Delay Threshold for Social Interaction in Volumetric eXtended Reality Communication

Cortés, Carlos; Viola, Irene; Gutiérrez, Jesús; Jansen, Jack; Subramanyam, Shishir; Alexiou, Evangelos; Pérez, Pablo; García, Narciso; César, Pablo

DOI
10.1145/3651164

Publication date
2024

Document Version
Final published version

Published in
ACM Transactions on Multimedia Computing, Communications and Applications

Delay Threshold for Social Interaction in Volumetric eXtended Reality Communication

CARLOS CORTÉS, Universidad Politécnica de Madrid, Madrid, Spain
IRENE VIOLA, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
JESÚS GUTIÉRREZ, Universidad Politécnica de Madrid, Madrid, Spain
JACK JANSEN, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
SHISHIR SUBRAMANYAM, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
EVANGELOS ALEXIOU, Netherlands Organisation for Applied Scientific Research, Den Haag, The Netherlands
PABLO PÉREZ, eXtended Reality Labs, Madrid, Spain
NARCISO GARCÍA, Universidad Politécnica de Madrid, Madrid, Spain
PABLO CÉSAR, Centrum Wiskunde & Informatica and the Delft University of Technology, Amsterdam, The Netherlands

Immersive technologies like eXtended Reality (XR) are the next step in videoconferencing. In this context, understanding the effect of delay on communication is crucial. This article presents the first study on the impact of delay on collaborative tasks using a realistic Social XR system. Specifically, we design an experiment and evaluate the impact of end-to-end delays of 300, 600, 900, 1,200, and 1,500 ms on the execution of a standardized task involving the collaboration of two remote users that meet in a virtual space and construct block-based shapes. To measure the impact of the delay in this communication scenario, objective and subjective data were collected. As objective data, we measured the time required to execute the tasks and computed conversational characteristics by analyzing the recorded audio signals. As subjective data, a questionnaire was prepared and completed by every user to evaluate different factors such as overall quality, perception of delay, annoyance using the system, level of presence, cybersickness, and other subjective factors associated with social interaction. The results show a clear influence of the delay on the perceived quality and a significant negative effect as the delay increases. Specifically, the results indicate that the acceptable threshold for end-to-end delay should not exceed 900 ms. This article additionally provides guidelines for developing standardized XR tasks for assessing interaction in Social XR environments.

This work was partially supported by projects HORIZON-IA-101070250 (XRECO) and HORIZON-IA-101070109 (TRANSMIXR) funded by the European Union, by project PID2020-115132RB (SARAOS) funded by MCIN/AEI/10.13039/501100011033 of the Spanish Government, and by projects UNICO-5G I+D TSI063000-2021-79 (B5GEMiNI-AIUC) and RED.ES 2021/C005/00144164 (COMODIA) funded by the Ministry of Digital Transformation of the Spanish Government and the NextGenerationEU (Recovery, Transformation and Resilience Plan—PRTR).
Authors' addresses: C. Cortés, Universidad Politécnica de Madrid, Madrid, 28040, Spain; e-mail: carlos.cs@upm.es; I. Viola, J. Jansen, and S. Subramanyam, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands; e-mails: irene.viola@cwi.nl, jack.jansen@cwi.nl, s.subramanyam@cwi.nl; J. Gutiérrez and N. García, Universidad Politécnica de Madrid, Madrid, 28040, Spain; e-mails: jesus.gutierrez@upm.es, narciso.garcia@upm.es; E. Alexiou, Xiaomi Communications Co Ltd, The Hague, NL 2595 AM, The Netherlands, e-mail: alexiou@xiaomi.com; P. Pérez, eXtended Reality Labs, Madrid, 28050, Spain, e-mail: pablo.perez@nokia.com; P. César, Centrum Wiskunde & Informatica and the Delft University of Technology, Amsterdam, The Netherlands; e-mail: p.s.cesar@cwi.nl.

ACM 1551-6857/2024/04-ART206
https://doi.org/10.1145/3651164

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

Additional Key Words and Phrases: eXtended reality, volumetric Social XR, delay

ACM Reference Format:
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 Trans. Multimedia Comput. Commun. Appl.* 20, 7, Article 206 (April 2024), 22 pages. https://doi.org/10.1145/3651164

1 INTRODUCTION

The use of immersive technologies has aroused interest in several telecommunications-based applications, such as industrial training [42, 54], telecare [10], and telemeetings [48]. However, 2D videoconferencing is still the most widely used technology for teleconferences, although it presents certain drawbacks that affect the user experience. According to Skowronek et al. [48], prolonged videoconferencing can strain human interaction factors in telemeetings, causing fatigue and increased cognitive load due to the unnatural communication, reduced mobility, and the added effort of non-verbal communication (known as videoconferencing fatigue). Therefore, 2D videoconferencing presents inherent limitations due to its 2D visual representation and the lack of user free movement.

To overcome the limitations of 2D videoconferencing, Social **eXtended Reality (XR)** has emerged as a promising solution by offering a more natural and immersive communication. This is because of the inherent 3D nature of XR technology, which allows users to freely move around and interact with each other in a way that is more realistic and engaging than ever before [20, 24, 31]. In addition, under the XR paradigm, local and distant physical realities can be blended with virtual assets to offer realistic interactions in 6 **Degrees of Freedom (DoF)** that enhance the user experience. Within the possibilities offered by this paradigm, Social XR communications are called to be the next step in immersive communications [24, 31, 48].

However, despite the increasing popularity of XR communications, the effects of system factors on user experience and performance have not been widely studied yet, with delay being among the most important. On the contrary, the influence of delay in 2D videoconference is a well-studied field [2, 6, 11, 45]. Previous studies show that delay has different ways of affecting users. On the one hand, desynchronization and echo cause severe damage to the perceived quality of users with respect to the system. On the other hand, by mitigating these effects and making the delay synchronous, users are able to withstand higher delays [45]. This is the most common and studied aspect of delays in videoconferencing.

In earlier studies, the influence of delay on the adoption of videoconferencing technology has been examined through subjective experiments [6–8, 43, 46, 50]. Together with objective metrics, these experiments have identified acceptable delay thresholds for videoconferencing [35, 37, 41]. The recommended delay threshold for avoiding user annoyance is below 600 ms [37], but recent studies have suggested higher values, exceeding 900 ms [6, 43]. While these values apply to 2D videoconferencing, they may not be applicable to richer Social XR communication scenarios. However, to the best of our knowledge, there are still no similar studies to establish the limits of delay for videoconferencing in Social XR. Moreover, there is still no established methodology for the evaluation of interactive videoconferencing in Social XR.

This article addresses the challenge of determining new appropriate delay limits to guarantee the user's acceptance in collaborative Social XR. For this purpose, a subjective experiment was

conducted with remote users communicating verbally and visually using photorealistic 3D representations [25] within a shared virtual environment, under different delay conditions. Moreover, we present a new methodology for evaluation of interactive videoconferences in XR adapted from the standard for evaluation in 2D videoconferences. Our results show an impact of the delay on the user experience and conversation flow above 900 ms. These values are related to previous studies on video-based conferences that pointed to delay acceptance values above 600 ms [6, 43]. Therefore, this study contributes to the following:

— Set an acceptance limit at 900-ms end-to-end delay for Social XR.
— Provide a new evaluation protocol for interactive teleconferencing in Social XR.

2 RELATED WORK

The objective of this study is to evaluate the impact of interaction delay in immersive teleconferencing environments using a photorealistic Social XR system. Delay can be defined as the elapsed time between the transmission of a signal and its reception at the destination. In the context of videoconferencing, end-to-end delay refers to the delay between the movement of a user and the moment when the remote user sees that movement. Delay can have a detrimental effect on the communication process, leading to a decrease in the quality of interaction [41].

Audiovisual communication systems, including videoconferencing and streaming services, are highly reliant on user experience in terms of system acceptance [48, 49]. Besides, one of the key factors that can impact user acceptance is delay [1, 2, 11]. In particular, the delay is crucial for real-time applications such as videoconferencing.

This section analyses the delays in other non-immersive environments to provide an overview of the recommended values for more classical communications. In addition, the current state of immersive communications in Social XR is described along with examples of systems, and the current methodologies to assess the influence of technical factors on user acceptance.

2.1 Delay on 2D Communications

The conventional approach to videoconferencing involves the use of at least, a display, a camera, a microphone, and an audio playback device for each participant. The transmission of audiovisual data may cause delays that affect the videoconference. Research on the acceptance of videoconferencing systems establishes that the threshold for synchronization between video and audio signals can vary between +90 ms and −185 ms on average, respectively [32]. Although synchronization issues may be resolved through the use of synchronizers, reducing the overall end-to-end delay within communication systems is not a straightforward task.

The perception of delay, as well as its effects on interaction, is a field of study within the areas of user experience, conversation, and interaction [7, 14, 37, 47]. With respect to the factors used to evaluate user experience, we found analyses of both subjective factors through questionnaires and objective data [37]. The relevant data gathered by the standards include the perceived **Quality of Experience (QoE)**, the annoyance using the system, the perception of the interruptions, and whether users notice the delay [37, 41]. According to the evaluation of these factors, the overall delay tolerance for maintaining an acceptable experience is said to be under 600 ms [37]. Furthermore, recent studies set user's acceptable delay in higher values (above 900 ms) [44, 46]. From another point of view, some studies have analyzed the impact of delay in video-mediated interactions by assessing the impact on how conversations flow [15, 45–47]. In this sense, video delay in video-mediated interaction has significant implications for communication and user experience. High delay can result in a disjointed and unnatural conversation, with participants experiencing delays between their actions and corresponding responses. This delay can hinder the flow of

conversation, disrupt the natural turn-taking process, and negatively influence non-verbal cues and gestures [47]. Nevertheless, the results of these studies point out that the users somehow adapt to them and attribute these technical difficulties to the poor fluency of the other users [45].

However, these results are intended for 2D videoconferencing. Due to the higher DoF and different imaging modalities that are being used in XR communications, the thresholds in delay may be different; therefore, these delay thresholds need to be revised. For this purpose, in this article, we conducted a comprehensive study of latencies in a Social XR environment.

2.2 Tasks for Evaluating XR Systems

The various protocols established to assess the impact of the system on interaction in videoconferencing environments are varied and are reflected in different international recommendations [36–38, 41]. These protocols involve the performance of a task with more than one user. While such protocols are well established and standardized, there is a lack of protocols for the evaluation of Social XR systems. In the literature, several evaluation studies of Social XR systems can be found that include tasks such as watching a movie [25], collaborating to achieve a common pattern [22], and playing a game [17].

In this work, we have replicated a collaborative block-building task described in a standard recommendation on interactive test methods for 2D audiovisual communications [41]. To adapt the task as faithfully as possible, we present a Social XR environment that mimics the building block task using photorealistic representations of the users and the figures. Additionally, for the experimental design we have followed the recommendation for immersive video evaluation ITU-T P.919 [19]. All of these contributions together form a new protocol designed according to different international standards for interaction evaluation in Social XR.

2.3 XR Communications System

Social XR refers to a paradigm where individuals can interact with each other and their surroundings through the use of XR technologies. Therefore, Social XR systems enable remote and synchronous communication, providing an immersive experience that goes beyond 2D videoconferencing [31].

The main difference between Social XR and 2D videoconferencing is the DoF for user exploration and interaction [48]. DoF signifies how freely a user can view different angles of media content. The level of DoF in Social XR systems ranges from 3 DoF, which involves head movements (pitch, yaw, and roll) to 6 DoF, including translational coordinates (x, y, z). Therefore, Social XR should allow video viewing from different points of view.

In the literature, we can find different Social XR systems with different DoF capabilities. For example, Kachach et al. [21] present a virtual environment where users can interact with a distant environment in 3 DoF using a 360-degree camera. Another example is the work of Becher et al. [4], which presents an environment with purely virtual avatars where users interact with 6 DoF using their voice and controllers. However, this 6-DoF environment does not use video for user representation. Finally, Viola et al. [52] present a 6-DoF Social XR system using volumetric video through a set of color and depth coordinated cameras. Therefore, volumetric video is a promising approach for Social XR because it enables users to see each other in photorealistic detail from multiple perspectives.

Volumetric video is an emerging technology that further enhances the user experience in XR environments. Unlike 2D video formats, which offer fixed viewpoints, volumetric video enables users to see each other from various perspectives within the virtual space. This means that users can explore and interact with one another from different angles, providing a more natural and engaging way to communicate in virtual environments. This capability adds an extra layer of

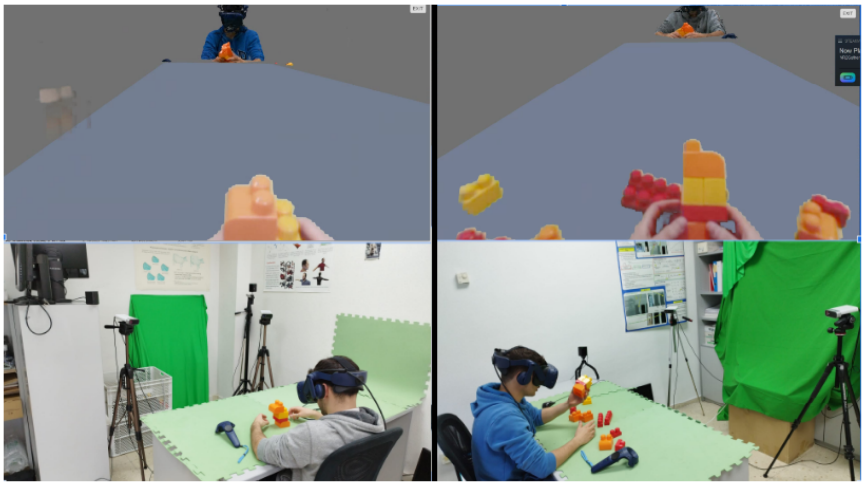


Fig. 1. Two users sitting in two different physical rooms and meeting in the same Social XR environment during the experience.

realism and interactivity to XR experiences, making them feel even more like face-to-face interactions [31, 48]. With respect to volumetric video, we can find two representation techniques. One approach is mesh-based techniques. These techniques generate a set of dependent triangles that are positioned and colored according to the information received by the depth and color cameras. Some examples of mesh-based volumetric videoconferencing systems can be found in other works [5, 27, 55]. Although these techniques have been shown to provide good performance under loose grid conditions, the triangle generation process requires complex processing that can affect system delay [51].

Another approach to represent volumetric video is point clouds. Point clouds are generated by giving an independent volume in space to each color and depth pixel set provided by the cameras. The fact that they are independent and derive directly from the camera streams makes their implementation for real-time systems more suitable [51]. In addition to the real-time requirement, the use case for videoconferencing in Social XR requires systems that are adapted to immersive technologies. Some state-of-the-art systems that use volumetric video in Social XR are Free Viewpoint Video Live [9], Holoportation in Microsoft Mesh [29], and VR2Gather [52].

In this work, the VR2Gather Social XR system [52] has been selected because it is a point cloud based volumetric videoconferencing system prepared for immersive environments. Moreover, it allows symmetric communication in terms of visualization between users. In other words, users see themselves and others in a reciprocal manner (Figure 1). Another decisive factor was that it is open source [52], allowing modifications to be made to introduce artificial latencies. In addition, it allows the replicability of the experiment allowing the protocol described in this article to be included as part of the tasks of a forthcoming recommendation for the evaluation of volumetric Social XR systems.

3 SOCIAL XR VIDEOCONFERENCE ENVIRONMENT

The objective of the system is to enable interactive videoconferencing using immersive technology. To achieve this, different modules are linked together, allowing users to see themselves in an XR environment where they can manipulate objects from their physical reality. Additionally, the system needs to be able to represent and display the remote user in the shared environment.

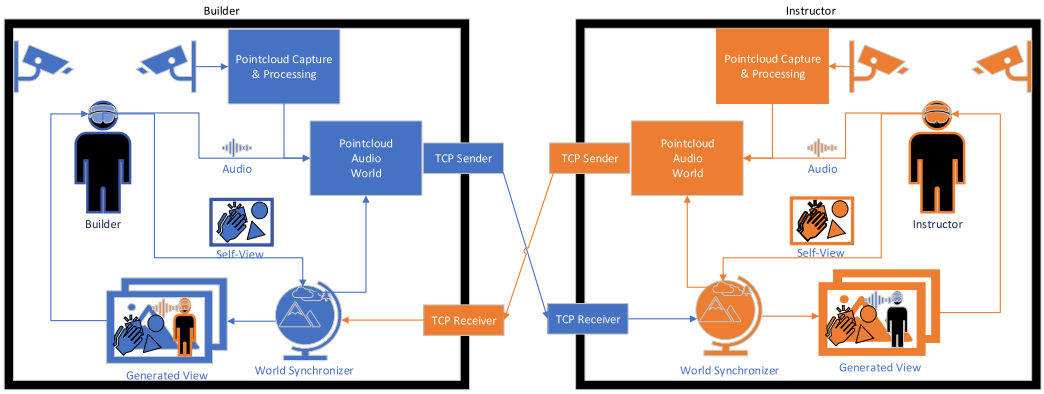


Fig. 2. Diagram of volumetric XR communications.

Therefore, the system must capture aspects of two physical realities, namely where the two remote users are located, and position all of that information in a Social XR environment. As an illustration, Figure 1 shows two users placed in two different physical rooms (bottom), with each user wearing a **Head-Mounted Display (HMD)**, and corresponding snapshots of the views generated from their HMDs (top). In this figure, it can be seen that both users are immersed in a virtual world with a virtual table that mimics the physical one while hands and physical blocks are visible. Additionally, the volumetric representation of the remote user is visible at the end of the virtual table.

3.1 Social XR System

The different elements that make up the Social XR system are defined here. The two roles related to the collaborative task, namely the instructor and the builder, are presented in Figure 2. Furthermore, each color (blue and orange) represents the flow of information from each role. The black border boxes represent the elements contained in each physical reality. In other words, the physical room where each user is located. In this study, we use a room with a table (see Figure 1). In each black frame of Figure 2, it can be seen a user wearing an HMD being captured by surrounding cameras. The cameras surrounding the users capture color and depth information from the physical reality to generate a point cloud representation. Besides, the HMD generates two types of information. It captures the user's voice with the built-in microphone and, through the integrated camera, captures the physical reality from an egocentric perspective (self-view). The audio and the point cloud are combined with information about the world and then encoded and transmitted to the remote user via TCP transmission protocol. It is at this point that the remote user integrates this information into their virtual world to generate the view of the Social XR environment that will be reproduced by their HMD.

According to the diagram described previously, there are two information loops in the system: one for the generation of the self-view and another for the generation of the volumetric avatar (point cloud, audio or voice, and world position).

For the generation of self-view, the XR environment should represent the physical environment that usually includes the user's body and real objects. In our case, we capture the physical environment using egocentric cameras that are attached to the HMD, and by using image segmentation algorithms to crop the image, only the body of the user and some real objects are included within the Social XR environment (Figure 3)

Fig. 3. Local environment self-view without distant user.

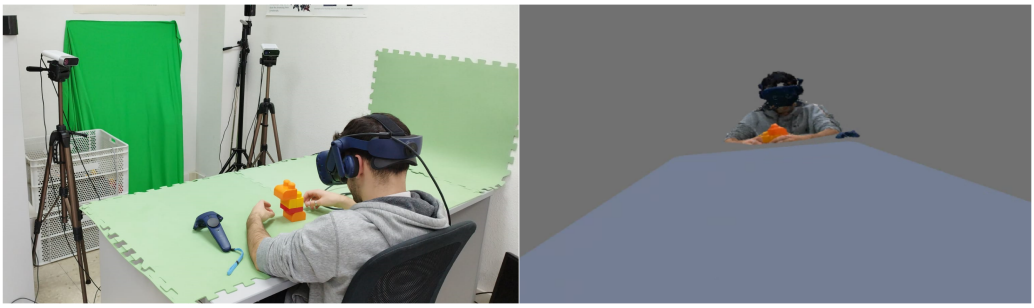


Fig. 4. Physical environment of the instructor and the generated viewport of the builder in the Social XR environment.

Table 1. Summary of the Different Delay Components

Delay Component	Meaning
τ_{cap}	Time elapsed between the user's motion and the time it takes for the camera to capture it
τ_{proc}	Processing time of the camera frames (including avatar segmentation and composition)
τ_{disp}	Time from the end of processing until the user can see the processing result on the display.
τ_{tx}	Transmission time between the environments of each user
τ_{sync}	Synchronization time of audio, pointcloud, and virtual environment streams

For the generation of the user volumetric avatar, the system includes an acquisition setup that uses multiple cameras with depth sensors to capture volumetric data of the user from different angles [25]. In addition, the voice is captured by the HMD's built-in microphone. The captured data is then processed, transmitted, and integrated into the shared environment (Figure 4). An analysis of the different processes that contribute to the end-to-end delay is presented in the next subsection.

3.2 System Delay

The system has numerous sequential processes, each of which can add an intermediate delay that will affect the total end-to-end delay. Table 1 summarizes the different components that consist of delays related to capturing, processing, display, transmission, and synchronization.

In the XR communication system, there are two different information loops that are sensitive to delay. The first one is the self-view. The Social XR system uses the egocentric camera for capturing the physical environment; then, this image is processed to include only the user's hands and some objects of the physical environment (see Figure 3). After that, the result is rendered in the virtual world and displayed in the HMD. In Figure 4, this loop is illustrated in the self-view element that traverses through the world synchronizer to add the hands and some real objects into the generated view. Therefore, the elements that contribute to the composition of the self-view delay are

$$\textit{self-view delay} = \tau_{cap} + \tau_{proc} + \tau_{disp}. \tag{1}$$

In Equation (1), the τ_{cap} stands for the time the HMD camera frames are available in the processor memory. The τ_{proc} includes the transformation of the camera to adapt to virtual reality and the segmentation process. The τ_{disp} stands for the time that the XR engine takes to show the result of the processing in the HMD.

To generate the user representation, the process is more elaborated. First, a set of color and depth cameras should be placed around the user to cover its volume. Then, the captured information of each camera is processed with a common reference in real space (calibration). With this information, the system generates a point cloud representation of the user. Then, the point cloud is coded and transmitted to the remote user together with the microphone audio and the world information through a TCP connection. Then, the remote user server should receive, synchronize the audio and video, and render the point cloud into the remote user XR environment according to the world information. Therefore, the elements that contribute to the composition of the Social XR delay are

$$\textit{XR delay} = \tau_{cap} + \tau_{proc} + \tau_{tx} + \tau_{sync} + \tau_{disp}. \tag{2}$$

In Equation (2), the τ_{cap} stands for the time the HMD camera frames are available in the processor memory. The τ_{pro} includes the transformation of the point cloud generation. The τ_{tx} stands for the transmission time of the volumetric avatar. The τ_{sync} stands for the time of world synchronization—that is, audio and video synchronization plus world positioning. Finally, the τ_{disp} stands for the time XR engine takes to show the result of the processing in the HMD.

Although the local user client and remote user server capturing and display delays can be determined and stabilized, the transmission and processing delays are subject to network variables and computer capabilities. As a result, these delays can have an unexpected impact on the user experience. In the experiment, the delay under consideration represents the duration between the local camera capture and their rendering on the remote display.

4 EXPERIMENTAL DESIGN

The aim of this study is to assess the impact of interaction delay on immersive teleconferencing environments for Social XR, by utilizing photorealistic user representations. To accurately evaluate the effects of delay, a task was selected from the standard for interaction assessment in videoconferencing: the ITU-T P.920 [41]. This task involves collaborating to construct block-based figures, with one participant designated as the instructor and the other as the builder. The objective is for the instructor to guide the builder to reproduce the complete figure. Communication and interaction take place through both audio and visual channels, as the teleconferencing environment is audiovisual in nature. However, the task was originally intended for 2D videoconference using a basic camera and a 2D monitor, and thus modifications were necessary to adapt it to the immersive environment. Specifically, egocentric capture with chroma-based physical environment segmentation was employed to represent the local environment, whereas multi-camera-based volumetric capture was used to represent distant users. These adaptations are illustrated in Figure 1.

The Social XR system under consideration encompasses two distinct delays: the self-view delay and the XR delay. An assessment of the impact of the self-view delay on the block-building task's performance was conducted on a previous study [12], using an identical system configuration. To eliminate the effect of additional parameters, in this experiment, there was no remote user involved (typically responsible for providing instructions on the building process), but we incorporated a pre-reconstructed 3D figure into the setup that was serving as a reference. The study determined the minimum latency of the system self-view to be 190 ms. Moreover, we tested the user's experience under different self-view delays of up to 587 ms that were artificially introduced. Our results showed that for delays lower than 338 ms, the user experience was unaffected. As a result, it is concluded that the self-view delay introduced by the system (190 ms) yields very good results in terms of user experience and does not influence the Social XR study presented in the current article.

This section introduces the methodology employed in the current study. The research involved the adaptation of the standardized ITU-T P.920 task, which entailed the collaborative construction of block-based figures within the Social XR environment. A description of the software utilized for synchronizing the virtual environments of two users and artificially manipulating delays is provided. Furthermore, the hardware configuration for each room, signifying distinct task roles, is expounded upon. Moreover, the process of experimental design, encompassing task adaptation, administration of subjective quality questionnaires, and collection of objective data during experimental sessions are outlined. Finally, it should be mentioned that the experimental process was refined based on pilot studies that were conducted with a limited participant pool, which are briefly reported.

4.1 Hardware

The experimental hardware utilized in this study encompassed a range of functionalities, namely local reality capture, point cloud capture and transmission, synchronization, and Social XR environment display, allocated per user. Local reality capture and environment display were achieved through the use of the HMD HTC Vive Pro, whereas point cloud capture and generation were facilitated by using the CWIPC system [25], utilizing the Kinect Azure color and depth cameras. The synchronization of social worlds was managed by VR2Gather software [25], installed on Windows 10 PCs with an Intel Core i7-4790 with a clock speed of 3.6 GHz, boasting eight cores, alongside an NVIDIA TITAN Xp GPU.

4.2 Software

The predominant software used was VR2Gather, a socially immersive software platform designed by the Centrum Wiskunde & Informatica (CWI) using the Unity engine, which enables audiovisual communication in XR settings. To assess diverse delay circumstances, a software component was adapted that was tasked with synchronizing the audio and video components of an avatar—that is, the synchronizer. The synchronizer is responsible for matching the audio and volumetric video received by each user. In addition, it has the option of storing this information so that the total delay is controlled (taking into account the time it took to receive the audio and video from its capture). Therefore, the synchronizer makes the experiment possible, allowing the delay to be artificially varied. Additionally, we use OBS [28] software to capture the audio of the conversations. This software was configured to capture the microphone and headphones integrated into the HMD. Each of these sources was stored in a channel of an audio file to facilitate further analysis. The MIRO360 [13] application was used to conduct the questionnaires within the virtual environment.

Table 2. Questionnaire Used in the Experiment

Category	Factor	Question	Reference
Subjective performance	Global QoE	How would you rate the quality of the experience globally?	[41]
	System Annoyance	How easy did you find it to communicate using the system?	[37]
	Delay Perception	Did you perceive any reduction in your ability to interact during the conversation due to delay?	[37]
	Interruptions	How would you judge the effort needed to interrupt the other party?	[37]
Presence	Involvement	How much did your experiences in the virtual environment seem consistent with your real-world ones?	[12, 30]
	Adaption	How well could you concentrate on the assigned tasks rather than on the mechanisms used to perform them?	[12, 30]
	Accomplishment	I am confident that we completed the task correctly.	[12, 30]
Social Factors	Social Presence	I felt connected with my partner.	[18, 53]
	Social Annoyance	I was able to understand my partner's message.	[18, 53]
	Social Adaptation	My partner and I worked together well.	[18, 53]
	Collaboration	Information from my partner was helpful.	[18, 53]

4.3 Objective Data

During the experiment, objective data was captured to analyze the impact of delay on user performance. The time required by each pair of users to complete the task was recorded using a data log from Unity. Furthermore, the audio of the conversations was captured to identify the number of interventions and the activity time of each user.

4.4 Questionnaire

To evaluate the influence of interaction delay, a combination of objective and subjective measures was employed. Subjective quality questionnaires were selected based on their previous use in assessing interaction quality. Table 2 presents the subjective factors evaluated in conjunction with their respective questions. Subjective factors analysis included global quality, system annoyance, delay perception, and interruption perception, derived from international standards and specifically aimed at assessing the impact of delay on system acceptance [34, 37, 41]. Additionally, to evaluate the effect of delay on the perception of interaction with the local environment, a validated questionnaire for this type of environment was used [30]. This questionnaire was also used for the self-view delay experiment [12]. To further examine the impact on subjective social quality, questions from the work of Gupta et al. [18] used in an experiment with a similar task [53] were included to assess subjective social factors.

4.5 Experimental Conditions

The experimental conditions comprised the pairing of delay values and block-based figures. A pilot test was conducted to select the different delay conditions, by which a proposal of figures and delays was presented. The delay intervals were anchored at 300 ms, which was deemed to be the base. To evaluate the effectiveness of the proposed experimental conditions, a pilot test was conducted with 10 participants who evaluated the system using four figures with four different delays. The pilot test established that quality degradation ranged from 600 to 1,000 ms and that the degradation was more significant for the builder role. Additionally, the feedback from the participants suggested that the figures were relatively complex. Consequently, for the actual experiment, the number of latencies surrounding 600 and 1,000 was increased by reducing the number of blocks for each figure. The following delay values were selected: 300 ms (minimum), 600 ms, 900 ms,

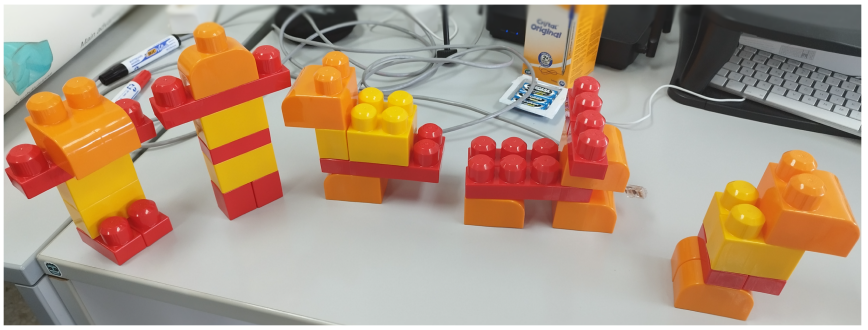


Fig. 5. Selected block based figures, from right to left: *Mazinger*, *Rocket*, *Bird*, *Dog*, and *TRex*.

1,200 ms, and 1,500 ms. In addition, the selected block-based figures are shown in Figure 5. Each figure is composed of seven blocks.

An essential consideration when establishing experimental conditions is randomization and balancing [33]. To ensure that conditions were balanced, the Graeco-Latin distribution was used to organize the delay and figure conditions [23]. In this way, we ensured that the same number of pairs of conditions existed for each possible combination. In addition, the order of the conditions were randomized.

4.6 Experiment Workflow

The experimental procedure involves several sequential steps. First, the participants are informed about the collaborative task and instructed to disregard any visual effects arising from egocentric capture and volumetric avatars. Subsequently, the roles of instructor and builder are assigned to the participants and they are located in separate rooms. Participants are informed of a training session during which they can familiarize themselves with the system. In the training session, users must complete two buildings under the best (300 ms) and worst (1,500 ms) delay conditions. This methodology is in line with the conventional practices in subjective experiments [33, 40]. A 10-minute break follows the training session before the start of the actual experiment. The experiment consists of a repetition of five tasks with different delay conditions and figures. Figure 6 shows a flow diagram of the experiment. Each "task" involves the collaborative process between an instructor and a builder, utilizing an immersive videoconferencing system to construct a figure. At the start of each task, the instructor begins with a perfectly constructed figure, whereas the builder starts with a set of loose parts. The users then collaborate to enable the builder to replicate the figure held by the instructor. Once the users determine they have completed the task, the experimenter initiates a virtual environment where the users can respond to the questionnaire outlined in Table 2. After both users complete their questionnaires, they wait in an empty environment for the experimenter to disassemble the builder's constructed piece and replace the instructor's reference figure, preparing for the next iteration.

4.7 Participants

We conducted an experiment with 60 subjects (29 female and 31 male; ages between 20 and 33 years, mean: 22.8, standard deviation: 2.1). None of them were experts in the use of virtual reality. All users reported no vision problems in terms of color perception, and the HMD was adjusted in the training phase to assure the best visual experience.

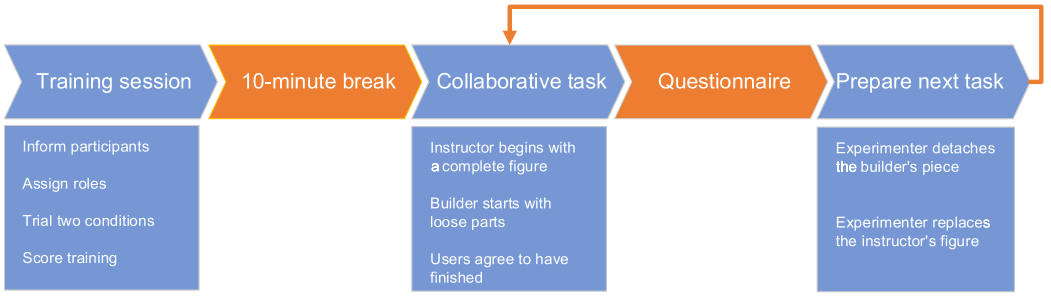


Fig. 6. Experiment workflow diagram.

Table 3. Subjective Performance Analysis

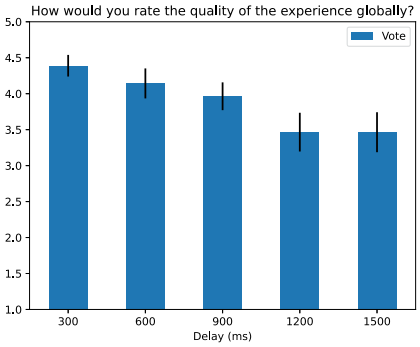
Factor	Variable		ANOVA			Significantly Different
			F	p	η^2	
Global QoE	Role		$F_{1,230} = 2.781$	0.097	0.008	
	Delay		$F_{4,230} = 12.484$	**<0.001**	**0.152**	(≤900) vs (≥1200)
	Figure		$F_{4,230} = 2.759$	**0.029**	0.034	(Bird) vs (Mazinger)
System Annoyance	Role		$F_{1,230} = 2.207$	0.169	0.007	
	Delay		$F_{4,230} = 10.890$	**<0.001**	**0.136**	(≤600) vs (≥1200) (900) vs (1500)
	Figure		$F_{4,230} = 1.626$	0.139	0.020	
Delay Perception	Role		$F_{1,230} = 9.957$	**0.002**	0.026	–
	Delay	Builder	$F_{4,115} = 4.548$	**0.002**	0.118	(≤600) vs (1500)
		Instructor	$F_{4,115} = 5.744$	**<0.001**	**0.300**	(≤900) vs (≥1200)
	Figure	Builder	$F_{4,115} = 1.452$	0.222	0.038	–
		Instructor	$F_{4,115} = 1.442$	**0.001**	0.064	(Bird) vs (Mazinger)
Interruptions	Role		$F_{1,230} = 7.067$	**0.008**	0.020	–
	Delay	Builder	$F_{4,115} = 7.155$	**< 0.001**	**0.167**	(≤900) vs (≥1200)
		Instructor	$F_{4,115} = 15.528$	**< 0.001**	**0.200**	(≤900) vs (≥1200)
	Figure	Builder	$F_{1,115} = 1.388$	0.242	0.032	
		Instructor	$F_{4,115} = 3.319$	0.083	0.046	

5 RESULTS

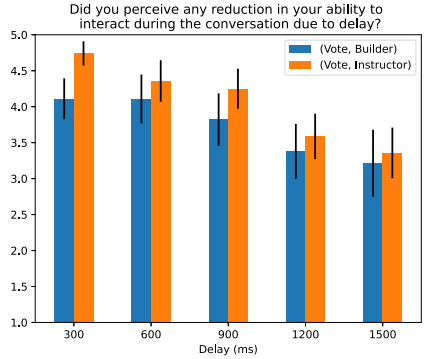
This section presents the results of the various factors assessed in the experiment. Each subsection comprises a normality test to assess the distribution of scores; an ANOVA to examine the impact of delay, figure, and role on voting outcomes; and a bar graph of the average score for each role and delay value. In addition, Tukey's HSD (honestly significant difference) post hoc analysis was performed to evaluate the differences between the delay values.

5.1 Subjective Performance Factors

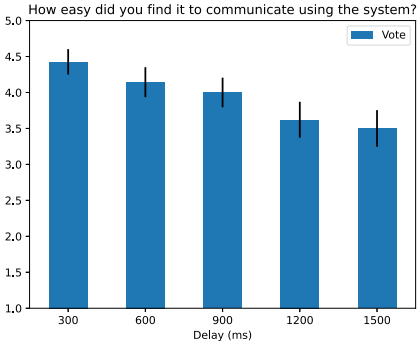
Initially, normality was confirmed for each of the factors either by a Kolmogorov-Smirnov normality test or by checking that both skew and kurtosis were in the range (–2, 2) as established by George [16]. Table 3 shows the statistical results for each factor of the subjective performance of the system. This table shows for each factor an analysis of the statistical significance (by means of an ANOVA analysis) of the different variables of the experiment (role, delay, and figure). If it is established that the role had an influence on the scores, an analysis by role is performed for this



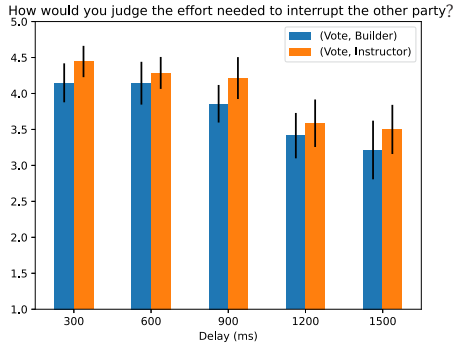
(a) Mean score values of the Global QoE with 95% confidence intervals



(b) Mean score values of the perceived delay with 95% confidence intervals



(c) Mean score values of the System Annoyance with 95% confidence intervals



(d) Mean score values of the perception of Interruptions with 95% confidence intervals

Fig. 7. Subjective performance results.

factor. In addition, for variables showing significance ($p < 0.05$), Tukey's HSD post hoc analysis was performed to identify statistically different delay pairs.

According to the results, the role was significant for the influence factor of delay perception and interruptions, which is why for these factors the analysis is done individually by role. Furthermore, the study examined the impact of different figures on the voting results and found that while certain figures significantly influenced Global QoE and the instructor's perception of delay influence, the effect was relatively small ($\eta^2 < 0.06$). Tukey's HSD analysis revealed significant differences between only two figures (Mazinger and Bird). On the contrary, the delay was found to have a significant impact on voting for all factors ($p < 0.05$), with a large effect size ($\eta^2 > 0.14$) in general.

Figure 7 shows the average scores for each factor and delay with their 95% confidence intervals. It can be observed that for the factors of perceived delay and interruptions, we can find differences between roles, with the builders being more sensitive to delay (i.e., they notice it earlier). Moreover, we can find significant differences from 600 ms of delay for the two conditions and for the two roles. At the level of averages, we also find for the perception of delay and interruptions that the quality values drop significantly from 900-ms delay onward. For overall quality and system annoyance, no differences were found between the roles, but differences were also found for the two factors from 900 ms, with the two worst delays (1,200 ms and 1,500 ms) reaching levels on average of 3.5. At the

Table 4. Presence Analysis

Factor	Variable		ANOVA			Significantly Different
			F	p	η^2	
Involvement	Role		$F_{1,230} = 1.585$	0.209	0.005	–
	Delay		$F_{4,230} = 7.318$	**<0.001**	0.096	(≤600) vs (≥1200) (900) vs (1500)
	Figure		$F_{4,230} = 2.769$	**0.028**	0.036	(Bird) vs (Mazinger)
Adaption	Role		$F_{1,230} = 5.221$	**0.023**	0.017	–
	Delay	Builder	$F_{4,115} = 4.602$	**0.002**	0.119	(≤600) vs (1500)
		Instructor	$F_{4,115} = 4.281$	**0.003**	0.113	(300) vs (≥1200)
	Figure	Builder	$F_{4,115} = 0.442$	0.778	0.011	
		Instructor	$F_{4,115} = 0.634$	0.639	0.017	–
Accomplishment	Role		$F_{1,230} = 0.252$	0.616	<0.001	–
	Delay		$F_{4,230} = 1.641$	0.165	0.024	–
	Figure		$F_{4,230} = 2.186$	0.071	0.031	–

level of QoE in the system, we could establish 900 ms as a threshold that guarantees an acceptable delay. This result is higher than that established in the recommendation [37]; however, it is in line with later studies [6, 44].

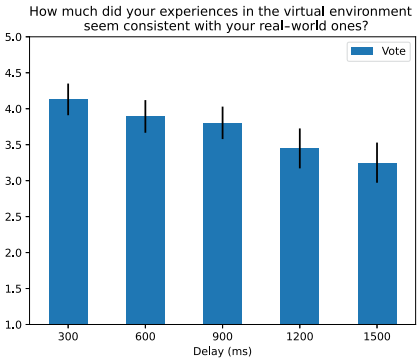
5.2 Presence

The study examined the presence of the adaptation factor. First, we verified the normality of the skew and kurtosis ratings, which were found to have absolute values less than 2. The results of the analysis of variance are presented in Table 4, which includes the role, delay, and figure variables for the presence factors under consideration, namely involvement, adaptation, and task. Additionally, Tukey's HSD post hoc analysis was performed to identify significant differences between pairs. After examining the influence of the role variable, it was determined that it only impacted the adaptation factor. Therefore, a separate analysis of the variables by roles was conducted for this factor. Results indicate that the delay and task factors had a significant impact with a medium effect ($\eta^2 > 0.06$) observed. The significant differences column reveals that differences between delays (1,200 and 1,500 ms) and delays of 600 ms or longer were observed. For the feeling of having completed the task correctly, we can observe that the delay did not have a significant effect.

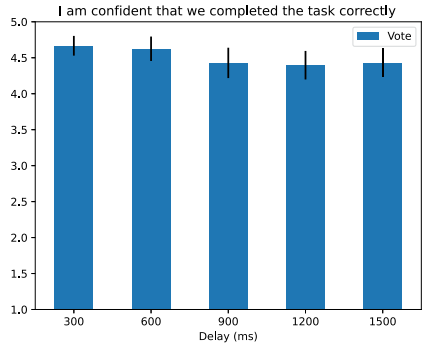
According to the average results in Figure 8, we only found differences between the roles in adaptation factor. Here, we can observe that the builders suffered more from the delay than the instructors. This is in line with the idea that builders notice the delay earlier and that it is more difficult for them to adapt to the task since they need to interrupt the other user. For instructors, this effect is smaller, although it also affects them. The last factor of presence refers to whether users feel that they have completed the task. This result is good for all delays. It was probably influenced by the fact that they needed to agree on the completion of the task to move on to the next figure.

5.3 Social Factors

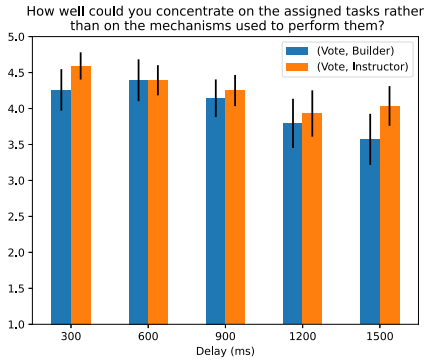
The present study examined some social factors. First, we verified the normality of the skew and kurtosis ratings, which were found to have absolute values less than 2. Utilizing an ANOVA (Table 5), it was determined that, for most of the social factors, only the delay factor had a significant impact on the ratings ($p < 0.05$), whereas the role and figure factors were deemed insignificant ($p > 0.05$). With respect to role, only the social annoyance factor shows statistically different results between instructors and constructors ($p = 0.01$). For the social presence factor, we can see



(a) Mean score values of the Involvement with 95% confidence intervals



(b) Mean score values of the Adaptation with 95% confidence intervals



(c) Mean score values of the Adaptation with 95% confidence intervals

Fig. 8. Presence factors results.

an effect of the figure on the results, but it is at the limit of statistical significance ($p = 0.048$) and the effect size is small ($\eta^2 < 0.06$).

Tukey's HSD post hoc analysis was subsequently conducted between delay pairs, revealing statistically significant differences between 600 ms with 1,200 ms and 1,500 ms.

According to the average results from Figure 9, social collaboration and adaptation have similar behavior to the task completion factor for presence. Users have the feeling that they finished the task correctly, both from the self and the whole point of view. Social presence, however, suffered a clear impact of delay, degrading similarly on average to those obtained for the Global QoE values. Finally, for the social annoyance factor, instructors were able to understand the users' message better than builders for higher delay values. The average results of the builder were significantly influenced by the delay (on average) from 900 ms, whereas the instructors kept their averages relatively stable.

5.4 Duration

This section presents an analysis of the impact of completion time for each experimental condition, namely delay and figure. First, a normality test was conducted to determine the distribution of the data, which indicated a non-normal distribution with kurtosis that exceeded an absolute value of 2. Subsequently, a more detailed examination of the results was performed, revealing a significant

Table 5. Social Factors Analysis

Factor	Variable		ANOVA			Significantly Different
			F	p	η^2	
Social Presence	Role		$F_{1,230} = 3.549$	0.061	0.0117	–
	Delay		$F_{4,230} = 7.761$	**<0.001**	0.102	(≤600) vs (≥1200) (900) vs (1500)
	Figure		$F_{4,230} = 2.440$	**0.048**	0.032	(Bird) vs (Rocket)
Social Annoyance	Role		$F_{1,230} = 5.714$	**0.001**	0.033	–
	Delay	Builder	$F_{4,115} = 3.310$	**0.001**	0.086	(≤600) vs (1500) (900) vs (1200)
		Instructor	$F_{4,115} = 3.027$	**0.020**	0.07	(≤600) vs (≥1200)
	Figure	Builder	$F_{4,115} = 2.188$	0.075	0.057	–
		Instructor	$F_{4,115} = 2.138$	0.081	0.050	–
Social Adaptation	Role		$F_{1,230} = 0.224$	0.637	<0.001	–
	Delay		$F_{4,230} = 3.986$	**0.004**	0.053	(300) vs (≥1200)
	Figure		$F_{4,230} = 1.315$	0.265	0.017	–
Collaboration	Role		$F_{1,230} = 0.774$	0.380	0.003	–
	Delay		$F_{4,230} = 3.425$	**0.010**	0.046	–
	Figure		$F_{4,230} = 1.394$	0.237	0.019	–

variation in the data. Following the identification of outliers with $|zscore| > 3$, two outliers of the conditions were identified and removed. Upon the elimination of these outliers, a normality test was conducted once again, which confirmed the normal distribution of the data with kurtosis and skew being less than 2 in absolute value.

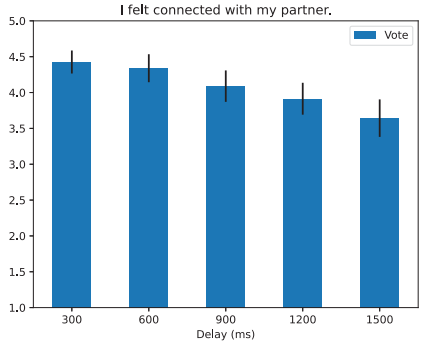
To investigate the influence of figures and delay on task completion time, an ANOVA was performed. The results revealed that the figure had a significant effect on task completion time, but the delay value did not. Subsequently, Tukey's HSD post hoc analysis was performed that revealed significant differences between two pairs of figures, namely the Dog with Rocket and TRex figures. The mean times for each delay value are presented in Figure 10, and it was observed that the confidence intervals were wide and no significant differences were found between the delay values. In particular, the average completion time was found to be 160 seconds for delays ranging from 300 to 1,200 ms, whereas for the worst condition, an average of 190 was obtained, representing ∼ 19% increase.

5.5 Audio

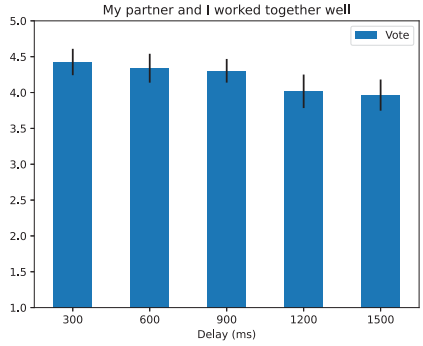
During the experimental sessions, the conversations of the participants for each condition (delay and figure) were captured using OBS [28] software, which enabled the recording of both the microphone channel (representing the voice of the local subject) and the headphone channel (representing the voice of the remote user). These audio channels were recorded in an audio file, where the left and right channels represented local and remote audio, respectively.

To ensure uniformity and standardization of the audio signals, the audio files were normalized to −26 dBov according to ITU P.56 [39]. The activity time of each user was then determined by calculating the squared mean amplitude of each 200-ms audio segment and comparing it against a threshold value of −16 dBFS. Any audio segment with a dBFS that exceeded the threshold value was classified as active. In Figure 11(a), an example of the audio signal (in blue) can be observed, with a running average of 200 ms (in orange) and a threshold of −16 dBFS (in red).

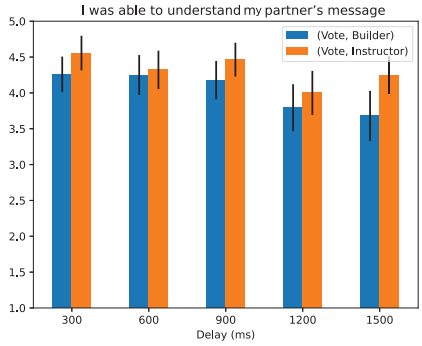
Once the threshold has been applied, we can see in Figure 11(b) the average time taken to finish the different figures for each role and delay. According to this graph, we can see that the average



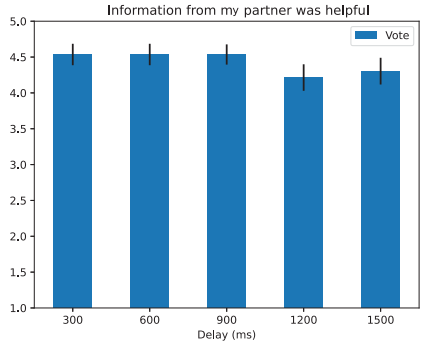
(a) Mean score values of the Social Presence with 95% confidence intervals



(b) Mean score values of the Social Adaptation with 95% confidence intervals



(c) Mean score values of the Social Annoyance with 95% confidence intervals



(d) Mean score values of the Collaboration with 95% confidence intervals

Fig. 9. Social factor results.

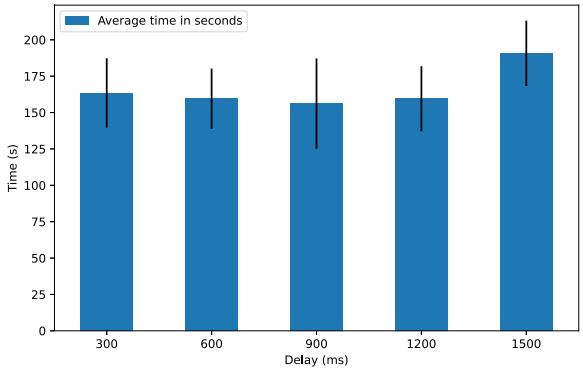
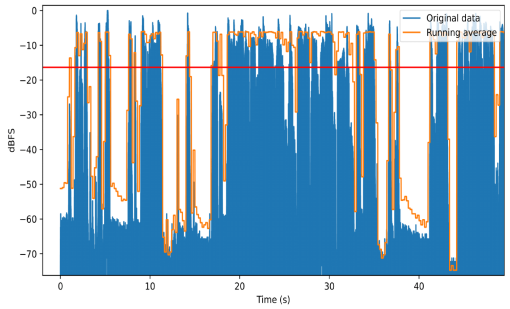
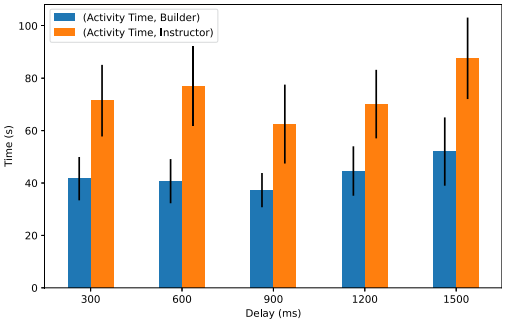


Fig. 10. Mean score values of the task duration in seconds with 95% confidence intervals.

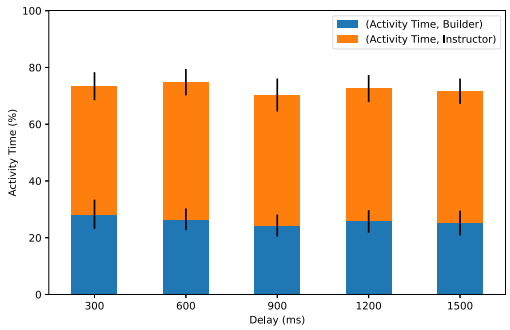
values increase by 1,500 ms for the instructors and from 1,200 ms for the builder. To check if this increase in activity is due to longer interventions or if there are more interventions, we calculate the percentage of time occupied by each of the roles in the conversation. In Figure 11(d), it can be seen the average of the activity times of each construction divided by the total time of that construction. In addition, we calculated the average number of interventions of each role by counting



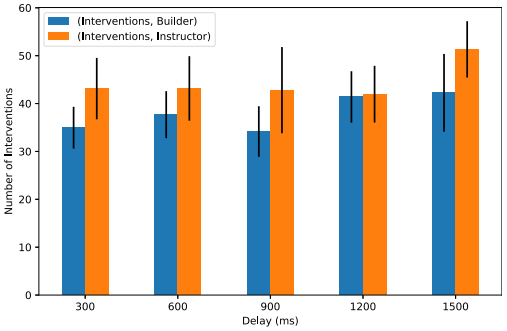
(a) Audio segment in decibels with the signal in blue, running squared average of 200 ms in orange and threshold for activity in red



(b) Average activity time in seconds per role and delay



(c) Average activity time per user and delay in percentage



(d) Average number of interventions per role and delay

Fig. 11. Audio results.

each intervention as the time between two silences of more than 200 ms following ITU-T P. 1305 [37]. The results of the number of interventions show similar results to those of the activity time per role. Together with the results shown in Figure 11(c), everything seems to indicate that for delays above 900 ms, the builder had to intervene more times than for shorter delays. Similarly, this effect can be seen for instructors at 1,200 ms and higher. However, the distribution of activity time was not altered. This indicates that users had to intervene more times to perform the same task from 900 ms onward.

6 DISCUSSION

We have analyzed subjective and objective factors varying the end-to-end delay of a photorealistic Social XR communication system. To do so, we have conducted an experiment on a system validated in terms of user experience, to which we have artificially introduced audiovisual delay in a collaborative Social XR task. Additionally, we have carried out an exhaustive analysis of the results for each subjective factor evaluated as well as of the possible elements that may introduce noise to the measures of the impact of delay on user experience. A discussion of the results follows.

The results of the experiment can be examined from a dual perspective: subjective and objective. Subjective results can be categorized into three distinct dimensions: overall perceived quality, presence, and social factors.

Although we could observe a reduction in the overall perceived quality as the delay increases, it is not too pronounced. The existing literature on conversations with delay [45, 46] suggests that users partially attribute the delay to the inoperability of their peers, thus absolving the system of blame. This attribute allows for greater delays in synchronous environments, as observed in the presented experiment. In absolute terms, and taking into account the data obtained for the

subjective assessment, we can recommend not to exceed 900 ms of end-to-end delay for collaborative videoconference Social XR systems. This value is higher than the threshold established by the recommendations for 2D videoconferences (600 ms), but is in line with more recent 2D videoconference studies [6, 44].

From an objective standpoint, the impact of delay on task completion time was analyzed. According to the results, an increase in the mean time required to construct the figures is evident. However, this increase is not statistically significant or as apparent as in the case of subjective results. This is attributed to the users' ability to adapt to the degraded environment, with their subjective perceptions of task performance remaining relatively unaffected by the deleterious effects of delay [12, 15]. In the experiment, we conducted further analysis on the influence of delay on users' recorded conversations. Our observations indicate that the instructor's role accounted for most of the conversation time (~45%), whereas the builder spoke for ~25% of the time (see Figure 11(b)). The remaining 30% of the time corresponds to silence. This silence is attributed to the time required to assemble the figures. Importantly, this distribution of conversation time was not altered with increasing delay. Although, as mentioned previously, the interactions were prolonged with higher delays, an examination of the number of interventions made by each role in relation to delay reveals that there were more interventions with longer delays while still maintaining the distribution consistent with the respective roles. In other words, there was an increased frequency of interventions, but the pace of the conversation remained unchanged. This fact supports the user adaption hypothesis.

Nevertheless, according to the factors that compose the perception of delay [2] (prior experience, task complexity, and expectations), we can find a great influence of the type of task [3]. In particular, the block-building task represents the most common form of interactive collaboration in videoconferencing—in other words, a conversation between two users who collaborate to perform a task [26]. However, other tasks could have a component that encourages users to interact as fast as possible. In this sense, the maximum acceptable delay value could vary. Therefore, further studies on the influence of delay are needed to set thresholds with respect to the specific use case.

Another aspect that has been addressed during this work is the adaptation of 2D videoconferencing protocols to the Social XR paradigm. In the same way that the first recommendations proposed tasks for telephone calls, there was *a posteriori* work to adapt these tasks and to propose different ones to evaluate the user experience in the field of videoconferencing. In this work, we have gone a step further and adapted a task for interactive videoconferencing to the Social XR paradigm. In this case, the differentiating element with respect to usual videoconferencing standards is that we consider 3D environments. At system level, Social XR still faces a number of challenges associated with the 3D environment in which users are immersed. While in 2D videoconferencing environments the remote user occupies the entire screen, in Social XR environments the other user's avatar must be located in a shared space. This adds an extra dimension in that the shared virtual elements must be synchronized. Moreover, the Social XR system should guarantee that the two users can interact between them and have a twin behavior in the shared space. For the building block task, it was crucial to configure the immersive environment in such a way that users can visually perceive the form of the figures that the remote user had in their hands without the ability to replicate them without asking the partner, while still maintaining sufficient proximity to prevent the task from becoming solely reliant on audio communication. Another important aspect regarding the social task is that the role of the builder was more sensitive to the delay even though he was the one who spoke the least. It is reasonable to believe that in the future we can centralize the analysis only on the builder part and use some kind of confederate user that always repeats the instructor role. In this way, we can increase the number of conditions at the same time even if we lose the information related to the role (but it has already been analyzed in this study).

7 CONCLUSION

To the best of our knowledge, we have presented the first analysis of delay for collaborative tasks in realistic Social XR environments. The main contribution is that the end-to-end delay should not exceed 900 ms if user acceptance is to be guaranteed. Another relevant contribution is the analysis of the adaptation of standardized tasks for evaluation that allows a correct comparison of new forms of videoconferencing with previous studies. We have also provided an evaluation protocol for interactive teleconferencing in Social XR. Therefore, a basis is established for different studies on the quality of collaboration in different use cases within the XR paradigm. As a future research direction, we consider assessing the influence of delay in different tasks that demand tighter delays, such as competitive environments and tasks involving translational movements.

REFERENCES

[1] Maha Abdallah, Carsten Griwodz, Kuan-Ta Chen, Gwendal Simon, Pin-Chun Wang, and Cheng-Hsin Hsu. 2018. Delay-sensitive video computing in the cloud: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, (June 2018), Article 54, 29 pages. https://doi.org/10.1145/3212804

[2] Christiane Attig, Nadine Rauh, Thomas Franke, and Josef F. Krems. 2017. System latency guidelines then and now—Is zero latency really considered necessary? In *Engineering Psychology and Cognitive Ergonomics: Cognition and Design*, Don Harris (Ed.). Springer International Publishing, Cham, 3–14.

[3] Leilani Battle, R. Jordan Crouser, Audace Nakeshimana, Ananda Montoly, Remco Chang, and Michael Stonebraker. 2020. The role of latency and task complexity in predicting visual search behavior. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 1246–1255. https://doi.org/10.1109/TVCG.2019.2934556

[4] Armin Becher, Jens Angerer, and Thomas Grauschopf. 2020. Negative effects of network latencies in immersive collaborative virtual environments. *Virtual Reality* 24, 3 (Sept. 2020), 369–383. https://doi.org/10.1007/s10055-019-00395-9

[5] Stephan Beck, André Kunert, Alexander Kulik, and Bernd Froehlich. 2013. Immersive group-to-group telepresence. *IEEE Transactions on Visualization and Computer Graphics* 19, 4 (2013), 616–625. https://doi.org/10.1109/TVCG.2013.33

[6] Gunilla Berndtsson, Mats Folkesson, and Valentin Kulyk. 2012. Subjective quality assessment of video conferences and telemeetings. In *Proceedings of the 2012 19th International Packet Video Workshop (PV'12)*. 25–30. https://doi.org/10.1109/PV.2012.6229740

[7] Kjell Brunnström, Elijs Dima, Tahir Qureshi, Mathias Johanson, Mattias Andersson, and Morten Sjöström. 2020. Latency impact on quality of experience in a virtual reality simulator for remote control of machines. *Signal Processing: Image Communication* 89 (2020), 116005. https://doi.org/10.1016/j.image.2020.116005

[8] Kjell Brunnström, Katrien De Moor, Ann Dooms, Sebastian Egger-Lampl, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Chaker Larabi, Bob Lawlor, Patrick Le Callet, Sebastian Möller, Fernando Pereira, Manuela Pereira, Andrew Perkis, Antonio Pinheiro, Ulrich Reiter, Peter Reichl, Raimund Schatz, and Andrej Zgank. 2013. *Qualinet White Paper on Definitions of Quality of Experience*. Qualinet.

[9] Pablo Carballeira, Carlos Carmona, César Díaz, Daniel Berjón, Daniel Corregidor, Julián Cabrera, Francisco Morán, Carmen Doblado, Sergio Arnaldo, María Mar Martín, and Narciso García. 2022. FVV Live: A real-time free-viewpoint video system with consumer electronics hardware. *IEEE Transactions on Multimedia* 24 (2022), 2378–2391. https://doi.org/10.1109/TMM.2021.3079711

[10] Fernando Cassola, Manuel Pinto, Daniel Mendes, Leonel Morgado, António Coelho, and Hugo Paredes. 2021. A novel tool for immersive authoring of experiential learning in virtual reality. In *Proceedings of the 2021 IEEE Conference on Virtual Reality and 3D User Interfaces (VRW'21)*. 44–49. https://doi.org/10.1109/VRW52623.2021.00014

[11] Gregory W. Cermak. 2005. Multimedia quality as a function of bandwidth, packet loss, and latency. *International Journal of Speech Technology* 8, 3 (Sept. 2005), 259–270. https://doi.org/10.1007/s10772-006-6368-3

[12] Carlos Cortés, Jesús Gutiérrez, Pablo Pérez, Irene Viola, Pablo César, and Narciso García. 2022. Impact of self-view latency on quality of experience: Analysis of natural interaction in XR environments. In *Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP'22)*. 3131–3135. https://doi.org/10.1109/ICIP46576.2022.9897983

[13] C. Cortés, P. Pérez, and N. García. 2019. Unity3D-based app for 360VR subjective quality assessment with customizable questionnaires. In *Proceedings of the 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin'19)*. 281–282. https://doi.org/10.1109/ICCE-Berlin47944.2019.8966170

[14] Carlos Cortés, Pablo Pérez, Jesús Gutiérrez, and Narciso García. 2020. Influence of video delay on quality, presence, and sickness in viewport adaptive immersive streaming. In *Proceedings of the 2020 12th International Conference on Quality of Multimedia Experience (QoMEX'20)*. 1–4. https://doi.org/10.1109/QoMEX48832.2020.9123114

[15] Shrikant Garg, Ayushi Srivastava, Mashhuda Glencross, and Ojaswa Sharma. 2022. A study of the effects of network latency on visual task performance in video conferencing. In *Extended Abstracts of the 2022 CHI Conference on Human*

Factors in Computing Systems (CHI EA'22). ACM, New York, NY, USA, Article 213, 7 pages. https://doi.org/10.1145/3491101.3519678

[16] Darren George. 2011. *SPSS for Windows Step by Step: A Simple Study Guide and Reference, 17.0 Update, 10/e*. Pearson Education India.

[17] Simon Gunkel, Hans Stokking, Martin Prins, Omar Niamut, Ernestasia Siahaan, and Pablo Cesar. 2018. Experiencing virtual reality together: Social VR use case study. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video (TVX'18)*. ACM, New York, NY, USA, 233–238. https://doi.org/10.1145/3210825.3213566

[18] Kunal Gupta, Gun A. Lee, and Mark Billinghurst. 2016. Do you see what I see? The effect of gaze tracking on task space remote collaboration. *IEEE Transactions on Visualization and Computer Graphics* 22, 11 (2016), 2413–2422. https://doi.org/10.1109/TVCG.2016.2593778

[19] Jesús Gutiérrez, Pablo Pérez, Marta Orduna, Ashutosh Singla, Carlos Cortés, Pramit Mazumdar, Irene Viola, Kjell Brunnström, Federica Battisti, Natalia Cieplinska, Dawid Juszka, Lucjan Janowski, Mikolaj Leszczuk, Anthony Adeyemi-Ejeye, Yaosi Hu, Zhenzhong Chen, Glenn Van Wallendael, Peter Lambert, Cesar Diaz, John Hedlund, Omar Hamsis, Stephan Fremerey, Frank Hofmeyer, Alexander Raake, Pablo Cesar, Marco Carli, and Narciso Garcia. 2022. Subjective evaluation of visual quality and simulator sickness of short 360° videos: ITU-T Rec. P.919. *IEEE Transactions on Multimedia* 24 (2022), 3087–3100. https://doi.org/10.1109/TMM.2021.3093717

[20] Thorsten Hennig-Thurau, Dorothea N. Aliman, Alina M. Herting, Gerrit P. Cziehso, Marc Linder, and Raoul V. Kübler. 2022. Social interactions in the metaverse: Framework, initial evidence, and research roadmap. *Journal of the Academy of Marketing Science*. Published Online, December 7, 2022. https://doi.org/10.1007/s11747-022-00908-0

[21] Redouane Kachach, Sandra Morcuende, Diego Gonzalez-Morin, Pablo Perez-Garcia, Ester Gonzalez-Sosa, Francisco Pereira, and Alvaro Villegas. 2021. The Owl: Immersive telepresence communication for hybrid conferences. In *Proceedings of the 2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct'21)*. 451–452. https://doi.org/10.1109/ISMAR-Adjunct54149.2021.00104

[22] Louise Lawrence, Arindam Dey, and Mark Billinghurst. 2018. The effect of video placement in AR conferencing applications. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction (OzCHI'18)*. ACM, New York, NY, USA, 453–457. https://doi.org/10.1145/3292147.3292203

[23] James R. Lewis. 1989. Pairs of Latin squares to counterbalance sequential effects and pairing of conditions and stimuli. *Proceedings of the Human Factors Society Annual Meeting* 33, 18 (1989), 1223–1227. https://doi.org/10.1177/154193128903301812

[24] Jie Li and Pablo Cesar. 2023. Social virtual reality (VR) applications and user experiences. In *Immersive Video Technologies*, Giuseppe Valenzise, Martin Alain, Emin Zerman, and Cagri Ozcinar (Eds.). Academic Press, 609–648. https://doi.org/10.1016/B978-0-32-391755-1.00028-6

[25] 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'21)*. 284–293. https://doi.org/10.1109/ISMAR52148.2021.00044

[26] Lifesize. 2019. 2019 Impact of Video Conferencing Report. Industrial Report. Retrieved March 15, 2024 from https://blog.tmcnet.com/blog/rich-tehrani/wp-content/uploads/2019/09/2019-Impact-of-Video-Conferencing-Report-Lifesize_FINAL.pdf

[27] 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 (MMSys'13)*. ACM, New York, NY, USA, 24–35. https://doi.org/10.1145/2483977.2483980

[28] OBS. n.d. OBS Studio. Retrieved March 15, 2024 from https://github.com/obsproject/obs-studio

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

[30] Pablo Pérez, Ester Gonzalez-Sosa, Redouane Kachach, Francisco Pereira, and Alvaro Villegas. 2021. Ecological validity through gamification: An experiment with a mixed reality escape room. In *Proceedings of the 2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR'21)*. 179–183. https://doi.org/10.1109/AIVR52153.2021.00040

[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), Article 917684, 22 pages. https://doi.org/10.3389/frsip.2022.917684

[32] ITU. 2021. *Relative Timing of Sound and Vision for Broadcasting*. Rec. ITU-R BT.1359. ITU.

[33] ITU. 2023. *Methodology for the Subjective Assessment of the Quality of Television Pictures*. Rec. ITU-R BT.500-15. ITU.

[34] RITU. 2021. *Influencing Factors on Quality of Experience for Virtual Reality Services*. Rec. ITU-T G.1035. ITU.
[35] ITU. 2015. *The E-Model: A Computational Model for Use in Transmission Planning*. Rec. ITU-T G.107. ITU.
[36] ITU. 2017. *Subjective Quality Evaluation of Audio and Audiovisual Multiparty Telemeetings*. Rec. ITU-T P.1301. ITU.
[37] ITU. 2016. *Effect of Delays on Telemeeting Quality*. Rec. ITU-T P.1305. ITU.
[38] ITU. 2016. *Method for the Measurement of the Communication Effectiveness of Multiparty Telemeetings Using Task Performance*. Rec. ITU-T P.1312. ITU.
[39] ITU. 1993. *Objective Measurement of Active Speech Level*. Rec. ITU-T P.56. ITU.
[40] ITU. 2008. *Subjective Video Quality Assessment Methods for Multimedia Applications*. Rec. ITU-T P.910. ITU.
[41] ITU. 2000. *Interactive Test Methods for Audiovisual Communications*. Rec. ITU-T P.920. ITU.
[42] Juan Sanchez-Margallo, Carlos Plaza de Miguel, Roberto A. Fernandez Anzules, and Francisco M. Sanchez-Margallo. 2021. Application of mixed reality in medical training and surgical planning focused on minimally invasive surgery. *Frontiers in Virtual Reality* 2 (2021), Article 692641, 11 pages. https://doi.org/10.3389/FRVIR.2021.692641
[43] Marwin Schmitt, Judith Redi, Dick Bulterman, and Pablo S. Cesar. 2018. Towards individual QoE for multiparty videoconferencing. *IEEE Transactions on Multimedia* 20, 7 (2018), 1781–1795. https://doi.org/10.1109/TMM.2017.2777466
[44] Marwin Schmitt, Simon Gunkel, Pablo Cesar, and Dick Bulterman. 2014. The influence of interactivity patterns on the quality of experience in multi-party video-mediated conversations under symmetric delay conditions. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia (SAM'14)*. ACM, New York, NY, USA, 13–16. https://doi.org/10.1145/2661126.2661135
[45] Katrin Schoenenberg, Alexander Raake, and Judith Koeppe. 2014. Why are you so slow?—Misattribution of transmission delay to attributes of the conversation partner at the far-end. *International Journal of Human-Computer Studies* 72, 5 (2014), 477–487. https://doi.org/10.1016/j.ijhcs.2014.02.004
[46] K. Schoenenberg, A. Raake, and P. Lebreton. 2014. Conversational quality and visual interaction of video-telephony under synchronous and asynchronous transmission delay. In *Proceedings of the 2014 6th International Workshop on Quality of Multimedia Experience (QoMEX'14)*. 31–36. https://doi.org/10.1109/QoMEX.2014.6982282
[47] Lucas M. Seuren, Joseph Wherton, Trisha Greenhalgh, and Sara E. Shaw. 2021. Whose turn is it anyway? Latency and the organization of turn-taking in video-mediated interaction. *Journal of Pragmatics* 172 (Jan. 2021), 63–78.
[48] Janto Skowronek, Alexander Raake, Gunilla H. Berndtsson, Olli S. Rummukainen, Paolino Usai, Simon N. B. Gunkel, Mathias Johanson, Emanuël A. P. Habets, Ludovic Malfait, David Lindero, and Alexander Toet. 2022. Quality of experience in telemeetings and videoconferencing: A comprehensive survey. *IEEE Access* 10 (2022), 63885–63931. https://doi.org/10.1109/ACCESS.2022.3176369
[49] Jiarun Song, Fuzheng Yang, Yicong Zhou, Shuai Wan, and Hong Ren Wu. 2016. QoE evaluation of multimedia services based on audiovisual quality and user interest. *IEEE Transactions on Multimedia* 18, 3 (2016), 444–457. https://doi.org/10.1109/TMM.2016.2520090
[50] Jennifer Tam, Elizbeth Carter, Sara Kiesler, and Jessica Hodgins. 2012. Video increases the perception of naturalness during remote interactions with latency. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems (CHI EA'12)*. ACM, New York, NY, USA, 2045–2050. https://doi.org/10.1145/2212776.2223750
[51] Irene Viola and Pablo Cesar. 2023. Volumetric video streaming: Current approaches and implementations. In *Immersive Video Technologies*, Giuseppe Valenzise, Martin Alain, Emin Zerman, and Cagri Ozcinar (Eds.). Academic Press, 425–443. https://doi.org/10.1016/B978-0-32-391755-1.00021-3
[52] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, and Pablo Cesar. 2023. VR2Gather: A collaborative social VR system for adaptive multi-party real-time communication. *IEEE MultiMedia* 30, 2 (2023), 48–59. https://doi.org/10.1109/MMUL.2023.3263943
[53] Peng Wang, Xiaoliang Bai, Mark Billinghurst, Shusheng Zhang, Weiping He, Dechuan Han, Yue Wang, Haitao Min, Weiqi Lan, and Shu Han. 2020. Using a head pointer or eye gaze: The effect of gaze on spatial AR remote collaboration for physical tasks. *Interacting with Computers* 32, 2 (2020), 153–169. https://doi.org/10.1093/iwcomp/iwaa012
[54] Yue Wang, Chen Chen, Kun Sha, and Jinxiu Wang. 2021. Research on the application of medical teaching based on XR and VR. In *Proceedings of the 2nd International Conference on Big Data and Informatization Education (ICBDIE'21)*. 688–691. https://doi.org/10.1109/ICBDIE52740.2021.00162
[55] 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'16)*. 365–369. https://doi.org/10.1109/ICIP.2016.7532380

Received 5 October 2023; accepted 15 February 2024