

From Flat Screens to Full Immersion: Social Interaction in VR2Gather, Horizon Workrooms, and MS Teams

POLINA ORHUNOVA, Computer Science, Texas State University, San Marcos, Texas, USA

ASHUTOSH SINGLA, Distributed and Interactive Systems, Centrum Wiskunde en Informatica, Amsterdam, Netherlands

WENHAN TAO, Computer Science, Texas State University, San Marcos, Texas, USA

JACK JANSEN and **IRENE VIOLA**, Distributed and Interactive Systems, Centrum Wiskunde en Informatica, Amsterdam, Netherlands

PABLO CESAR, Distributed and Interactive Systems, Centrum Wiskunde en Informatica, Amsterdam, Netherlands and INSy, TU Delft, Delft, Netherlands

MYLÈNE C. Q. FARIAS, Computer Science, Texas State University, San Marcos, Texas, USA

Recent advancements in immersive communication technologies have facilitated the exploration of novel modalities of social interaction, utilizing platforms that offer a spectrum of representations from simplified avatars to photorealistic volumetric and point-cloud reconstructions. This study focused on how these different representations and degrees of immersion impact communication processes and collaborative task performance. The primary objective was to examine potential variances in performance and interaction quality across different systems during a basic communication task. To achieve this, we employed Meta Horizon Workrooms as a three degrees of freedom (3DoF) system, VR2Gather as a six degrees of freedom (6DoF) system, and the MS Teams meeting application as a comparative benchmark. A user study was conducted whereby participants engaged in a charade game, alternating between the roles of “mimic” and “guesser.” Throughout the experimental session, data were collected to assess comfort, presence, task load, and interaction quality, with performance quantified by the number of words guessed per minute. The findings indicated no significant differences in workload, presence, or simulator sickness across platforms; however, significant differences were observed in audio–visual quality, communication effort, and enjoyment. Performance metrics revealed that participants exhibited the highest task performance within MS Teams, followed by VR2Gather, and subsequently Horizon Workrooms. These findings underscore the influence of embodiment fidelity and system familiarity on communication quality within immersive environments. We discuss system-level constraints and propose considerations for the design of future social extended reality (XR) platforms that effectively balance realism with usability and performance.

Authors’ Contact Information: Polina Orhunova (corresponding author), Computer Science, Texas State University, San Marcos, Texas, USA; e-mail: sih56@txstate.edu; Ashutosh Singla, Distributed and Interactive Systems, Centrum Wiskunde en Informatica, Amsterdam, Netherlands; e-mail: Ashutosh.Singla@cwi.nl; Wenhan Tao, Computer Science, Texas State University, San Marcos, Texas, USA; e-mail: qoa10@txstate.edu; Jack Jansen, Distributed and Interactive Systems, Centrum Wiskunde en Informatica, Amsterdam, Netherlands; e-mail: jack.jansen@cwi.nl; Irene Viola, Distributed and Interactive Systems, Centrum Wiskunde en Informatica, Amsterdam, Netherlands; e-mail: irene.viola@cwi.nl; Pablo Cesar, Distributed and Interactive Systems, Centrum Wiskunde en Informatica, Amsterdam, Netherlands and INSy, TU Delft, Delft, Netherlands; e-mail: p.s.cesar@cwi.nl; Mylène C. Q. Farias, Computer Science, Texas State University, San Marcos, Texas, USA; e-mail: mylene@txstate.edu.



This work is licensed under Creative Commons Attribution International 4.0.

© 2026 Copyright held by the owner/author(s).

ACM 1551-6865/2026/9-ART254

<https://doi.org/10.1145/3839363>

CCS Concepts: • **Human-centered computing** → **User studies; Mixed / augmented reality; Virtual reality**; • **Information systems** → **Multimedia streaming**;

Additional Key Words and Phrases: Social XR, Point Cloud, Microsoft Teams, Charade, Communication, eXtended Reality, Avatar, Horizon

ACM Reference format:

Polina Orhunova, Ashutosh Singla, Wenhan Tao, Jack Jansen, Irene Viola, Pablo Cesar, and Mylène C. Q. Farias. 2026. From Flat Screens to Full Immersion: Social Interaction in VR2Gather, Horizon Workrooms, and MS Teams. *ACM Trans. Multimedia Comput. Commun. Appl.* 22, 9, Article 254 (September 2026), 25 pages. <https://doi.org/10.1145/3839363>

1 Introduction

In recent decades, rapid advances in imaging, networking, and computational technologies have enabled increasingly precise digital representations of the physical world, fostering immersive and interactive experiences that closely approximate reality [1, 31]. As a result, **Virtual Reality (VR)** and mixed reality applications have gained significant interdisciplinary attention, supporting new forms of multimodal human–computer interaction where users can seamlessly engage with or within hybrid physical–virtual environments [7, 31]. These immersive experiences span the *virtuality continuum*, from augmented and mixed reality to fully virtual environments.

Despite their promise, current VR systems still face substantial challenges, including visual discomfort, motion sickness, latency, limited spatial awareness, and inconsistent interaction fidelity. These limitations affect not only system usability and task performance but also broader adoption across domains such as healthcare, education, training, entertainment, and remote collaboration. Addressing these issues requires a deeper understanding of how system performance, perceptual factors, and interaction design jointly influence user experience.

A useful framework for studying these issues is provided by the concept of **Quality of Experience (QoE)**. Qualinet [4] defines QoE as the degree of delight or annoyance experienced during the use of an application or service, shaped by user expectations, context, and personal factors. In immersive media, QoE encompasses perceptual and cognitive elements such as presence, immersion, and motion-induced discomfort [32]. These are affected by three main **Influential Factors (IF)**: the system, the environment, and the human user. Although much prior work has emphasized system factors, such as display resolution, latency, and field of view [20, 21], the interplay between these and human perceptual mechanisms is not yet understood adequately.

In this article, we focus on understanding how variations in system-level properties affect user performance and perceived experience in VR-based communication tasks. Specifically, we investigate whether changes in a System IF (independent variable) influence QoE (dependent variable) as participants perform a fundamental communication task under controlled conditions. To this end, we conducted subjective experiments involving a visual charade game, comparing user experiences across three platforms: MS Teams (a traditional 2D conferencing baseline), Meta Horizon Workrooms (hereafter Horizon Workrooms or Horizon) (**Three Degrees of Freedom (3DoF)**), and VR2Gather (**Six Degrees of Freedom (6DoF)**). The experiments were carried out at Texas State University, USA, and **Centrum Wiskunde and Informatica (CWI)**, The Netherlands. This research is conducted in collaboration with the Video Quality Experts Group, particularly within the Immersive Media Group¹ project, which aims to evaluate paradigms that assess various aspects of QoE in **Extended Reality (XR)** communications.

¹<https://vqeg.org/projects/img/>.

The article is organized as follows. Section 2 describes the current similar work, while Section 3 describes the experimental setup used in this research. Section 4 details the collected experimental data and Section 5 shows the results. Finally, Section 6 presents a discussion on the results, while Section 7 presents the conclusions of the research.

2 Related Work

The interest in remote communication and collaboration is increasingly gaining traction, particularly within the realm of Social XR environments. Human representations in these environments can be avatar-based representations [6, 8, 11, 13, 27, 33, 38] or photorealistic representations [3, 26, 34, 41, 43, 44], with some Social XR systems incorporating both representations. The systems for representing users with avatars vary in their levels of fidelity and embodiment [6, 8, 11, 13, 27, 33, 35, 38]. For instance, the ImmerTai system, as documented by Chen et al. [8], employs pre-recorded videos of Taichi experts for educational purposes; however, it lacks the capacity for real-time interaction and feedback. In contrast, various social XR platforms, as referenced in the literature [6, 11, 13, 27, 33], provide real-time communication capabilities. These platforms utilize avatars that predominantly depict upper body and hand movements, thereby facilitating collaborative activities among participants. A common flaw in these systems is their simplistic user representations and the lack of facial expression tracking. The embodVR system [38] improves this by utilizing motion capture suits and the OptiTrack system to trace full-body movements, but still ignores facial expressions.

In other systems [14], local users are represented by photorealistic volumetric capture, while remote users appear as full-body avatars, allowing a shared sense of presence. Shimamoto et al. [35] propose a hybrid solution with users represented as 3D avatars complemented by live 360° video streams, fostering a shared experience of presence. Participants can also participate in real-time audio communication. The XR system in [18] explores various avatar realism levels, including point clouds, personalized avatars, cartoon-style, and abstract avatars, used in tasks such as verbal explanation and self-disclosure. The system presented in [29] features photorealistic volumetric reconstructions for all participants, allowing real-time communication. It is designed for applications such as virtual debates and immersive educational sessions. An asymmetric communication system is presented in [43], where a photorealistic representation is used for the teacher, and a viewpoint avatar is used for the student. In [34], users are represented using live-streamed 2D video, which is seamlessly integrated into the shared VR environment to allow real-time interaction. Some systems use mesh representations for users [3, 26, 44], while VR2Gather represents users using point clouds [17, 41]. For our experiment, we selected VR2Gather because it is open source, supports symmetric communication, and allows for easy modifications.

Social VR systems are frequently designed for collaborative engagements, including but not limited to block construction [9], physiotherapy applications [40], photo-sharing interactions [23], and communal viewing of films [10, 22]. The effective implementation of these systems requires a comprehensive understanding of their performance, which can be achieved through an evaluation of the user experience. This evaluation is conducted predominantly through subjective experiments that involve voluntary human participants and aim to quantify various attributes of these systems, such as social presence [2], presence [42], **Quality of Interaction (QoI)** [23], simulator sickness [19, 36], and audio-visual quality [15, 16]. Numerous studies systematically assess the presence [8–10, 13, 23, 24, 27], simulator sickness [6, 27], and QoI [10, 13, 23, 24, 27, 40] to elucidate participants' perception of immersion and the efficacy of user communication. In addition, visual quality and comprehensive QoE are evaluated to determine users' perceptions of visual representations [9, 24, 40]. Certain studies have also measured users' task-related experiences during task execution within social VR environments [6, 24, 40]. In specific cases, researchers have

developed and validated customized questionnaires tailored to their research objectives [22, 23]. Audio-specific experiences are also scrutinized within social VR contexts. For example, Immohr et al. [13] illustrated user preferences for spatial audio in increasing social presence. The aforementioned studies adapt testing protocols in alignment with their respective research questions or tasks.

Several works in the literature compare two or more platforms for various tasks [10, 14, 23–25, 37]. Li et al. [23] compared three platforms, **Face-to-Face (F2F)**, Skype, and **Facebook Spaces (FS)** for photo-sharing experiences. Li et al. [24] compared two platforms, **Head Mounted Display (HMD)** and 2D screens, to watch a virtual movie together. De Simone et al. [10] examined video-watching experiences on three platforms: F2F, FS, and a social VR system. Irlitti et al. [14] compared the ShareAR System with the traditional video conferencing system for a block-building task. Macchi and De Pisapia [25] investigated the F2F, 2D videoconferencing, and Horizon Workrooms for idea generation and decision-making task. Singla et al. [37] compared a Social XR platform with 2D video communication for a physiotherapy use case. De Simone et al. [10] and Li et al. [23] found that the social VR system provides a comparable experience to F2F communication for their respective use cases. Several studies concluded that XR supports better remote collaboration than 2D platforms [14, 25, 37], while others found that XR provides a greater sense of presence [24] than 2D platforms.

Kaiser et al. [18] executed a thorough examination concerning the influence of various user representations on social presence and user comfort. The results indicate that point-cloud and photo-realistic avatars deliver a superior degree of realism relative to alternative representation modalities. Merz et al. [28] investigated the delineation between auditory and non-auditory communication within XR-based collaborative environments. The findings indicate that voice communication significantly enhances task performance and user experience, as expected. The study also finds that asymmetric XR communication reduces cognitive load.

The study presented in this manuscript highlights the critical role of non-verbal communication and bodily gestures in evaluating the efficacy of communication across three distinct platforms: VR2Gather (point cloud-based XR), MS Teams (two-dimensional video conferencing), and Horizon Workrooms (avatar-based XR). To the best of our knowledge, extant literature lacks a systematic comparison of these platforms within the framework of charade-based interaction.

3 Experimental Setup

In this section, we describe the experimental setup used in the experiments carried out in this work, including the implemented charade game, the hardware and software setup, and the experimental methodology.

3.1 The Charade Game

The charade game is generally played by two teams that alternate between guessing and mimicking words. For simplicity, in this work, only two people played the game in the experiments, which are referred to as Player A and Player B. Players A and B alternated between being the *mimic* and the *guesser*; both were standing during the experiment and were able to gesticulate. The mimic employed gestures, facial expressions, and body language to express the word's meaning, entirely without speaking or producing any sound. The guesser was asked to speak when trying to guess the mimicked word.

The charade words were selected with the awareness that some participants were not native English speakers. To ensure accessibility, we chose words that were simple, intuitive, and broadly recognizable. We first reviewed word lists commonly used in online charade games as an initial source of candidate terms. We then conducted a pilot trial to identify words that were either difficult to convey through gestures or likely to be poorly captured by the HMD-based tracking system. To minimize confusion and reduce cultural bias, we excluded movie titles, book names, idiomatic

expressions, and other complex references. Instead, we focused on common words that could be easily mimed. The final word set was organized into five broad categories: animals, objects, activities, professions, and food. Two separate lists of 30 words each were created; the full lists are available online.² Each of the 60 words was printed on an individual slip of paper and placed in a designated box. Each player was assigned a box containing 30 unique words, with no overlap between the two sets to avoid repetition or learning effects. Before each experimental session, the paper slips were thoroughly shuffled to ensure randomization.

At the beginning of the session, the experimenter randomly picked one of the participants as the mimic (Player A) and the other as the guesser (Player B). Then, the experimenter started the game by giving the first word to Player A, who had the option to accept or skip the word and ask for another word. After Player A accepted a word, they began to gesticulate the word without using sounds or voice. Player B could begin guessing by speaking (out loud) the words that they thought better represented the gesticulation. Should Player B fail to guess the word within 2 minutes, Player A would receive a new word to mimic, restarting the process. After 15 minutes, the roles of Players A and B were swapped, with Player A now being the guesser and Player B the mimic. The process was repeated in the next 15 minutes. The experimenter recorded the number of words guessed in each 15-minute interval and noted which words were skipped, guessed, or not attempted. Words were categorized as “skipped” when the participant chose to move on before the word was guessed or when the time limit for a word (2 minutes) was reached; as “correctly guessed” when the exact word was guessed within the time limit; and as “not attempted” for all words that were not reached within the overall session time limit (15 minutes).

3.2 Hardware and Software Setup

The experimental study focused on the same task, the charade game, carried out using three different communication platforms. As noted by Osborne et al. [30], there are numerous VR applications available; however, many lack comprehensive hand tracking or only provide limited tracking for a few fingers (e.g., Spatial and RecRoom). For the proposed experimental task, it is important to have hand tracking capabilities to allow players to gesticulate and visualize their hands accurately. Subsequent to preliminary pilot tests, Meta Horizon Workrooms and VR2Gather were selected as VR meeting platforms for this study. MS Teams was used as a baseline for a comparative experimental analysis. In addition, these three platforms are known for their extensive popularity, availability, and user-friendliness. The implementation of the charade task in the chosen VR applications (Horizon Workrooms and VR2Gather) required the use of two HMDs that offer a minimum of 3DoF to enable the two participants to execute movements of their upper body, encompassing both head and hand motions.

Horizon Workrooms is a free VR social platform developed by Meta (formerly known as Facebook). It was designed as a space where users can interact, socialize, play games, and create content in VR. Figure 1(a) shows an example of a virtual working environment in Horizon Workrooms. Notice that in this environment, which is meant for work collaboration, users are represented by full-body avatars whose upper-body movements are tracked. Similar to other platforms, in Horizon Workrooms, users can either pick a pre-made avatar or customize an avatar using Meta’s avatar creation system. The avatar in Horizon Workrooms shows the hand movements as long as the hands are in front of the HMD visor. Therefore, while playing a charade game in Horizon Workrooms, players can gesticulate with the movements of the hands and of each of the five fingers being displayed correctly. Notice in Figure 1(a) that the avatars available in Horizon Workrooms look somewhat realistic. For Horizon Workrooms, we have used the Meta Quest 2 HMD, which

²https://docs.google.com/document/d/1zwQCXtDI2qtXojCGvupVj6a-zdL-M3ncbHQ-_C6pHm0/edit?usp=sharing.

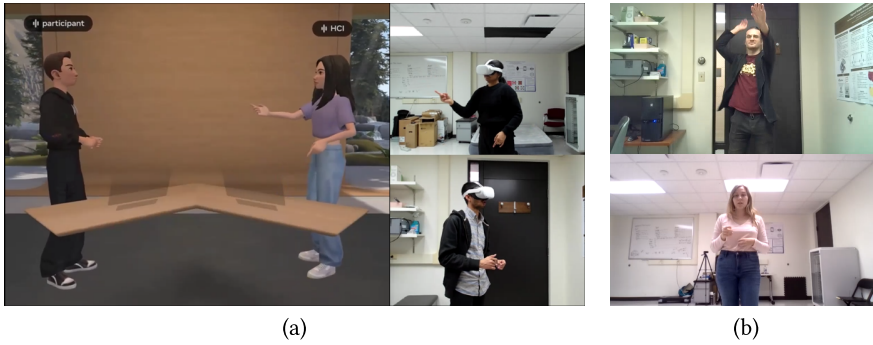


Fig. 1. (a) Meta Horizon Workrooms meeting application, available at meta.com/horizon-workrooms. (b) Screenshot of a teleconference with multiple users at MS Teams.

has eye, head, and hand tracking capabilities and is recognized by many VR meeting applications. To minimize bias and lag associated with rendering, we opted for a simple room setup (shown in Figure 1(a)), which allows the two players to maintain clear visibility of each other. The sessions were run using two Meta Quest 2 devices, two Windows-based desktops, and two large displays. The Quest 2 devices were used only by the two participants playing the game. A third Quest 2 device was made available in the laboratory for backup purposes. The use of desktops had the goal of providing the experimenter with an easy way to control VR and to allow the sessions to be recorded. The audio-visual recording had the goal of helping to capture the exact time the words are guessed. They were performed using the cameras on the two desktops and showed a lateral view of the players' interaction, capturing the audio of the guesser.

VR2Gather [17, 41] is a socially immersive XR platform that enables real-time audio-visual communication using volumetric representations. It relies on the **CWI Point Clouds (CWIPC)** system³ to capture and generate point clouds using RGB-D cameras such as Azure Kinect or Intel RealSense. In our experimental setup, point clouds were captured using four Azure Kinect cameras arranged in a circular configuration around the participant, as illustrated in Figure 2(a). The cameras were positioned at a distance of approximately 1.5 meters from the participant and mounted at a height of 1.80 meters. To ensure stability throughout the experiment, the cameras were fixed using tape and remained unchanged across sessions. Prior to data collection, a multi-camera calibration was performed using the CWIPC calibration pipeline to ensure accurate spatial alignment and fusion of the point clouds.

Point clouds were generated at 15 frames per second from synchronized RGB-D streams, with RGB images captured at a resolution of $2,048 \times 1,548$ and depth images at 640×576 . The CWIPC pipeline was configured with subject isolation to reduce noise and improve participant segmentation. No additional compression was applied to the point cloud data, thereby preserving high-fidelity volumetric reconstruction. An example of the resulting point cloud representation is shown in Figure 2(b). The virtual environment was designed as a minimalistic room with gray walls and a door, ensuring reduced visual distractions and allowing participants to focus on the interaction task. Participants were positioned at a distance of 2 meters in the virtual space, providing sufficient room for natural movement while maintaining clear visibility of each other's actions.

Each participant used a Meta Quest Pro HMD, wirelessly connected to a dedicated workstation. The built-in microphone of the HMD was used for audio capture. The VR2Gather application was executed on two identical Windows 10 machines, each equipped with an Intel i9 CPU (3.6 GHz),

³<https://github.com/cwi-dis/cwipc>.

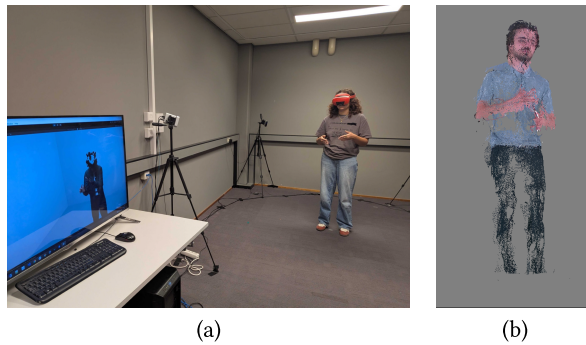


Fig. 2. (a) Experimental setup used to capture participants' movements and the participant's view on the screen. (b) Example of the corresponding point cloud representation.

64 GB RAM, and an NVIDIA GeForce RTX 2080 Ti GPU. Communication between the two systems was established over a 1 Gbit per second Ethernet connection using a transmission control protocol-based transport protocol. No artificial delay was introduced in either the audio or point cloud streams, ensuring that the system operated under natural latency conditions.

MS Teams is a collaborative platform developed by MS and utilized for purposes such as chat, video conferencing, and file storage. Its secure design makes it suitable for use by companies and organizations. Within this platform, users have the ability to create and join teams and channels, facilitating organized management of conversations and projects. In addition, they can participate in meetings and engage in real-time document collaboration. MS Teams' support for video and audio conferencing makes it particularly suitable for the experiment at hand. Figure 1(b) provides a visual representation of a teleconference conducted through Teams. For the MS Teams experiment, the sessions required only the computer and displays. The audio-visual recording was done on the Teams platform and showed both participants.

3.3 Experimental Methodology

Before conducting the experiment, several experimental parameters were specified. With respect to the virtual environment, as previously described, a minimalistic VR room was used to reduce potential biases and limit computational complexity. The participants were positioned facing each other within this virtual space. Furthermore, to minimize preparation time, participants were assigned preconfigured avatars by the experimenters within Horizon Workrooms. The experimental sessions were conducted at two locations: the Texas State University Computer Science Research Laboratory in the United States and the CWI in the Netherlands.

In both locations, participants were recruited from university communities. The recruitment was conducted on a voluntary basis through institutional mailing lists and local announcements. The eligibility criteria required participants to be at least 18 years of age, have sufficient English proficiency to understand the task instructions and charade words, and report no known motor or visual impairments that could reasonably be expected to interfere with the performance of the task in the game. Prior experience with VR was not a prerequisite for participation; however, self-reported familiarity with VR systems and charade games was obtained through a pre-experimental questionnaire for subsequent analytical purposes.

The study utilized a between-subjects design, which consisted of different groups of subjects performing the experiment with specific communication platforms, thus minimizing learning effects and ensuring consistency in the vocabulary used across all environments. In this design, each pair

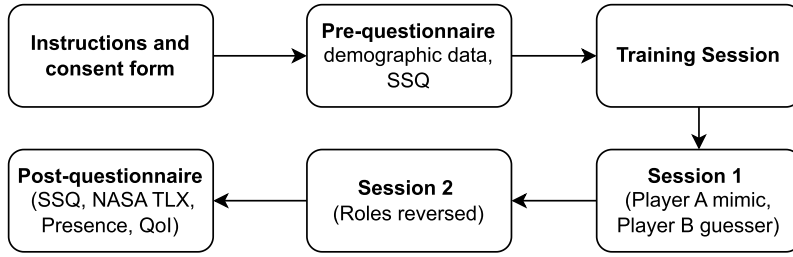


Fig. 3. Flowchart of the experimental procedure, representing the order of stages of the experiment.

of participants was assigned to one single communication condition (MS Teams, VR2Gather, or Meta Horizon Workrooms). This approach was chosen to minimize learning and carry-over effects between platforms and to ensure that participants encountered the charade task in only one communication setting. The participants' pairs were assigned conditions to maintain comparable group sizes and balanced demographic characteristics in the three scenarios. The general experimental procedure is summarized in Figure 3.

At the start of all three experiments, participants were welcomed by the experimenter. They were instructed to complete a Pre-Questionnaire and were required to read and sign a Consent Form. This form provided a short overview of the experiment, outlining potential risks and benefits, as well as the procedure. It clarified that collected data would be dedicated solely to research. Participants received a copy of the signed consent form, while the experimenter retained another copy. The Pre-experiment Questionnaire included numerous questions organized into distinct sections. The initial section gathered participants' personal information, such as gender, education level, age, experience with VR, HMDs, charades, native language, and so on. The subsequent section consisted of a **Simulator Sickness Questionnaire (SSQ)** [19], intended to assess sickness in the context of motion-induced discomfort in VR. The pre-questionnaire is available online.⁴

The participants were then placed in two different rooms to limit audio and visual communication to be exclusively through the VR environment or the MS Teams. Next, a training session was held with the goal of getting the participants familiarized with the HMD, the VR/teleconference environment, and the Charade game. For VR2Gather, it was also necessary to verify point cloud alignment before each session. The training session consisted of a small set of trials in which the experimenter first explained the charade game orally, detailing the task, how the words are communicated to them, and what they are allowed to do. The participants were then asked to try out the interface and gesticulate to check the setup. Finally, they were asked to mimic/guess at least one word (each) to make sure they understood the task. The participants were told that they could ask the experimenter questions during the game to clarify any doubts. Most importantly, participants were also informed that they could request a pause or withdraw from the experiment at any time if they experienced discomfort or simulator sickness. During the training phase, participants were given sufficient time to adapt to the system and confirm that they felt comfortable before starting the main task. Short breaks were allowed whenever needed, and the experimenters continuously monitored participants for signs of discomfort throughout the session.

After finishing the training, the experimenter asked if the participants felt comfortable playing the game. If so, they proceeded to the main session and played the Charade game (see Section 3.1). As mentioned earlier, the experimental session was recorded for better precision of the metrics, since the player's performance was measured in terms of words completed per time unit. For the

⁴https://drive.google.com/file/d/1_Fnx7m9SoGJf_tzJpJVpX38hJxZ33J/view?usp=drive_link.

main session, we randomly selected the order of the two word sets (Sessions 1 and 2) for each pair of participants, so that either set could be played first. The complete list of words used in the two sets is provided in Appendix A. The total duration of each experimental session was approximately 60 minutes, including setup, training, gameplay, questionnaires, and optional short breaks when needed. Each experimental session consisted of two rounds of charades. In the first round, one participant acted as the mimic, while the other was the guesser; in the second round, the roles were reversed. This fixed role-reversal structure ensured that both participants experienced both roles equally. To reduce learning effects, different word sets were used for each round, and the article slips were shuffled before the start of each session.

After finishing the experimental session, the participants were asked to complete the post-experiment questionnaire, which was also divided into several sections. The first section was a repetition of the SSQ questionnaire answered before the experiment. This allowed determining whether the experiment caused motion sickness problems. In the second section, participants filled out the NASA **Task Load Index (TLX)** questionnaire [5], which provided multidimensional ratings of overall workload based on a weighted average of six subscales: mental demands, physical demands, temporal demands, performance, effort, and frustration. In the third and last section, participants answered a few questions about presence and social interaction.

The complete post-questionnaire can be found online,⁵ while the list of all questions asked (Q1-Q19) is shown in Table 1. The simulator sickness and workload questionnaires were based on the validated SSQ [19] and NASA-TLX [5] instruments, respectively. The elements of presence and social interaction were adapted from previous validated VR collaboration and presence studies, selected to capture the most relevant aspects to gesture-based remote communication. Most subjective questions used a 7-point Likert scale to allow finer-grained responses. Questions 7 and 8 used a 5-point Likert scale because they assessed broader evaluative judgments, for which a simpler response format was considered more appropriate.

4 Experimental Data

4.1 Data Collection Instruments

To evaluate the experience of participants in three environments, we collected data on motion sickness, comfort, presence, subjective workload, and social interaction, using widely used questionnaires for these factors, along with some custom-modified questions:

- (1) *NASA TLX*: The NASA TLX measures the overall workload, covering both mental and physical aspects. The participants rated six domains: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. All questions were assessed using a scale ranging from 0 to 20, with higher values indicating an increased workload. However, for Performance, which initially utilized higher scores to signify better outcomes, we inverted this scale for consistency in plotting and analysis. Once the data was gathered, we adjusted all scores to fall within a 0-100 scale. This was done to facilitate comparison and interpretation across different subscales, adhering to the standard reporting practices for NASA TLX.
- (2) *SSQ*: The SSQ [19] is a widely used tool to evaluate the symptoms of sickness that may occur during VR use. It includes 16 different symptoms such as dizziness, eye fatigue, general discomfort, and others. After participants rated all symptoms, the scores were weighted using standard formulas and grouped into four categories: Nausea, Oculomotor, Disorientation, and Total Score.
- (3) *Presence and QoI*: The presence was measured using questions on a Likert scale from 1 to 7, where 7 represented an immersive experience, and 1 represented no immersion at all. For

⁵https://drive.google.com/file/d/1EJRYpCqetLX_nw3Lx2ftE5orpPaW8jh-/view?usp=drive_link.

Table 1. Survey Questions, Response Scales, and References for Presence (Question 1–6) and QoI (Questions 7–19) Sections

No.	Question	Scale	Ref.
1	Please rate your sense of being in the space on the following scale from 1 to 7, where 7 represents your normal experience of being in a place.	1–7	Q1
2	To what extent were there times during the experience when the space was the reality for you?	1–7	Q2
3	When you think back about your experience, do you think of the space more as images that you saw, or more as somewhere that you visited?	1–7	Q3
4	During the time of the experience, which was strongest on the whole: your sense of being in the space or of being elsewhere?	1–7	Q4
5	Consider your memory of being in the space. How similar in terms of the structure of the memory is this to the structure of the memory of other places you have been today? By “structure of the memory,” consider things like the extent to which you have a visual memory of the office space, whether that memory is in color, the extent to which the memory seems vivid or realistic, its size, location in your imagination, the extent to which it is panoramic in your imagination, and other such structural elements.	1–7	Q5
6	During the time of the experience, did you often think to yourself that you were actually in the office space?	1–7	Q6
7	How would you rate the overall audio–visual quality?	1–5	Q7
8	How easy did you find it to communicate using the system?	1–5	Q8
9	I felt connected with my partner.	1–7	Q9
10	I felt I was present with my partner.	1–7	Q10
11	My partner was able to sense my presence.	1–7	Q11
12	My partner could tell when I needed assistance.	1–7	Q12
13	I enjoyed the experience.	1–7	Q13
14	I was able to focus on the task activity.	1–7	Q14
15	I am confident that we completed the task correctly.	1–7	Q15
16	My partner and I worked together well.	1–7	Q16
17	I was able to express myself clearly.	1–7	Q17
18	I was able to understand partner’s message.	1–7	Q18
19	Information from partner was helpful.	1–7	Q19

QoI, two questions target the technical quality of the system, while the remainder assessed the interaction between participants using the system. Details of these questions can be found in Table 1.

- (4) *Performance Metrics*: The performance metric involved the measurement of the time participants allocated to elucidate each word, alongside the classification of each word as Guessed, Skipped, or Not Attempted, in order to furnish an objective assessment of task performance across different environments.

The next section presents the analysis of all scores for these metrics, which we used to identify differences between the three environments examined in this study.

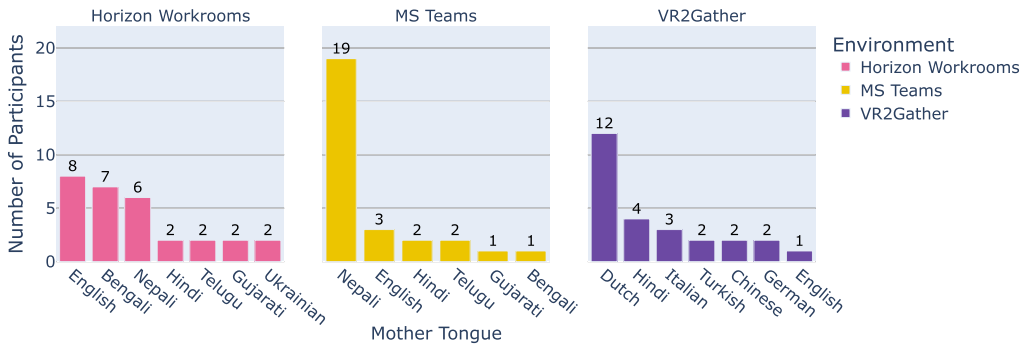


Fig. 4. Native language distribution across three environments (≥ 2 participants overall across datasets).

4.2 Demographics

Following the exclusion of data from seven experimental sessions due to non-compliance with the experimental protocol, incomplete data, and internet connectivity issues, the final dataset comprised 90 samples from healthy adult participants, with 30 samples recorded per environment. A total of 61 males, 28 females, and one individual categorized as “Other” participated in the experiment. Participants’ ages for the Horizon Workrooms VR platform ranged from 18 to 31 years, with an average age of 22.7 years (SD 4.1, median 20). For the VR2Gather VR platform, ages varied from 23 to 64 years, averaging 34.4 years (SD 12.6, median 28.5). In contrast, for the MS Teams platform, ages spanned 18 to 27 years, with an average age of 20.1 years (SD 2.0, median 20). Participants reported a diverse range of native languages, with Nepali being the most common, followed by English and Dutch. Other languages included Bengali, Hindi, and several others represented by one or two participants each. Figure 4 illustrates the participants’ native language distribution. The majority of participants (67 out of 90) had previous experience with the game of Charades. Among these individuals, 55.6% reported being somewhat familiar, 20% indicated being very familiar, 18.9% were a little familiar, and 5.6% were not familiar at all with the rules of the game.

5 Experimental Results

5.1 Subjective Results

To determine whether the workload varies in different settings, we used the NASA TLX questionnaire [5], which evaluates mental, physical, and temporal demands, as well as performance, effort, and frustration. Figure 5 presents the TLX Scores in all three experiments. We conducted a Shapiro-Wilk normality test to assess the distribution of overall workload, revealing that in one of the environments the distribution is not normal. As a result of this **Assumption Violation for the One-Way Analysis of Variance (ANOVA)**, we proceeded with a non-parametric test. The Kruskal-Wallis test indicated that there was no significant difference in workload between the three environments ($H(2) = 5.75, p = 0.057, \epsilon^2 = 0.043, 95\% \text{ CI } [0.00, 0.19]$).

Questions 7-19 (Q7-Q19), which evaluate QoI, were analyzed either individually or as composite constructs, depending on the conceptual similarity and Cronbach’s alpha. The final set of variables is as follows: Q7, Q8, Q15, and Q16 were analyzed separately, as they are testing different aspects; the remaining questions were grouped as follows: social presence (Q9-Q12, $\alpha = 0.653$), engagement (Q13, Q14, $\alpha = 0.912$), communication (Q17-Q19, $\alpha = 0.838$). The Shapiro-Wilk test demonstrated that most variables significantly deviated from normality ($p < 0.001$), with the exception of social presence ($p = 0.13$). Therefore, we decided to use nonparametric tests for consistency. Kruskal-Wallis H test was utilized as a nonparametric alternative to the one-way ANOVA. As illustrated in Table 2, the test revealed significant differences in scores in different settings. Dunn’s test, adjusted

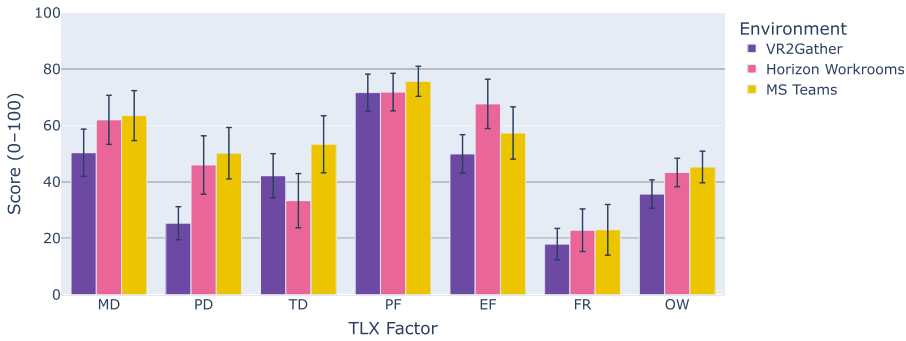


Fig. 5. NASA TLX scores (mean $\pm$ SD) by environment, grouped by the following factors: MD, PD, TD, PF, EF, FR, and OW. Reported higher values indicate greater perceived task load, except for Performance, where higher means better. Error bars represent 95% confidence intervals. EF, effort; FR, frustration; MD, mental demand; OW, overall workload; PD, physical demand; PF, performance; TD, temporal demand.

Table 2. Results of the Kruskal–Wallis Test for QoI Variables (Questions 7–19), Including Individual Items and Composite Ores, Showing the H-Statistic, p -Values, and Effect Sizes (ϵ^2) across Environments

Variable	H-Statistic	p -Value	ϵ^2
Q7 (Audio-visual Quality)	13.529	0.001	0.133
Q8 (Communication Effort)	9.987	0.007	0.092
Q15 (Task Confidence)	1.352	0.509	0.000
Q16 (Collaboration)	0.379	0.828	0.000
Social Presence (Q9–Q12)	0.681	0.712	0.000
Engagement (Q13 and Q14)	10.300	0.006	0.095
Communication (Q17–Q19)	1.655	0.437	0.000

Bold p -values indicate statistically significant differences between environments ($p < 0.05$).

with the Holm-Bonferroni correction, was executed to further examine these differences, with significant results observed for $p < 0.05$.

Figure 6 shows the bar plots for the results of Q7, Q8, and the composite score for Q13 and Q14 (engagement). Notice that the audio-visual quality ratings (Q7) provided by the participants for VR2Gather were significantly lower than those of Horizon ($p = 0.028$) and MS Teams ($p < 0.001$), as illustrated in Figure 6(a). No significant differences were observed between the Horizon and MS Teams ($p = 0.251$). According to the responses to Q8, shown in Figure 6(a), the participants reported that communication was perceived to require a greater degree of effort in the Horizon Workrooms compared to VR2Gather ($p = 0.018$) and MS Teams ($p = 0.013$). The engagement level, illustrated in Figure 6(b), was assessed to be markedly lower in VR2Gather compared to Horizon ($p = 0.0498$) and MS Teams ($p = 0.0056$).

Figure 7 illustrates the presence scores (Q1–Q6) derived from the set of three experiments, ascertained by calculating the mean of all participants' responses. A Shapiro-Wilk test verified that these composite scores adhered to a normal distribution across all environments. In contrast, Levene's test demonstrated a significant discrepancy in variances among the environments ($p = 0.025$). Consequently, Welch's ANOVA was applied, which did not detect any statistically significant differences in presence scores among the environments ($F(2, 53.83) = 0.71, p = 0.496, \eta_p^2 = 0.021, 95\% \text{ CI } [0.00, 0.15]$).

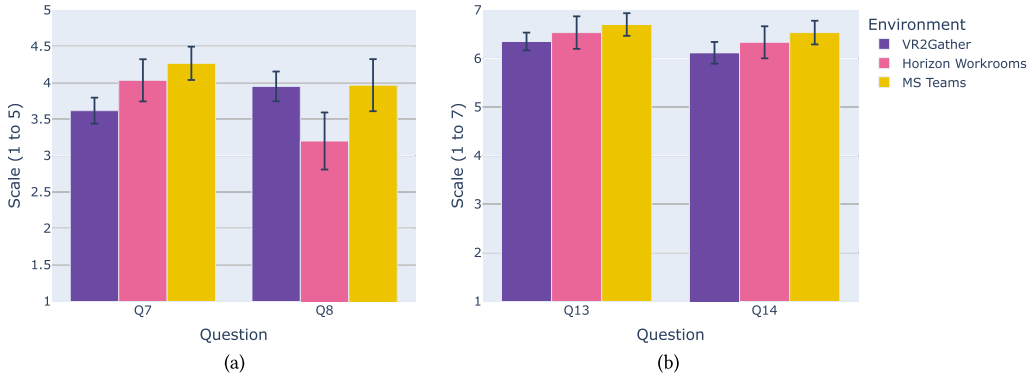


Fig. 6. QoI evaluation: audio-visual quality (Q7), communication effort (Q8), and engagement (Q13 and Q14) across environments. Error bars represent 95% confidence intervals. For visual consistency, the Q8 scale was inverted so that higher values indicate better ratings, as in all other plots in this figure. The analysis was performed on the original scale.

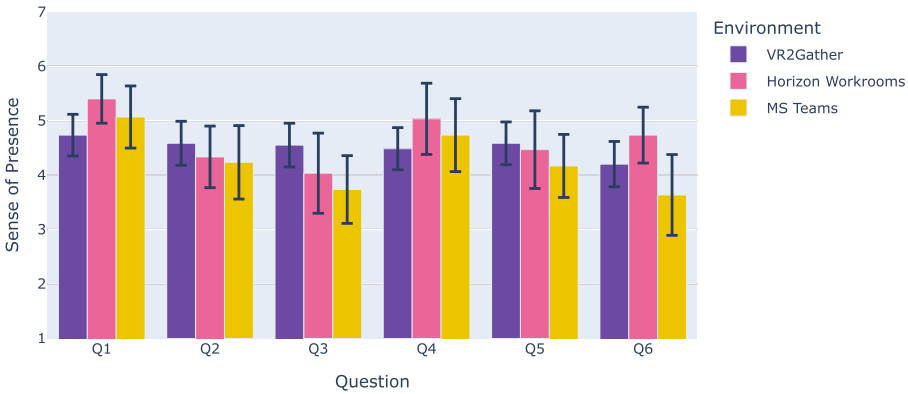


Fig. 7. Mean sense of presence scores for each of the six questions (Q1–Q6), grouped by environment, where 1 indicates low presence, and 7 indicates high presence. Error bars represent 95% confidence intervals.

Figure 8 shows the average SSQ scores captured from the three experimental environments after all sessions. In order to evaluate the SSQ scores among the three environments, the Shapiro-Wilk test was used to assess the normality assumption. The assumption was found to be violated across all groups, requiring the adoption of the Kruskal-Wallis H-test as an alternative to one-way ANOVA. The analysis did not reveal statistically significant differences between environments in terms of Nausea (N) ($H(2) = 5.06, p = 0.080, \epsilon^2 = 0.035, 95\% \text{ CI } [0.00, 0.18]$), Oculomotor (O) ($H(2) = 1.35, p = 0.509, \epsilon^2 \approx 0, 95\% \text{ CI } [0.00, 0.11]$), Disorientation (D) ($H(2) = 4.92, p = 0.085, \epsilon^2 = 0.034, 95\% \text{ CI } [0.00, 0.18]$) or the Total SSQ score ($H(2) = 1.66, p = 0.436, \epsilon^2 \approx 0, 95\% \text{ CI } [0.00, 0.11]$). All observed effect sizes were small or negligible. While the SSQ scores depicted in Figure 8 suggest that VR2Gather (N mean = 9.67 ± 12.83 , O mean = 10.70 ± 16.42 , D mean = 9.77 ± 20.04 , Total mean = 11.53 ± 16.70) exhibited relatively increased values in comparison to Horizon (N mean = 3.82 ± 5.93 , O mean = 8.09 ± 11.77 , D mean = 6.03 ± 14.48 , Total mean = 7.11 ± 10.13) and MS Teams (N mean = 6.36 ± 12.10 , O mean = 5.81 ± 7.36 , D mean = 1.86 ± 4.81 , Total mean = 5.86 ± 7.96), these observed discrepancies did not attain statistical significance.

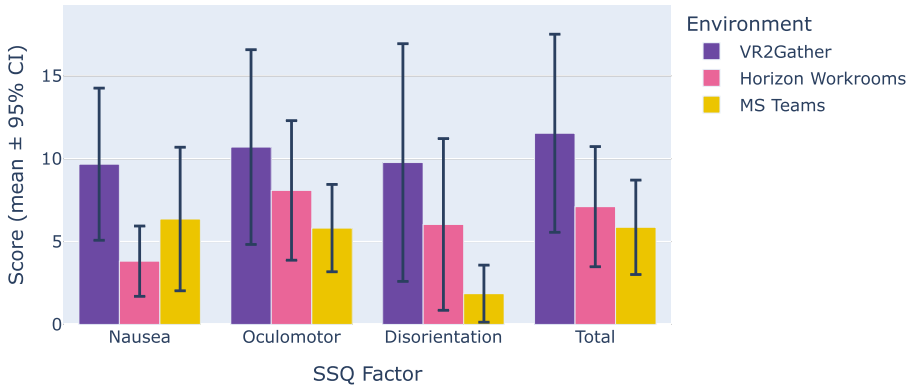


Fig. 8. SSQ mean scores for three environments, grouped by factors: Nausea, Oculomotor, Disorientation, and Total. Error bars show 95% confidence intervals for the mean. The higher values indicate higher discomfort.

5.2 Objective Results

Figure 9 illustrates the distribution of word statuses categorized by session and environmental context, with each stacked bar depicting the count of words that were correctly guessed, skipped, or not attempted by each participant. Initially, statistical analyses were conducted to ascertain whether discrepancies existed in the number of guessed words across different environmental configurations. Each group conformed to a normal distribution as verified by the Shapiro-Wilk normality test; however, Levene's test indicated that the homogeneity of variances assumption was violated ($p = 0.0085$). Consequently, Welch's ANOVA was employed, revealing a significant disparity in the number of guessed words among different environments, $F(2, 54.57) = 11.21$, $p < 0.001$, accompanied by a large effect size ($\eta^2 = 0.204$, 95% CI [0.09, 0.38]).

Post hoc analyses employing the Games-Howell test revealed that participants utilizing MS Teams identified a significantly greater number of words compared to those utilizing the Horizon platform (mean difference = 5.87, 95% CI [3.32, 8.42], $p = 0.00008$, $g = 1.18$). Participants employing the VR2Gather platform similarly outperformed those utilizing Horizon (mean difference = 5.67, 95% CI [2.49, 8.84], $p = 0.0021$, $g = 0.91$). Nonetheless, the comparison between the VR2Gather and MS Teams environments did not yield a statistically significant difference (mean difference = 0.20, 95% CI [-2.54, 2.94], $p = 0.988$, $g = 0.037$).

Our next research question focused on whether a learning effect occurred between Sessions 1 and 2. To investigate, we analyzed whether there were notable differences in the number of words correctly guessed over sessions within each setting. Due to factors such as a breached normality assumption in one group and the limited sample size per group ($n = 15$), we utilized the Mann-Whitney U test. The findings indicated that participants correctly guessed a significantly greater number of words in the second session for both the VR2Gather ($U = 53.0$, $p = 0.014$, $r_{rb} = 0.53$, 95% CI [0.17, 0.84]) and MS Teams ($U = 59.5$, $p = 0.028$, $r_{rb} = 0.47$, 95% CI [0.08, 0.80]) platforms, whereas no significant difference was observed between Sessions 1 and 2 for Horizon Workrooms ($U = 111.0$, $p = 0.967$, $r_{rb} = 0.01$, 95% CI [-0.40, 0.45]).

We used the Chi-Square test to see if response status (Correct/Skipped/Not Attempted) for specific words differs significantly between environments. Figure 10 provides a visual overview of the response distribution for each word across the three environments and the two experimental sessions. We could not test the words *Airplane*, *Basketball*, *Book*, and *Swimming* because there were no observations for these words in one of the three categories (Correct/Skipped/Not Attempted). All other results are presented in Table 3. Significant results are marked in bold. Several words



Fig. 9. Participant response distribution by session and environment. It illustrates the number of words that were correctly guessed (purple), skipped (orange), or not attempted (light orange) by each of 15 participants in two separate sessions and three environments ((a) MS Teams, (b) Horizon Workrooms, and (c) VR2Gather).

showed high significant differences across environments, including words *Dinosaur*, *Mustache*, and *Soccer*, which showed the strongest disparities ($p < 0.001$). These results indicate that performance for certain words varied meaningfully across the three environments.

We evaluated the performance of the participants using the **Words per Minute (WPM)** metric. The WPM was calculated as the total number of words correctly guessed by a participant divided by the total time (in minutes) spent guessing words, excluding any time spent on skipped or unattempted words. Before analysis, we identified outliers using the **Interquartile Range (IQR)** rule, which defines outliers as values below $Q1 - 1.5 \cdot IQR$ or above $Q3 + 1.5 \cdot IQR$. Values that were out of this range were excluded from visualizations and further analysis.

As before, we checked data distribution across the three environments for each session, and the Shapiro-Wilk test confirmed that the data were normally distributed in all groups. Levene's test showed that the variances were equal in both Session 1 ($p = 0.352$) and Session 2 ($p = 0.267$). Therefore, we used One-way ANOVA to determine whether there were significant differences in performance between the three platforms. The results revealed a significant effect of the environment on WPM in both Session 1 ($F = 20.31$, $p < 0.001$, $\omega^2 = 0.47$, 95% CI [0.28, 0.68]) and Session 2 ($F = 37.88$, $p < 0.001$, $\omega^2 = 0.63$, 95% CI [0.52, 0.78]). Subsequently, pairwise comparisons were executed employing Tukey's **Honestly Significant Difference (HSD)** test, with the outcomes detailed in Table 4.

For referential purposes, we have depicted the overall efficiency calculated by the ratio of accurately guessed words to the total duration of the session in minutes. Although no statistical analyses were performed on this measure, observations from Figure 12 suggest that the distribution of values aligns with the WPM depicted in Figure 11, with MS Teams demonstrating superior outcomes in both sessions.

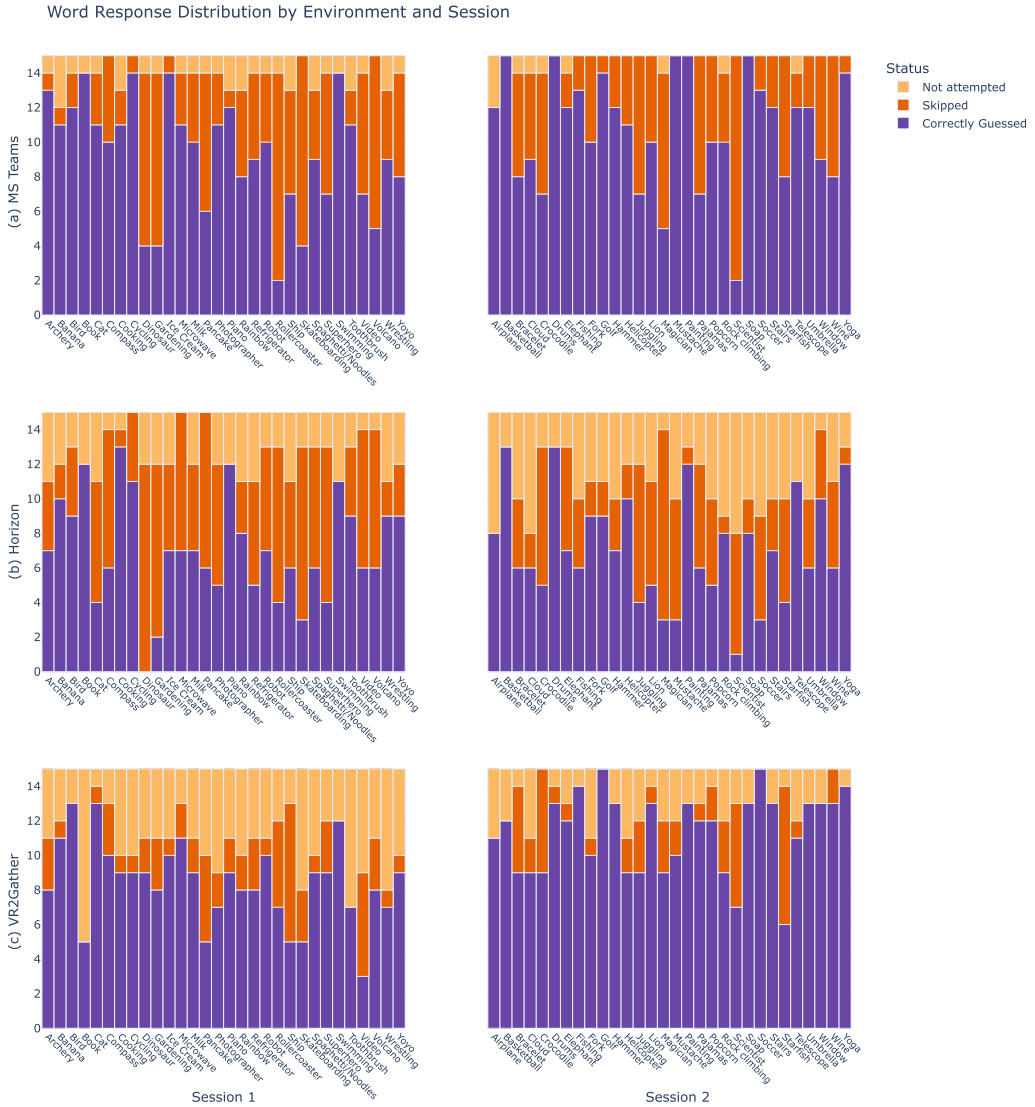


Fig. 10. Word response distribution by environment and session. It illustrates the number of times each unique word was correctly guessed (purple), skipped (orange), or not attempted (light orange) by each of 15 participants in two separate sessions and three environments ((a) MS Teams, (b) Horizon Workrooms, and (c) VR2Gather).

6 Discussion

The main goal of this article was to investigate whether there are differences in the performance and QoI when performing a simple communication task in three environments. Although Horizon Workrooms are technically built on 6DoF VR hardware, in practice the system constrains avatar movement to 3DoF, allowing participants to use upper-body gestures and head turns to interact, which aligned with the protocol of our experiment. VR2Gather represents the 6DoF system, and MS Teams was used as a benchmark. We did not find statistically significant differences between

Table 3. Chi-Square Test Results for (a) Session 1 and (b) Session 2 Words

(a) Session 1			(b) Session 2		
Word	Chi-Square	<i>p</i> -Value	Word	Chi-Square	<i>p</i> -Value
Archery	5.9643	0.2018	Airplane	—	—
Banana	0.5625	0.9671	Basketball	—	—
Bird	5.1647	0.2708	Bracelet	5.5801	0.2328
Book	—	—	Cloud	7.2500	0.1233
Cat	12.8766	0.0119*	Crocodile	3.4286	0.4888
Compass	5.6058	0.2306	Drums	4.1951	0.3802
Cooking	4.4773	0.3453	Elephant	6.6796	0.1538
Cycling	14.1176	0.0069**	Fishing	14.4545	0.0060**
Dinosaur	18.1346	0.0012***	Fork	7.3190	0.1200
Gardening	10.0109	0.0402*	Golf	11.6316	0.0203*
Ice Cream	10.6728	0.0305*	Hammer	10.3661	0.0347*
Microwave	7.8727	0.0964	Helicopter	4.9143	0.2962
Milk	3.5612	0.4686	Juggling	7.5316	0.1103
Pancake	8.2995	0.0812	Lion	12.2000	0.0159*
Photographer	9.7348	0.0451*	Magician	9.4159	0.0515
Piano	3.2121	0.5230	Mustache	21.2024	0.0003***
Rainbow	2.6727	0.6140	Painting	4.3500	0.3607
Refrigerator	4.1818	0.3820	Pajamas	10.4800	0.0331*
Robot	6.1212	0.1903	Popcorn	11.3889	0.0225*
Rollercoaster	6.7692	0.1486	Rock climbing	5.7722	0.2168
Ship	2.0702	0.7229	Scientist	18.1744	0.0011**
Skateboarding	13.9167	0.0076**	Soap	11.5952	0.0206*
Spaghetti/Noodles	7.2500	0.1233	Soccer	27.0000	<0.0001***
Superhero	5.8474	0.2108	Stairs	10.3661	0.0347*
Swimming	—	—	Starfish	8.6190	0.0714
Toothbrush	10.8889	0.0278*	Telescope	3.8088	0.4325
Video	8.1607	0.0859	Umbrella	11.9171	0.0180*
Volcano	9.6511	0.0467*	Window	8.4125	0.0776
Wrestling	5.2431	0.2633	Wine	13.6032	0.0087**
Yoyo	6.5436	0.1621	Yoga	3.2000	0.5249

Bold *p*-values indicate significant differences in response distributions across environments. DoF = 4 for all tests.

p* < 0.05, *p* < 0.01, ****p* < 0.001.

Table 4. Tukey's HSD Pairwise Comparisons of WPM across Environments after One-Way ANOVA

Session 1							Session 2						
Group 1	Group 2	Δ Mean	<i>p</i> -Adj	Lower	Upper	Hedges' <i>g</i>	Group 1	Group 2	Δ Mean	<i>p</i> -Adj	Lower	Upper	Hedges' <i>g</i>
Horizon	MS Teams	0.7281	0.0085	0.1651	1.2911	1.020	Horizon	MS Teams	1.6001	<0.001	1.1221	2.0782	2.896
Horizon	VR2Gather	-0.7220	0.0091	-1.2850	-0.1590	1.380	Horizon	VR2Gather	0.2903	0.3001	-0.1797	0.7603	0.604
MS Teams	VR2Gather	-1.4501	<0.001	-2.0033	-0.8969	2.168	MS Teams	VR2Gather	-1.3098	<0.001	-1.7798	-0.8398	2.305

Session 1 (left) and Session 2 (right). Hedges' *g* is reported as an absolute value; the sign of Δ mean indicates direction.

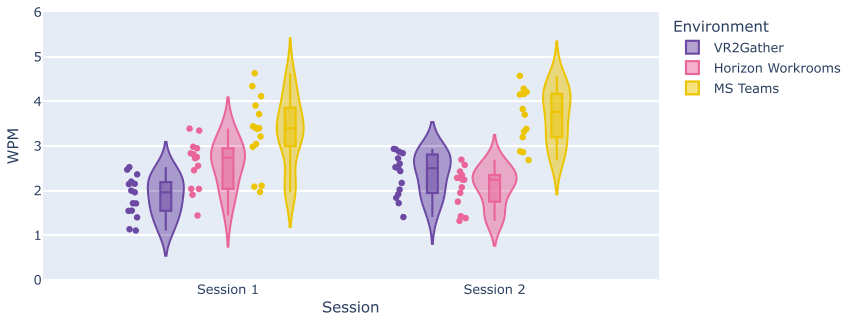


Fig. 11. WPM average across all participants.

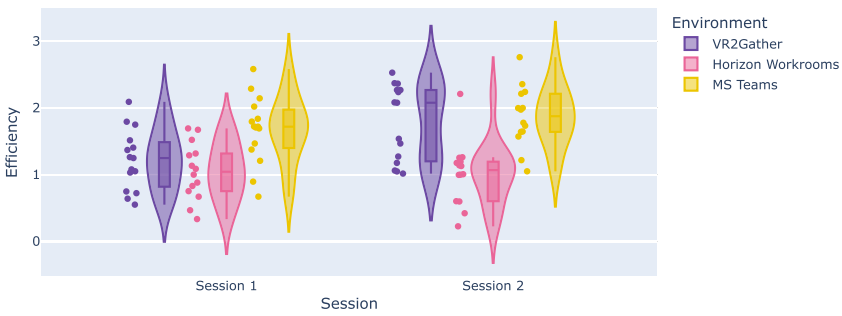


Fig. 12. Words per total time average across all participants.

the systems for NASA TLX, SSQ, or Presence measures, suggesting that levels of workload, motion sickness, and sense of being in space are relatively similar in three environments for the charade game task. This finding contrasts with previous work, which has reported a higher presence in VR/XR platforms compared to 2D systems [24, 39]. However, our results showed differences in other aspects of user experience, such as QoI, enjoyment, and task performance. In this section, we focus on interpreting these findings, considering possible system-level and user-level explanations and system limitations that could be addressed in future work.

6.1 Interpreting Subjective Results

Audio-visual quality was rated lower for VR2Gather compared to Horizon Workrooms and MS Teams. Point-cloud representations in VR2Gather are sensitive to camera positioning, lighting conditions, and errors in the surface reconstruction process. As a result, avatars often appeared fragmented or messy for participants, which could result in lower audio-visual quality ratings compared to Horizon's avatar-based system, where rendering is more consistent and stable 2D video from MS Teams. Similar issues have been reported in the study of point-cloud avatars, where users noted that while motion fidelity was high, visual fidelity was reduced compared to pre-constructed avatars [12].

Communication required more effort in Horizon Workrooms (avatar-based) when compared to VR2Gather (point-cloud). This might be because point-cloud representations provide people with all the details captured from the person itself, while avatar-based representations have smoother (less detailed) movement patterns and, in our case, no facial expressions that are often crucial for communication. MS Teams may have felt easier to interact with because it is a familiar 2D video format that preserves expressions without the abstraction of avatars. Interestingly, Macchi and De

Pisapia [25] found that overall interaction quality was higher in VR (Horizon Workrooms) than in video conferencing, which contrasts with our results, possibly because in their experimental setting the participant benefited from the virtual environment because it reduced distraction, while in our case, the lack of facial cues in avatars outweighed that advantage.

The engagement experienced in VR2Gather was slightly inferior compared to that in Horizon and MS Teams. A plausible explanation for this discrepancy may relate to the visual appearance of point clouds, which, despite their realistic appearance, occasionally manifested as fragmented, potentially serving as a distraction and disrupting the immersive experience. Previous research has shown that while point-cloud avatars score high in realism, technical flaws such as pixelation or missing eye contact can reduce user comfort [18], which may explain lower enjoyment and reduced ability to maintain focus. On the other hand, Horizon Workrooms has a consistent avatar-based representation, and MS Teams has a standard 2D view that participants are used to. In addition, Horizon's default office environment with a landscape background may have been perceived as more visually pleasant compared to the standard virtual office room used in VR2Gather. At the same time, engagement is a subjective construct, which can be constrained not only by personal preferences but also by cultural differences among participants, rather than just system-level limitations.

In summary, MS Teams is a widely used 2D interface, with which many people work. That is why it created fewer distractions than fully immersive environments such as VR2Gather and Horizon Workrooms, which are still novel to many users. The divergence in participants' ability to maintain focus in Horizon Workrooms was likely minimal owing to its advanced user interface architecture, which is purposefully engineered to emulate an office environment geared toward enhancing productivity.

6.2 Interpreting Objective Results

In terms of the total number of words guessed in both sessions, participants performed better in VR2Gather than in Horizon, and also better in MS Teams than in Horizon, while no significant differences were observed between VR2Gather and MS Teams. Given that gestures and facial expressions are crucial for recognizing words in a charades task, these results suggest that point-cloud representations provide users with more precise gesture depiction and overall representation than pre-made avatars. These results are consistent with previous findings that show that gestures are more easily recognizable in point-cloud systems compared to avatar-based [12]. Similarly, for the MS Teams vs. Horizon Workrooms comparison, the higher number of guessed words compared to Horizon may be caused by the limited details provided in the Horizon avatar representations.

To exclude the potential impact of the learning effect, we checked the difference in the number of words guessed between sessions for each of the three environments. The results showed that for the VR2Gather and MS Teams environments the number of guessed words increased in the second session, which might suggest that the results improved through practice, while for Horizon the difference was not significant. This indicates that Horizon Workrooms might have had some limitations in operation, such as the inability to track lower body movements or facial expressions and reliance on only 3DoF, which participants could not overcome through practice. These issues become clearer when looking at performance on individual words, which are discussed in the following paragraphs.

The words "Cat," "Dinosaur," "Ice Cream," "Lion," "Mustache," and "Wine" were more frequently guessed in VR2Gather and MS Teams than in Horizon Workrooms. This discrepancy is likely attributable to Horizon Workrooms' inability to display facial expressions, which are crucial for interpreting such words. Furthermore, in Horizon Workrooms, the tracking of hand movements near the face is obstructed if the hands approach too closely to the VR headset; consequently,

“Mustache” was correctly identified only 3 out of 10 attempts. In VR2Gather, despite the direct capture of the participant’s image, the VR headset likely inhibited the participants’ ability to mimic facial expressions effectively, which might explain the lower frequency of correctly guessed “Mustache” compared to MS Teams. Additionally, “Soccer” and “Stairs” were significantly less frequently guessed in Horizon Workrooms relative to the other two platforms, due to the absence of full-body tracking, leading to the omission of leg-related mimetic actions necessary for these terms.

In Horizon Workrooms, swift movements might have been affected by latency issues. For example, during mimicking words “Fishing” (characterized by the rapid winding of a fishing line), “Golf” (involving the swinging of a club and striking a ball), or “Hammering,” some hand representations exhibited lag, erratic motion, and failed to render these movements with sufficient smoothness to enable recognition by the observer. Consequently, there was a reduced frequency of correctly identifying these words compared to their performance in VR2Gather and MS Teams.

In reference to the word “Soap,” the primary issue was the insufficiency in the precision of hand movement depiction. Although Horizon Workrooms offered superior hand and finger tracking compared to other social interaction platforms assessed during the initial phase of the experiment, it nevertheless failed to accurately capture intricate movements, particularly when the hands and fingers overlapped in the real world. This occasionally resulted in delays or tracking inaccuracies, thereby hindering participants from effectively mimicking the gestures and accurately deducing the word.

The word “Toothbrush” was guessed with lower frequency in VR2Gather in contrast to Horizon Workrooms and MS Teams. This discrepancy is plausibly attributed to the avatar’s point-cloud representation, which experienced occlusions caused by the VR headset or the user’s hand, potentially complicating the guesser’s ability to discern the gesture. On the other hand, the words “Dinosaur,” “Gardening,” and “Scientist” were guessed more often in VR2Gather than in Horizon Workrooms or MS Teams. This may be because VR2Gather allowed participants to use full-body movements in 6DoF, while in Horizon they were limited to 3DoF without detailed facial or hand actions, and in MS Teams fine gestures may have been harder to interpret through a standard webcam.

The WPM metric indicates that participants operating within the MS Teams environment attained the highest guessing rate compared to the other two environments throughout both sessions. These findings imply that performance improved within the MS Teams platform, which could be attributed to several factors. Primarily, the aspect of familiarization deserves consideration: prior to beginning the experiment, 80% of the participants involved in the MS Teams settings expressed a degree of familiarity with the platform, while this familiarity was reported by 3.3% of the participants for Horizon Workrooms and 20% for VR2Gather. Additionally, the disparity in WPM between these platforms may be rationalized by the fact that 2D platforms such as MS Teams are devoid of the complexities inherent to immersive systems, such as the requirement to wear a cumbersome HMD for the entire session, the limitation in facial mimicry due to approximately 50% of the face being obscured by the HMD, and the issues of motion-tracking inaccuracies or latency, which can impair the perception of gestures.

In the context of Horizon Workrooms, some constraints were identified that may have contributed to diminished performance. These constraints include avatar latency related to inaccurately detected user height (for example, when the avatar is depicted as seated whereas the participant is standing in reality), inadequate representation of hand occlusions, and the lack of facial expressions, body tracking, and leg tracking. Although the design of the VR2Gather platform mitigates the majority of the limitations observed in Horizon Workrooms, imperfections in point-cloud representation persist. These imperfections may result in avatars appearing fragmented due to surface discontinuities,

thereby disrupting immersion, complicating the visual interpretation of gestures, and ultimately leading to decreased performance relative to MS Teams.

The results of this study indicate that MS Teams, as a two-dimensional conferencing platform, demonstrated superior task performance in contexts where the guesser's capability to decode body language, facial expressions, and hand gestures of the other participant was crucial. Concurrently, the study underscores the limitations inherent in immersive platforms, which must be addressed to equate their performance with the human performance benchmarks set by established two-dimensional systems.

6.3 Limitations

There are several limitations in this study that should be considered when interpreting the results. First, participants completed the experiment in pairs, which is why they might have had similar experiences, and their responses therefore cannot be considered fully independent. Second, the study uses between-subjects design, with each participant pair evaluating only one scenario. Therefore, results might be influenced by different prior experience, platform familiarity, guessing ability, language proficiency, and other personal factors, which come with the use of a between-subjects design. Finally, we want to acknowledge the demographic differences among participants who were assigned to evaluate the tested environments. The age of the participants for VR2Gather platform is on average higher than the age of participants in the other conditions, which may have introduced additional variability into the results.

7 Conclusions and Future Work

This work studied how user experience differs across immersive and non-immersive communication platforms, with a focus on social interaction and task performance. Although the results showed that workload, sickness, and presence remained statistically similar in the three environments, we found differences in the QoI and task performance. VR2Gather provides more detailed hand and gesture representation through point-cloud avatars, but shows lower results in audio-visual quality and enjoyment. Horizon Workrooms has smoother visuals due to avatar-based representation but lacks facial expressions, therefore requiring a higher communication effort, as reported by participants. MS Teams achieved the highest (objective) performance results, likely due to users' familiarity with the platform and full facial and body cues.

The findings highlight both potential and limitations of immersive systems compared to a well-established 2D conference platform for a simple communication task on the example of the charade game. Based on this finding, future work might focus on a broader range of quantitative data collection methods, using techniques such as eye-tracking, motion capture, and physiological signal monitoring, for a more comprehensive comparison between immersive and non-immersive communication platforms. Furthermore, future work can address participant-related limitations of this study, such as the number of participants or demographic differences between groups, as in our work older participants participated in the VR2Gather experiment compared to MS Teams and Horizon Workrooms experiments.

References

- [1] John G. Apostolopoulos, Philip A. Chou, Bruce Culbertson, Ton Kalker, Mitchell D. Trott, and Susie Wee. 2012. The road to immersive communication. *Proceedings of the IEEE* 100, 4 (2012), 974–990.
- [2] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. 2001. Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators and Virtual Environments* 10, 6 (2001), 583–598.

- [3] Stephan Beck, André Kunert, Alexander Kulik, and Bernd Froehlich. 2013. Immersive group-to-group telepresence. *IEEE Transactions on Visualization and Computer Graphics* 19 (2013), 616–625. Retrieved from <https://api.semanticscholar.org/CorpusID:1031304>
- [4] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, et al. 2013. *Qualinet White Paper on Definitions of Quality of Experience: Output from the Fifth Qualinet Meeting*. Novi Sad, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003).
- [5] Alex Cao, Keshav K. Chintamani, Abhilash K. Pandya, and R. Darin Ellis. 2009. NASA TLX: Software for assessing subjective mental workload. *Behavior Research Methods* 41 (2009), 113–117.
- [6] Eunhee Chang, Yongjae Lee, Mark Billinghurst, and Byoungyun Yoo. 2024. Efficient VR-AR communication method using virtual replicas in XR remote collaboration. *International Journal of Human-Computer Studies* 190 (2024), 103304.
- [7] Weiling Chen, Fengquan Lan, Hongan Wei, Tiesong Zhao, Wei Liu, and Yiwen Xu. 2024. A comprehensive review of quality of experience for emerging video services. *Signal Processing: Image Communication* 128 (2024), 117176.
- [8] Xiaoming Chen, Zhibo Chen, Ye Li, Tianyu He, Junhui Hou, Sen Liu, and Ying He. 2019. ImmerTai: Immersive motion learning in VR environments. *Journal of Visual Communication and Image Representation* 58 (2019), 416–427.
- [9] Carlos Cortés, Irene Viola, Jesús Gutiérrez, Jack Jansen, Shishir Subramanyam, Evangelos Alexiou, Pablo Pérez, Narciso García, and Pablo César. 2024. Delay threshold for social interaction in volumetric eXtended reality communication. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 7 (2024), 1–22.
- [10] Francesca De Simone, Jie Li, Henrique Galvan Debarba, Abdallah El Ali, Simon N. B. Gunkel, and Pablo Cesar. 2019. Watching videos together in social virtual reality: An experimental study on user's QoE. In *Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 890–891.
- [11] Sebastian J. Frisdon, Ben J. Congdon, David Swapp, Lisa Izzouzi, Klara Brandstätter, Daniel Archer, Otto Olkkonen, Felix Johannes Thiel, and Anthony Steed. 2021. Ubiq: A system to build flexible social virtual reality experiences. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, 1–11.
- [12] Guillaume Gamelin, Amine Chellali, Samia Cheikh, Aylen Ricca, Cedric Dumas, and Samir Otmene. 2021. Point-cloud avatars to improve spatial communication in immersive collaborative virtual environments. *Personal and Ubiquitous Computing* 25, 3 (2021), 467–484.
- [13] Felix Immohr, Gareth Rendle, Annika Neidhardt, Steve Göring, Rakesh Rao Ramachandra Rao, Stephanie Arevalo Arboleda, Bernd Froehlich, and Alexander Raake. 2023. Proof-of-concept study to evaluate the impact of spatial audio on social presence and user behavior in multi-modal VR communication. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences (IMX '23)*. ACM, New York, NY, 209–215. DOI: <https://doi.org/10.1145/3573381.3596458>
- [14] Andrew Irlitti, Mesut Latifoglu, Qiushi Zhou, Martin N. Reinoso, Thuong Hoang, Eduardo Velloso, and Frank Vetere. 2023. Volumetric mixed reality telepresence for real-time cross modality collaboration. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–14.
- [15] ITU-R BS.1284-2. 2019. Recommendation ITU-R BS.1284-2: General Methods for the Subjective Assessment of Sound Quality. Retrieved from <https://www.itu.int/rec/R-REC-BS.1284-2-201901-I>
- [16] ITU-T P.910. 2008. Recommendation ITU-T P.910: Subjective Video Quality Assessment Methods for Multimedia Applications. Retrieved from <https://www.itu.int/rec/T-REC-P.910-200804-S>
- [17] Jack Jansen, Thomas Röggl, Silvia Rossi, Irene Viola, and Pablo Cesar. 2024. Open-sourcing VR2Gather: A collaborative social VR system for adaptive multi-party real time communication. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. ACM, New York, NY, 11210–11213. DOI: <https://doi.org/10.1145/3664647.3685515>
- [18] Jonah-Noël Kaiser, Simon Kimmel, Eva Licht, Eric Landwehr, Fabian Hemmert, and Wilko Heuten. 2025. Get real with me: Effects of avatar realism on social presence and comfort in augmented reality remote collaboration and self-disclosure. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.
- [19] Robert S. Kennedy, Norman E. Lane, Kevin S. Berbaum, and Michael G. Lienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology* 3, 3 (1993), 203–220.
- [20] Hind Kharoub, Mohammed Lataifeh, and Naveed Ahmed. 2019. 3D user interface design and usability for immersive VR. *Applied Sciences* 9, 22 (2019), 4861.
- [21] Yong Min Kim, Ilsun Rhiu, Minshik Rhie, Hye Seon Choi, and Myung Hwan Yun. 2019. Current state of user experience evaluation in virtual reality: A systematic review from an ergonomic perspective. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63, Sage Publications, Los Angeles, CA, 1274–1275.
- [22] Sueyoon Lee, Irene Viola, Silvia Rossi, Zhirui Guo, Ignacio Reimat, Kinga Lawicka, Alina Striner, and Pablo Cesar. 2024. Designing and evaluating a VR lobby for a socially enriching remote opera watching experience. *IEEE Transactions on Visualization and Computer Graphics* 30, 5 (2024), 2055–2065.

- [23] Jie Li, Yiping Kong, Thomas Röggl, Francesca De Simone, Swamy Ananthanarayan, Huib de Ridder, Abdallah El Ali, and Pablo Cesar. 2019. Measuring and understanding photo sharing experiences in social virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.
- [24] Jie Li, Shishir Subramanyam, Jack Jansen, Yanni Mei, Ignacio Reimat, Kinga Ławicka, and Pablo Cesar. 2021. Evaluating the user experience of a photorealistic social VR movie. In *Proceedings of the 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 284–293. DOI: <https://doi.org/10.1109/ISMAR52148.2021.00044>
- [25] Gregorio Macchi and Nicola De Pisapia. 2024. Virtual reality, face-to-face, and 2D video conferencing differently impact fatigue, creativity, flow, and decision-making in workplace dynamics. *Scientific Reports* 14, 1 (2024), 10260.
- [26] Rufael Mekuria, Michele Sanna, Stefano Asioli, Ebroul Izquierdo, Dick C. A. Bulterman, and Pablo Cesar. 2013. A 3D tele-immersion system based on live captured mesh geometry. In *Proceedings of the 4th ACM Multimedia Systems Conference*, 24–35.
- [27] Christian Merz, Christopher Göttfert, Carolin Wienrich, and Marc Erich Latoschik. 2024. Universal access for social XR across devices: The impact of immersion on the experience in asymmetric virtual collaboration. In *Proceedings of the 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 859–869.
- [28] Christian Merz, Carolin Wienrich, and Marc Erich Latoschik. 2024. Does voice matter? The effect of verbal communication and asymmetry on the experience of collaborative social XR. In *Proceedings of the 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 1127–1136.
- [29] Mario Montagud Climent, Álvaro Egea Benavente, Marc Martos Cabré, Javier Montesa, Francisco Ibañez, and Sergi Fernández. 2025. Social eXtended reality (XR) and virtual production: Toward new engaging immersive experiences. In *Proceedings of the 2025 ACM International Conference on Interactive Media Experiences*, 458–461.
- [30] Anya Osborne, Sabrina Fielder, Joshua Mcveigh-Schultz, Timothy Lang, Max Kreminski, George Butler, Jialang Victor Li, Diana R. Sanchez, and Katherine Isbister. 2023. Being social in VR meetings: A landscape analysis of current tools. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 1789–1809.
- [31] Pablo Pérez, Ester Gonzalez-Sosa, Jesús Gutiérrez, and Narciso García. 2022. Emerging immersive communication systems: Overview, taxonomy, and good practices for QoE assessment. *Frontiers in Signal Processing* 2 (2022), 917684.
- [32] Andrew Perkiš, Christian Timmerer, Sabina Baraković, Jasmina Baraković Husić, Søren Bech, Sebastian Bosse, Jean Botev, Kjell Brunnström, Luis Cruz, Katrien De Moor, et al. 2020. QUALINET white paper on definitions of immersive media experience (IMEx). arXiv:2007.07032. Retrieved from <https://arxiv.org/abs/2007.07032>
- [33] Thammathip Piumsomboon, Youngho Lee, Gun Lee, and Mark Billinghurst. 2017. CoVAR: A collaborative virtual and augmented reality system for remote collaboration. In *Proceedings of the SIGGRAPH Asia 2017 Emerging Technologies*, 1–2.
- [34] Martin J. Prins, Simon N. B. Gunkel, Hans M. Stokking, and Omar A. Niamut. 2018. TogetherVR: A framework for photorealistic shared media experiences in 360-degree VR. *SMPTE Motion Imaging Journal* 127, 7 (2018), 39–44.
- [35] Hironori Shimosato, Nozomi Hayashida, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. 2022. XR communication system for remote control wheelchairs. In *Proceedings of the 12th International Conference on the Internet of Things*, 219–223.
- [36] Ashutosh Singla, Steve Göring, Dominik Keller, Rakesh Rao Ramachandra Rao, Stephan Fremerey, and Alexander Raake. 2021. Assessment of the simulator sickness questionnaire for omnidirectional videos. In *Proceedings of the 2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 198–206.
- [37] Ashutosh Singla, Irene Viola, Jack Jansen, and Pablo Cesar. 2025. QoE evaluation of remote physiotherapy in volumetric video and video-based real-time communication. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*.
- [38] Harrison Jesse Smith and Michael Neff. 2018. Communication behavior in embodied virtual reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, 1–12. DOI: <https://doi.org/10.1145/3173574.3173863>
- [39] Frank Steinicke, Nale Lehmann-Willenbrock, and Annika Luisa Meinecke. 2020. A first pilot study to compare virtual group meetings using video conferences and (immersive) virtual reality. In *Proceedings of the 2020 ACM Symposium on Spatial User Interaction*, 1–2.
- [40] Shishir Subramanyam, Irene Viola, Jack Jansen, Evangelos Alexiou, Alan Hanjalic, and Pablo Cesar. 2022. Evaluating the impact of tiled user-adaptive real-time point cloud streaming on VR remote communication. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3094–3103.
- [41] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, and Pablo Cesar. 2023. VR2Gather: A collaborative, social virtual reality system for adaptive, multiparty real-time communication. *IEEE MultiMedia* 30, 2 (2023), 48–59. DOI: <https://doi.org/10.1109/MMUL.2023.3263943>

- [42] Bob G. Witmer and Michael J. Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments* 7 (1998), 225–240. Retrieved from <https://api.semanticscholar.org/CorpusID:2825590>
- [43] Jason W. Woodworth, Sam Ekong, and Christoph W. Borst. 2017. Virtual field trips with networked depth-camera-based teacher, heterogeneous displays, and example energy center application. In *Proceedings of the IEEE Virtual Reality (VR)*, 471–472. DOI: <https://doi.org/10.1109/VR.2017.7892384>
- [44] Nikolaos Zioulis, Dimitrios Alexiadis, Alexandros Doumanoglou, Georgios Louizis, Konstantinos Apostolakis, Dimitrios Zarpalas, and Petros Daras. 2016. 3D tele-immersion platform for interactive immersive experiences between remote users. In *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, 365–369. DOI: <https://doi.org/10.1109/ICIP.2016.7532380>

Appendix

A.1 List of Words

The word list used in the study is presented in Table A1. Please note that the order of the sessions was randomized during the experiment.

Table A1. Charades Word List Used in the Study, Separated by Session

Session 1	Session 2
Archery	Airplane
Banana	Basketball
Bird	Bracelet
Book	Cloud
Cat	Crocodile
Compass	Drums
Cooking	Elephant
Cycling	Fishing
Dinosaur	Fork
Gardening	Golf
Ice Cream	Hammer
Milk	Helicopter
Microwave	Juggling
Pancake	Lion
Photographer	Magician
Piano	Mustache
Rainbow	Painting
Refrigerator	Pajamas
Robot	Popcorn
Rollercoaster	Rock climbing
Ship	Scientist
Skateboarding	Soap
Spaghetti/Noodles	Soccer
Superhero	Stairs
Swimming	Starfish
Toothbrush	Telescope
Video	Umbrella
Volcano	Window
Wrestling	Wine
Yo-yo	Yoga

Received 12 November 2025; revised 22 June 2026; accepted 26 July 2026