Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Investigating Activity-Linked Audio for Conveying Partner Activity Information in Collaborative VR

Miki Hasegawa

hasegawa@iplab.cs.tsukuba.ac.jp

University of Tsukuba

Yasu Nakamizo

University of Tsukuba

Ikkaku Kawaguchi

University of Tsukuba

Research Article

Keywords: Collaborative Virtual Reality, Workspace Awareness, Social Presence, Activity-Linked Audio, Sonification

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11297899/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Investigating Activity-Linked Audio for Conveying Partner Activity Information in Collaborative VR

Abstract

In collaborative virtual reality (VR), users need information about a partner's presence, location, and activity, but visual access is limited to the head-mounted display's field of view and sounds arising from bodily activity are not necessarily reproduced. We investigated activity-linked audio, defined as non-speech auditory cues linked to avatar activity, including sounds triggered by specific activities and sonification varying with activity state. We compared three conditions: Silent, with no activity-linked audio; Footstep, which presented footstep sounds during walking; and Noise, which mapped an avatar's overall activity level to red-noise volume. Twenty-four participants in 12 pairs performed a collaborative block-assembly task while alternating between Director and Executor roles. We evaluated social presence, affective evaluation of the partner avatar, task performance, head orientation, body movement, and interpersonal distance. Behavior was analyzed across five task phases and by whether the target block was Front, Back, or Lateral relative to the Director. No consistent condition differences were found in condition-level subjective evaluations, task performance, or head orientation toward the partner. Executor body movement showed Condition × Phase × Direction interactions. During the search-and-selection phase, when the Executor worked behind the Director, the proportion of time moving was higher in Footstep and Noise than in Silent, while movement distance per trial was greater in Footstep than in both Silent and Noise. Thus, differences among auditory conditions depended on task phase and spatial relationship, particularly during search behind the Director.

Keywords: Collaborative Virtual Reality, Workspace Awareness, Social Presence, Activity-Linked Audio, Sonification

1 Introduction

In recent years, social VR, in which multiple users interact through avatars in a shared virtual environment, has become increasingly common. In such environments, it is important not only to perceive others' presence but also to understand where a partner is and what they are doing. These aspects have been studied from two distinct perspectives: social presence (SP) [5, 27] and workspace awareness (WA) [14]. SP refers to the degree to which a user perceives another person as present in interactions mediated by technologies such as VR. WA refers to an up-to-the-moment understanding of a partner's location and activities in a shared workspace and supports anticipating the partner's actions, coordinating work, and providing assistance when needed. Supporting both SP and WA is therefore important, particularly in collaborative VR environments. However, immersive VR limits visibility to the field of view of the head-mounted display (HMD), requiring users to turn their heads toward partners outside the current field of view. In addition, sensory cues that naturally arise from bodily activity in face-to-face settings, such as footsteps, incidental sounds, and vibrations, are not necessarily reproduced in VR. Consequently, collaborative VR can lack some of the cues to others' presence and activity that are naturally available face to face, making information that supports SP and WA more difficult to obtain.

To address this problem, prior work in VR has explored visually supplementing information about others through detailed representations of avatar appearance and body motion and through

visual cues such as gaze, pointers, rays, and view frustums [2, 28]. These methods can explicitly convey information such as a partner's location, gaze target, and the object being manipulated. However, because these approaches rely on visual presentation, access to such information may still be limited by the user's current field of view. Auditory cues provide another means of conveying information about a partner without requiring direct visual observation. In face-to-face interactions, sounds generated by activity, such as footsteps, incidental sounds, typing sounds, and chair creaks, provide cues to a partner's presence and activity. Such information can reach others as a by-product of action without being intentionally communicated; Gutwin and Greenberg characterize this form of information transmission as consequential communication [14]. Prior work in VR and remote collaborative environments has likewise explored using sounds associated with user actions, spatial audio, object-manipulation sounds, and footsteps to convey information about others' presence and activity [7, 15, 25, 29, 31]. Because such auditory cues can be perceived without directly viewing the partner, they can complement visual cues; accordingly, this study focuses on presenting information about others' activity through sound.

Auditory presentation of activity information can broadly follow two approaches: sounds that semantically correspond to specific activities, such as footsteps, and abstract sonification that maps activity levels to sound parameters [13, 19]. Sounds tied to specific activities can provide cues that are readily associated with those activities, whereas sonification can continuously represent multiple activity states, including those without a single corresponding real-world sound. However, evidence comparing these approaches within the same collaborative VR environment remains limited.

Accordingly, we use activity-linked audio as an umbrella term for non-speech auditory cues linked to avatar activity, encompassing both sounds tied to specific activities and sonification that changes according to activity state. Based on these two approaches, we implemented two types of activity-linked audio: Footstep and Noise. Footstep presents footstep sounds in response to avatar walking. Noise computes an overall activity level from avatar locomotion, hand movements, and other bodily activities and continuously maps this value to the volume of red noise that follows the avatar's position. Rather than reproducing a specific real-world sound, Noise was designed as an abstract sonification of the avatar's overall activity state, which has no single corresponding real-world sound. Before the user study, we compared candidate sounds in a preliminary study of directional perception in VR and acceptability during continuous presentation, and selected red noise for Noise based on the results.

To investigate activity-linked audio in collaborative VR, we examined whether differences among auditory conditions appeared in condition-level subjective ratings and task performance and what behavioral differences emerged depending on role, task phase, and spatial relationship. We compared three conditions: Silent, in which no activity-linked audio was presented; Footstep, in which footstep sounds were presented in response to walking; and Noise, in which red noise was presented according to the overall activity level. Twenty-four participants forming 12 pairs performed a collaborative block-assembly task in VR while alternating between the Director and Executor roles. For each block, the Director verbally instructed the Executor, who searched the four peripheral tables for the target block, carried it to the central table, and placed it at the instructed location. In addition to condition-level measures such as social presence, evaluations of the partner avatar, and task performance, we recorded time-series behavioral logs of head orientation toward the partner, body movement, and interpersonal distance. Because the information needed about a partner during collaboration may vary with task progress and spatial configuration, we divided each block into five task phases and classified the peripheral table containing the target block as Front, Back, or Lateral relative to the Director. Using these classifications, we analyzed the behavioral logs by role, task phase, and Director-relative direction, alongside condition-level subjective ratings and task performance. No consistent condition differences were found in condition-level subjective evaluations, task performance, or head orientation toward the partner; however, Executor body movement differed among conditions in specific combinations of task phase and Director-relative direction.

2 Related Work

This section reviews research on acquiring and presenting information about others in collaborative VR. We first discuss visual and auditory information presentation and their effects on subjective evaluations and collaborative behavior, and then focus on auditory presentation of activity information, reviewing auditory icons and sonification to position the present study within this literature.

2.1 Partner Information in Collaborative VR

In collaborative VR environments, users need to keep track of a partner's state as needed while continuing their own work rather than continuously monitoring the partner. Gutwin and Greenberg describe collaboration involving shifts of attention between individual and shared work as mixed-focus collaboration and show that information about a partner, such as Presence, Action, and Location, is used to coordinate work and anticipate the partner's actions [14]. Accordingly, the information needed about a partner during collaboration and how it is acquired may vary with the user's role and task progress.

In collaborative XR, visual awareness support using avatars, gaze, pointers, rays, and view frustums has been explored to supplement information about partners [2, 28]. Auditory approaches have likewise used sounds generated by activity to convey others' states without direct visual observation. Gutwin et al. mapped sounds to drawing actions in a distributed drawing environment and used them as cues to activity occurring outside the user's view [15]. Cidota et al. compared audio and visual notifications in a remote collaborative environment and investigated information presentation for supporting workspace awareness [7]. Schubert et al. associated friction sounds with object movement in collaborative VR and investigated their use for detecting surrounding changes and locating them [31]. Lee et al. presented other users' footsteps and floor vibrations and examined their effects on social presence in a VR environment without speech [25]. Together, these studies suggest that auditory information associated with activity can supplement information about a partner during collaboration in addition to visual presentation.

The effects of such information presentation may appear not only in condition-level subjective evaluations but also in behavior during collaboration. Kimmel et al. reported no clear differences in self-reported social presence across avatar body-movement presentation conditions in collaborative VR, whereas interpersonal distance differed between conditions [23]. Brument et al. likewise showed that differences in collaborative MR system configurations were reflected in spatial behaviors such as interpersonal distance, body orientation, and looking at the partner [6]. These findings suggest that, in addition to subjective evaluations, analyzing behavior while considering users' roles and task contexts is important for assessing the effects of presenting information about others.

2.2 Auditory Presentation of Activity Information

Existing auditory approaches in collaborative environments have used sounds that semantically correspond to specific activities, such as footsteps and friction sounds associated with object movement [15, 25, 31]. Gaver characterizes non-speech displays that draw on familiar relationships between real-world events and the sounds they produce as auditory icons [13]. Such displays allow users to draw on prior experience to interpret the activity that produced a sound, while the information they can convey depends on the corresponding event. For example, footsteps can convey the occurrence of walking but cannot represent activities that do not involve walking, such as hand movements performed while stationary.

Sonification, in contrast, represents information that has no inherent real-world sound by mapping numerical values or changes in state to sound parameters [19]. Movement sonification, which maps bodily movement to sound, has also been studied for supporting motor learning and interaction in virtual environments [8, 12]. Schubert et al. further showed that presenting changes associated with virtual-object movement through sound can provide cues to the occurrence, location, and movement direction of events that are not directly observed visually [31].

These findings suggest that sonification could supplement activity information about a partner avatar that is difficult to obtain visually, allowing users to perceive the occurrence and changes of the partner's activity without directly viewing the partner. In particular, integrating multiple activities, such as locomotion and hand movements, into a single continuous auditory representation can also represent activities for which no natural corresponding sound exists, such as hand movements performed while stationary. Making the sound source follow the partner avatar could also provide supplementary information about the partner's approximate location together with their activity state. However, the use of sonification that integrates multiple bodily activities of a partner avatar as a cue for conveying the partner's activity information has not been sufficiently explored in collaborative VR.

2.3 Our Approach

Accordingly, this study focuses on two design approaches for presenting activity information through sound: sounds that semantically correspond to specific activities and abstract sonification that integrates multiple activities. We implemented Footstep, which corresponds to walking, and Noise, which presents the avatar's overall activity level as a continuous abstract sound, and compared them within the same collaborative VR environment. In addition to condition-level subjective evaluations, we analyze behavior while considering collaborators' roles, task phases, and spatial relationships to their partners to examine the situations in which differences among the auditory conditions emerge during collaboration.

3 Preliminary Study

To select the abstract sound for the subsequent user study, we conducted a preliminary study comparing directional perception and subjective evaluations of four auditory stimuli in VR. In the user study, the abstract sound is presented as spatial audio from both the participant's own avatar and the partner avatar, with its volume varying continuously according to each avatar's overall activity level. We therefore considered both the ability to perceive the approximate direction of the sound source and the acceptability of the sound for continuous presentation during collaboration.

For the abstract sound, we used broadband noise that does not semantically correspond to a specific real-world event and whose volume can be varied continuously according to activity level. We compared white noise, pink noise, and red noise, which differ in the slope of their power spectral density as a function of frequency. A 1-kHz single-frequency tone was also used as a reference stimulus. Because 1 kHz is included in standard pure-tone audiometry and is also used as an initial test frequency [1], we selected it as a reference stimulus that avoids the extreme low- and high-frequency ranges.

A preliminary study previously reported in [17] is presented here as the sound-material selection process for the present study.

3.1 Method

Eight undergraduate and graduate students participated in the study (all male; mean age = 23.0 years, SD = 0.93, range = 22–24 years). The four auditory-stimulus conditions were a 1-kHz single-frequency tone, white noise, pink noise, and red noise. All stimuli were generated using Audacity. For the noise stimuli, we used the Audacity Noise Generator to generate white noise, pink noise, and the noise implemented in Audacity as Brownian noise [3]. In this paper, we refer to this Brownian noise as red noise. All stimuli were mono with a sampling rate of 48 kHz, and no normalization or amplification was applied after generation.

The study used a within-subjects design. In each condition, the auditory stimulus was presented twice from each of eight directions around the participant, resulting in 16 trials per condition. Direction order was randomized within each condition, and condition order was counterbalanced using a Latin square.

The experimental system was implemented using Unity 2022.3.7f1 [32], with Meta Quest 2 used as the HMD [26]. The auditory stimuli were presented as spatial audio using Steam Audio [34]. An invisible sound source was placed in one of eight directions around the participant and moved toward them at a constant speed. Participants listened while facing forward and pressed a controller button when they perceived the direction of arrival. They then selected the perceived direction

from eight directional options presented in the VR environment.

After each sound condition, participants rated six impressions of the sound—comfort, relaxation, naturalness, liveliness, calmness, and interest—on a five-point scale. After completing all conditions, they ranked the four conditions in terms of ease of directional perception and preference.

For directional perception, we calculated exact eight-direction accuracy as the proportion of trials in which the presented and reported directions matched. Inspection of the response distributions revealed front–back confusions, in which participants reported a direction rotated by 180 degrees, such as confusing front and back. To further characterize the error pattern, we also calculated axis-level accuracy, treating directions separated by 180 degrees as belonging to the same directional axis. In the subsequent user study, the abstract sound is used as a supplementary cue that follows the partner avatar's position rather than as a navigation signal for precisely identifying the partner's location from sound alone. Therefore, sound-material selection considered not only exact eight-direction accuracy but also coarse directional perception along the directional axis.

Subjective ratings and rankings were compared across conditions using the Friedman test. When a significant difference was found, post-hoc pairwise comparisons were conducted using Wilcoxon signed-rank tests with Holm correction for multiple comparisons.

3.2 Results

The main results of the preliminary study are shown in Table 1. Exact eight-direction accuracy was 29.7% for the single-frequency tone, 64.1% for white noise, 60.9% for pink noise, and 56.3% for red noise. Axis-level accuracy, in which directions separated by 180 degrees were treated as belonging to the same directional axis, was 48.4%, 75.8%, 76.6%, and 75.0%, respectively, with similar values across the three noise conditions.

For red noise, all response errors along the directional axis were within 45 degrees. In addition, in 97% of trials in which the presented direction was either front or back, the reported direction also lay on the front–back directional axis. This 97% does not indicate the proportion of trials in which front and back were correctly distinguished, but rather the agreement when they were treated as belonging to the same directional axis.

Table 1 Main results of the preliminary study. Values indicate means or mean ranks. Lower rank values indicate more favorable evaluations.

Measure	Single	White	Pink	Red
8-direction accuracy (%)	29.7	64.1	60.9	56.3
Axis-level accuracy (%)	48.4	75.8	76.6	75.0
Impression score (1–5)	1.85	3.48	3.46	3.54
Perception Rank	3.13	2.25	1.88	2.75
Preference Rank	3.63	2.50	2.13	1.75

The mean score across the six impression items differed significantly across conditions ($\chi^2(3) = 15.61$, $p = .0014$, $W = .650$). Post-hoc comparisons with Holm correction showed that white noise, pink noise, and red noise each received significantly higher scores than the single-frequency tone (all $p = .047$), whereas no significant differences were found among the three noise conditions.

Preference rankings also differed significantly across conditions ($\chi^2(3) = 9.45$, $p = .0239$, $W = .394$), although none of the pairwise comparisons was significant after Holm correction. Descriptively, red noise had the best mean preference rank at 1.75, and all eight participants ranked it first or second. For ease of directional perception, pink noise had the best mean rank, followed by white noise, red noise, and the single-frequency tone.

3.3 Selection of Red Noise

Based on these results, we selected red noise as the abstract sound for the user study. Although red noise did not yield the highest exact eight-direction accuracy among the noise conditions, its axis-level accuracy was comparable to those of the other noise conditions, and all response errors along the directional axis were within 45 degrees. It also had the highest mean score across the six impression items among the noise conditions, and all eight participants ranked it first or second in terms of preference.

We therefore adopted red noise as a design decision that jointly considered its ability to provide supplementary directional information as a

sound following the partner's position and its acceptability for continuous presentation. This selection does not indicate that red noise was statistically superior to white noise or pink noise. Because the preliminary study did not involve collaborative work with another avatar, it also did not examine the effects of red noise on social presence or WA.

4 Experimental System

To investigate how differences among the auditory conditions are reflected in subjective evaluations and collaborative behavior in collaborative VR, we conducted a user study comparing three auditory conditions: Silent, Footstep, and Noise. For this study, we implemented a collaborative VR system in which two users perform a shared task while activity-linked audio is manipulated across conditions. Voice communication was available in all conditions, while the additional non-speech activity-linked audio varied according to the auditory condition. The following sections describe the collaborative VR environment, the activity-linked audio conditions, the activity-level sonification used in the Noise condition, and data logging.

4.1 System Overview

The collaborative VR environment was implemented using Unity 6000.4.7f1 [33], with Meta Quest 2 used as the HMD [26]. Photon Fusion 2 [10] was used to synchronize avatars and task states between the two HMDs, while Photon Voice 2 [11] was used for voice communication between participants. The two participants connected to the same VR environment and were represented to each other as full-body avatars.

The poses of the full-body avatars were constructed from tracking data obtained from the HMDs and controllers. To stabilize avatar rendering and pose synchronization over the network, normalized reference points for the head and both hands were generated from the tracking data and shared between the two clients. Final IK 2.5 [30] was used to control the full-body avatar poses.

Participants used the controller thumbsticks to move and rotate horizontally in the VR environment and manipulated blocks using rays extending from the controllers. The environment contained a central table for block assembly and four peripheral tables holding candidate blocks. During the collaborative task, one participant was presented with information about the block to be placed, while the other searched the peripheral tables for the corresponding block based on this explanation and carried it to the central table. Details of the collaborative task, including role assignment and alternation, are described in Section 5.

4.2 Activity-Linked Audio Conditions

We implemented three conditions: Silent, Footstep, and Noise. Table 2 summarizes the three conditions. Voice communication was available in all conditions, while the additional non-speech activity-linked audio varied across conditions. In the Footstep and Noise conditions, activity-linked audio was presented for both the participant's own avatar and the partner avatar.

In the Silent condition, no non-speech auditory cues linked to avatar bodily activity were presented.

In the Footstep condition, footstep sounds were presented from the avatar's position as a 3D sound source in response to avatar walking. The footsteps were implemented as a looped audio source that followed the avatar and was enabled when the avatar's movement speed exceeded 0.10 m/s or when the system's walking state was active. This condition was designed as an auditory icon that semantically corresponds to the specific activity of walking [13].

In the Noise condition, to continuously represent the avatar's activity state beyond the presence or absence of walking, we calculated an activity level integrating avatar locomotion and positional changes of multiple reference points and varied the volume of red noise following the avatar's position according to this value.

4.3 Activity Level and Volume Control in the Noise Condition

In the Noise condition, we calculated an overall activity level from positional changes of the avatar as a whole, the head reference point, and the left- and right-hand reference points in Unity to represent bodily activity through continuous changes in volume. Frame-to-frame velocities were calculated

Table 2 Three auditory conditions compared in the user study.

Condition	Activity	Audio presentation	Primary information
Silent	None	None	—
Footstep	Avatar walking	Footsteps from the avatar position while walking	Walking occurrence and spatial location
Noise	Avatar locomotion and reference-point position changes	Red-noise volume at the avatar position varies with overall activity	Activity occurrence, intensity, and spatial location

from these positions, and velocities below 0.05 m/s were excluded from the activity-level calculation to avoid treating small tracking fluctuations as activity. The activity level combined whole-avatar movement and hand movements, with weighting that emphasized whole-body movement during walking and hand activity when the avatar was not walking. The resulting activity level was normalized to a range from 0 to 1 and continuously mapped such that larger values produced higher red-noise volume. The head reference point used in this study was generated for pose synchronization and therefore did not directly represent the raw translational motion of the HMD.

In the Footstep and Noise conditions, activity-linked audio was presented from both the participant's own avatar and the partner avatar. In the Noise condition, a higher gain was assigned to audio from the partner avatar to make it more perceptible than audio from the participant's own avatar. The weights used for activity-level and volume control were empirically tuned after the initial implementation by three researchers through listening in the VR environment while checking the volume changes associated with bodily activity and the perceptibility of audio from the partner avatar.

Activity-linked audio was presented as a 3D sound source that followed the avatar's position. The boundaries for distance attenuation were determined using Hall's interpersonal distance zones in proxemics as design references. Hall defined personal distance as 1.5–4 ft and social distance as 4–12 ft [16]. We used 4 ft (approximately 1.22 m) and 12 ft (approximately 3.66 m) as design reference values and set the implementation boundaries to 1.2 m and 3.6 m, respectively. Although proxemic patterns associated with interpersonal relationships have also been observed in VR, their absolute distances may differ from those in physical space [22]. We therefore did not treat Hall's distance zones as universal thresholds for VR, but used them as design references for determining the audible range and distance attenuation of the activity-linked audio. To maintain perceptibility near 3.6 m while avoiding excessively high volume at close distances, we used 1.2 m as the maximum-volume boundary and kept the volume constant at shorter distances. Specifically, the volume was set to its maximum within 1.2 m, attenuated linearly between 1.2 and 3.6 m, and not presented at distances of 3.6 m or greater. The same distance attenuation was applied in the Footstep condition, with footstep sounds presented only while the avatar was walking. Temporal smoothing was applied to reduce abrupt changes in Noise volume. The volume values used in this study were relative digital control values in Unity and do not represent calibrated sound pressure levels or perceived loudness.

4.4 Data Logging

Each HMD recorded an event log and a time-series log sampled at approximately 0.1-s intervals. The event log contained the experimental condition and the start and end times of each block.

The time-series log contained the positions and velocities of both avatars' roots, head reference points, and left- and right-hand reference points, as well as the head forward vectors, walking states, experimental roles, interpersonal distance, and volume-control values in the Footstep and Noise conditions. For the Noise condition, the normalized activity level ranging from 0 to 1 was also recorded.

Because Player 1 and Player 2 in the network did not represent fixed experimental roles, the data were later reconstructed by role based on the Director and Executor IDs recorded for each block. The evaluation measures derived from these logs and the corresponding analysis methods are described in Section 5.

5 User Study

We conducted a within-subjects experiment using a two-person collaborative task to evaluate the three auditory conditions described in Section 4.2.

5.1 Participants

Twenty-four undergraduate and graduate students of local university participated in the study (12 women and 12 men). Their ages ranged from 18 to 24 years (mean = 20.79, SD = 2.04), and all participants were right-handed. Participants formed 12 same-gender pairs. They were recruited through a university student information website and received 2,460 JPY as compensation for participation. According to the pre-experiment questionnaire, all participants reported a visual acuity of at least 0.7, with correction if applicable, and individually confirmed that they had no hearing problems that would interfere with participation in the study.

5.2 Collaborative Task

Participants performed a collaborative task in which they sequentially assembled eight blocks in the VR environment. The environment contained a central table for assembling the completed structure and four peripheral tables holding candidate blocks. The layout of the shared workspace is shown in Fig. 1(a).

For each block, one participant acted as the Director and the other as the Executor. Only the Director was presented with a hint indicating the shape, color, and placement of the next block. An example of the Director's view is shown in Fig. 1(b). The Director verbally conveyed the information to the Executor, who searched the peripheral tables for the corresponding block based on the explanation. An example of the Executor's view while searching for a candidate block is shown in Fig. 1(c). The Executor then carried the selected block to the central table and placed it at the instructed location.

5.3 Experimental Design

We used a three-level repeated-measures design with auditory condition as a within-subjects factor, and each pair experienced all three conditions: Silent, Footstep, and Noise. Condition order was counterbalanced at the pair level such that each condition appeared in each ordinal position for four pairs. Each condition consisted of eight blocks, with the Director and Executor roles alternating across blocks. Thus, each participant served as the Director four times and as the Executor four times in each condition.

The positions of the central table and four peripheral tables were kept constant across all conditions. However, in the experimental program, Layout A was fixed to Silent, Layout B to Footstep, and Layout C to Noise. Across the three layouts, the number of blocks, the number of times each color was used, and the assembly order were kept constant; only the block shapes and completed structures differed. Because the block shapes and their arrangement in the completed structure differed across layouts, the effects of auditory condition and layout could not be statistically separated.

5.4 Procedure

The experiment was conducted with pairs of participants, with each participant performing the task in a separate room. Communication between participants was conducted through Photon Voice, and the physical environment prevented them from directly seeing or hearing each other.

Before the experiment, the study and experimental procedure were explained to the participants, and written informed consent was obtained. After completing a pre-experiment questionnaire, participants put on the HMDs and practiced moving in the VR environment, manipulating blocks, and performing both the Director and Executor roles. After practice, participants performed the collaborative task under the Silent, Footstep, and Noise conditions. After each condition, they completed the subjective evaluations, and breaks of up to 5 minutes were provided between conditions as needed. After completing all conditions, participants ranked the conditions, provided open-ended comments, and participated in a semi-structured interview. The entire experiment lasted up to 120 minutes.

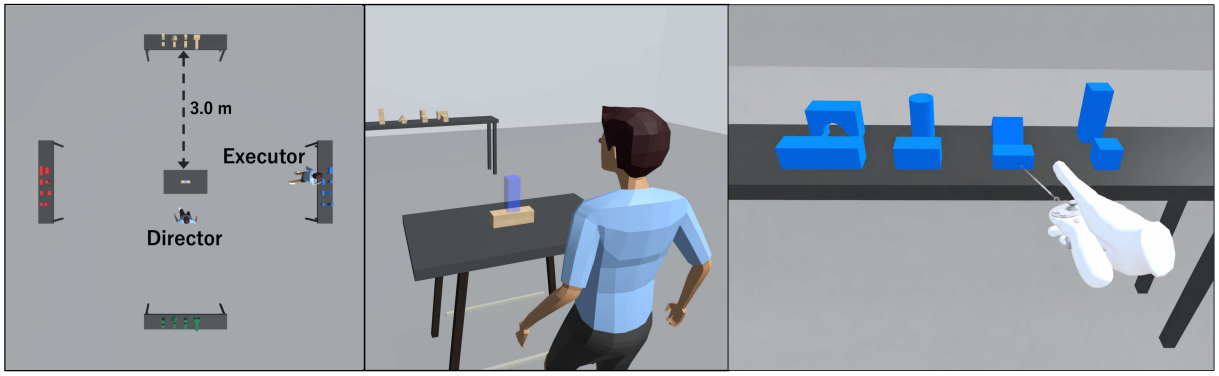


Fig. 1 Overview of the collaborative VR task. (a) Top view of the shared workspace, consisting of a central assembly table and four peripheral tables. The centers of the central and peripheral tables were separated by 3.0 m. (b) Example Director view. Information about the target block and its placement was presented only to the Director. (c) Example Executor view while searching for a candidate block on a peripheral table based on the Director's verbal instructions.

5.5 Measures

5.5.1 Subjective Evaluations

Social presence was assessed using the Networked Minds Social Presence Inventory (NMSPI) [5]. The Japanese wording of the questionnaire items was based on Ide et al. [21]. The subscale related to Emotion was excluded because it was considered to have low content relevance to the collaborative block-assembly task. We analyzed four subscales: Co-presence, Perceived attentional engagement, Perceived comprehension, and Perceived behavioral interdependence. Negatively worded items were reverse-coded, and the mean of the items comprising each subscale was used as the scale score. Cronbach's α for the four analyzed subscales ranged from .776 to .972 across the three conditions.

Affective evaluation of the partner avatar was assessed using Affective Attraction, consisting of five items: pleasantness, warmth, positivity, friendliness, and psychological closeness [18]. Each item was rated on a bipolar scale from -6 to $+6$, and the mean of the five items was used as the scale score. After completing all conditions, participants ranked the three conditions from first to third on two criteria: ease of noticing the partner's presence and activity, and ease of task performance and preference.

5.5.2 Semi-Structured Interview

We conducted a deductive content analysis of the transcripts from the semi-structured interviews with all 24 participants. Statements were classified into seven categories: awareness of the partner's activity and location, reduced visual checking, sound-triggered visual checking, changes in conversation and collaborative coordination, backgrounding and low salience of the sound, interpretations of Footstep and Noise, and context-dependent utility. Each participant could be coded into multiple categories, but a category was assigned only when explicit supporting evidence was identified in their statements. When a statement clearly referred to either Footstep or Noise, the corresponding condition was coded separately. One researcher reviewed the classification results and supporting statements for all 24 participants. The number of participants in each category was used only to describe the distribution of participant experiences and not for statistical comparisons between conditions.

5.5.3 Task Performance

As measures of task performance, we calculated the completion time of each block in each condition. We also calculated the time required for the Executor's first round trip and the duration of each task phase.

5.5.4 Head Orientation Toward the Partner

Because eye gaze was not recorded in this experiment, we used head orientation as a proxy for orientation toward the partner. At each time point, we calculated the angular difference between the horizontal forward vector of the participant's

HMD and the horizontal vector from the participant to the partner avatar's head reference point. An angular difference of 0° indicated that the partner avatar was directly in front of the participant's head, with larger differences indicating positions farther from the forward head direction. The mean angular difference to the partner was calculated for each task phase.

We also calculated the proportion of time during which the angular difference was below 48° to evaluate how long the partner remained near the participant's forward head direction. This threshold was set to approximately half of the reported horizontal field of view of the Meta Quest 2, which is approximately 97° [24]. Thus, a higher proportion of time below 48° indicates that the partner avatar remained longer within the horizontal field of view defined relative to the participant's forward head direction. Because the vertical component of the head forward direction could not be reconstructed from the recorded logs, the vertical field of view was not used in this determination. These measures were calculated separately for the Director and Executor.

5.5.5 Body Movement and Interpersonal Distance

To evaluate body movement, we integrated over time the planar movement speed recorded in each participant's time-series log to calculate the movement distance within each task phase of each block. Each block was treated as one trial, and this value was used as the movement distance per trial for each phase. To evaluate the proportion of time moving, we divided the time during which the walking state was active by the valid recorded duration of the corresponding phase. These body-movement measures were calculated separately for the Director and Executor.

To evaluate the spatial relationship between participants, interpersonal distance was calculated as the horizontal distance between the avatar roots of the two participants. The mean interpersonal distance was calculated for each task phase and treated as a pair-level measure because it was shared by both participants.

Table 3 Definitions of the task phases in the collaborative task.

Phase	Name	Definition
A	Initial dwell	From the start of the block until the Executor first leaves the vicinity of the central table
B	Outbound movement	From leaving the vicinity of the central table until reaching the work area near a peripheral table
C	Peripheral work	Period during which the Executor searches for and selects a block near a peripheral table
D	Return movement	From the end of peripheral work until the Executor first returns to the vicinity of the central table
E	Post-return	From the first return until the end of the block

5.6 Behavioral Log Processing and Task Phases

5.6.1 Task Phases

To compare collaborative behavior across task progress, each block was divided into five phases, Phase A through Phase E, based on the Executor's horizontal position and movement speed. The task-phase definitions are shown in Table 3. Phase A represented the initial dwell near the central table, Phase B the outbound movement toward a peripheral table, Phase C the search and selection period near a peripheral table, Phase D the return movement to the central table, and Phase E the period from the first return until the end of the block. Phases A, C, and E included small movements such as positional adjustments within the corresponding work area. Auditory condition, head orientation, and subjective evaluations were not used to determine phase boundaries.

To extract phase boundaries, we first identified low-speed dwell intervals from the Executor's horizontal position and movement speed. For each interval, we calculated the median horizontal position and identified dwell intervals near the central table, near the peripheral tables, and after the return based on their relative movement distance from the first low-speed dwell interval and their temporal order. Additional round trips in which the Executor departed again from the vicinity of

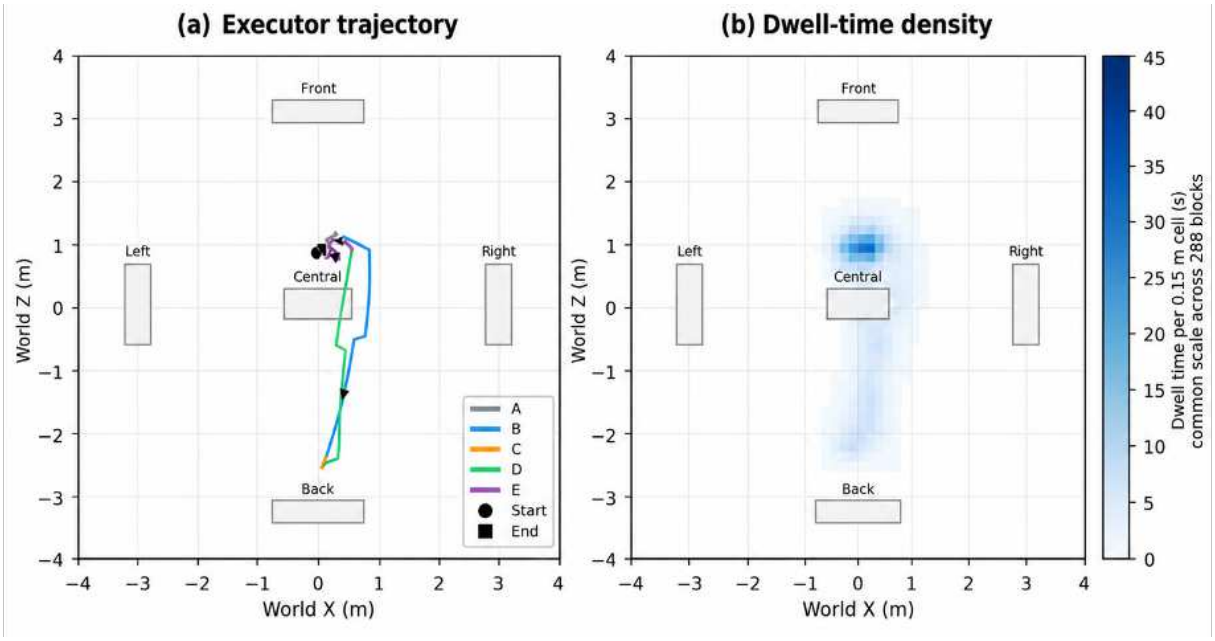


Fig. 2 Example of the Executor's spatial behavior during one block. (a) Executor movement trajectory colored by task phase from Phase A to Phase E. Arrows indicate the direction of temporal progression, and the circle and square indicate the start and end positions, respectively. (b) Executor dwell-time density for the same block. Darker blue indicates longer dwell time. Gray rectangles indicate the central and peripheral tables. This example shows Pair 08, Block 08 in the Noise condition (Director-relative Back); the same color scale was used for the density plots of all blocks.

the central table after Phase D and reached the work area near a peripheral table were excluded from the phase-based analyses. In contrast, movements near the central table that did not reach a peripheral-table work area were included in Phase E. An example of the Executor's movement trajectory and dwell-time density is shown in Fig. 2.

5.6.2 Director-Relative Task Direction

To examine the relationship between task phase and spatial configuration, the peripheral table containing the correct block in each block trial was classified as Front, Back, or Lateral relative to the Director for that block. Because the two participants were positioned at opposing reference locations across the central table, the classification was based on a fixed reference direction in which each participant faced the central table. The two side peripheral tables were combined into the Lateral category. The Director's instantaneous head orientation and the Executor's actual movement trajectory were not used for this classification. Thus, Front and Back represent directions relative to the Director in each block rather than absolute directions in the VR environment.

5.6.3 Analyzed Blocks

Following quality checks, two blocks were excluded from the phase-based analyses involving Direction because the direction of the initial round trip did not match the direction of the correct table and the Executor subsequently moved in the correct direction during an additional round trip. One additional block was excluded because its starting position differed substantially from those of the other trials, causing initial movement that could not be treated as part of the task phases to be included in Phases A and B. Thus, 285 of the 288 blocks were included in the main analyses involving Phase and Direction.

5.7 Statistical Analysis

Subjective evaluations were analyzed at the participant level using a one-way repeated-measures analysis of variance with Condition (Silent, Footstep, and Noise) as the within-subjects factor. When the assumption of sphericity was violated,

the Greenhouse–Geisser correction was applied. Condition rankings were analyzed using the Friedman test; when a significant Condition effect was found, pairwise comparisons between conditions were conducted with Holm correction for multiple comparisons [20].

Task-progress and overall behavioral measures were evaluated using all 288 blocks, separately from the main analyses involving Direction, which used 285 blocks. Mixed models accounting for repeated observations within pairs were used for block completion time, first round-trip time, and task-phase duration. Condition order was included in each model. Pair random intercepts were included to account for repeated measurements, and participant random intercepts were used for participant-level measures. For task-phase duration, phase-specific Condition effects were tested using Wald tests derived from the fitted mixed model, with Benjamini–Hochberg correction applied across the five Phases. Pairwise condition comparisons within each Phase were adjusted using Holm correction. In addition, 29 measures of overall behavior across the first sequence from Phase A through Phase E, covering head orientation toward the partner, body movement, head activity, interpersonal distance, and pair coordination, were examined in exploratory analyses. The Benjamini–Hochberg procedure [4] was used to correct for multiple tests across these measures, while Holm correction was used for pairwise condition comparisons.

For the main analyses considering task phase and Director-relative direction, mean angular difference to the partner and the proportion of time below 48° were used as measures of head orientation toward the partner, while movement distance per trial and the proportion of time moving were used as measures of body movement. These measures were analyzed separately for the Director and Executor. Because interpersonal distance is shared by both participants, it was analyzed at the pair level without separating the two roles.

For each measure, data were aggregated within each pair into Condition × Phase × Direction cells. A three-way analysis of variance using the aligned rank transform (ART) was conducted for each measure, with Condition (Silent, Footstep, Noise), Phase (A, B, C, D, E), and Direction (Front, Back, Lateral) as factors [35]. Each measure was analyzed as an independent outcome, and no multiple-comparison correction was applied across measures.

Analyses proceeded from higher-order interactions to lower-order effects. We first tested the three-way Condition × Phase × Direction interaction. When the three-way interaction was significant, hierarchical post-hoc analyses were conducted by testing the Condition × Phase interaction within each Direction, followed by the Condition effect within each Phase and pairwise comparisons between conditions. When multiple tests were conducted at a given stage, Holm correction was applied, and ART-C [9] was used for pairwise condition comparisons.

When the three-way interaction was not significant, we tested the two-way Condition × Phase, Condition × Direction, and Phase × Direction interactions. For significant two-way interactions, the corresponding simple main effects were tested, followed by pairwise comparisons between levels when necessary. Holm correction was also applied to the number of tests conducted at each stage of these follow-up analyses. When none of the two-way interactions was significant, the main effect of each factor was examined, and significant main effects for factors with three or more levels were followed by pairwise comparisons with Holm correction.

6 Results

6.1 Subjective Evaluations and Task Performance

Table 4 presents the participant-level subjective evaluations ($n = 24$). No significant condition differences were found for any of the four NMSPI subscales: Co-presence, Perceived attentional engagement, Perceived comprehension, and Perceived behavioral interdependence. Descriptively, the Noise condition had the highest mean score for all four subscales. For Perceived comprehension, the mean scores were 5.826 for Silent, 5.715 for Footstep, and 6.076 for Noise, but the condition effect was not significant ($F(2, 46) = 2.672, p = .080$). No significant condition difference was found for Affective Attraction ($F(2, 46) = 0.490, p = .616$).

No significant condition differences were found in the post-experiment rankings for either ease of noticing the partner's presence and activity

Table 4 Results of the subjective evaluations. Values indicate means (standard deviations).

Measure	Silent	Footstep	Noise	$F(2, 46)$	p
Co-presence	5.271 (1.038)	5.135 (1.127)	5.323 (1.043)	0.729	.446
Perceived attentional engagement	5.021 (1.652)	4.868 (1.455)	5.139 (1.707)	1.033	.364
Perceived comprehension	5.826 (0.927)	5.715 (0.932)	6.076 (0.985)	2.672	.080
Perceived behavioral interdependence	5.326 (1.159)	5.368 (1.194)	5.431 (1.207)	0.148	.863
Affective Attraction	3.208 (1.846)	3.300 (1.839)	3.508 (1.553)	0.490	.616

($\chi^2(2) = 1.00, p = .607, W = .021$) or ease of task performance and preference ($\chi^2(2) = 3.25, p = .197, W = .068$). Fig. 3 shows the distribution of participants across ranking positions.

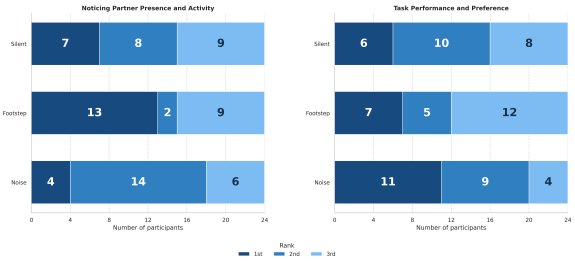


Fig. 3 Distribution of participants across ranking positions after completing all conditions. (a) Ease of noticing the partner's presence and activity. (b) Ease of task performance and preference. Each colored segment indicates the number of participants who ranked the corresponding condition first, second, or third.

No significant condition differences were found in block completion time ($p = .473$) or first round-trip time ($p = .867$). No phase-specific Condition effects on task-phase duration were significant after Benjamini–Hochberg correction. Similarly, no consistent condition differences were found in the overall behavioral measures aggregated across roles and task phases after correction for multiple comparisons.

In the semi-structured interviews, 23 of the 24 participants described differences between Footstep and Noise in terms of meaning, naturalness, salience, or utility. Thirteen participants associated Footstep with specific activities such as walking, stopping, or moving. For Noise, five participants described it as a cue to the partner's presence or activity, whereas 15 described it as an environmental sound, background sound, or sound without a clear meaning. These counts were not mutually exclusive because individual participants could mention multiple aspects.

Table 5 ART results for head orientation toward the partner. The p values for the three-way interactions are uncorrected. When the three-way interaction was not significant, the p values for the three tested two-way interactions were corrected using the Holm method within each role and measure.

Role	Measure	Effect	F	p
Director	Mean angular difference	$C \times P \times D$	1.948	.083
		$C \times P$	1.299	.280
		$C \times D$	3.527	**.042**
		$P \times D$	66.955	$< .001$
	Proportion of time $< 48°$	$C \times P \times D$	1.777	.111
		$C \times P$	1.667	.174
		$C \times D$	3.489	**.048**
		$P \times D$	45.528	$< .001$
Executor	Mean angular difference	$C \times P \times D$	1.859	.089
		$C \times P$	0.677	1.000
		$C \times D$	0.382	1.000
		$P \times D$	38.478	$< .001$
	Proportion of time $< 48°$	$C \times P \times D$	1.477	.200
		$C \times P$	0.789	.776
		$C \times D$	1.041	.776
		$P \times D$	42.269	$< .001$

C: Condition, P: Phase, D: Direction. Degrees of freedom were Greenhouse–Geisser corrected. Bold indicates $p < .05$.

6.2 Behavioral Analysis by Role, Task Phase, and Direction

The main analyses involving task phase and Direction included 285 blocks. The following sections report results for head orientation toward the partner, body movement, and interpersonal distance.

6.2.1 Head Orientation Toward the Partner

Mean angular difference to the partner and the proportion of time below 48° were analyzed separately for the Director and Executor. The ART

results for these head-orientation measures are shown in Table 5.

For the Director, no significant Condition × Phase × Direction interaction was found for mean angular difference to the partner ($F(6.25, 68.70) = 1.948$, $p = .083$, rank partial $\eta^2 = .150$). No significant three-way interaction was found for the proportion of time below 48° either ($F(6.43, 70.71) = 1.777$, $p = .111$, rank partial $\eta^2 = .139$). Because the three-way interactions were not significant, we tested the Condition × Phase, Condition × Direction, and Phase × Direction interactions. For the Director's mean angular difference, the Condition × Direction interaction (Holm-corrected $p = .042$) and Phase × Direction interaction (Holm-corrected $p < .001$) were significant, whereas the Condition × Phase interaction was not (Holm-corrected $p = .280$). For the proportion of time below 48°, the Condition × Direction interaction (Holm-corrected $p = .048$) and Phase × Direction interaction (Holm-corrected $p < .001$) were also significant, whereas the Condition × Phase interaction was not (Holm-corrected $p = .174$).

For the significant Condition × Direction interactions observed for the Director, we tested the simple effect of Condition within each Direction. For mean angular difference, no significant Condition effect remained after Holm correction in Front, Back, or Lateral ($p = .154$, $p = 1.000$, and $p = 1.000$, respectively). Similarly, for the proportion of time below 48°, no significant Condition effect was found in any Direction ($p = .498$, $p = 1.000$, and $p = 1.000$, respectively).

For the Executor, no significant three-way interaction was found for either mean angular difference ($F(6.91, 75.96) = 1.859$, $p = .089$, rank partial $\eta^2 = .145$) or the proportion of time below 48° ($F(5.98, 65.76) = 1.477$, $p = .200$, rank partial $\eta^2 = .118$). For mean angular difference, neither the Condition × Phase nor Condition × Direction interaction was significant (both Holm-corrected $p = 1.000$), whereas the Phase × Direction interaction was significant (Holm-corrected $p < .001$). The same pattern was found for the proportion of time below 48°: the Condition × Phase and Condition × Direction interactions were not significant (both Holm-corrected $p = .776$), whereas the Phase × Direction interaction was significant (Holm-corrected $p < .001$). Overall, head orientation toward the partner varied with task phase and direction, but no consistent changes associated with auditory condition were observed.

6.2.2 Body Movement

For body movement, we analyzed movement distance per trial and the proportion of time moving. For the Director, neither measure showed a significant Condition × Phase × Direction interaction.

For the Executor, however, movement distance per trial showed a significant Condition × Phase × Direction interaction ($F(6.84, 75.29) = 2.411$, $p = .029$, rank partial $\eta^2 = .180$). A significant three-way interaction was also found for the proportion of time moving ($F(6.70, 73.74) = 2.608$, $p = .020$, rank partial $\eta^2 = .192$). Table 6 summarizes the hierarchical ART analyses of Executor body movement.

We therefore tested the Condition × Phase interaction separately for each Direction. For movement distance per trial, a significant Condition × Phase interaction was found only in the Back direction ($F(4.49, 49.35) = 3.822$, Holm-corrected $p = .020$). No significant interactions were found in Front or Lateral (both Holm-corrected $p = 1.000$). For the proportion of time moving, a significant Condition × Phase interaction was likewise found only in the Back direction ($F(4.99, 54.92) = 3.823$, Holm-corrected $p = .015$). The interactions were not significant in Front or Lateral (Holm-corrected $p = .812$ and $p = .691$, respectively).

We next tested the Condition effect within each Phase in the Back direction. For movement distance per trial, significant Condition effects were found in Phase C ($F(1.58, 17.43) = 8.032$, Holm-corrected $p = .021$) and Phase D ($F(1.69, 18.54) = 8.467$, Holm-corrected $p = .018$). No significant Condition effects were found in Phases A, B, or E.

In Phase C of the Back direction, mean movement distance per trial was 0.393 m in Silent, 1.247 m in Footstep, and 0.522 m in Noise. ART-C pairwise comparisons showed that movement distance per trial was significantly greater in Footstep than in Silent (Holm-corrected $p = .002$, $d_z = 1.052$) and Noise (Holm-corrected $p = .025$, $d_z = .950$). No significant difference was found between Silent and Noise (Holm-corrected $p = .248$).

Table 6 Hierarchical ART analyses of Executor body movement. Holm-corrected p values are reported for (a) Condition × Phase interactions across the three Directions, (b) Condition effects across the five Phases in the Back direction, and (c) the three ART-C pairwise condition comparisons within each Phase. GG df denotes Greenhouse–Geisser-corrected degrees of freedom, and pt denotes percentage points. Bold indicates $p_{\text{Holm}} < .05$.

(a) Condition × Phase interactions by Direction

Measure	Direction	F	df_1	df_2	p_{Holm}
Movement distance per trial [m]	Front	0.764	4.34	47.73	1.000
	Back	3.822	4.49	49.35	**.020**
	Lateral	0.680	4.75	52.21	1.000
Proportion of time moving [%]	Front	0.411	4.28	47.13	.812
	Back	3.823	4.99	54.92	**.015**
	Lateral	1.151	4.20	46.25	.691

(b) Condition effects by Phase in the Back direction

Measure	Phase	F	df_1	df_2	p_{Holm}
Movement distance per trial [m]	A	0.398	1.57	17.28	.811
	B	0.921	1.81	19.87	.811
	C	8.032	1.58	17.43	**.021**
	D	8.467	1.69	18.54	**.018**
	E	2.392	1.51	16.64	.395
Proportion of time moving [%]	A	1.644	1.57	17.32	.482
	B	2.555	1.68	18.46	.447
	C	7.027	1.92	21.09	**.025**
	D	2.070	1.64	17.99	.482
	E	1.442	1.89	20.76	.482

(c) ART-C pairwise condition comparisons

Measure	Phase	Comparison	Difference	p_{Holm}	d_z
Movement distance per trial [m]	C	Footstep − Silent	+0.854	**.002**	1.052
		Noise − Silent	+0.129	.248	0.195
		Footstep − Noise	+0.725	**.025**	0.950
Movement distance per trial [m]	D	Footstep − Silent	−0.724	**.002**	−1.130
		Noise − Silent	−0.482	**.018**	−1.041
		Footstep − Noise	−0.243	.271	−0.392
Proportion of time moving [pp]	C	Footstep − Silent	+22.7	**.006**	0.877
		Noise − Silent	+17.1	**.019**	0.915
		Footstep − Noise	+5.6	.494	0.227

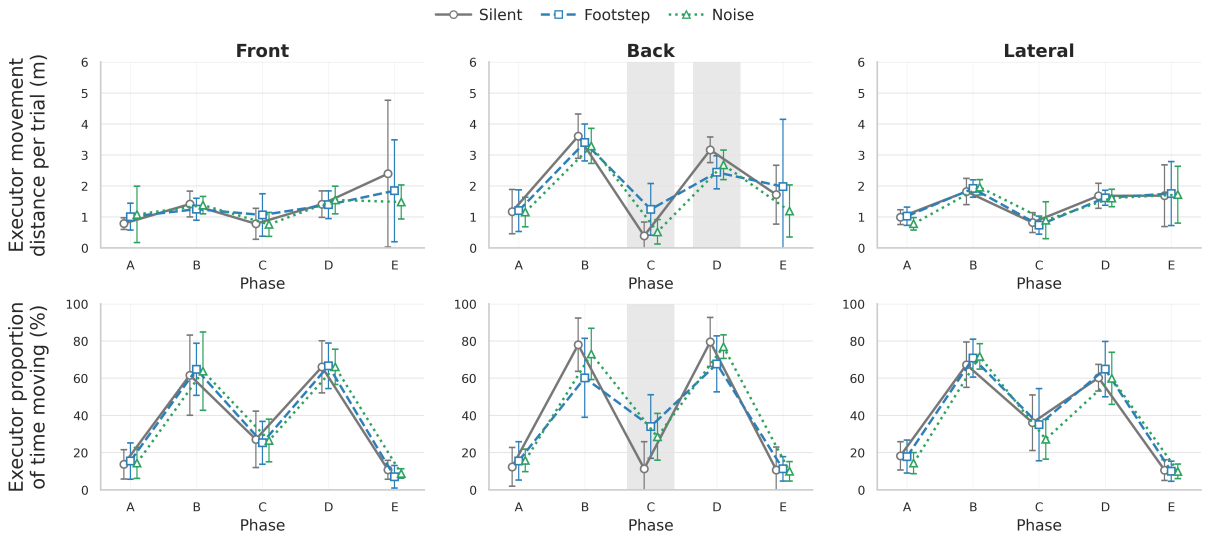


Fig. 4 Executor body movement across auditory conditions, task phases, and Director-relative directions. The top and bottom rows show movement distance per trial and the proportion of time moving, respectively; columns correspond to Front, Back, and Lateral. Points show means across 12 pairs, and error bars show ±1 SD. Gray shading marks phases with significant Condition effects in the hierarchical post-hoc analyses: Phases C and D in the Back direction for movement distance per trial, and Phase C for the proportion of time moving.

In Phase D of the Back direction, mean movement distance per trial was 3.166 m in Silent, 2.441 m in Footstep, and 2.684 m in Noise. Movement distance per trial was significantly lower in Footstep than in Silent (Holm-corrected $p =$

.002, $d_z = -1.130$), and significantly lower in Noise than in Silent (Holm-corrected $p = .018$, $d_z = -1.041$). No significant difference was found between Footstep and Noise (Holm-corrected $p = .271$).

For the proportion of time moving, a significant Condition effect was found only in Phase C of the Back direction ($F(1.92, 21.09) = 7.027$, Holm-corrected $p = .025$). No significant Condition effects were found in Phases A, B, D, or E.

In Phase C of the Back direction, the mean proportion of time moving was 11.4% in Silent, 34.1% in Footstep, and 28.5% in Noise. The proportion of time moving was significantly higher in Footstep than in Silent (Holm-corrected $p = .006$, $d_z = .877$) and significantly higher in Noise than in Silent (Holm-corrected $p = .019$, $d_z = .915$). No significant difference was found between Footstep and Noise (Holm-corrected $p = .494$). Fig. 4 shows movement distance per trial and the proportion of time moving across task phases for each Direction.

In the semi-structured interviews, 18 of the 24 participants described at least one form of activity-linked audio as a cue for understanding the partner's movement, location, direction, distance, presence, or activity state. Sixteen participants reported becoming aware of or understanding the presence or state of a partner outside their field of view through sound. Among them, 14 specifically described information obtained from sound, such as the partner's location, direction, movement, or task state. In addition, five participants mentioned changes in conversation or collaborative coordination associated with activity-linked audio, including speech, conversation initiation, instructions, listening, or confirmation behavior.

6.2.3 Interpersonal Distance

Mean interpersonal distance showed a significant Condition × Phase × Direction interaction ($F(5.99, 65.93) = 2.294$, $p = .045$, rank partial $\eta^2 = .173$). However, none of the Direction-specific Condition × Phase interactions remained significant after Holm correction, so no further analyses of Phase-specific simple effects or pairwise condition comparisons were conducted.

7 Discussion

In this study, we examined the effects of activity-linked audio in collaborative VR through condition-level subjective evaluations and task performance, as well as behavior considering role, task phase, and spatial relationship to the collaborative partner. As shown in Section 6, no consistent condition differences were found in condition-level subjective evaluations, task performance, or head orientation toward the partner, whereas context-specific condition differences in Executor body movement emerged primarily in Phase C of the Back direction. Here, context-specific refers to condition differences that emerged only in particular combinations of role, task phase, and spatial relationship, rather than consistently throughout collaboration. These results suggest that differences associated with the auditory conditions may depend on the collaborative context in which they are examined. The following sections first discuss head orientation toward the partner, Executor movement behavior, and subjective evaluations, and then integrate these findings with the semi-structured interview results to discuss the role of activity-linked audio and its design implications for collaborative VR.

7.1 Head Orientation Toward the Partner

Activity-linked audio could potentially affect head orientation toward the partner because it can convey information about the partner's location and activity even when the partner is not directly visible. However, as shown in Section 6.2.1 and Table 5, neither the Director nor the Executor showed a consistent pattern in which head orientation toward the partner increased or decreased under a particular condition. Thus, the present results do not indicate that Footstep or Noise systematically changed how participants oriented their heads toward the partner.

At the same time, head orientation toward the partner varied with the combination of task phase and direction for both the Director and Executor. This result indicates that head orientation reflected not only the auditory condition but also the current task activity and spatial relationship to the partner. In this experiment, the Executor moved between the central and peripheral tables

while searching for and transporting blocks, causing the relative spatial relationship between the participants to change across task phases. These differences in task structure and spatial configuration likely contributed to the observed variation in head orientation across phases and directions.

Of particular relevance to the movement analysis in Section 7.2, during Phase C in the Back direction, the Executor worked behind the Director, making direct visual confirmation more likely to require the Director to turn toward the partner. Nevertheless, no significant Condition effect on the Director's head orientation toward the partner was found in any Direction, including Back. Thus, activity-linked audio did not consistently increase or decrease the Director's visual checking behavior as measured by head orientation. Because eye gaze was not recorded in this study, head orientation toward the partner does not directly represent actual visual attention. Accordingly, these results cannot be used to draw direct conclusions about changes in visual attention; this measurement limitation is discussed further in Section 8.3.

7.2 Executor Movement by Task Phase and Spatial Relationship

Executor body movement differed depending on the combination of auditory condition, task phase, and direction, with the hierarchical post-hoc analyses showing that condition differences emerged primarily in Phase C of the Back direction (Table 6, Fig. 4). In this phase, the proportion of time moving was higher in Footstep and Noise than in Silent, while movement distance per trial was greater in Footstep than in both Silent and Noise. Thus, the condition differences in body movement did not appear uniformly throughout collaboration, but were limited to particular combinations of task phase and spatial relationship.

As defined in Section 5.6, Phase C was the period in which the Executor searched for and selected candidate blocks near a peripheral table. Unlike Phases B and D, which primarily involved movement toward or away from the peripheral table, Phase C could include movements within the peripheral-table work area and positional adjustments while inspecting candidate blocks. In the Back direction, unlike the Front and Lateral directions, the Executor worked at a peripheral table behind the Director, making direct visual confirmation more likely to require the Director to turn toward the partner. This combination of search-related movement within the peripheral-table work area and the Back spatial configuration provides the task and spatial context for interpreting the movement differences observed in Phase C.

Because no consistent Condition effect on the Director's head orientation toward the partner was found, including in the Back direction (Section 7.1), these movement differences are difficult to explain solely as a consequence of increased visual checking by the Director. One possibility is that, when working behind the Director, the Executor increased movements that generated activity-linked audio, thereby making their presence or activity more perceptible to the Director without requiring direct visual confirmation. However, because we did not measure whether participants intentionally used movement to communicate their presence or activity through sound, this interpretation remains a hypothesis.

The different movement patterns observed between Footstep and Noise may also be related to how bodily activity was mapped to sound in each condition. In Footstep, generating activity-linked audio required walking, whereas in Noise, the audio could also change in response to positional changes of multiple avatar reference points, including the hands. If Executors used self-generated activity-linked audio to make their presence or activity perceptible to the Director, Footstep would have required locomotion that generated footstep sounds, whereas Noise could have been modulated by a broader range of bodily activity. This difference could be consistent with the finding that, in Footstep, both the proportion of time moving and movement distance per trial were greater than in Silent, whereas in Noise, only the proportion of time moving was greater than in Silent. Nevertheless, the present measures cannot determine whether participants intentionally generated activity-linked audio in this way or whether the differences between Footstep and Noise resulted from these different sound mappings.

In Phase D of the Back direction, the pattern for movement distance per trial was reversed: movement distance per trial was greater in Silent than in Footstep and Noise. This result suggests that the auditory conditions were not associated

with a uniform increase or decrease in movement, but with different behavioral patterns across the distinct task phases of search and return. However, because we did not analyze the relationship between the Executor's position at the end of Phase C and the return path in Phase D, these two phases cannot be interpreted as a single continuous behavioral change.

These interpretations should be treated cautiously because auditory condition and task layout were fixed in combination in this experiment, and activity-linked audio originating from the participant's own avatar was also presented in the Footstep and Noise conditions. Therefore, the observed movement differences cannot be interpreted as causal effects of activity-linked audio alone, or specifically of auditory information originating from the partner.

7.3 Subjective Evaluation of Activity-Linked Audio

As shown in Section 6.1, no significant condition differences were found for the four NMSPI subscales, Affective Attraction, or the condition rankings. Therefore, the present results do not indicate that Footstep or Noise improved social presence, affective evaluation of the partner, ease of noticing the partner's presence and activity, or ease of task performance and preference at the condition level.

One possible reason why no clear differences emerged in these condition-level subjective evaluations is that voice communication between participants was available in all conditions. In the collaborative task, the Director verbally conveyed the shape, color, and placement of each block to the Executor, who searched for and placed the block based on these instructions. Thus, regardless of the presence of activity-linked audio, participants could continuously obtain information related to the partner's presence and task progress through voice communication. The availability of this verbal information source across all conditions may have made the additional information provided by activity-linked audio less apparent in subjective evaluations summarizing the entire condition. However, because we did not quantitatively analyze speech content, amount, or timing, we cannot directly determine the extent to which voice communication reduced differences between conditions in subjective evaluations.

Although condition-level subjective evaluations showed no clear differences, condition differences in Executor body movement were observed in the specific context of Phase C in the Back direction. This discrepancy may also reflect differences in the target and temporal scale of the measures. NMSPI, Affective Attraction, and condition rankings summarize participants' experience across an entire condition, whereas the observed differences in body movement were limited to a specific combination of task phase and spatial relationship. Such context-specific behavioral differences may therefore not have been clearly reflected in subjective evaluations of the condition as a whole.

Furthermore, NMSPI does not directly measure WA, and the absence of condition differences in subjective evaluations should not be equated with an absence of differences in the acquisition of information about the partner. In the semi-structured interviews, some participants reported becoming aware of or understanding the presence or state of a partner outside their field of view through activity-linked audio. Condition-level subjective evaluations and partner information used in specific collaborative situations should therefore be considered separately. The lack of a direct measure of WA and the relationship with voice communication are discussed further in Section 8.2.

7.4 Design Implications for Activity-Linked Audio in Collaborative VR

7.4.1 Activity-Linked Audio as a Peripheral Cue

Taken together, the context-specific behavioral differences and interview reports suggest that activity-linked audio may be better understood as a peripheral auditory cue that supplements partner information when visual access is limited, rather than as a presentation method expected to uniformly change condition-level outcomes. This interpretation is consistent with the concept of mixed-focus collaboration, in which people obtain information about others while shifting attention

between individual and shared work [14]. Accordingly, activity-linked audio could be designed not as a direct replacement for visual information, but as an information source that users can draw on when needed while continuing their primary task.

7.4.2 Balancing Interpretability and Activity Coverage

Footstep and Noise differed in both the range of bodily activities they represented and how participants interpreted the resulting sounds. Footstep was directly linked to the specific activity of walking, and participants often described it in relation to concrete states such as walking, stopping, or moving. Noise, in contrast, integrated multiple bodily activities into a continuous auditory representation and was interpreted either as a cue to the partner's presence or activity or as environmental or background sound. These differences illustrate a trade-off between interpretability and activity coverage when mapping bodily activity to sound.

Footstep draws on a familiar semantic correspondence and can therefore provide a relatively interpretable cue, but the activities, but it primarily represents walking. Noise can represent a broader range of bodily activities within a single auditory cue, but the relationship between activity state and auditory change is less explicit. Designing activity-linked audio therefore requires considering both which bodily activities should be represented and whether users can understand the resulting auditory changes as information about those activities.

Activity coverage and interpretability should also be considered alongside the salience of the auditory cue. The fact that Noise was sometimes perceived as environmental or background sound is not necessarily undesirable for a peripheral cue intended to remain available without disrupting the primary task. Such a cue need not remain highly salient at all times, but it should still be perceptible when partner information is needed. Thus, abstract activity-linked audio should balance the range of bodily activities it represents with both the interpretability and salience of the resulting auditory cue.

7.4.3 Context-Dependent Presentation

The utility of activity-linked audio should also be considered in relation to the contexts in which partner information may be lacking, rather than simply whether sound is continuously added. For example, auditory presentation could be reduced when the partner is easy to observe visually and emphasized when the partner is outside the field of view or working in a different area. Similarly, rather than sonifying all bodily activities with equal importance, meaningful state changes for the collaborator, such as the onset or cessation of walking or a change in the object being manipulated, could be presented selectively. Such context-dependent presentation represents one possible design direction for supplementing partner information while limiting unnecessary auditory information.

7.4.4 Evaluating Activity-Linked Audio in Collaborative Contexts

The present results also have implications for how activity-linked audio should be evaluated. Evaluation of auditory presentation in collaborative VR may benefit from combining overall measures such as social presence and preference with analyses of bodily behavior and information acquisition in specific situations where partner information is needed. In particular, distinguishing situations in which the partner is within or outside the user's field of view, or situations involving search versus locomotion, may provide a more detailed understanding of when activity-linked audio is used.

Overall, the design of activity-linked audio for collaborative VR should consider not only what information is mapped to which sounds, but also who needs information about whom, during which task phase, and under what spatial relationship. Rather than serving as a continuously presented notification that replaces visual information, activity-linked audio may be more effectively designed as an auditory cue that supplements partner information when it may otherwise be insufficient during collaboration. However, because auditory condition and task layout were fixed in combination in this experiment, these design implications should be regarded as hypotheses based on the observed condition differences rather than as causal effects of activity-linked audio alone.

8 Limitations and Future Directions

This study has several limitations concerning causal interpretation of the results, the range of information that could be measured, and generalizability. The following sections discuss the fixed assignment of auditory conditions to task layouts, measurement of WA and communication, limitations of the behavioral measures and auditory presentation, and generalizability to other evaluation settings.

8.1 Fixed Assignment of Auditory Conditions and Task Layouts

The most important limitation of this study is that the auditory conditions were fixed to specific task layouts. Layout A was assigned to Silent, Layout B to Footstep, and Layout C to Noise. Although condition order was counterbalanced, auditory condition and task layout were not manipulated independently. Therefore, the causal effect of activity-linked audio alone cannot be separated from the observed condition differences.

This limitation is particularly important when interpreting the differences in Executor body movement observed during Phase C in the Back direction. Body movement during Phase C may have been influenced not only by the auditory presentation but also by layout-specific factors, such as the shapes and arrangement of candidate blocks and how easily the target block could be found. Although the number of blocks, the number of times each color was used, and the assembly order were kept constant across layouts, this control does not eliminate the confounding between auditory condition and task layout. The observed differences in body movement should therefore be interpreted as differences among experimental conditions that included activity-linked audio, rather than as effects of the auditory presentation alone.

Future studies should independently manipulate auditory condition and task layout. Furthermore, balancing multiple task layouts across auditory conditions would allow examination of whether the context-specific behavioral differences observed in this study can be replicated independently of a particular spatial configuration or search target.

8.2 Measurement of WA and Communication

This study did not directly measure WA through measures such as accuracy in estimating the partner's location or identifying the partner's activity. Although the results allow us to discuss the possibility that activity-linked audio was used to obtain information about the partner, they do not establish that activity-linked audio improved WA itself. Future studies should more directly assess the acquisition of Presence, Action, and Location information by measuring how accurately users can identify the partner's location, activity state, manipulated object, or the onset and cessation of activity. However, frequently inserting questions or judgment tasks during collaboration could alter participants' allocation of attention and collaborative behavior. Measurement methods that do not excessively interrupt collaboration are therefore needed.

Voice communication was also available in all conditions, while speech content, amount, and timing were not quantitatively analyzed. We therefore cannot separate the contributions of information obtained through activity-linked audio from those of information obtained through voice communication. It also remains unclear whether the differences in body movement observed during Phase C in the Back direction were associated directly with understanding the partner's state or arose through changes in communication, such as confirmation utterances or repeated instructions. Future work should analyze speech turns, confirmation utterances, clarification requests, repair, and the temporal relationship between instructions and action onset in synchronization with the behavioral logs to examine how activity-linked audio and voice communication contribute to collaborative coordination.

8.3 Limitations of Behavioral Measures and Auditory Presentation

We used head orientation as a proxy for orientation toward the partner, but did not record actual eye gaze. The analysis also considered only horizontal head orientation and did not include the vertical field of view or the actual object of gaze. Thus, a small angular difference to the partner

does not necessarily indicate that the participant was actually looking at the partner avatar, nor does a large angular difference indicate that no information about the partner was being obtained. Future studies should use HMDs with eye tracking to directly measure gaze allocation to the partner avatar, task objects, peripheral tables, and other targets, allowing a more detailed examination of the relationship between activity-linked audio and visual information acquisition.

In the Noise condition, the overall activity level was calculated by integrating positional changes of the avatar root, head reference point, and left- and right-hand reference points. The present results therefore cannot separate the contributions of individual bodily activity components, such as walking, hand operations, and whole-body movement, to the perception of Noise or collaborative behavior. Future work should manipulate or analyze these activity components independently to determine which bodily activities are useful to sonify for understanding the partner's state.

Furthermore, Footstep and Noise differed not only in semantic meaning but also in the bodily activities to which they responded, sound duration, sound material, and the volume settings for self- and partner-generated audio. The parameters used to control activity level and volume in Noise were empirically determined through listening by the authors in the VR environment, and Footstep and Noise were not matched for perceived loudness during the experiment. The observed differences between Footstep and Noise therefore cannot be attributed solely to differences in semantic structure between an auditory icon and sonification or to differences in activity mapping. The red noise used for Noise was also only one sound material selected on the basis of our preliminary study and does not represent abstract activity-linked audio in general. Future studies should manipulate sound material, loudness, temporal characteristics, and activity mapping more independently to examine their respective effects on partner-information acquisition and collaborative behavior.

8.4 Generalizability and Future Evaluation Settings

The experiment comprised 12 independent pairs (24 participants), consisting of same-gender pairs of undergraduate and graduate students performing a single collaborative block-assembly task. It also involved a single short-term session in an environment where the partner avatar could, in principle, always be visually checked and voice communication was available when needed. The behavioral patterns observed in this study therefore may not directly generalize to collaboration in larger groups, professional collaborative work, long-term use, collaborators with different social relationships, or different task structures.

Future studies should evaluate activity-linked audio in environments with different costs of directly checking the partner, such as larger VR spaces with occlusions, parallel work across multiple work areas, loosely coupled collaboration, and remote assistance. In multi-user collaboration, the amount of available partner information and auditory load would also differ from those in two-person collaboration, making it necessary to examine whether similar behavioral differences emerge.

As a next step, the reproducibility of the context-specific behavioral differences observed in this study should first be examined in an experiment that independently manipulates auditory condition and task layout. Subsequently, combining direct measures of WA with eye tracking and conversation analysis could provide a more direct examination of what information about the partner is obtained through activity-linked audio and how that information contributes to collaborative coordination through visual attention, communication, and bodily behavior.

9 Conclusion

This study investigated auditory presentation for supplementing information about a partner's activity in collaborative VR. We implemented Footstep, which presents footstep sounds corresponding to walking, and Noise, which continuously maps the avatar's overall activity level to red-noise volume, and compared them with a Silent condition without activity-linked audio. Twenty-four participants forming 12 pairs performed a collaborative block-assembly task, and we evaluated subjective ratings, task performance, head orientation toward the partner, body movement, interpersonal distance, and semi-structured interviews.

No consistent condition differences were found in condition-level subjective evaluations, task performance, or head orientation toward the partner. In contrast, context-specific condition differences in Executor body movement emerged depending on task phase and spatial relationship to the Director. In particular, during Phase C in the Back direction, the proportion of time moving was higher in Footstep and Noise than in Silent, while movement distance per trial was greater in Footstep than in both Silent and Noise. The semi-structured interviews also included participants who described activity-linked audio as a cue for understanding the location or activity of a partner outside their field of view.

These results suggest that differences among the auditory conditions may not emerge uniformly throughout collaboration, but rather in behavior under task phases and spatial relationships in which partner information is needed. Although causal interpretation is limited because auditory condition and task layout were fixed in combination, the present study highlights the importance of considering not only condition-level evaluations but also users' roles, task phases, and spatial relationships to their partners when evaluating and designing auditory presentation in collaborative VR. Viewing activity-linked audio as a context-dependent auditory cue that supplements partner information that is difficult to obtain visually offers one direction for designing information presentation in collaborative VR.

References

[1] American Speech-Language-Hearing Association (2005) Guidelines for manual pure-tone threshold audiometry. Guidelines, https://doi.org/10.1044/policy.GL2005-00014

[2] Assaf R, Mendes D, Rodrigues R (2024) Cues to fast-forward collaboration: A survey of workspace awareness and visual cues in XR collaborative systems. Computer Graphics Forum 43(2):e15066. https://doi.org/10.1111/cgf.15066

[3] Audacity Team (n.d.) Noise. Audacity Manual, URL https://manual.audacityteam.org/man/noise.html, accessed 17 September 2026

[4] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[5] Biocca F, Harms C, Burgoon JK (2003) Toward a more robust theory and measure of social presence: Review and suggested criteria. Presence: Teleoperators and Virtual Environments 12(5):456–480. https://doi.org/10.1162/105474603322761270

[6] Brument H, De Pace F, Podkosova I (2025) Does mixed reality influence joint action? impact of the mixed reality setup on users' behavior and spatial interaction. Journal on Multimodal User Interfaces 19:119–138. https://doi.org/10.1007/s12193-024-00445-w

[7] Cidota M, Lukosch S, Datcu D, et al (2016) Comparing the effect of audio and visual notifications on workspace awareness using head-mounted displays for remote collaboration in augmented reality. Augmented Human Research 1(1):3. https://doi.org/10.1007/s41133-016-0003-x

[8] Effenberg AO, Fehse U, Schmitz G, et al (2016) Movement sonification: Effects on motor learning beyond rhythmic adjustments. Frontiers in Neuroscience 10:219. https://doi.org/10.3389/fnins.2016.00219

[9] Elkin LA, Kay M, Higgins JJ, et al (2021) An aligned rank transform procedure for multifactor contrast tests. In: Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology. ACM, pp 754–768, https://doi.org/10.1145/3472749.3474784

[10] Exit Games GmbH (2026) Photon Fusion 2: Fusion 2 introduction. Photon Engine Documentation, URL https://doc.photonengine.com/fusion/v2/fusion-intro, accessed 17 September 2026

[11] Exit Games GmbH (2026) Photon Voice 2: Voice intro. Photon Engine Documentation, URL https://doc.photonengine.com/voice/v2/getting-started/voice-intro, accessed 17 September 2026

[12] Frid E, Moll J, Bresin R, et al (2019) Haptic feedback combined with movement sonification using a friction sound improves task performance in a virtual throwing task. Journal on Multimodal User Interfaces 13:279–290. https://doi.org/10.1007/s12193-018-0264-4

[13] Gaver WW (1986) Auditory icons: Using sound in computer interfaces. Human–Computer Interaction 2(2):167–177. https://doi.org/10.1207/s15327051hci0202_3

[14] Gutwin C, Greenberg S (2002) A descriptive framework of workspace awareness for real-time groupware. Computer Supported Cooperative Work 11(3–4):411–446. https://doi.org/10.1023/A:1021271517844

[15] Gutwin C, Schneider OS, Xiao R, et al (2011) Chalk sounds: The effects of dynamic synthesized audio on workspace awareness in distributed groupware. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work. ACM, pp 85–94, https://doi.org/10.1145/1958824.1958838

[16] Hall ET (1966) The Hidden Dimension. Doubleday, Garden City, NY

[17] Hasegawa M, Kawaguchi I (2026) Investigation of a method for enhancing social presence through noise presentation in VR (in japanese). IPSJ SIG Technical Report, Collaboration and Network Services (CN) 2026-CN-127(54):1–8

[18] Herbst KC, Gaertner L, Insko CA (2003) My head says yes but my heart says no: Cognitive and affective attraction as a function of similarity to the ideal self. Journal of Personality and Social Psychology 84(6):1206–1219. https://doi.org/10.1037/0022-3514.84.6.1206

[19] Hermann T, Hunt A, Neuhoff JG (eds) (2011) The Sonification Handbook. Logos Publishing House, Berlin, Germany, URL https://sonification.de/handbook/

[20] Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2):65–70. https://doi.org/10.2307/4615733

[21] Ide M, Oshima S, Mori S, et al (2020) Effect of avatars' gesture expression in VR meeting (in japanese). IPSJ SIG Technical Report 2020-GN-110(1):1–8

[22] Kim I, Sung J (2024) New proxemics in new space: Proxemics in VR. Virtual Reality 28(2):85. https://doi.org/10.1007/s10055-024-00982-5

[23] Kimmel S, Landwehr E, Heuten W (2024) Kinetic connections: Exploring the impact of realistic body movements on social presence in collaborative virtual reality. Proceedings of the ACM on Human-Computer Interaction 8(CSCW2):1–30. https://doi.org/10.1145/3686910

[24] Koch D, Medlar A, Głowacka D (2024) Read the fine print: Assessing the user experience of reading in high resolution virtual reality. In: Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems. ACM, pp 1–7, https://doi.org/10.1145/3613905.3650916

[25] Lee M, Bruder G, Welch GF (2017) Exploring the effect of vibrotactile feedback through the floor on social presence in an immersive virtual environment. In: 2017 IEEE Virtual Reality (VR). IEEE, pp 105–111, https://doi.org/10.1109/VR.2017.7892237

[26] Meta Platforms, Inc. (2026) Hardware and software requirements. Meta Horizon OS Developers, URL https://developers.meta.com/horizon/documentation/unity/unity-development-requirements/, accessed 17 September 2026

[27] Oh CS, Bailenson JN, Welch GF (2018) A systematic review of social presence: Definition, antecedents, and implications. Presence: Teleoperators and Virtual Environments 27(1):1–35. https://doi.org/10.1162/pres_a_00340

[28] Piumsomboon T, Dey A, Ens B, et al (2019) The effects of sharing awareness cues in collaborative mixed reality. Frontiers in Robotics and AI 6:5. https://doi.org/10.3389/frobt.2019.00005

[29] Rasmussen T, Feuchtner T, Huang W, et al (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 89:103655. https://doi.org/10.1016/j.jvcir.2022.103655

[30] RootMotion (2026) Final IK. Unity Asset Store, version 2.5, URL https://assetstore.unity.com/packages/tools/animation/final-ik-14290, released 10 April 2026; accessed 17 September 2026

[31] Schubert G, Bratoev I, Strelchenko V, et al (2019) I hear, what you are doing! workspace awareness in collaborative virtual environments. In: Proceedings of the 24th International Conference of the Association for Computer-Aided Architectural Design Research in Asia, vol 1. CAADRIA, pp 583–592, URL https://cdp.ai.ed.tum.de/publications/Caadria_2019_I_hear_what_you_are_doing.pdf

[32] Unity Technologies (2023) Unity 2022.3.7f1 release notes. Unity Releases, URL https:

//unity.com/releases/editor/whats-new/2022.3.7f1, accessed 17 September 2026

[33] Unity Technologies (2026) Unity 6000.4.7f1 release notes. Unity Releases, URL https://unity.com/releases/editor/whats-new/6000.4.7f1, accessed 17 September 2026

[34] Valve Corporation (n.d.) Steam Audio Unity integration. Steam Audio Documentation, URL https://valvesoftware.github.io/steam-audio/doc/unity/index.html, accessed 17 September 2026

[35] Wobbrock JO, Findlater L, Gergle D, et al (2011) The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp 143–146, https://doi.org/10.1145/1978942.1978963