. This may indicate that auditory user representations are rather appreciated for practical reasons, as *Spatial Audio* nevertheless ranked higher in preference than *Mono Audio*. This is in line with prior work reporting higher perceived overall quality of spatial audio [2, 37, 43, 50].

Design Perspective: As *Spatial Audio* was generally preferred for its realism and support of the tasks – excluding speech intelligibility issues – we recommend using *Spatial Audio* to support collaboration scenarios. To further improve speech intelligibility, we suggest investigating additional refinement of the volume roll-off curve, for example, to further reduce its impact within shorter distances.

6.4 Perceived Social Presence

In RQ2, we asked how the visual and auditory representations influence the users' subjective perception of social presence. Our data from the Networked Minds Social Presence Inventory revealed no significant differences between conditions and participants had a strong sense of co-presence across all conditions. This was further confirmed in interview statements, e.g., participant D7B described the presence of an avatar as "having naturally a bit more of a feeling that another person is there" (cf. RQ2). This is in line with research reporting high social presence for even abstract avatars [51], as well as inconclusive results regarding the impact of avatar representations on social presence [55].

However, our interviews reveal subtle differences in how the avatars were perceived: Users explicitly highlighted a sense of social presence with the *Rich Avatar* (e.g., "it was like sticking notes to a wall with a friend" [D1A], or "[I had] already forgotten [...] that the person is not really there" [D6B]). Collaboration with the *Rich Avatar* was perceived as more personal and intimate: "I directly felt at ease. I felt more connected with that person" [D23A], "[it] created more personal connection" [D19B], or "[I had] the feeling I work with them and not with something like a robot" [D24A]. In contrast, "with the *Simple Avatar* it was obvious that is not a human" [D2B], and participants reported it limited the personal connection (e.g., "[it was] strange, because he seemed very anonymous, very cold, very distant" [D23A]). Hence, we might interpret from our interview findings that the *Rich Avatar* indeed contributes to the sense of social presence. However, such differences cannot be detected with the employed questionnaires, as these do not cover the affective aspects of social presence.

Reflecting on auditory user representation characteristics, participants' comments about *Spatial Audio* reveal that it supports spatial awareness, not only for environmental objects (i.e., the virtual smartphones) but also for (moving) remote collaborators. This may suggest an increased sense of spatial presence. Importantly, the aforementioned speech intelligibility issues with *Spatial Audio* did not affect "Perceived Message Understanding" (see Figure 5).

Design Perspective: Although our questionnaire results on perceived social presence show no significant differences between conditions, the interview reveals benefits in affective qualities of social presence for the *Rich Avatar* and an increased sense of spatial awareness for *Spatial Audio*. Therefore, if the collaboration scenario aims at facilitating intimate human-to-human connections, we recommend using *Rich x Spatial* representations. Examples are meetings with family, friends, or significant others using social mixed reality applications. In other scenarios, such as efficient work meetings, *Simple x Spatial* should suffice to provide a high sense of co-presence.

6.5 Proxemics

The participants' interpersonal distance was similar across conditions. Overall, participants spent most time within the "Personal" space (44 % of the time) and "Intimate" space (27 %). This was expected due to the nature of the task and layout of the collaboration space: The workspaces containing sticky notes (i.e., table and whiteboard surfaces) were shared by both users, so they

would need to stand close together to engage with the materials simultaneously. Additionally, all dyads were familiar with each other, which has been shown to reduce the comfortable distance between people [19]. However, although the work area at the table was rather small, the avatars overlapped only about 3 % of the time on average. The interviews reveal that participants felt "at the table, the other's avatar was in my way" [D23B] and that they did not want to reach or walk through it. These results are strong indicators that both avatars provided a high sense of spatial presence (cf. RQ2).

We can leverage this knowledge for future designs of a proxemics-adaptive user representation that, for example, fades, scales, or changes its constitution or abstractness as users approach or reach through it. This may increase the efficiency of collaborating with avatars and avoid limiting remote collaboration in AR by physical constraints known from co-located collaboration, e.g., the other person standing in the way, invading personal space, or obstructing the view.

> **Design Perspective:** Our results show that participants avoided overlapping with the collaborator's user representations, limiting the efficiency of virtual collaboration with (simulated) physical constraints. Therefore, we propose a proxemics-adaptive user representation that changes its visual representation when in close proximity to others (e.g., by fading, scaling, or abstracting) for social mixed reality applications in general, regardless of the scenario.

6.6 Gaze Behavior and Attention

Both avatars reflected the user's body language to the same degree – with the exception of the fingers – as the facial expressions (eye blinking, lip-syncing) were only simulated. Nevertheless, analysis of participants' gaze revealed that participants looked significantly longer at the head of the *Rich Avatar* than at the one of the *Simple Avatar* (cf. Figure 7). This is in line with previous work showing that realistic avatars lead to a more human-like gaze behavior [55]. This was mirrored in the interviews, as participants reported that they liked looking at the *Rich Avatar* (cf. section 6.2). Our results suggest that the simulated facial expressions when speaking created the illusion of richer non-verbal communication cues and may have had a positive effect on social presence (cf. RQ2). However, the results could not be confirmed with subjective ratings on "Attentional Allocation", which show a slight increase for *Rich x Spatial* but no significant differences. Hence, while some participants described the facial expressions as generic, the overall positive response suggests that simulated facial expressions offer an effective approach, if the used HMD does not provide the technical capabilities for face tracking. Further, a *Simple Avatar* could incorporate 'stylized' facial expressions, and their effects on the perceived effectiveness of communication should be investigated in future research. Remarkably, with the *Simple Avatar* the head was on average viewed longer when combined with *Mono Audio* (*Simple x Mono*) than with *Spatial Audio* (*Simple x Spatial*). Here, we assume that participants attempted to visually compensate for the lack of *Spatial Audio* cues compared to *Simple x Spatial*, despite the lack of facial expressions of the *Simple Avatar*.

> **Design Perspective:** Since our results suggest that even simulated facial expressions subjectively improve communication, possibly through the illusion of richer body language, we recommend supporting real facial expressions when tracking is available or simulation thereof based on other available information such as voice.

6.7 Overall Impact on Task Performance

Investigating RQ3, we analyzed the influence of visual and auditory user representation characteristics on users' performance. Task completion times from the collaborative sorting task was

comparable across conditions. In the interview, participants explained that the avatar was not essential for completing the task ($n = 20$, e.g., "I did not need to see her to do [the task]" [D4A]) or "[...] it was more about the voice [...], appearance did not matter [...]" [D2B]. A few participants even stated that a simpler visualization (e.g., the three spheres in the training) would be sufficient. Similar insights were reported by McDonnell et al. [29], who "found that the audio and animation contributed to the interpretation of the characters' intention rather than the render style."

However, the avatar visualization did impact participants' *perceived* performance: The *Simple Avatar* was described as efficient, since it communicated the remote person's position and activity for the sorting task, and the *Rich Avatar* as distracting, because "it attracts more attention to the avatar away from the task" [D17B]. Specifically, "[...] there are many more details you can observe, and it takes more time or effort" [D15B]. The eye gaze data confirm this perceived pattern of attention, as participants focused significantly longer on the *Rich Avatar*'s head (cf. Figure 7), yet our data do not allow us to attribute this increased attention to any of its specific features (e.g., lip-syncing or eye blinking). Further, some participants explained that they were less aware of the *Simple Avatar*'s presence, which appears to stand in contrast to participants perceiving the *Simple Avatar* as an obstacle because it was "in their way" – a problem that participants attributed to both avatars. A possible design solution could be a variation of the aforementioned adaptive avatar (see section 6.5), that additionally considers the type or phase of a task, e.g., using a richer visual representation during discussions and a reduced representation during focused parallel work. However, it is important to consider that these adaptations may, in turn, also affect perceived social presence and attention toward the collaborators and requires more in-depth research.

Design Perspective: While the visual aspect of the user representation did not affect objective task performance, the visually rich representation was perceived as distracting. Therefore, we suggest reduced avatar representations for collaborative tasks where efficiency and performance are important and hedonic aspects play only a subordinate role, as a feature-rich avatar can potentially lead to distraction from the task.

Reflecting on the auditory user representation, participants reported that *Spatial Audio* was helpful in solving the task efficiently because they were more aware of their surroundings, which is in line with previous research [9, 58]. These reports are supported by call response times, which were significantly shorter, i.e., participants were faster in responding with *Spatial Audio* compared to *Mono Audio* (see "Audio Aggregated" in Figure 8). Notably, *Spatial Audio* only improved performance in the secondary task of which the user was notified through environmental sounds, but no effect was apparent in the primary sticky note sorting task. Therefore, results may differ for tasks that require frequent shifts of attention or transitions between close collaboration and individual parallel work. Furthermore, participants found that *Mono Audio* increased their dependency on visual cues because the direction of audio sources could not be identified, which impacted teamwork as "you [directed at the other person] could not assess whether you are in my way" [D5A]. Arguably, the relevance of *Spatial Audio* might increase for collaborative scenarios with more than two users (e.g., for speaker identification – see section 6.3).

Design Perspective: As *Spatial Audio* improved participants' performance on the parallel task using environmental sounds, we recommend applying *Spatial Audio* for scenarios that include multiple task-related objects that require attention shifts, especially when they are spatially distributed. Arguably, this might also apply to non-environmental sounds (collaborator's action sounds and speech).

7 Limitations and Future Work

A limitation we encountered in our study was that some users found the volume too low (mostly for *Spatial Audio*). As we wanted to control this factor, they were not allowed to adjust the volume themselves. While no participant reported hearing deficiencies, in future work the volume should be calibrated to minor differences in hearing ability. Further, the volume roll-off curve, which decreases volume by distance to mimic sound propagation, could be refined based on how users modulate their speaking volume depending on their collaborator's proximity in co-located settings.

With regards to our study task, the sorting task [27] was deliberately chosen, as it was used to study collaborative activities before [1, 26] and could be adapted to our specific setup. Yet, for this particular task, the visual user representation was potentially less essential in terms of effectiveness compared to the auditory user representation, and some users even commented on preferring the minimal visual user representation with the spheres from the tutorial. However, the hedonic qualities of the *Rich Avatar* led to an improved user experience. Thus, the choice of visual user representations may be described as a trade-off between performance and user experience.

We cannot disentangle the contributions of individual features (e.g., eye blinking, facial expressions, or articulated hands) towards experience, and as prior research suggests that detailed hands do not provide any significant benefit for remote collaboration [46], a systematic investigation is needed to explore both their pragmatic and hedonic qualities. The value of visual richness could also be further studied using tasks that more strongly involve face-to-face communication [10], or joint assembly or puzzle-like tasks that more strongly encourage cooperation. Eventually, such tasks might also allow us to systematically study how visual and auditory user representations influence collaboration and communication (e.g., coupling styles).

8 Conclusion

This work explores symmetric remote collaboration in AR, leveraging increasingly detailed user representations enabled through current technology, which reflect the state-of-the-art in consumer applications. Specifically, we investigate the impact of different combinations of visual (*Simple Avatar* vs. *Rich Avatar*) and auditory (*Mono Audio* vs. *Spatial Audio*) user representations on users' experience, social presence and performance in a user study (n = 48, 24 dyads). Our findings reveal a preference for user representations with *Spatial Audio*, highlighting its positive influence on overall user experience. Additionally, the visually more detailed *Rich Avatar* significantly enhanced hedonic aspects of the user experience, and drew increased visual attention to the avatar's head. Notably, the visual representation did not affect performance, while *Spatial Audio* led to faster reaction times for parallel tasks not requiring continuous attention.

We condensed our findings into design perspectives for user representations that take the collaboration scenario and objective into account: When hedonic qualities of the user experience (e.g., playful or creative activities) or social interaction on a personal level (e.g., meetings with family, friends, or significant others) are important, a *Rich x Spatial* user representation is recommended. In scenarios where a personal connection is of lower priority (e.g., efficient work meetings), *Simple x Spatial* can already provide a high sense of co-presence and contribute to increased (perceived) task performance by reducing distractions.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2117 – 422037984 and Project-ID 251654672 – TRR 161, as well as the Innovation Fund Denmark (IFD grant no. 9090-00002B; MADE Fast Project). We thank Samar Abed for her great support in conducting the study.

A Full list of themes and codes

Table 1. Table listing all themes (in the order of appearance in the result section 5.5), the corresponding sub-themes and their associated RQs, and the corresponding codes, including the number of individual participants whose interviews were coded (and not the number of coded segments) with the respective code in the fourth column 'n'.

Theme			
Sub-theme	**RQ**	**Code**	**n**
Theme: Rich avatars are perceived as stimulating.			
Having an avatar is stimulating	RQ1	cool / fun	6
		Novelty Effekt	9
Rich Avatar is stimulating	RQ1	fun / cool / fascinating / entertaining	12
Theme: Visual avatar richness affects sense of being together with another person.			
Having an avatar creates a sense of (physical) co-presence	RQ2	joint activity	5
		communicates activity and position	16
		does not want to touch avatar	6
		person actually present	4
		prefers having an avatar	17
		in their way	22
Rich Avatar creates a personal connection	RQ2	more intimacy	8
		more 'face-to-face'	18
Rich Avatar creates a sense of (physical) co-presence	RQ2	joint activity with real person	12
		improved observability of position and actions	16
		does not want to touch avatar	5
		seems like person actually present	15
		in their way	9
Rich Avatar is less perceptible	RQ2	low / no co-presence	4
		hard to spot avatar	3
Simple Avatar limits personal connection	RQ2	anonymous / impersonal	7
Simple Avatar was perceived as an obstacle	RQ2	distracting	4
		bulky	4
		in their way	10
Theme: Visual and auditory richness impacts perceived realism.			
Rich Avatar resembles real person	RQ1	focus on the face	9
		rich in detail	19
		(more) human	19
		resembles real person	32
		(more) realistic	21
		face resembles real person	18
[*Rich Avatar*] Human characteristics		(recognizable) arms	6
		brows	3
		blinking	3
		hair	9
		similar skin color	5
		(recognizable) hands / fingers	25
		clothes	13
		moving mouth when speaking	21
		teeth	6
Rich Avatar mimics or creates the illusion of body language	RQ2	gestures	14
		facial expressions	17
Rich Avatar does not meet expectations regarding realism	RQ1	scary	4

Sub-theme	RQ	Code	n
		missing legs	8
		reduce level of detail	6
		discrepancy between expectations and visualization	8
		discrepancy between avatar and real person	11
		weird	12
		generic facial expressions	9
		not realistic	14
Simple Avatar reduces access to body language	RQ2	unfeeling / indifferent	4
		absence of gestures	2
		absence of facial expressions	7
		reduced / no communication of activity and position	5
Simple Avatar was not perceived as human	RQ2	abstract	24
		bright / glowing	1
		boring	5
		scary	2
		consisting of simple shapes	36
		marshmallow	12
		tic tac	12
		lego (ghost)	4
		Michelin Man	12
		not human	17
[*Simple Avatar*] Reduced humanda characteristics		missing legs	4
		lacking character	2
		no face	8
		no mouth	6
		no hair	2
		spheres instead of (real) hands	11
		dots as eyes	17
		unnatural	6
		confusing	5
		resembles a human	4
Spatial Audio increases perceived realism	RQ1	prefers realistic sound	2
		(more) realistic	13
Theme: Spatial audio support orientation and task completion but can impact speech intelligibility.			
Spatial Audio facilitates task completion	RQ3	(more) pleasant	7
		efficient	26
		helpful	27
Spatial Audio supports the orientation	RQ2	easier to perceive the space and its content	16
		high accuracy	28
		person actually present	1
		position (of person) recognizable	12
		direction (of sound source) recognizable	44
		direction (of sound source) recognizable in training	14
Spatial Audio impacts speech intelligibility	RQ3	good/better intelligibility	10
		low perceived latency	15
		faint (sound)	3
		not clear (sound)	3
		low(er) intelligibility	9
Spatial Audio increases perceived task load	RQ3	more demanding / distracting	6

Sub-theme	RQ	Code	n
		identifying direction of sound source hard during task completion	13
		relying on visual indicators	6
Mono Audio can facilitate task completion	RQ3	(more) efficient	6
		helpful	2
Mono Audio increases the dependency on visual cues	RQ3	distracting	7
		reduced / no sense of space	5
		direction (of sound source) not recognizable during training	6
		direction (of sound source) not recognizable	34
		sound origin on the right	5
		less efficient / more difficult	12
		relying on visual indicators	28
Theme: Avatars can support specific tasks, while reduced visual features can lower distractions.			
Having an avatar was not essential for task completion	RQ3	task needs little interaction	11
		irrelevant for task	20
Full (upper) body avatar impacts task completion	RQ3	gestures	4
		helpful	25
		not helpful	8
		prefers training avatar (3 spheres)	6
		disturbing	8
		unclear what exactly is transmitted to remote person	4
Rich Avatar impacts task completion	RQ3	(more) distracting	14
		(more) efficient / supporting	11
Simple Avatar is less distracting	RQ3	(more) efficient	20
		low distraction due to low co-presence	10
		detailed enough	23
		communicates activity and position	13

References

[1] Ignacio Avellino, Cédric Fleury, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2017. CamRay: Camera Arrays Support Remote Collaboration on Wall-Sized Displays. 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, 6718–6729. https://doi.org/10.1145/3025453.3025604

[2] Jessica J. Baldis. 2001. Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. Association for Computing Machinery, New York, NY, USA, 166–173. https://doi.org/10.1145/365024.365092

[3] Paul Bertelson and Monique Radeau. 1976. Ventriloquism, Sensory Interaction, and Response Bias: Remarks on the Paper by Choe, Welch, Gilford, and Juola. *Perception & Psychophysics* 19, 6 (Nov. 1976), 531–535. https://doi.org/10.3758/BF03211222

[4] Michael Bonfert, Anke V. Reinschluessel, Susanne Putze, Yenchin Lai, Dmitry Alexandrovsky, Rainer Malaka, and Tanja Döring. 2023. Seeing the Faces Is so Important—Experiences from Online Team Meetings on Commercial Virtual Reality Platforms. *Frontiers in Virtual Reality* 3 (Jan. 2023), 945791. https://doi.org/10.3389/frvir.2022.945791

[5] Virginia Braun and Victoria Clarke. 2012. Thematic Analysis. In *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological.*, Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf, and Kenneth J. Sher (Eds.). American Psychological Association, Washington, 57–71. https://doi.org/10.1037/13620-004

[6] Robert L. Brennan and Dale J. Prediger. 1981. Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement* 41, 3 (1981), 687–699. https://doi.org/10.1177/001316448104100307

[7] SungIk Cho, Seung-wook Kim, JongMin Lee, JeongHyeon Ahn, and JungHyun Han. 2020. Effects of Volumetric Capture Avatars on Social Presence in Immersive Virtual Environments. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Atlanta, GA, USA, 26–34. https://doi.org/10.1109/VR46266.2020.00020

[8] Daniel Immanuel Fink, Johannes Zagermann, Harald Reiterer, and Hans-Christian Jetter. 2022. Re-locations: Augmenting Personal and Shared Workspaces to Support Remote Collaboration in Incongruent Spaces. *Proc. ACM Hum.-Comput. Interact.* 6, ISS, Article 556 (Nov. 2022), 30 pages. https://doi.org/10.1145/3567709

[9] William W. Gaver. 1989. The SonicFinder: An Interface That Uses Auditory Icons. *Human–Computer Interaction* 4, 1 (1989), 67–94. https://doi.org/10.1207/s15327051hci0401_3

[10] Georges Grinstein, Theresa O'Connell, Sharon Laskowski, Catherine Plaisant, Jean Scholtz, and Mark Whiting. 2006. VAST 2006 Contest - A Tale of Alderwood. In *2006 IEEE Symposium On Visual Analytics Science And Technology*. 215–216. https://doi.org/10.1109/VAST.2006.261420

[11] 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. https://doi.org/10.1145/3411764.3445698

[12] Jens Emil Sloth Grønbæk, Ken Pfeuffer, Eduardo Velloso, Morten Astrup, Melanie Isabel Sønderkær Pedersen, Martin Kjær, Germán Leiva, and Hans Gellersen. 2023. Partially Blended Realities: Aligning Dissimilar Spaces for Distributed Mixed Reality Meetings *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 456, 16 pages. https://doi.org/10.1145/3544548.3581515

[13] Jens Emil Sloth Grønbæk, Juan Sánchez Esquivel, Germán Leiva, Eduardo Velloso, Hans Gellersen, and Ken Pfeuffer. 2024. Blended Whiteboard: Physicality and Reconfigurability in Remote Mixed Reality Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 798, 16 pages. https://doi.org/10.1145/3613904.3642293

[14] Edward T Hall. 1966. *The hidden dimension*. Vol. 609. Anchor.

[15] Gunnar Harboe and Elaine M. Huang. 2015. Real-World Affinity Diagramming Practices: Bridging the Paper-Digital Gap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/2702123.2702561

[16] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. In *Seventh annual international workshop: Presence*, Vol. 2004. Universidad Politecnica de Valencia Valencia, Spain.

[17] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909

[18] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[19] Leslie A. Hayduk. 1983. Personal Space: Where We Now Stand. *Psychological Bulletin* 94, 2 (Sept. 1983), 293–335. https://doi.org/10.1037/0033-2909.94.2.293

[20] Daniel Hepperle, Christian Felix Purps, Jonas Deuchler, and Matthias Wölfel. 2022. Aspects of visual avatar appearance: self-representation, display type, and uncanny valley. *The Visual Computer* 38, 4 (April 2022), 1227–1244. https://doi.org/10.1007/s00371-021-02151-0

[21] Sebastian Hubenschmid, Daniel Immanuel Fink, Johannes Zagermann, Jonathan Wieland, Harald Reiterer, and Tiare Feuchtner. 2023. Colibri: A Toolkit for Rapid Prototyping of Networking Across Realities. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 9–13. https://doi.org/10.1109/ISMAR-Adjunct60411.2023.00010

[22] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 794, 14 pages. https://doi.org/10.1145/3544548.3581085

[23] IBM Corp. 2023. *IBM SPSS Statistics for Windows*. Armonk, NY.

[24] British Standards Institution. 2003. BS 4274-1: 2003: Visual acuity test types–Part 1: Test charts for clinical determination of distance visual acuity–Specification.

[25] Mikkel R. Jakobsen and Kasper Hornbæk. 2015. Is Moving Improving? Some Effects of Locomotion in Wall-Display Interaction. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 4169–4178. https://doi.org/10.1145/2702123.2702312

[26] Can Liu, Olivier Chapuis, Michel Beaudouin-Lafon, and Eric Lecolinet. 2016. Shared Interaction on a Wall-Sized Display in a Data Manipulation Task. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2075–2086. https://doi.org/10.1145/2858036.2858039

[27] Can Liu, Olivier Chapuis, Michel Beaudouin-Lafon, Eric Lecolinet, and Wendy E. Mackay. 2014. Effects of Display Size and Navigation Type on a Classification Task. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. ACM, New York, NY, USA, 4147–4156. https://doi.org/10.1145/2556288.2557020

[28] David Mal, Erik Wolf, Nina Döllinger, Mario Botsch, Carolin Wienrich, and Marc Erich Latoschik. 2024. From 2D-Screens to VR: Exploring the Effect of Immersion on the Plausibility of Virtual Humans. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 167, 8 pages. https://doi.org/10.1145/3613905.3650773

[29] Rachel McDonnell, Martin Breidt, and Heinrich H. Bülthoff. 2012. Render me real? investigating the effect of render style on the perception of animated virtual humans. *ACM Trans. Graph.* 31, 4, Article 91 (jul 2012), 11 pages. https://doi.org/10.1145/2185520.2185587

[30] Rachel McDonnell, Michéal Larkin, Benjamín Hernández, Isaac Rudomin, and Carol O'Sullivan. 2009. Eye-Catching Crowds: Saliency Based Selective Variation. *ACM Trans. Graph.* 28, 3, Article 55 (July 2009), 10 pages. https://doi.org/10.1145/1531326.1531361

[31] Harry Mcgurk and John Macdonald. 1976. Hearing Lips and Seeing Voices. *Nature* 264, 5588 (Dec. 1976), 746–748. https://doi.org/10.1038/264746a0

[32] Masahiro Mori, Karl MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100. https://doi.org/10.1109/MRA.2012.2192811

[33] Fares Moustafa and Anthony Steed. 2018. A longitudinal study of small group interaction in social virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology* (Tokyo, Japan) *(VRST '18)*. Association for Computing Machinery, New York, NY, USA, Article 22, 10 pages. https://doi.org/10.1145/3281505.3281527

[34] 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. https://doi.org/10.1145/2556288.2557000

[35] Jens Müller, Roman Rädle, and Harald Reiterer. 2017. Remote Collaboration With Mixed Reality Displays: How Shared Virtual Landmarks Facilitate Spatial Referencing. 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, 6481–6486. https://doi.org/10.1145/3025453.3025717

[36] Jens Müller, Johannes Zagermann, Jonathan Wieland, Ulrike Pfeil, and Harald Reiterer. 2019. A Qualitative Comparison Between Augmented and Virtual Reality Collaboration with Handheld Devices. In *Proceedings of Mensch Und Computer 2019* (Hamburg, Germany) *(MuC'19)*. Association for Computing Machinery, New York, NY, USA, 399–410. https://doi.org/10.1145/3340764.3340773

[37] 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. https://doi.org/10.1145/3596671.3598578

[38] Patrick Aggergaard Olin, Ahmad Mohammad Issa, Tiare Feuchtner, and Kaj Grønbæk. 2021. Designing for Heterogeneous Cross-Device Collaboration and Social Interaction in Virtual Reality. In *Proceedings of the 32nd Australian Conference on Human-Computer Interaction* (Sydney, NSW, Australia) *(OzCHI '20)*. Association for Computing Machinery, New York, NY, USA, 112–127. https://doi.org/10.1145/3441000.3441070

[39] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-Time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 741–754. https://doi.org/10.1145/2984511.2984517

[40] Thammathip Piumsomboon, Arindam Day, Barrett Ens, Youngho Lee, Gun Lee, and Mark Billinghurst. 2017. Exploring Enhancements for Remote Mixed Reality Collaboration. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications* (Bangkok, Thailand) *(SA '17)*. Association for Computing Machinery, New York, NY, USA, Article 16, 5 pages. https://doi.org/10.1145/3132787.3139200

[41] Thammathip Piumsomboon, Gun A. Lee, Jonathon D. Hart, Barrett Ens, Robert W. Lindeman, Bruce H. Thomas, and Mark Billinghurst. 2018. Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration. In *Proceedings of*

the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173620

[42] Troels Rasmussen, Tiare Feuchtner, Weidong Huang, and Kaj Grønbæk. 2022. Supporting workspace awareness in remote assistance through a flexible multi-camera system and Augmented Reality awareness cues. *Journal of Visual Communication and Image Representation* 89 (2022), 103655. https://doi.org/10.1016/j.jvcir.2022.103655

[43] Gilbert L. Ricard and Susan L. Meirs. 1994. Intelligibility and Localization of Speech from Virtual Directions. *Human Factors* 36, 1 (1994), 120–128. https://doi.org/10.1177/001872089403600107

[44] Loïc Rosset, Hamed Alavi, Sailin Zhong, and Denis Lalanne. 2021. Already It Was Hard to Tell Who's Speaking Over There, and Now Face Masks! Can Binaural Audio Help Remote Participation in Hybrid Meetings?. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 403, 7 pages. https://doi.org/10.1145/3411763.3451802

[45] Anca Salagean, Eleanor Crellin, Martin Parsons, Darren Cosker, and Danaë Stanton Fraser. 2023. Meeting Your Virtual Twin: Effects of Photorealism and Personalization on Embodiment, Self-Identification and Perception of Self-Avatars in 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 499, 16 pages. https://doi.org/10.1145/3544548.3581182

[46] Alexander Schäfer, Gerd Reis, and Didier Stricker. 2022. A Survey on Synchronous Augmented, Virtual, And Mixed Reality Remote Collaboration Systems. *ACM Comput. Surv.* 55, 6, Article 116 (dec 2022), 27 pages. https://doi.org/10.1145/3533376

[47] Martin Schrepp, Jörg Thomaschewski, and Andreas Hinderks. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4, 6 (12/2017 2017), 103–108. https://doi.org/10.9781/ijimai.2017.09.001

[48] 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. https://doi.org/10.1109/MCG.2018.193142628

[49] Richard Skarbez, Missie Smith, and Mary C. Whitton. 2021. Revisiting Milgram and Kishino's Reality-Virtuality Continuum. *Frontiers in Virtual Reality* 2 (March 2021), 647997. https://doi.org/10.3389/frvir.2021.647997

[50] Janto Skowronek and Alexander Raake. 2015. Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls. *Speech Communication* 66 (2015), 154–175. https://doi.org/10.1016/j.specom.2014.10.003

[51] Harrison Jesse Smith and Michael Neff. 2018. Communication Behavior in Embodied Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173863

[52] Antonella Maselli Sofia Seinfeld, Tiare Feuchtner and Jörg Müller. 2021. User Representations in Human-Computer Interaction. *Human–Computer Interaction* 36, 5-6 (2021), 400–438. https://doi.org/10.1080/07370024.2020.1724790

[53] Theresa Jean Tanenbaum, Nazely Hartoonian, and Jeffrey Bryan. 2020. "How Do I Make This Thing Smile?": An Inventory of Expressive Nonverbal Communication in Commercial Social Virtual Reality Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376606

[54] VERBI Software. 2021. *MAXQDA 2022*. Berlin, Germany.

[55] Florian Weidner, Gerd Boettcher, Stephanie Arevalo Arboleda, Chenyao Diao, Luljeta Sinani, Christian Kunert, Christoph Gerhardt, Wolfgang Broll, and Alexander Raake. 2023. A Systematic Review on the Visualization of Avatars and Agents in AR & VR displayed using Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2596–2606. https://doi.org/10.1109/TVCG.2023.3247072

[56] Emily Wong, Jens Emil Sloth Grønbæk, and Eduardo Velloso. 2024. The Jamais Vu Effect: Understanding the Fragile Illusion of Co-Presence in Mixed Reality. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 2227–2246. https://doi.org/10.1145/3643834.3661574

[57] Haijun Xia, Sebastian Herscher, Ken Perlin, and Daniel Wigdor. 2018. Spacetime: Enabling Fluid Individual and Collaborative Editing in Virtual Reality. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 853–866. https://doi.org/10.1145/3242587.3242597

[58] Jing Yang, Prasanth Sasikumar, Huidong Bai, Amit Barde, Gábor Sörös, and Mark Billinghurst. 2020. The Effects of Spatial Auditory and Visual Cues on Mixed Reality Remote Collaboration. *Journal on Multimodal User Interfaces* 14, 4 (Dec. 2020), 337–352. https://doi.org/10.1007/s12193-020-00331-1

[59] Boram Yoon, Hyung-il Kim, Gun A. Lee, Mark Billinghurst, and Woontack Woo. 2019. The Effect of Avatar Appearance on Social Presence in an Augmented Reality Remote Collaboration. In *2019 IEEE Conference on Virtual Reality and 3D*

User Interfaces (VR). IEEE, Osaka, Japan, 547–556. https://doi.org/10.1109/VR.2019.8797719

[60] Leonard Yoon, Dongseok Yang, Jaehyun Kim, ChoongHo Chung, and Sung-Hee Lee. 2022. Placement Retargeting of Virtual Avatars to Dissimilar Indoor Environments. *IEEE Transactions on Visualization and Computer Graphics* 28, 3 (2022), 1619–1633. https://doi.org/10.1109/TVCG.2020.3018458

[61] Johannes Zagermann, Ulrike Pfeil, Philipp von Bauer, Daniel Fink, and Harald Reiterer. 2020. "It's in My Other Hand!" – Studying the Interplay of Interaction Techniques and Multi-Tablet Activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376540

[62] Katja Zibrek and Rachel McDonnell. 2019. Social Presence and Place Illusion Are Affected by Photorealism in Embodied VR. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games* (Newcastle upon Tyne, United Kingdom) *(MIG '19)*. Association for Computing Machinery, New York, NY, USA, Article 13, 7 pages. https://doi.org/10.1145/3359566.3360064

Received 2024-07-01; accepted 2024-09-20