UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Telecomunicación

Interaction in Social eXtended Reality: A Quality of Experience Approach

D O C T O R A L T H E S I S

Submitted for the degree of Doctor by:

Carlos Cortés Sánchez
Máster Universitario en Ingeniería de Sistemas Electrónicos

Madrid, 2024


UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Telecomunicación

Doctoral Degree in Communication Technologies and Systems

Interaction in Social eXtended Reality: A Quality of Experience Approach

D O C T O R A L T H E S I S

Submitted for the degree of Doctor by:

Carlos Cortés Sánchez
Máster Universitario en Ingeniería de Sistemas Electrónicos

Under the supervision of:
Dr. Narciso García Santos
Dr. Pablo Pérez García

Madrid, 2024

Title: Interaction in Social eXtended Reality: A Quality of Experience Approach

Author: Carlos Cortés Sánchez

Doctoral Programme: Communication Technologies and Systems

Thesis Supervision:

Dr. Narciso García Santos, Professor, Universidad Politécnica de Madrid (Supervisor)

Dr. Pablo Pérez García, Lead Scientist, Nokia Extended Reality Lab (Supervisor)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This thesis has been partially supported by the Ministerio de Ciencia, Innovación y Universidades under the grant Programa de Ayudas para la Formación del Profesorado Universitario FPU18/05067.

En primer lugar quiero agradecer a mis directores de tesis Narciso y Pablo por la gran dedicación para que esta tesis salga adelante. No sólo por aguantarme sino por su dedicación activa. En un mismo orden, agradecer a mi familia a Alicia, a José Luís, a Víctor y a Jésica por haberme aguantado hasta llegar aquí.

En segundo orden de agradecimiento, quiero agradecer a mis amigos, que han sido el oasis de mi diáspora.

Finalmente, agradecer a todos aquellos que no se sientan incluidos en los apartados anteriores y crean que merecen ser objeto de agradecimiento.

Abstract

The rise of immersive technologies has led to an increase in the number of use cases that adapt this type of technology within the telecommunications area. Some examples are: industrial training, multimedia content consumption and tele-training. Among all the immersive technologies, eXtended Reality through the use of Head-Mounted Displays (HMD) is the one that focuses the majority of current developments. Specifically, the Social XR paradigm frames the use of immersive technologies in a multi-user or social context. Among the decisive factors for using immersive technology in communications use cases, two stand out: the possibility of making the user believe that they has been transported to another place (sensation of presence) and the possibility of increasing interactions by allowing displacements through space (6 degrees of freedom) as well as the possibility of interacting in a more natural way. Such improvements are ultimately improvements in user experience (UX). Therefore, UX evaluation is crucial for effective XR development. In a telecommunications context, this is known as quality of experience (QoE) evaluation.

In the initial stages of the thesis development, the focus was primarily on exploring possible areas of scientific contribution. The first significant area that emerged was the proposal of a methodology for evaluating the QoE of immersive environments based on 360 video. To this end, an inter-laboratory experiment was conducted within the video quality expert group (VQEG) of the International Telecommunications Union (ITU). As a result of this experiment, the ITU-T P.919 Recommendation was published.

As the thesis progressed, another key area of exploration was the development and evaluation of natural user interfaces (NUI) in the context of industrial training. Within a public-private partnership, we developed a training environment for fiber optic review with specific object manipulation requirements. In this section of the thesis, NUI-based manipulation solutions with subjective evaluation by subject matter experts are presented. Thanks to these contributions, we have been able to confirm that such natural interfaces allow the development of training that reduce cost and environmental impact while maintaining high user satisfaction values.

As we performed interaction development for Social XR, we identified that delay appeared to be a key element in guaranteeing QoE. Therefore, the third area of scientific contribution focused on investigating the impact of latency in different processing loops within the Social XR domain. In this sense the thesis presents two major contributions, a first contribution that focuses on the study of the different delays perceptible by users and how these affect them differently. Within this same contribution, a processing framework common to different existing Social XR systems is presented. Finally, a state of the art of different studies that identify allowable latencies in different use cases involving XR communication is presented. Using these values, a QoE prediction model adapted from an ITU recommendation is presented in order to be flexible to new use cases. The second major contribution presents three novel QoE studies investigating the impact of delays on: environment updates, self-view perception, and video conferencing within Social XR environments. This doctoral thesis has significantly advanced our understanding of immersive video-based environments. We can now effectively assess the QoE within these environments using novel methods. Furthermore, the thesis

explores the development of natural interfaces for interaction in XR, allowing us to evaluate XR interaction environments from a QoE perspective. This includes pinpointing the impact and location of delays within Social XR systems. By understanding how different delay values influence UX for various use cases, we can establish acceptable delay thresholds for optimal QoE in video-based Social XR.

Resumen

El auge de las tecnologías inmersivas ha impulsado su uso en el ámbito de las telecomunicaciones para diversos fines, como la formación industrial, el consumo de contenido multimedia y la teleformación. Entre estas tecnologías, la Realidad Extendida (XR) mediante gafas de realidad virtual (HMD) es la que concentra la mayor parte del desarrollo actual. En concreto, el paradigma de la XR Social plantea el uso de tecnologías inmersivas en un contexto multiusuario o social. Dos factores decisivos para el empleo de la tecnología inmersiva en las comunicaciones son: la sensación de presencia (ser transportado a otro lugar) y la posibilidad de incrementar las interacciones permitiendo desplazamientos (6 grados de libertad) e interacciones más naturales. Estas mejoras se traducen, en última instancia, en una mejor experiencia de usuario (UX). Por tanto, la evaluación de la UX resulta crucial para un desarrollo eficaz de la XR. En el contexto de las telecomunicaciones, esto se conoce como evaluación de calidad de experiencia (QoE).

Al comenzar la tesis, el objetivo principal fue explorar posibles áreas de contribución científica. La primera área destacada fue la propuesta de una metodología para evaluar la QoE de entornos inmersivos basados en vídeo 360°. Para ello, se llevó a cabo un experimento interlaboratorio dentro del grupo de expertos en calidad de vídeo (VQEG) de la Unión Internacional de Telecomunicaciones (UIT). Como resultado de este experimento, se publicó la Recomendación UIT-T P.919.

Otra área fundamental del trabajo de tesis fue el desarrollo y la evaluación de interfaces naturales de usuario (NUI) en el contexto de la formación industrial. Mediante una colaboración público-privada, se desarrolló un entorno de formación con requisitos específicos de manipulación de objetos. En esta sección de la tesis, se presentan soluciones de manipulación basadas en NUI con una evaluación subjetiva por parte de expertos en la materia. Gracias a estas aportaciones, se ha podido confirmar que dichas interfaces naturales permiten desarrollar formaciones que reducen costes e impacto medioambiental, manteniendo a la vez altos niveles de satisfacción del usuario.

Durante el desarrollo de la interacción para la XR Social, se identificó el retardo como un elemento clave para garantizar la QoE. Por lo tanto, la tercera área de contribución científica se centró en investigar el impacto de la latencia de distintos procesos en la XR Social. En este sentido, la tesis presenta dos contribuciones principales: un primer estudio sobre los distintos retardos perceptibles por los usuarios y cómo les afectan de manera diferente. Dentro de esta misma contribución, se presenta un marco de procesamiento común a diferentes sistemas de XR Social existentes. Por último, se ofrece un análisis del estado del arte sobre estudios que identifican las latencias admisibles en diferentes casos de uso que involucran comunicación por XR. Utilizando estos valores, se presenta un modelo de predicción de la QoE adaptado de una recomendación de la UIT para ser flexible ante nuevos casos de uso. La segunda contribución sobre retardos presenta tres nuevos estudios de QoE que investigan el impacto de los retardos en: actualizaciones del entorno, percepción de la autoimagen y videoconferencia dentro de entornos de XR Social.

Esta tesis doctoral ha supuesto un avance significativo en la comprensión de los entornos

inmersivos basados en vídeo. Ahora podemos evaluar eficazmente la QoE dentro de estos entornos. Este trabajo sienta las bases para la evaluación de la QoE en entornos de interacción natural. Además, también se incluye la identificación del impacto y la ubicación de los retardos dentro de los sistemas de XR Social. Al comprender cómo los diferentes valores de retardo influyen en la UX para diversos casos de uso, hemos identificado los umbrales de retardo aceptables en entornos de XR Social basados en vídeo.

Table of Contents

List of Figures

List of Tables

Abbreviations and acronyms

UPM Universidad Politécnica de Madrid

ACR Absolute Category Rating

AI Artificial Intelligence

ANOVA Analysis Of Variance

AR Augmented Reality

AV Augmented Virtuality

CI Confidence Interval

DMOS Differential Mean Opinion Score

DoF Degrees of Freedom

FoV Field of View

GDPR General Data Protection Regulation

HCI Human Computer Interaction

HEVC High Efficiency Video Coding

HMD Head Mounted Display

ITU International Telecommunication Union

MOS Mean Opinion Score

MR Mixed Reality

PQ Presence Questionnaire

PVS Processed Video Sequence

QoE Quality of Experience

QoS Quality of Service

QP Quantization Parameter

RQ Research Questions

SD Standard Deviation

sPQ Subsampling of the Presence Questionnaire

SRC Source

SSQ Simulator Sickness Questionnaire

SS Single Stimulus

SSCQE Single Stimulus Continuous Quality Evaluation

SSDQE Single Stimulus Discrete Quality Evaluation

SUS System Usability Scale

TPI Temple Presence Inventory

UX User eXperience

VQEG Video Quality Experts Group

VR Virtual Reality

XR Extended Reality

Chapter 1

Motivation

1.1 Introduction

The fast evolution of immersive technologies has led the communications community to experience a explosion of new use cases related to these technologies. Some examples use immersive technology to perform remote operations, industrial training, or new forms of entertainment based on video games or video consumption. Specifically, eXtended Reality (XR) has emerged as the paradigm for interaction in varying degrees of blending physical and virtual realities. XR encompasses a broad range of technologies that enable users to interact with and experience virtual or augmented environments. These technologies include virtual reality (VR), which creates a completely virtual environment, augmented reality (AR), which overlays digital elements onto the real world, and mixed reality (MR), which combines both VR and AR to create a seamless blend of the virtual and physical worlds. It can be understood as a continuum between seeing the complete physical reality or a synthetic world at any angle of vision as presented in Fig. 1.1.

Such immersive technologies typically make use of a display device called a Head Mounted Display (HMD). These devices allow a rendered virtual world to be displayed in front of our eyes. In addition, they are designed to isolate us from the outside world, creating the sensation of being transported to another place. These technologies also allow users to interact and exist within the same virtual world, opening the door to multi-user or social experiences. During the research work of the thesis, we have focused on a specific paradigm of multi-user environments, the social XR. As shown in Fig. 1.2, Social XR communications involve two or

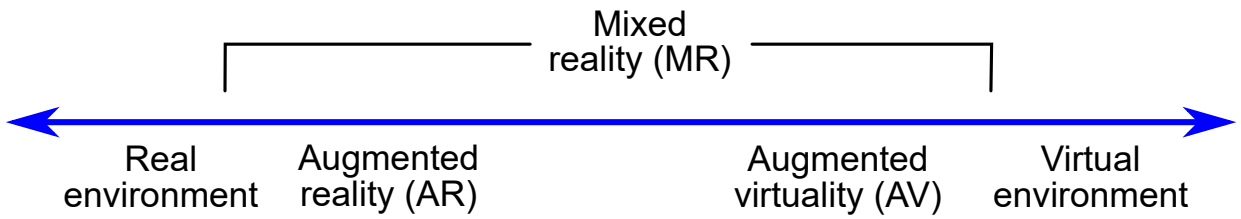


Figure 1.1: Reality–virtuality continuum [1].

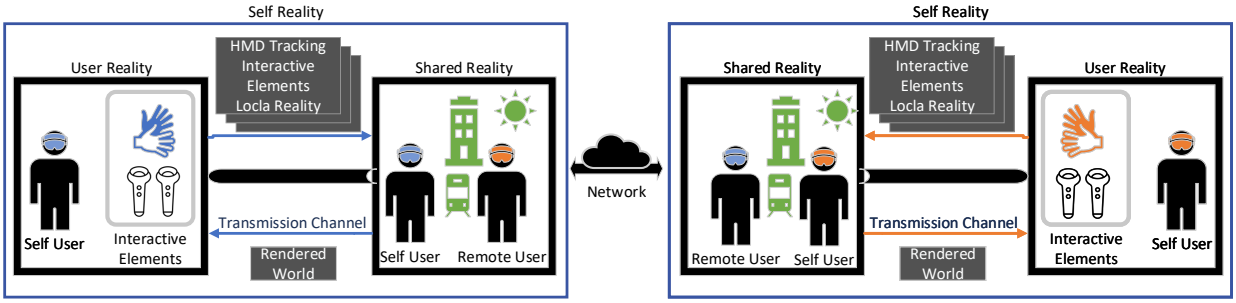


Figure 1.2: Social XR diagram.

more users in different physical spaces, who, through a visual representation, are transported to an interactive shared space.

According to [2], Social XR systems can be understood through the lens of presence. To feel "there" in a virtual world (spatial presence), the system should provide a sense of self-location with immersive visuals and accurate movement tracking. Users should also feel able to interact with the physical environment while feeling the self-perception of their own body (self-presence). For social interaction (social presence), features like avatars and real-time communication create a sense of co-presence.

Following Fig. 1.2, the Social XR can be decomposed into three aspects, which are related to the types of presence described above:

1. The physical reality of each user (user reality in the figure): Here we find the user in his own space (self-presence), they are able to interact with elements of their physical reality that will be represented in the Social XR environment.

2. Shared reality or shared world, users are able to interact through their physical world interactions to perceive the shared environment in a different way. According to the presence schema, this relates to spatial presence.

3. With information from remote realities, the shared world incorporates distant elements allowing users to interact with other people (social presence).

In the same way, this thesis work proposes a scheme of different interactions in Social XR based on this classification. Understanding, therefore, that we can find three types of interactions: spatial, personal, and social.

Furthermore, when introducing any new technology, it is crucial to evaluate the reasons why it's worth using. While the context of XR offers numerous justifications for its adoption, this thesis focuses on two key benefits:

- The possibility of interacting more naturally with the environment, for example, using our own body to move around or our hands to interact with visual interfaces

- The improved user experience while using these technologies.

These features are ultimately related to subjective perception. Thus, it seems reasonable to propose that one of the pillars when evaluating XR in the different use cases should

focus on measuring user experience. In a communications context, this type of evaluation is known as “Quality of Experience" (QoE). According to [3], QoE is defined as the degree of satisfaction of the user with a certain application or service. The influencing factors of QoE are 1) system-influencing factors (SIFs), 2) human influencing-factors (HIFs), and 3) context-influencing factors (CIFs). Human factors refer to the specific characteristics of the user, including their socioeconomic and demographic background, health status, or emotional state. System factors that influence quality are those properties and characteristics of the system that affect the user's perceived quality. Some examples of system-influencing factors in communication systems are capture, transcoding, storage, and playback. Finally, context factors are related to the user's environment in terms of physical, temporal, social, economic, and task. In order to establish a common framework in this regard, there are international recommendations that establish QoE evaluation methodologies [4], [5], recommendations on technical characteristics [6] and even predictive models of the level of satisfaction [7]. However, these recommendations were proposed for non-immersive communications systems. Therefore, immersive communications, such as Social XR, use features not covered by international recommendations.

1.2 Motivation

Initially, the lack of a validated methodology for evaluating QoE in immersive settings was identified as a major research gap. Specifically, on the evaluation of the representation of shared environments generated using 360° video. Drawing inspiration from former ITU's recommendations, an inter-laboratory QoE study was conducted to propose and validate an evaluation methodology. This study ultimately led to the development of the ITU P.919 recommendation.

Another aspect that the XR introduces in the field of communications is the possibility of interaction with the environment. For example, in XR users are able to move around, allowing interactions with 6 degrees-of-freedom (6DOF). Furthermore, this interactivity allows us to modify the environment in which users find themselves by augmenting virtual environments through our actions. With the aim of exploring different methods of interaction in XR, we studied natural interaction methods for developing immersive industrial training environments. This part of the thesis was driven by a concrete use case: training fiber-optic reviewers. Throughout the development of the training scenarios, the end-users – the instructors – actively participated in identifying the interaction requirements. One key requirement was to enable hand-based interaction with objects. A significant percentage of the systems use controllers replacing user hands in virtual reality [8]. Nevertheless, using controllers lead to hand gestures like grabbing being switched for buttons, which obstructs realistic interaction [9]. To avoid this disruption of realism we propose an immersive training environment using natural interfaces. In general, previous QoE assessment recommendations do not take into account the influence of such interactions on QoE. In addition to the developments, under this line of research we have validated questionnaires for such realistic manipulation environments. Thanks in part to these studies, VQEG is working on a new standard that takes into account personal interaction in Social XR.

Another aspect that was identified as an area of research was the influence of SIFs on Social XR systems. Currently, the various existing systems have been tested under ideal laboratory conditions and with specific hardware. After the various technology developments for Social XR, we determined that, among all the SIFs, delay was the most important when scaling Social XR systems to a real-world environment.

With respect to latency in Social XR, in this thesis I have made contributions in two ways. First, I have studied the state of the art to analyze the maximum latency values to guarantee QoE in different use cases, as well as the different types of impact on the user when exceeding these limits. Along with this analysis, I propose a QoE prediction model with respect to two key latency values: the value at which it is perceived and the value from which users do not accept it. Second, three latency experiments have been carried out that analyze three types of latency that were little/not studied in the literature: environment update latency, self-view latency, and conversational latency in volumetric video for Social XR. This latest study that takes into account the full spectrum of interaction in Social XR is being used, in part, as the basis for the ITU's future recommendation for the evaluation of interactive tasks in immersive communications environments.

Thus, the thesis has made significant contributions to these gaps of immersive technology:

- Validation of standard QoE methodologies for immersive technology.
- Development and evaluation of natural interaction techniques in XR.
- Latency analysis in Social XR from a QoE point of view.
- Latency studies regarding the influence of delay on QoE in video-based Social XR.

Figure 1.3: Scheme of the contributions located in the Social XR interaction classification.

The contributions of the thesis are framed within the three forms of interaction in Social XR proposed in the introduction. Fig. 1.3 summarizes the extent to which contributions have been made to the full spectrum of interaction in Social XR. Blue bars represent published

contributions covering an entire category of interaction while red bars represent work in progress.

1.3 Thesis Outline

The thesis is structured into six chapters, that are briefly described in the following paragraphs:

Chapter 2 focuses on the development of a methodology for QoE evaluation in immersive use cases. Specifically, an inter-laboratory QoE study is presented to validate QoE methods applied to the visualization of 360° video. This study led to the ITU recommendation: P919. In addition, a published tool is presented to facilitate the development of XR experiments in a public manner.

Chapter 3 focuses on the development of new interactive experiences based on natural interaction. During this chapter, the EPSILON system is presented, an industrial training system that requires interacting with the hands with different physical objects with virtual representation. During the development of the project, two positive evaluations were carried out with experts in the field that validated the developed natural interaction methods as a form of immersive training.

Chapter 4 Addresses the problem of latency in Social XR environments. Specifically, a common framework is proposed that details the different processes that enable interactive immersive communications. With the processes divided, the different process loops are analyzed, which start with an action by the user and end with the visualization of the immersive environment. For each loop, different technological challenges that can cause delays in them are exposed, along with a state-of-the-art study on the impact on users of these delays. Once the different delays have been classified, different use cases sensitive to each delay are exposed. In addition, this work proposes a methodology for classifying the key delay values based on two points, the point of perception and acceptability of the delay. Through a review of the literature, threshold values are proposed for each use case. Finally, a model adapted from an ITU recommendation is presented.

Chapter 5 presents three QoE studies related to three delays that, until then, were little/not studied in the literature. The delay of the environment updating, the delay of the self-view and the delay for videoconference in Social XR. For each study, questionnaires are used, if appropriate, that evaluate as global factors: the overall quality, the spatial presence and the social presence. These use cases were integrated into the model of the previous chapter. Finally, it is worth highlighting that the study in Social XR is, to our knowledge, the first to be carried out adapting standard methodology, and, it is used as a reference for the ongoing VQEG work towards a new recommendation for the evaluation of the QoE in Social XR.

Chapter 6 presents the contributions, conclusions and future work in relation to the research work of the thesis.

Chapter 2

Evaluation of QoE in XR

2.1 Introduction

The evaluation of XR systems is strongly linked to the evaluation of QoE. This is because several psychological factors are involved in the success of this technology. During the development of the thesis, the absence of specific methodologies for the evaluation of QoE in immersive video technologies was identified.

This chapter presents the contributions of the thesis in terms of methodology for the evaluation of immersive video-based environments. The common framework of these contributions encompasses the evaluation of new technologies based on already standardized methodologies and the development of a tool for enabling scoring during the immersive experience.

The following sections present the contributions made during the thesis in the field of QoE evaluation in the field of immersive video. Section 2.2 presents the related work of methodology for assessing the QoE in immersive video-based environments. Section 2.3 describes the evaluation of different techniques to measure QoE through an inter-laboratory experiment in the context of 360° video. In addition, the development of an immersive scoring tool is presented. Section 2.4 describes the conclusions and future work.

2.2 Related Work

QoE assessment a crucial element of the technology development pipeline. QoE represents the last step in the technology development cycle, among others, it allows us to validate the proposed solutions from a human point of view. However, the fact of measuring human perceptions, which are ultimately subjective and user-dependent, makes it especially necessary to establish common frameworks for the evaluation of QoE. Specifically, the ITU has been responsible for the standardization of common frameworks for the evaluation of QoE in the area of communications. In this context, there are different different areas of standardization. For example, the ITU-R Rec. BT.500 proposes the methodologies for the subjective assessment of the quality of television images, which focuses on the evaluation of static images [10]. The ITU-T P.910 recommendation establishes subjective video quality assessment methods for

multimedia applications, which focus on 2D video [5]. Another area of special interest in the area of QoE is that of quality prediction [7], [11], [12]. This type of solutions attempt to predict the average quality that a final user will perceive given some video quality parameters, for example, latency, bitrate, or compression ratio. In this context, ITU-T Rec. G.107 establishes the e-model that takes different parameters of a video stream to estimate the QoE [7]. Although these methods have been well validated and tested in the context of traditional videoconferencing, there is still ongoing work to develop recommendations analogous for the evaluation of QoE in immersive communications and new forms of video.

2.3 Subjective evaluation of 360° video

This section presents an evaluation of 360° video quality for cross-lab tests arranged by the Immersive Media Group (IMG) of the VQEG. More than 300 participants from 10 laboratories evaluated audio-visual quality, simulator sickness symptoms, and exploration behavior in short (10-30 seconds) 360° sequences. The influence of various factors, including assessment methodology, sequence duration, HMD device, uniform and non-uniform coding degradations, and simulator sickness assessment methods, was also analyzed. The results show that Absolute Category Rating (ACR) and Degradation Category Rating (DCR) are valid for subjective tests with 360° videos. Short videos (10 seconds with or without audio) are sufficient for evaluating coding artifact quality. Any commercial HMD that meets minimum requirements can be used. More efficient methods than the long Simulator Sickness Questionnaire (SSQ) have been proposed. These results have been used to develop the ITU-T Recommendation P.919. The annotated dataset from the tests is publicly available for the research community [1].

2.3.1 Subjective Experiment

This section presents an interlaboratory experiment carried out during the doctoral thesis to establish a methodology for the evaluation of immersive video. Subjective experiments are usually used to evaluate the QoE of users of immersive media technologies. However, most existing methodologies are based on 2D video assessments, and until this experiment was carried out, there were no international recommendation for 360° video.

In addition, to foster research and development, it is important to access databases with appropriate content and users' data from subjective experiments. This allows researchers to reproduce studies, compare results, and build models to estimate QoE. IMG[2] of the VQEG with the following objectives:

- To validate and recommend test methodologies to evaluate the audiovisual quality of 360° videos, taking into account:
  - The duration of the test sequences, considering short ones (10-30 seconds). Longer sequences, which may entail the evaluation of other aspects such as presence, immersion, etc. are left for future work.

[1] www.gti.ssr.upm.es/ ccs/VQEG360Dataset
[2] www.its.bldrdoc.gov/vqeg/projects/immersive-media-group

Table 2.1: Distribution of the nine test conditions and participant laboratories.

ID	Test Condition	Methodology	Lab	HMDs	Num. of PVSs	PVSs' Length
A	Video duration	ACR	Wuhan	HTC Vive	64	10s & 20s
B	Video duration	ACR	AGH	Oculus Rift	40	20s & 30s
C	Video duration	DCR	Roma3	HTC Vive	40	10s & 20s
D	Video duration	DCR	CWI	Oculus Rift	30	20s & 30s
E	Video duration	ACR	Surrey	HTC Vive	48	10s & 30s
F	Influence of HMD (desktop/mobile, High/low resolution)	ACR	UPM & Nokia	GearVR vs. HTC Vive vs. HTC Vive Pro	48	20s
G	Influence of HMD (Tethered vs. untethered)	ACR	Ghent	HTC Vive Pro	48	20s
H	Influence of audio (Videos with vs. without audio)	ACR	RISE	HTC Vive	48	20s
I	Influence of scoring method (App. vs. voice)	ACR	TU Ilmenau	HTC Vive Pro	48	20s

  - Influence factors such as the HMD, the source content characteristics, and the impact of uniform and non-uniform artifacts.

- To recommend methods to assess simulator sickness, considering:
  - One multi-item questionnaire (SSQ or derivation from it), or one single-question item.
  - When/how to assess simulator sickness and how to process and analyze the results.
- To generate and publish a dataset of subjectively assessed 360° content for future research, which is available in the databases section of the VQEG website.

The fulfillment of these objectives has supported the development of the recent ITU-T Recommendation P.919 [13]. This recommendation provides guidelines for subjective test methodologies for 360° video on HMDs, in line with the recommendations ITU-R BT.500 [10], ITU-T P.910 [5], and ITU-T P.913 [14] for 2D video, and ITU-T P.915 [15] for 3D video. This section presents the details of the subjective experiment and the results that supported the majority of the guidelines included in the new recommendation.

Table 2.2: Properties of the source sequences, the ones marked with * were not considered to generate the test stimuli used in test conditions B (AGH), C (Roma3) and D (CWI).

Name (ID)	**NokiaDojo (ND)***	**NokiaFlamenco (NF)**	**CheerLeading (CL)***	**BrazilMusic (BM)**
Screenshot				
Resolution	3840x2160, 30fps	3840x2160, 30fps	4096x2048, 25fps	4096x2048, 25fps
Provider	Nokia	Nokia	TU Ilmenau	TU Ilmenau
Description	Video of an indoor sport course, with ambient audio. Contains stitching artifacts.	Indoor dance course, with ambient audio. Contains stitching artifacts.	Cheerleading session indoors, with ambient audio.	Indoor scene of a band playing Brazilian music. With audio.
Name (ID)	**VSenseLuther (VL)**	**VSenseVaude (VV)**	**OculusMotion (OM)**	**OculusBeach (OB)***
Screenshot				
Resolution	4096x2048, 30fps	4096x2048, 30fps	3840x1920, 30fps	3840x1920, 30fps
Provider	VSense	VSense	Oculus	Oculus
Description	Video with animation content and a main character. Contains various shots (indoors and outdoors) and audio.	Video where a girl speaks to the camera. Contains audio and various indoor and outdoor shots.	Camera moving in a city. Contains music and two shots: one in daylight and one at night.	Scene with music of a beach at sunset with people dancing and moving.

2.3.2 Test Conditions

According to the objectives reported in Section 2.3.1, the nine test conditions shown in Table 2.1 were established to be evaluated in the cross-lab tests, including: two test methodologies (ACR and DCR), test videos of 10, 20 and 30 seconds, and different HMDs (desktop, mobile, tethered, untethered, etc.), methods to collect observers' ratings, and using sequences with and without audio. The selected test conditions cover factors influencing the assessment of audiovisual quality, including the impact of spatial degradations (e.g., coding artifacts), which is commonly done with short sequences [14]. Several other factors influence the overall QoE of the users when watching 360° videos [16], such as immersion [17] or temporal degradations (e.g, transmission degradations [18], latency [19], etc.), which may require longer sequences to be properly evaluated [20]–[22], and were out of the scope of the test campaign presented in this thesis. In addition, given that even with short sequences the users may experience simulator sickness, different questionnaires were considered to analyze how and when to assess it during the test session. The following subsections provide details on these test conditions and the experimental setups used in the tests.

2.3.3 Test Stimuli

Eight 360° videos of 30 seconds were used as source sequences (SRCs) in the tests. They were all in equirectangular projection, monoscopic, and had at least a resolution of 3840x1920 pixels and 25 fps. Screenshots of these sequences and their main characteristics are shown in Table 2.2. The original videos were provided by Nokia, TU Ilmenau, VSense [23], and Meta. The selected sequences present a wide range of content characteristics, including one video with camera motion (OM), one with animation scenes (VL), and different spatial and temporal complexities, as shown by the Spatial Information (SI) and Temporal Information

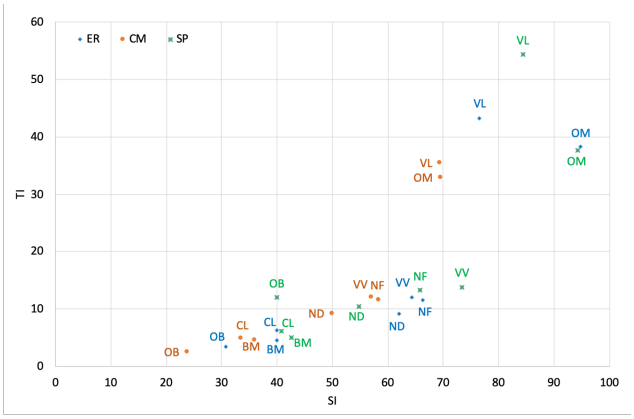


Figure 2.1: Scatter plot of SI and TI of the source sequences in ER, CM and SP projections.

#Tiles	Transition	ROI	QPs							
8x5**	Smooth	90°	42	37	32	22	22	32	37	42
6x3	Smooth	120°	42	32	22	22	32	42		
8x5**	Abrupt	180°	42	42	22	22	22	22	42	42
6x3	Abrupt	120°	37	37	22	22	37	37		

Figure 2.2: Settings for the non-uniform coding configurations.

(TI) indices [5] represented in Fig. 2.1. As it can be seen, SI and TI have been computed in three different projections, i.e., equirectangular (ER), cube-map (CM) and spherical (SP), to account for possible inaccuracies due to projection distortions [13], [24], [25]. Although small differences can be observed, the three domains' computations are highly correlated and show a wide distribution of spatial and temporal properties of the dataset.

Eight different HEVC coding configurations were applied to generate the test videos, including four uniform encodings (using homogeneous QPs) and four non-uniform encodings (using different configurations of tiles). For the uniform configurations, the following QPs were used: 15, 22, 32, 42, while Fig. 2.2 shows the settings for the non-uniform ones. As it can be seen, two different structures of tiles were used, and smooth and abrupt transitions between adjacent tiles were considered. The encoding of all the test sequences was done using the Kvazaar encoder, applying $period = 2s$, $gop = 0$ (structure disabled), and $ref = 1$ (forcing reference frames). Also, for each encoded video, three sequences were created with different duration using the first 10 seconds, the first 20 seconds, and the whole 30 seconds video[3].

2.3.4 Evaluation methodologies

The participants in the tests were asked to freely watch and explore the test contents and rate them in terms of audiovisual quality and simulator sickness according to the following methodologies.

[3]This dataset is publicly available at: https://www.its.bldrdoc.gov/vqeg/video-datasets-and-organizations.aspx

Audiovisual quality

To validate test methodologies for subjective quality assessment of 360° videos, two methodologies were implemented in different laboratories [14]:

- ACR: Single-stimulus method where the test videos are presented to the observers in random order, and they rate the stimuli independently on a five-grade category scale, from 5 (excellent) to 1 (bad).
- DCR (or DSIS): Double-stimulus method where, for each test video, the observers first watch the corresponding reference video, and they rate the degradations on a five-point scale, from 5 (imperceptible) to 1 (very annoying).

Questionnaire tools

The impact of two different ways to collect observers' ratings was investigated. On one side, the Unity-based application Miro360[4] [26] was used. During the course of the thesis I designed this tool under the Unity engine to be used in multiple HMDs. Furthermore, this tool can be configured for a variety of custom or predefined questions while the video sequence is played and/or after its ending. The head tracking is also registered in degrees with respect to the sphere that encloses the user and where the equirectangular video is projected. The voting method is gaze-based using joysticks keys for confirmation.

The features of the tool are:

- Labeling videos with custom ids.
- Custom questionnaires and scales.
- Custom start and duration of the video.
- Custom in-sequence questionnaires period.
- Randomize the contents according to the recommendation of the ITU-R BT.500-13.

The data that we can obtain from the application are: the position of the head and the scores of the users to each of the questions. The way to vote on the questionnaires is as follows. Users visualize the corresponding question in front of them in virtual reality in an aseptic grey environment following the recommendations of ITU-T Rec. P913 [14]. Under the question, the different options that the user has to answer appear. By moving their head, users can pre-select different answers, which, when pointed at, will light up in a red color. Once pre-selected, the vote is confirmed using the controller, giving way to the next question.

Fig. 2.3 illustrates the operation of the tool, showing on the left the configuration files that define the session, a sample of how it continuously transitions between the voting scene and the 360 video display scene, and finally, on the right, the session log that includes the head orientation information in each frame and the user's scores.

On the other side, one lab also collected the ratings that the observers provided verbally [27]. In this case, the participant had to say the number of the rating aloud, and the experimenter

[4]https://github.com/C-Cortes-spa/Miro360

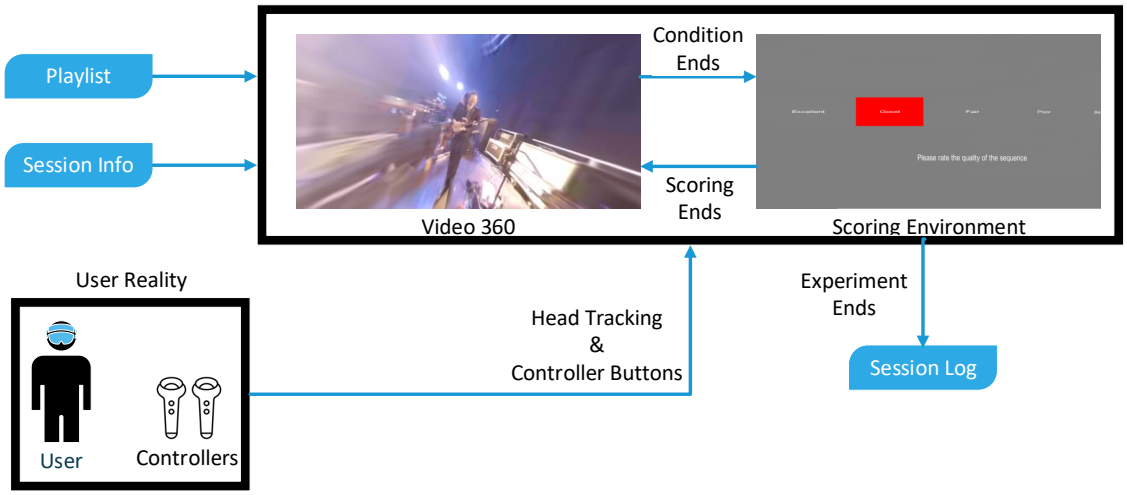


Figure 2.3: Miro 360 framework.

noted it down. In both cases, the rating scales were displayed in the HMD after each test video, and the observers were able to evaluate all the test videos without removing the HMD to rate.

Simulator sickness

In order to study appropriate methods to evaluate simulator sickness with 360° video, three different questionnaires were used in the cross-lab tests:

- Simulator Sickness Questionnaire (SSQ) [28]: The widely-used method by Kennedy, which evaluates 16 symptoms grouped in 3 factors: oculomotor, nausea, and disorientation. Each symptom is evaluated using a four-grade scale (0=none, 1=slight, 2=moderate, and 3=severe). In addition to global scores for each factor, a total score can be computed.
- Vertigo Scale [29]: The single-question method proposed by Pérez *et al.*, which evaluates simulator sickness stating the question "Are you feeling any sickness or discomfort now?" and using a five-grade scale (from "no problem" to "unbearable").
- Short-SSQ [30]: Another single-question method proposed by Tran *et al.*, which evaluates simulator sickness in terms of dizziness using the question "How is your level of dizziness or nausea?" and a five-grade scale (from "absolutely not dizzy" to "very dizzy").

These questionnaires were filled by the participants (not wearing the HMDs) in various moments during the test session (see details in Subsection 2.3.6), so it was possible to analyze the evolution of the symptoms. In all those moments, each participant filled the full SSQ and one of the single-item questionnaires (always the same), which were randomly assigned to obtain balanced samples.

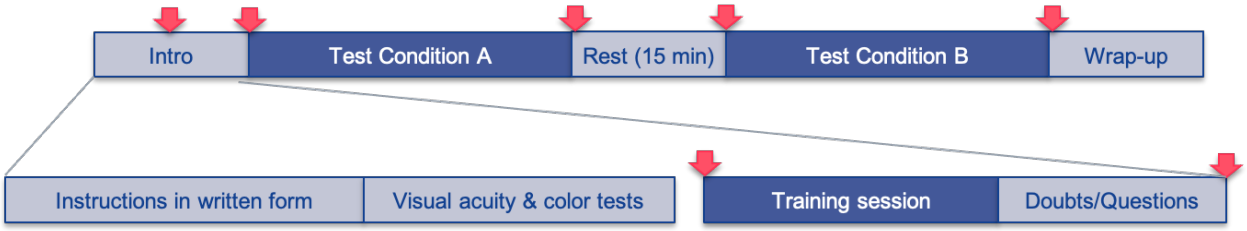


Figure 2.4: Diagram of the structure of the test session.

2.3.5 Environment and Equipment

The tests were carried out by ten laboratories at Universidad Politécnica de Madrid (Spain), Nokia Bell-Labs (Spain), Wuhan University (China), AGH University of Science and Technology (Poland), Roma TRE University (Italy), Centrum Wiskunde & Informatica (The Netherlands), Ghent University (Belgium), RISE Research Institutes of Sweden (Sweden), TU Ilmenau (Germany), and University of Surrey (United Kingdom). The tests were conducted in controlled environments in all laboratories, where the observers were seated in a swivel chair, so they could rotate freely to explore the 360° videos.

To study the influence of the HMD, four different devices were used in the cross-lab tests: Samsung GearVR, a mobile solution based on attaching a smartphone to an HMD support with a resolution of 1280x1440 pixels per eye and a refresh rate of 60Hz; Oculus Rift and HTC Vive, consumer desktop solutions with resolutions of 1080x1200 pixels per eye and refresh rates of 80Hz and 90Hz, respectively; and HTC Vive Pro, high-resolution (1440x1600 pixels per eye and 90Hz) solution available both tethered and untethered.

2.3.6 Session structure

As can been seen in Table 2.1, two test conditions were evaluated in each lab. The evaluation of the two test conditions was done by the same participants, following the session structure depicted in Fig. 2.4, which was followed by all laboratories. Firstly, an introductory session was conducted with the participants, where instructions for the test were provided, visual screening was performed, and training video samples were shown to appropriately adjust the HMD and familiarize them with the test methodology. Also, consent forms and background questionnaires were filled. At the end of this session, any doubts or questions from the participants were clarified. Then, the participants evaluated the test stimuli corresponding to the first test condition and, after a break of 15 minutes, they evaluated the corresponding ones for the second test condition. At the end of the test, the participants were requested to answer some more general questions about it.

As aforementioned, the participants were asked to fill questionnaires to evaluate simulator sickness, which was done various times during the test session. As depicted by the red arrows in Fig. 2.4, these questionnaires were filled before the training session (1), after the training session and just before starting the evaluation of the first test condition (2), after the evaluation of the first test condition (3), after the training and just before the evaluation of the second condition (4), and at the end of the test (5).

Table 2.3: Number, age distribution, and experience with VR/AR headsets of the observers. One participant from Roma3 did not report his/her experience.

Lab	Test ID	Number of Observers			Age			Experience with VR Headsets				
		Total	Female	Male	Min	Max	Avg	Times=1	Times<5	5<Times<20	Times> 20	Every day
Wuhan	A	30	15	15	20	30	24.5	8	15	7	0	0
AGH	B	40	13	27	18	79	28.5	13	17	8	2	0
Roma3	C	30	8	22	21	57	30.6	7	10	2	8	2
CWI	D	28	14	14	21	60	27.6	2	12	5	6	3
Surrey	E	31	10	21	19	44	25.9	13	12	3	2	1
UPM & Nokia	F	60	25	35	20	31	23.2	18	32	9	1	0
Ghent	G	30	4	26	23	45	31.6	3	14	7	5	1
RISE	H	28	16	12	22	66	41.6	3	16	8	1	0
TU Ilmenau	I	29	14	15	20	37	25.9	4	18	4	3	0
Total		306	119 38.9%	187 61.1%	18	79	28.8	71 23.20%	146 47.71%	53 17.32%	28 9.15%	7 2.29%

The test sessions lasted less than 90 minutes, and the evaluation of each test condition did not last more than 25 minutes, approximately. In those cases in which DCR methodology was used, and longer test sequences were evaluated, a subset of the test stimuli was considered to satisfy those time limits. In particular, the source contents NokiaDojo, CheerLeading and OculusBeach (marked with * in Table 2.2) were not considered to generate the test stimuli used in test conditions B (AGH), C (Roma3) and D (CWI), and in this last case, the non-uniform coding configurations using 8x5 tiling patterns were also not used (marked with ** in Fig. 2.2).

2.3.7 Observers

A total of 306 participants took part in the cross-lab test (38.9% women, 61.1% men), with ages ranging between 18 and 79 (average of 28.8). Vision screening was carried out before the tests, to assure that observers had a standard or corrected-to-normal vision in terms of visual acuity and colour vision. The participants were also asked to fill a background questionnaire in which they had to indicate their experience using VR/AR headsets. All details by lab and in total are reported in Table 2.3. A total of 60 participants performed the tests in UPM & Nokia (Test F), who were organized so that each observer evaluated two HMDs, thus, each HMD was evaluated by 40 participants.

2.3.8 Results of Nokia-UPM test on the Influence of the HMD

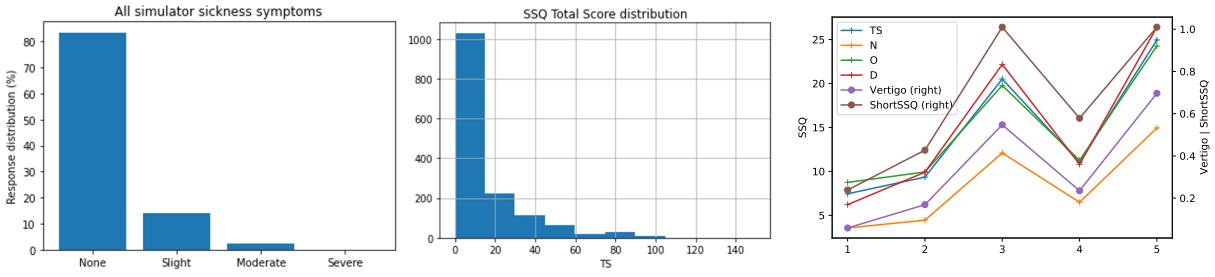
This subsection depicts the results of the experiments conducted at UPM and Nokia. Appendix B shows the results of those parameters of influence covered by of other laboratories and that are part of the data used to develop the recommendation. UPM conducted test F, where 60 participants evaluated 48 Processed Video Sequences (PVSs) of a fixed duration of 20 seconds. The ACR methodology was used for the evaluation. The test condition was to assess the influence of HMD on the results. For this purpose, three HMDs with different characteristics concerning resolution and possibility of wireless use were selected: Samsung GearVR HMD (mobile), HTC Vive and Vive Pro (both desktop-based). During the tests, each user evaluated two of the three HMDs. Thus, each HMD was evaluated 40 times.

Influence of HMD

On the one side, three different HMDs were compared in UPM & Nokia (Test F). The mixed-model analyses showed no significant differences comparing the mobile Samsung GearVR HMD with the desktop HMDs (Vive and Vive Pro) ($\chi^2(1) = 1.48$ and $p = 0.2230$ for Vive Pro, $\chi^2(1) = 2.57$ and $p = 0.1087$ for Vive), although, surprisingly, slightly higher MOSs were obtained with GearVR. However, significant differences were found comparing the HTC Vive and HTC Vive Pro ($\chi^2(1) = 10.16, p = 0.0014$), with better MOSs for the HTC Vive, which provides a lower resolution than HTC Vive Pro. However, the Wilcoxon Signed-Rank tests did not show any significantly different pair among all the possible comparisons (144) among the HMDs for all test videos.

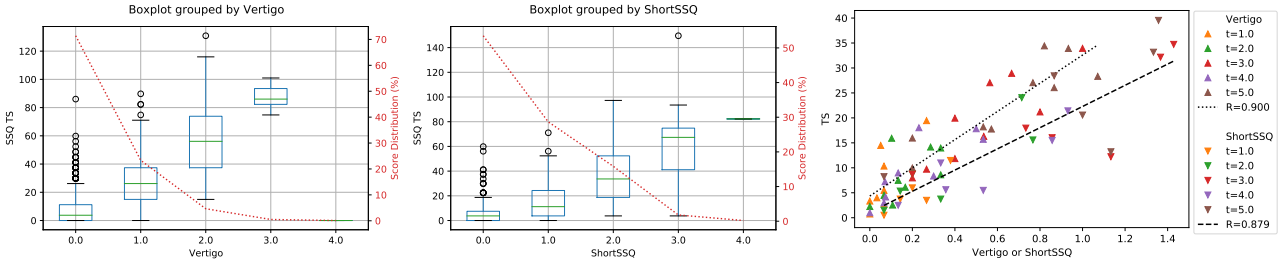
Moreover, the comparison between HTC Vive Pro with and without cables (Test G performed in Ghent) did not show any significant differences, neither from the mixed-model analysis nor from the post-hoc tests.

These results evidence that any commercial HMD (tethered or untethered) can be used in visual quality tests with 360° videos, provided that it has enough resolution and refresh rate to represent the content that is going to be tested, as included in the recommendation ITU-T P.919 [13].



(a) **(b)** **(c)**

Figure 2.5: Global results of simulator sickness: (a) Distribution of all symptoms, (b) Distribution of the total score, (c) Results on each measurement point.



(a) **(b)** **(c)**

Figure 2.6: Simulator sickness results from single-item questionnaires: (a) Boxplot of total scores grouped by the Vertigo scale [29], (b) Boxplot of total scores grouped by the Short-SSQ [30], (c) Total scores vs. Vertigo/Short-SSQ scores (average in each lab for each measurement point) and Pearson correlation coefficient.

Simulator Sickness

Test methodology

The scores collected from the widely-used SSQ [28] can be considered a ground truth for simulator sickness measurement. Thus, these results are used to analyze whether the implemented test methodologies are appropriate for simulator sickness. The distribution of all the symptoms shown in Fig. 2.5 (a), evidence that the simulator sickness of the participants was low, with only some slight/moderate symptoms. The distribution of the total scores also confirms it (computed form the evaluated symptoms according to [28]) shown in Fig. 2.5 (b), since mainly low scores were obtained. Regarding the evolution of simulator sickness during the test session, the results shown in Fig. 2.5 (c), demonstrate a positive effect of the break and no significant differences between the symptoms before and after the training.

Long vs. short questionnaires

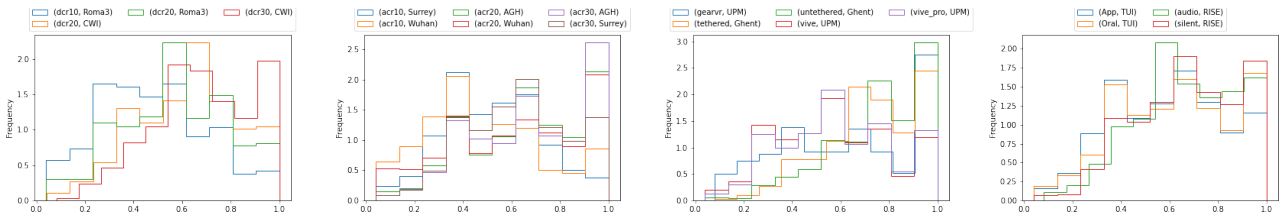
To analyze the performance of the single-item questionnaires used in the test, their results are compared to those obtained with the long SSQ, serving as ground truth. Fig. 2.6 (a) and (b) show the boxplots of the total scores (obtained from the long SSQ) grouped by the Vertigo scale [29] and by the Short-SSQ [30], respectively. In both cases, the differences among the single-item levels 0 to 3 are statistically significant ($p < 0.05$) after computing Kruskal-Wallis and post-hoc Mann-Whitney (with Bonferroni correction for multiple comparisons) tests. Also, the dotted lines represent the score distribution. As it can be seen, while the Short-SSQ provides a bit wider scores distribution (more scores in bins 1 and 2), the Vertigo scale covers a broader range of SSQ Total Score (bins 0-3 are more separated). Also, Fig. 2.6 (c) shows the correlation coefficient of the average total scores from the long SSQ with the Vertigo and Short-SSQ average scores (per lab and measurement point), 0.90 and 0.88, respectively. These results show that: *(i)* single-item questionnaires provide valid coarse-level information about simulator sickness; *(ii)* to compute the "Mean Sickness Score" for a test session (no individual scores needed), they can safely replace the full SSQ; and *(iii)* these two properties do not depend on the specific single-item questionnaire used.

To test whether all 16 symptoms of SSQ are needed to have a good understanding of simulator sickness for 360° video, three alternative sub-samplings were evaluated: the Virtual Reality Sickness Questionnaire (VRSQ) [31], the CyberSickness Questionnaire (CSQ) [32], and new factor analysis (New-FA) performed on the SSQ results of the cross-lab experiments to be used for benchmarking purposes. To obtain a similar number of items and factors as CSQ and VRSQ, New-FA considered 2-factor decomposition with *oblimin* rotation, keeping the eight symptoms with loadings greater than 0.5. The Pearson correlation coefficients between the SSQ total score and the rest of the total scores are greater than 0.9, as shown in the Table 2.4. The correlation coefficients between VRSQ and SSQ scores for the factors disorientation and oculomotor, and the total score are 0.910, 0.960 and 0.958, respectively. These results evidence that VRSQ can be a good shorter alternative to the SSQ for scenarios addressing 360° video.

Therefore, both Vertigo scale [29] and VRSQ [31] have been included in the recommendation ITU-T P.919 [13] as alternatives to the SSQ [28].

Table 2.4: Pearson correlation between SSQ total score and the rest of total scores.

Questionnaire	SSQ	VRSQ	CSQ	New-FA
SSQ	1.000	0.958	0.918	0.951
VRSQ	0.958	1.000	0.870	0.905
CSQ	0.918	0.870	1.000	0.878
New-FA	0.951	0.905	0.878	1.000



(a) **(b)** **(c)** **(d)**

Figure 2.7: Results of the participant's exploration (histograms of covered portions of the longitudinal range) of the test sequences.

2.3.9 Exploration behavior

The head rotation movements recorded through the HMD sensors while the participants watched the 360° videos allow the analysis of exploration behaviors depending on the different test conditions addressed in the experiment. The coverage results are shown Fig. 2.7, which provides information on the degree of horizontal exploration of the test contents by the participants. So, the abscissa axis represents the fraction of the sphere longitude that has been visited by them, while the ordinate axis represents how many times (as normalized frequencies) a certain portion of the sphere was visited, accounting for all participants and test videos. Thus, the right end of the abscissa axis (value "1.0") reflects the probability that the entire horizontal range is explored.

Fig. 2.7 (a) shows the coverage related to test conditions involving DCR methodology and different sequence duration. As expected, the participants explored more longer videos, as shown by the higher frequencies achieved for the exploration of the whole longitudinal range with 30-second sequences. On the contrary, with 10-second sequences the participants explored mainly less than half of the range. Generally, similar results can be seen with ACR methodology in Fig. 2.7 (b). Furthermore, the coverage related to conditions comparing different HMDs are depicted in Fig. 2.7 (c), showing that untethered devices (e.g., Samsung GearVR and HTC Vive Pro without cables) allow a wider exploration of the test sequences. Finally, Fig. 2.7 (d) shows the coverage related to test conditions involving sequences with and without audio and the two rating methods (i.e., rating app and verbal voting). On the one side, the participants explored more the silent sequences, which can be due to the fact that in those cases audio is not leading the participants' attention, especially in certain videos with characters speaking (e.g., VSenseVaude). On the other side, providing the ratings orally may allow a wider exploration of the sequences thanks to not holding the controllers, letting the participants to move more comfortably.

2.4 Conclusions and future work

This chapter presents a cross-lab study on subjective quality assessment of 360° video that was carried out within the IMG of the VQEG involving ten laboratories and more than 300 participants. The obtained results were instrumental on the development of the ITU-T Recommendation P.919. These tests allowed to analyze the influence on the visual quality ratings, simulator sickness, and exploration behavior of several factors. In particular, the tests conducted at the thesis project institution (UPM) were part of the test that sought to assess the influence of the type of HMD on subjective voting. In addition, the data collected during the sessions also form part of the analyses on the validity of a shorter version of the SSQ and of the influence of HMDs with respect to users exploratory patterns. The results with respect to the UPM test show the possibility of using any type of HMD (given minimum requirements) for 360° video evaluation. In addition, no differences in exploratory data were observed between users wearing a wired and wireless HMD. Also, methods to assess simulator sickness have been analyzed, recommending the most appropriate ones for tests with 360° videos. Finally, this work has resulted in the generation and publication of a dataset of subjectively assessed 360° content to foster future research. Future work will focus on: 1) obtaining more outcomes from the gathered subjective results with deeper analyses, 2) the study of the performance of objective metrics and the development of new models, and 3) the research on methodologies to assess other influencing factors not covered in these test, which require the use of longer 360° sequences for an appropriate evaluation, such as immersion and presence. This contribution falls within the spatial presence classification explored in Section 1.2.

Building upon this research, future work should explore new forms of video technology that offer 6DoF like volumetric video. This aligns with the need for methodologies to assess user experience in Social XR environments, as envisioned by the complete interaction diagram. In this regard, our Appendix C presents a validation study on ACR for volumetric avatar representation, which is a preliminary step towards this goal.

Chapter 3

Evaluation of Natural Interaction in eXtended Reality

3.1 Introduction

One of the advantages of using immersive technologies is the transformation of the forms of interaction. Specifically, XR technology allows us to interact with our physical environment mixed with the virtual world. One of the simplest examples is navigation. Usually, navigation in virtual worlds such as video games is done by keyboard and mouse (eg. video games). In contrast, navigation in XR is done by moving around in the physical world, which is then transferred to the virtual one. Such interfaces that map interactions as they happen in the physical world and translate them to the virtual world are called Natural User Interfaces (NUI). During the development of the thesis we decided to approach the development and evaluation of NUI from a performance and QoE perspective.

Among other activities, the thesis development involved a collaborative public-private project known as the EPSILON project. The project's core objective was to create a training platform for fiber optic construction reviewers using XR. Specifically, it aimed to replicate the training processes that reviewers typically undergo. This project posed challenges in recreating complex industrial scenarios, requiring substantial development efforts. Additionally, it provided an opportunity to evaluate XR interaction methods that are not commonly tested outside of laboratory settings.

This chapter presents the development of natural interaction methods under the industrial training paradigm. Specifically, photorealistic manipulation techniques based on video capture were developed to allow users to interact with objects in their physical reality while being immersed in a virtual environment. In addition, following this philosophy, all possible elements of the environment were adapted to manipulation without controllers while maintaining a photorealistic aesthetic. For example, the use of an augmented physical tablet, the use of floating buttons, as well as the use of photorealistic textures to generate the environment and certain objects. Finally, all developments were evaluated from the QoE point of view by using validated questionnaires to measure the overall quality, the feeling of spatial presence, and

the visual quality of the environment elements. It is worth noting that the validation was carried out in part by expert users in industrial training for this use case. Finally, as a future work, a proposal for an experiment is presented to explore the incorporation of the instructor within the immersive experience.

Here, the structure of the chapter is presented. Section 3.2 presents the related work of NUIs in XR. Section 3.3 describes the EPSILON XR training system. Section 3.4 describes an initial study conducted for validating the system and the assessment methodology. Section 3.5 describes all NUIs developed at EPSILON along with a QoE study, evaluating, among others, the impact on QoE of different forms of visual and physical representation of XR elements. Finally, Section 3.6 presents the conclusions and future work of this chapter.

3.2 Related Work

The area of XR has gained popularity in industrial training due to its immersive and interactive capabilities [33]. XR technology offers a unique opportunity for trainees to learn and practice complex tasks in a safe and controlled environment, reducing the risk of accidents and errors. Recent studies have explored the use of XR in various industrial training scenarios, including manufacturing, maintenance, and construction [34].

In our training scenario, several objects necessitate manipulation, requiring a coordinated interaction between real-world elements and their virtual counterparts in XR. Presently, methods for seamlessly blending these realities for object manipulation encompass neural networks [35] and specialized tools such as object trackers [36], [37], or haptic gloves. The visual representation of these objects can take the form of virtual entities, such as 3D models, or realistic representations like segmented images or point clouds.

Traditionally, VR setups rely on controllers for interaction within training environments, driven by the need for simplicity and cost-effectiveness [8]. However, a shift towards more intuitive and immersive interaction methods has emerged, characterized by natural interaction solutions [38], [39]. Natural interaction aims to emulate real-world scenarios, enabling more realistic and seamless user engagement. An example of this paradigm is the use of voice assistants, allowing users to interact with the system through spoken commands [40].

This thesis chapter outlines the creation of a training system tailored for fiber optic reviewers, with a focus on the incorporation of NUIs to simulate authentic training environments. Moreover, the chapter includes two QoE studies, leveraging domain experts to assess user experience and system performance. Furthermore, it proposes a future study to expand the EPSILON system into Social XR by examining how to incorporate instructor representations.

3.3 EPSILON System

This section describes the development of the EPSILON, a XR training tool designed to address the need for efficient and safe training in fiber optic construction review. Traditional training methods often require travel to various locations and may not effectively incorporate safety protocols. EPSILON tackles these challenges by providing an immersive virtual environment

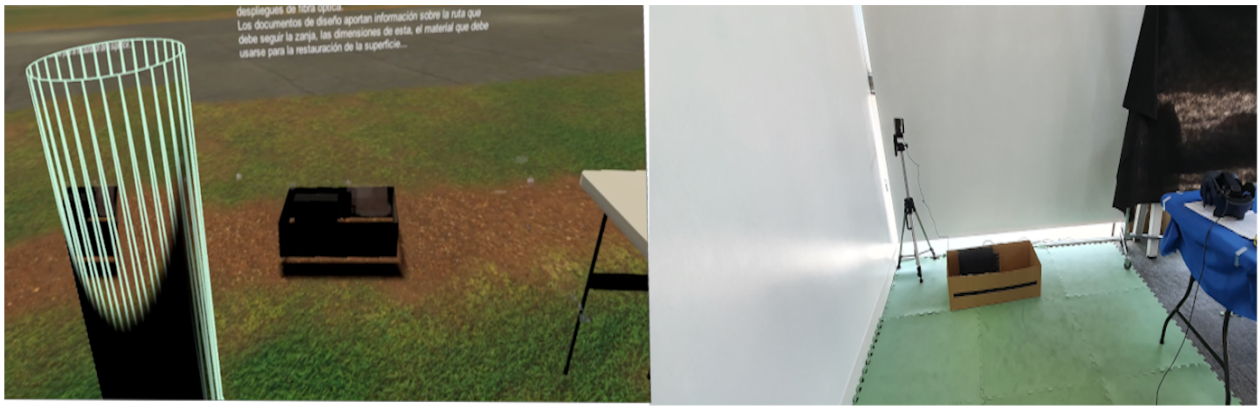


Figure 3.1: XR environment of fiber network construction area including a table with a design document and a fiber network handhole.

where users can perform training tasks without physical travel limitations. Additionally, the system integrates scenarios that simulate potential hazards, allowing users to practice safe work procedures in a controlled setting.

A core requirement for EPSILON's effectiveness was to enable NUI. This ensures users can intuitively manipulate various measuring devices commonly used in fiber optic construction, such as tape measures, GPS sticks, and a tablet, using their hands within the XR environment. The design of the interaction methods prioritizes a NUI approach to enhance the training experience and user comfort.

EPSILON's development involved two key phases:

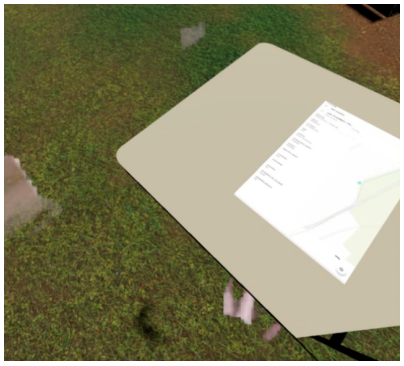
- XR Environment: This phase focused on creating virtual environments that realistically represent real-world construction spaces relevant to fiber optic installations.
- Natural interaction: This phase addressed the development of methods that facilitate natural interaction between the user and the virtual environment, allowing users to manipulate objects and perform actions intuitively using hand movements.

Figure 3.1 illustrates an example of a virtual environment within EPSILON. In this instance, users are immersed in a simulated field setting that incorporates elements commonly encountered in a physical workspace, such as a screwdriver table and a trench aligned with a box. These elements allow users to practice inspection and manipulation tasks within the virtual environment, replicating real-world scenarios. The following section details the specific hardware and software components that enable these interactions within the EPSILON XR system.

3.3.1 XR Setup

The XR architecture for the creation of the construction use case consists on hardware and software resources. Blender was used for the generation of the 3D models and the Unity engine was used for generating the XR environment as well as the task procedure. The architecture of the hardware resources is composed by the HTC Vive Pro HMD kit that includes the

(a) Scene calibration in physical reality mode.



(b) Scene calibration in virtual reality mode.

Figure 3.2: Calibration of a table using two different world representation.

HMD, the HTC controllers and the bases for tracking. In addition, to acquire the images from the physical reality, the integrated camera of the HMD is used. In the Table 3.1 there is a summary of the HTC Vive Pro specifications. The following subsections explain the use of these resources for generating the XR learning environment. Specifically, it explains how XR is addressed as well as the need for calibration of the two worlds to maintain coherence between them.

Table 3.1: HMD and Camera specifications

Screen	Dual AMOLED 3.5"
Resolution	1440 x 1600 pixels per eye
Refresh rate:	90 Hz
Field of View	110º
Camera Resolution	720×480 pixels
Camera Field of View	96º

3.3.2 Physical Environment Setup

The configuration of the actual environment is composed of two fundamental elements. The first are the green plates that will define the areas where the XR system will allow the visualization of physical environment elements through segmentation. The second will be the calibration of the physical environment with the virtual one. This calibration is carried out before starting the training. The calibration process requires a different Unity scenario with the following functionalities:

- Positioning of interactive objects at any point in the gaming area.
- Swap virtual reality for physical reality (see-thru).
- Save the calibrated positions in a JSON file.

Positioning the interactive objects into the calibrated gaming area requires a tracked device. For this purpose, we decided to use the controllers as the VR system gives you the position of

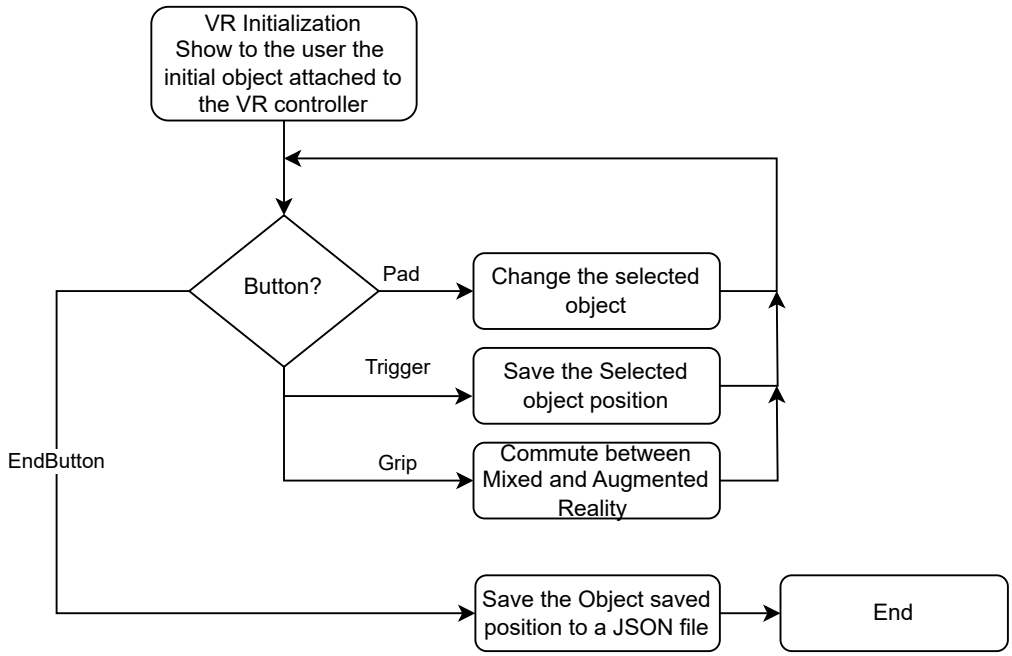


Figure 3.3: Diagram of the calibration scene process.

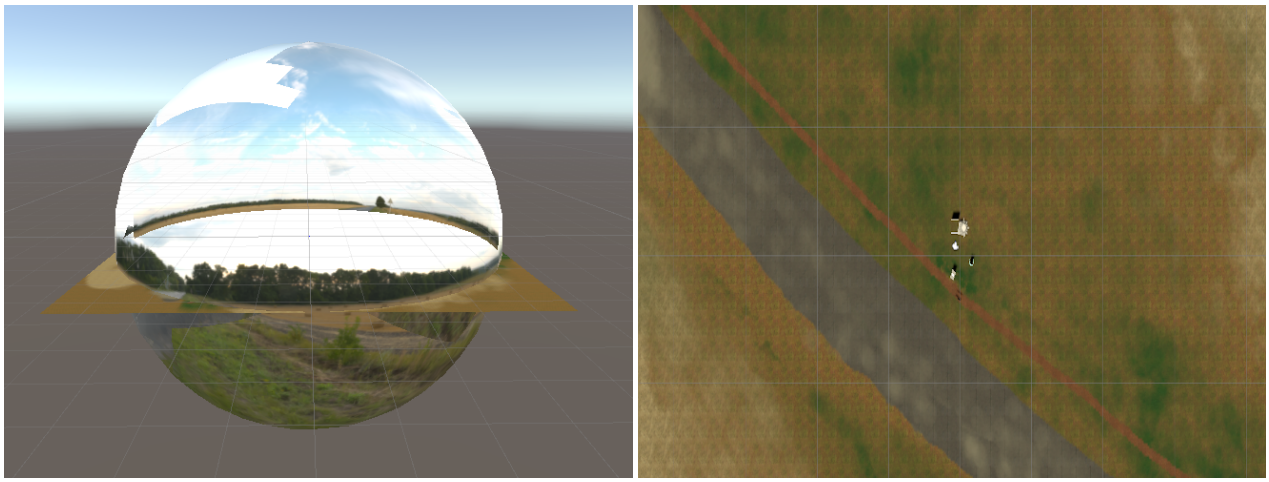
the controller with good precision. Thus, we developed a calibration scene where users can visualize the interactive element attached to the controller, save the current position of them, and switch between the virtual and real-world visualization. A diagram of the calibration scene possibilities is described in Fig. 3.3. In addition, an example of a calibrated table using real and virtual world visualization is shown in Fig. 3.2.

3.3.3 Virtual Environment Setup

The virtual world has been recreated in two different areas: the virtual representation of the training environment and the blending of the real and virtual worlds.

The virtual environment is generated using a 360° degree picture rendered on a virtual sphere. In addition, the field of the environment uses realistic textures. So both the world and the terrain are generated using real images as Fig. 3.4 shows. XR blendling uses a physical canvas placed in front of the camera that renders the virtual world. The real canvas shows the frames of the HMD integrated camera. Fig. 3.5 illustrates the relative position of the virtual camera according to the virtual world and the real canvas. This is how the physical reality is integrated into the virtual one. However, the camera frames have to be processed for including only hands and manipulable objects.

Hands and manipulable objects are included in the VR using a color-based segmentation algorithm. Specifically, the implemented chroma key is based on [9]. In that work, they used a chroma key based on red tonalities. However, due to the needs of the use case, we need to include elements without red tonalities as the measuring tape. For this purpose, we added to the physical evironment some green chroma carpets. In addition, the position of the carpets in the real and virtual world is known. Thus, during the chroma algorithm, the system checks



(a) Distant view of 360° image sphere. **(b)** Top view of the virtual scene.

Figure 3.4: Virtual world composition.

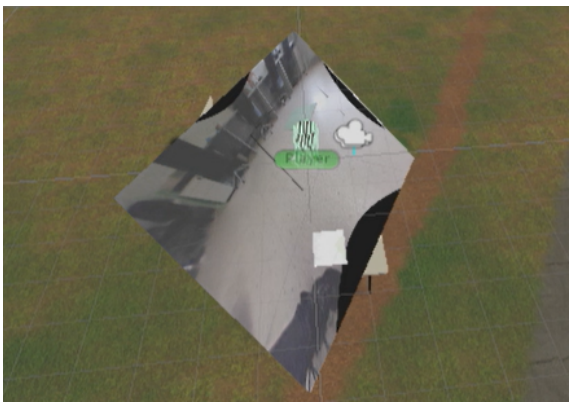


Figure 3.5: Distant view of the real canvas integrated (before chroma segmentation).

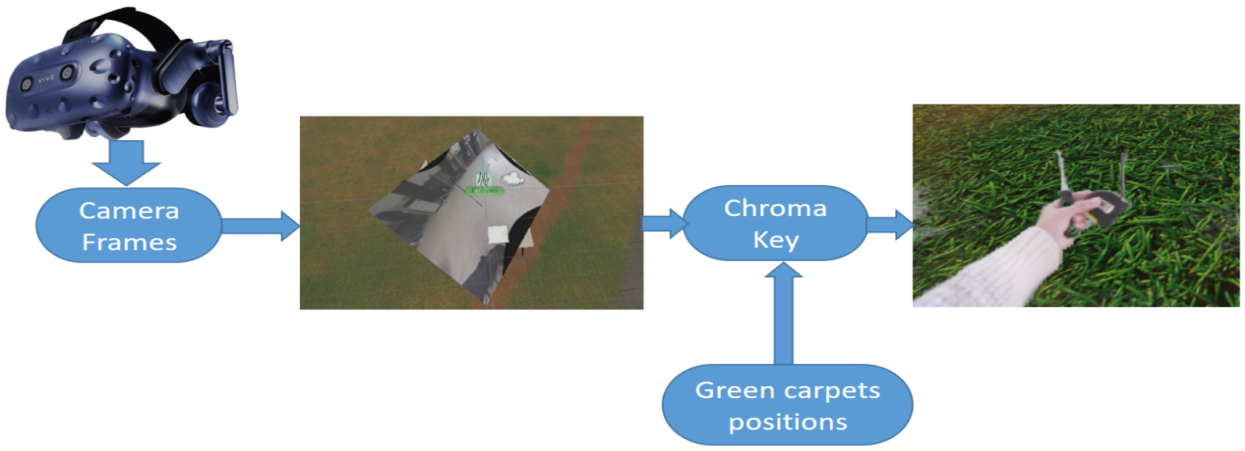


Figure 3.6: Chroma Key pipeline diagram.

whether each pixel of the real-world canvas is aligned with the calibrated green carpets or not. The aligned pixels will be filtered using green chroma segmentation and the rest with the red-based one. Fig. 3.6 illustrates the segmentation process based on the recorded position of the green carpets.

3.3.4 Complete XR Setup

The complete system setup consists of the previously described components, which are the calibration of the physical world, the design of the virtual environment and the technique for integrating frames from the cameras integrated in the headset. For more detail, Fig. 3.1 shows the complete setup of the calibrated environments for the first use case, the inspection of a trench. In the physical environment we can observe the green carpets on the ground, a cardboard box and a table. Both the box and the table are calibrated to match their positions in the virtual and physical world. In the virtual environment we can see the text that will guide the instruction and a portal that users will use to update the text.

In addition to the spatial interaction, the system incorporates several methods to allow for local interactions, the relationship between the various methods is illustrated in Fig. 3.7. An example is shown on the right in the Fig. 3.7 where a user is holding a tablet in the physical world, while in the XR environment they are able to visualize the tablet screen along with his segmented hands in the training environment. With the help of Fig. 3.7, the methods of self interaction are described below.

Pre-calibrated objects within the physical space serve a dual function. Firstly, they establish a precise spatial mapping between the real and virtual worlds, ensuring accurate virtual representations of static elements. In this respect, the position of the chroma key carpets facilitates segmentation. Then, the integrated camera captures the trainee's physical surroundings. Image processing algorithms then perform real-time segmentation, isolating objects of interest such as the trainee's hands.

Additionally, the system identifies ArUco markers attached to physical objects, like a tablet. This marker plays a crucial role – a screencasting application (http Screencast) running on

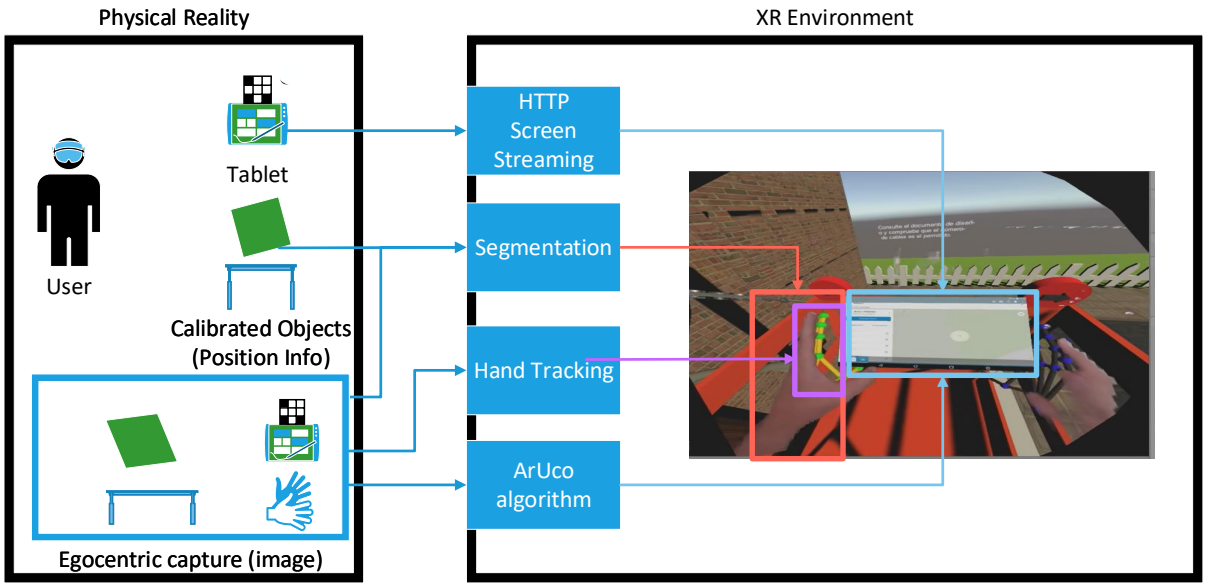


Figure 3.7: EPSILON framework showing the different processes involved in the interaction methods.

the tablet wirelessly transmits its display in MJPEG format. The ARuCO marker's position within the camera feed provides real-time spatial data, allowing the XR system to precisely overlay the casted tablet screen within the trainee's virtual environment. This integration of physical and virtual elements, coupled with real-time screencasting through marker-based positioning, allows to replicate the actual training of operators, who use tablets to view documents or perform measurements.

Furthermore, thanks to the SDK provided by HTC, the same image captured by the camera that integrates the HMD serves the Hand tracking information, which allows interaction with the XR environment. For instance, in the example of Fig. 3.7, it is possible to interact with the red buttons seen on the sides of the railing to modify the XR environment.

3.3.5 Discussion

In this section the EPSILON XR training system has been presented. Thanks to this system, training can be carried out in a specific area: the revision of fiber optic installations in different environments. The development of the tool has a special scientific interest as it proposed us challenges in the development of interaction methods. In particular, requiring natural manipulation led us to the development of egocentric segmentation methods. In addition, another challenging requirement was to incorporate smart devices (the tablet) into the scenario. All these developments were limited by the fact that the system has to work in real time. That is why the segmentation algorithm is so simple (and inaccurate) and needs help from the position of the chroma carpets, or, for example, the ArUco marker-based

Table 3.2: Questionnaire used in the experiment.

No.	Factor	Question
1	Involv.	How natural did your interactions with the environment seem?
2	Involv.	How compelling was your sense of objects moving through space?
3	Involv.	How much did your experiences in the virtual environment seem consistent with your real world ones?
4	Sens.Haptic	How well could you move or manipulate objects in the virtual environment?
5	Adapt.	How quickly did you adjust to the virtual environment experience?
6	Adapt.	How proficient in moving and interacting with the virtual environment did you feel at the end?
7	Adapt.	How well could you concentrate on the assigned tasks rather than on the mechanisms used to perform them?
8-16	Visual Quality	Please rate the perceived quality of the different scene elements
17	Sickness	Did you feel any sickness or discomfort during the experience? Please rate it
18	Global QoE	How would you rate the quality of the experience globally?
19	Usefulness	How useful this experience would be for training supervisors?

tracking system, which is very dependent on the lighting of the environment. Currently, there are solutions based on neural networks that offer better results in terms of precission than these algorithms without having to adapt the physical environment. However, it is only now that solutions such as MediaPipe, which offer fast and accurate processing, are beginning to appear. Nevertheless, it is not yet clear that these solutions can be applied with an acceptable delay for the user. During the development of the thesis, we have evaluated the validity of the complete training system by means of two subjective experiments.

3.4 QoE assessment of the first EPSILON pilot

This first pilot developed during the research work has only one form of natural interaction, the egocentric segmentation manipulation. Although we did not yet have the complete system, we decided to use this first approach to test the developments to date. The pilot study was designed taking into account the methodological guidelines presented in Chapter 2. The tool was evaluated by 14 people (7 female and 7 male), all of them working in construction of fiber networks (including civil works). Users had to perform a step-by-step tutorial in a construction environment. At the end of the experience, each user had to answer a questionnaire for measuring the sense of presence, the visual quality of elements, and their global opinion. The objectives of this first study were:

- Evaluate the satisfaction of training experts with the tool.
- Detect possible improvements in the system for the final environment.
- Test methodology more focused on interaction while maintaining the recommendations proposed in Chapter 2.

3.4.1 Experimental Setup

Before the experiment, the experimenter told them about the procedure of each task step. Specially, they were told how to enter to the application portal after each task so they would complete the experiment. Once the user finished a task, they had to enter to the application portal for updating the text canvas that told them the next step. The experiment had the following order:

1. The user approaches to a table to read the design document.
2. Then, the user will review the material used to reconstruct a handhole.
3. Finally, the user will have to measure the handhole width and check it according to the desgin doccument.

3.4.2 Questionnaire

After finishing the experience, users had to fill a questionnaire evaluating its more relevant QoE aspects: sense of presence, visual quality of the different elements, simulator sickness and global QoE [9]. The questionnaire included a total of 19 items, searching for a compromise between its extension and the diversity of QoE factors considered (Table 5.6).

Presence was measured using a sub-sampled version of Witmer's Presence Questionnaire [41], selecting 7 items which have shown good correlation with the results of the full version in total presence score and several sub-scales: involvement, adaptation, and haptic sensory fidelity [42]. All items were evaluated in a Likert-like 7-level scale.

Visual Quality questions requested subjects to individually assess the perceived quality of all the relevant elements in the scene. Most of them were virtual (grass, trench, road, handhold, table, document), one was a 360° picture (landscape), and the others were obtained from the HMD camera (tape, hands). All of them were evaluated using Absolute Category Rating (ACR) scale. Simulator sickness was evaluated using Vertigo Score Rating (VSR) scale [29]. Both ACR and VSR are Likert-like 5-level scales, recommended for the evaluation of immersive experiences by ITU-T P.919 [43].

Finally, two global QoE items were included: a global evaluation of the experience in ACR scale and an evaluation of the perceived usefulness of the method for its designed purpose: training construction supervisors. Both use Likert-like 5-level scales. Additionally, users could provide free-form written feedback about the experience.

3.4.3 Results

This subsection presents the results of the first study of the epsilon system. All values relating to questionnaire averages have been adapted to a scale of 1 to 5. For the overall quality and recommendation results, a histogram of the scores is presented, also from 1 to 5.

Presence

Presence results are shown in Fig. 3.8. The three presence factors covered in the studio show similarly high scores. Values are similar to the ones reported in [9] for the same measures, where also a high degree of sense of presence was reported (especially in comparison with not using natural interfaces for the interaction).

Visual Quality Fig. 3.9 shows the average visual quality of the evaluated elements. The elements that are stored together with the 3D engine are drawn in blue while those that are rendered from the camera images are presented in orange. Most of them are reported as *good* quality (4), with some of them just *fair* (3). None is reported as bad-or-worse (≤ 2). The

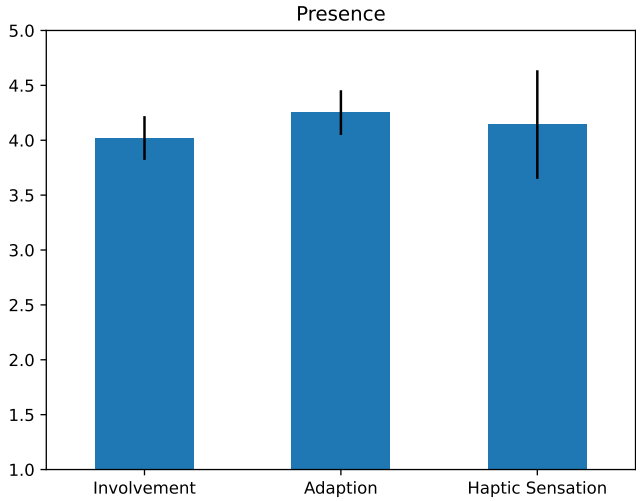


Figure 3.8: Presence results averaging the presence factors. Mean and 95% confidence interval are represented.

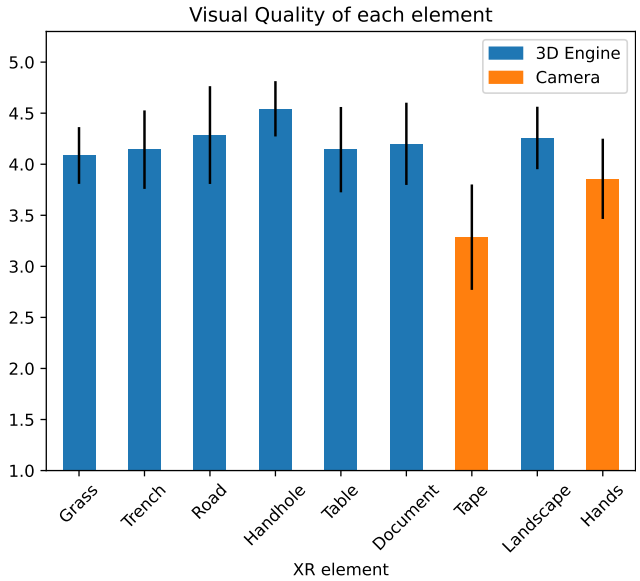
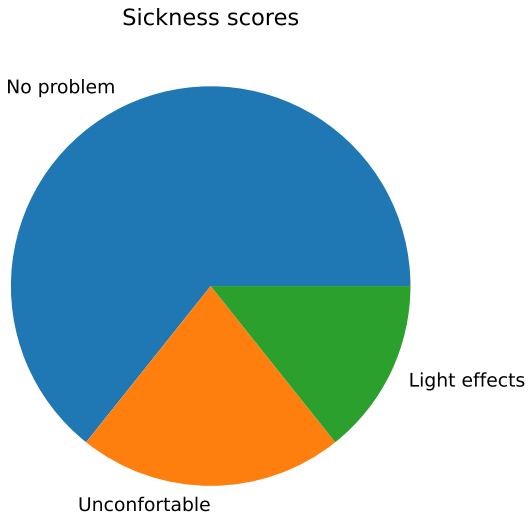
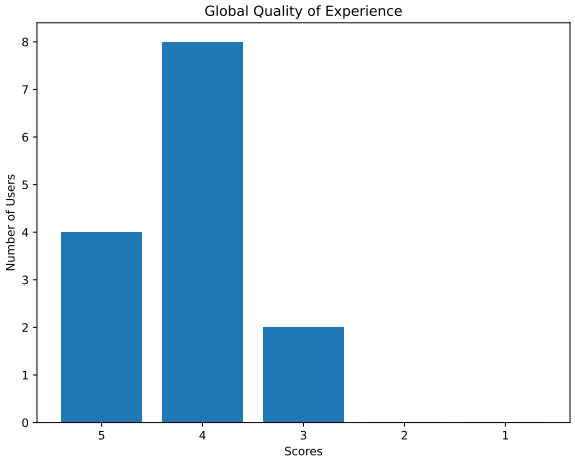


Figure 3.9: Average scores of each element visual quality. Mean and 95% confidence interval are represented.



(a) Pie chart of the simulator sickness question scores.



(b) Bar chart of the global QoE scores.

Figure 3.10: Results of the simulator sickness and global QoE scores.

worst quality was detected in the measuring tape, which was captured from the HMD camera. Surprisingly, the other captured element (hands) was perceived with similar quality as the virtual elements. This difference is probably explainable because the tape has a more detailed texture than the hands, and also because reading the tape measurements was part of the assigned task, therefore making users more sensitive to its visual quality.

Simulator Sickness From the 14 subjects participating in the experiment, 9 showed no simulator sickness symptom ($VSR = 5$), 2 reported light effects ($VSR = 4$) and 3 reported feeling uncomfortable ($VSR = 3$), as depicted in Fig. 3.10a. None reported severe symptoms ($VSR \leq 2$).

Global Quality and Recommendation

Figs. 3.10b and 3.11 show the distribution of scores of both QoE questions. Most users reported that the quality of the experience was good or excellent ($QoE \geq 4$) and none rated it as poor or worse ($QoE \leq 2$). Most users also found the method to be useful as a method to train construction supervisors. Since training construction supervisors is one of the main activities performed by their department and, in particular, the one for which the system has been designed, these results suggest that this technology can effectively be suitable for the purpose.

3.4.4 Conclusions and Future Work

In general, the measurements of the presence and the overall quality showed that the system had a good acceptance. In the light of the results shown in Fig. 3.8 of the factors measured with the questionnaire of Table 5.6, presence levels showed that the users felt a good adaption and haptic sensation, also, they felt quite involved in the experience.

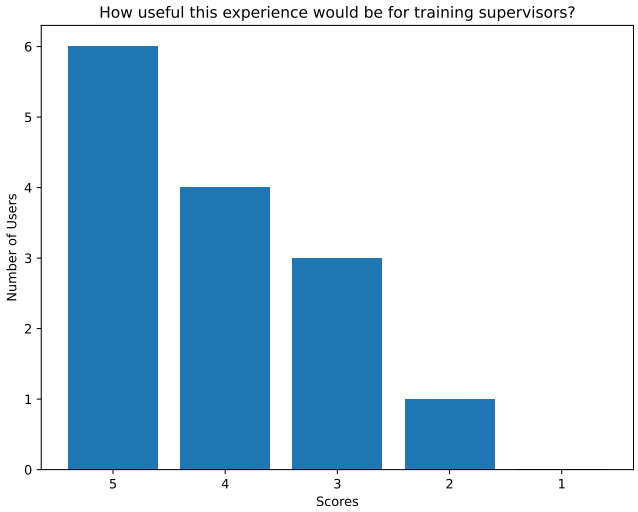


Figure 3.11: Scores of the recommendation question.

The visual quality of the objects was mainly ranked well. However, the measuring tape was the worst-ranked. Comparing the two objects that were introduced into the XR using the segmentation algorithm we can see that the hand scores better than the tape. The camera resolution is not good enough to see the details of the measuring tape. However, as the texture of the hand is very simple, the scores of the hand's visual quality weren't affected by this issue.

Our results suggest that the system can be used for its designed purpose: training of workers (particularly supervisors) in construction works. The need for some improvements was detected, mostly related with the visual quality of some virtual and real elements. Further research is also needed to validate the technology with more users and more scenarios. With this pilot test we were able to verify that the NUI-based manipulation provided good results in terms of QoE, both in terms of immersion and overall quality. In addition, we were able to validate that our setup did not produce simulator sickness. This pilot study laid the foundation for the next study that we developed during the thesis project. Specifically, our new study involved testing a wider range of interaction interfaces across various use cases, resulting in a more comprehensive training experience.

3.5 Natural Interfaces Evaluation

In the previous sections we have introduced how the egocentric segmentation system works along with a first experience to evaluate the training system. With the previous experiment we laid the groundwork for what would be the evaluation of the complete system.

The entire system involved different forms of interaction due to the requirements of the training session. These requirements included:

1. Handling of real segmented objects,
2. The use of segmented hands,

Figure 3.12: Virtual and real environment for industrial training using natural interaction.

3. The use of smart devices within the app, and finally
4. Users has to be able to realistically visualize the entire environment.

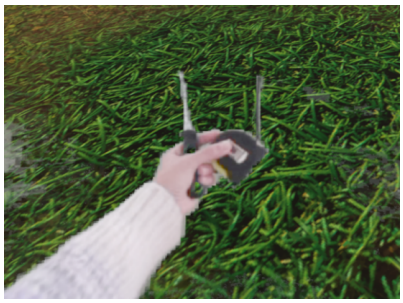
All these developments were done from the perspective of natural interfaces in XR.

As stated in the Fig. 1.1, the XR can be understood as an umbrella that brings together different mixes of virtual and physical realities. Under this context we can go from a purely virtual environment to one in which we perceive physical reality as it is. Our classification of XR interactions is based on whether elements exist in the physical reality and their visual representation in the virtual environment. This classification results in four categories: real objects with realistic representation, real objects with virtual representation, virtual objects with realistic representation, and virtual objects with virtual representation.

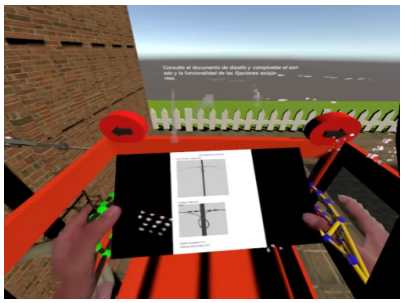
Interaction with Real Objects and Real Representation Manual interaction with real objects is achieved through egocentric image segmentation. Frames are captured from the camera's egocentric perspective, and a segmentation algorithm is applied based on color and user pose, as explained in detail in [9] and [44]. This chroma algorithm takes images captured from the front cameras that the HTC Vive Pro HMD incorporates. These frames are positioned appropriately in front of the user in the virtual environment. Afterwards, a chroma algorithm is applied that allows only the pixels of the frame related to the hands and manipulable objects to be displayed on the virtual environment. Fig. 3.6 illustrates this pipeline.

Interaction with Real Objects and Virtual Representation.

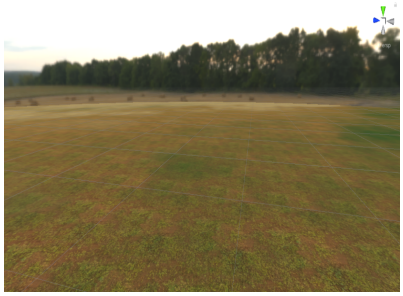
In some cases, interaction with elements in a simulation may require altering their physical appearance in reality due to cost or logistical constraints (e.g., changing the visual appearance of a VR controller to resemble a sword). To achieve this, real-world objects can be visually augmented and synchronized with virtual elements, enabling simulated tactile experiences by modifying the object's appearance in the VE. According to the virtuality continuum, this fits within the AV classification. Successful augmentation necessitates synchronization between real and virtual objects, particularly in the manipulation area. Synchronization methods include optical trackers like VIVE controllers and trackers [45] or fiducial markers such as



(a) Real Object and Realistic representation.



(b) Real Object and Virtual Representation



(c) Virtual Object and Realistic Representation.

(d) Virtual Object and virtual representation.

Figure 3.13: Examples of XR elements according to the visual representation.

ArUco for pose estimation [46]. The training system includes three objects in this category: the table, the GPS antenna, and a tablet. These objects are calibrated, coordinated, and synchronized between the real and virtual environments. The virtual tablet is even streamed to the VR engine, enabling interaction with the real tablet in the XR environment (see Fig. 3.12).

Interaction with Virtual Objects and Realistic Representation. Certain elements that exist only in VR are designed as realistically as possible in visual terms. This category includes photorealistic textures for elements such as terrain or background and 3D-captured objects. According to the virtuality continuum classification in [1], these elements fall under VR as they exist solely in the psychic environment. Fig. 3.13c provides an example of photorealistic textures used in the training system.

Interaction with Virtual Objects and Virtual Representation. The simplest case involves using virtual elements whose models are the result of 3D rendering. In the virtuality continuum, this is classified as VR. Examples of this category can be found in houses or buildings (see Fig. 3.13d).

3.5.1 Validation of Natural Interfaces for Local Interaction in XR

To ensure the success of immersive learning environments, user experience is a crucial element that needs to be evaluated. Therefore, we conducted an experiment to assess the level of the quality experienced by users in a construction-based learning environment. The study involved 8 participants, including 1 females and 7 males, all of whom had experience in reviewing fiber networks installations. The users were required to complete four use cases tasks (step-by-step

tutorials) in different construction environments, after which they were asked to fill out a questionnaire to measure their sense of presence, the visual quality of certain elements (see Table 3.3), and their overall opinion of the experience. This subjective assessment allowed us to validate the learning environment and ensure that it was conducive to a high level of immersion and performance. The objectives of the study were:

- Check whether natural interaction methods are appropriate for training tasks.
- Check the differences in terms of QoE regarding interaction methods
- Evaluate trainer satisfaction with the tool.

Table 3.3: Object classification according to its representation and physical being.

Object	Representation	Elements	Use cases	Virtuality
Real	Realistic Representation	Hands, Meter	1,2,3,4	Augmented
	Virtual Representation	Table, GPS, Tablet	1,2,3,4	Augmented
Virtual	Realistic Representation	Trench, Riser, Room, Box, Paviment, Grass, Road, Landscape	1,2,3,4	Virtual
	Virtual Representation	House, Design Documment, Buildings, Telescopic Arm	1,2,4	Virtual

Equipment

The equipment for this experiment consists of a PC running Windows 10 and an editor of the unity engine, a HTC Vive Pro HMD, a HTC Vive Pro controller, a Samsung Galaxy Tab A8 tablet, a table with a blue tablecloth and some green plates for the floor. The PC is in charge of rendering the virtual world, processing also the images captured by the front camera of the HMD. The tablet is used for some use cases to train in the use of a real tablet. The table and plates are used to assist the segmentation algorithm.

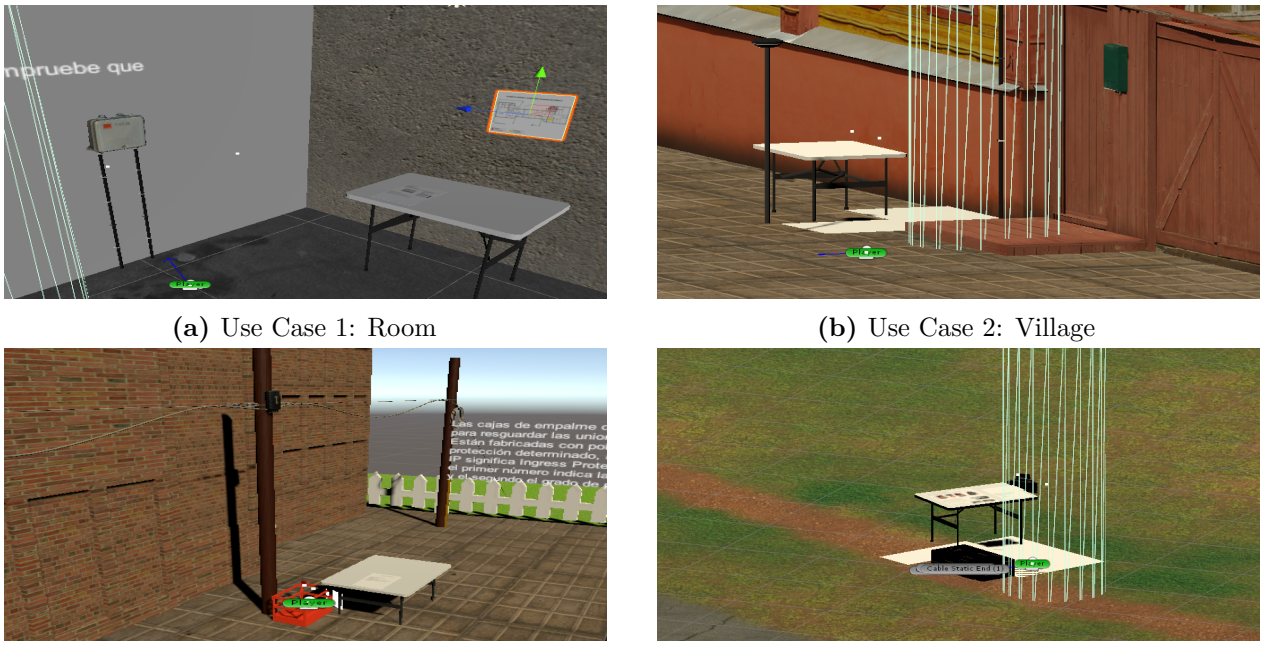
Experimental Conditions

The experimental conditions consist of a user familiar with the area of fiber construction review who will perform four use cases. Each use case is performed sequentially and in the same order for all users. Although normally in subjective evaluation it is necessary to alter the order to avoid bias between conditions, in this case, it was decided to order the use cases from least to most difficult to replicate the actual training that was being performed without immersive technology.

During the experiment, users were accompanied in the physical reality by an instructor, who tried to solve any doubts the user had during the task. Once the user had completed a use case, a quality questionnaire was completed in which different aspects of visual quality, interaction and performance were evaluated.

Stimuli

The experiment stimuli consist of four distinct use cases, each of which takes place in a different virtual environment. These use cases serve as central scenarios in which participants interact with various interaction methods. Each use case is designed according to real training situations. Adapting each interaction modality to the actual training needs. This diversity



(a) Use Case 1: Room

(b) Use Case 2: Village

(c) Use Case 3: Urban

(d) Use case 4: Field

Figure 3.14: Use cases.

of virtual environments allows a thorough evaluation of these methods in scenarios where they are actually needed. The four scenarios developed are detailed below, together with a description of the training tasks to be performed by users in each use case.

Use Case 1: Room. In this scenario, users find themselves immersed in a room containing a table, a splice box, and a design document (refer to Fig. 3.14a). Their task is to ensure that the installation of the splice box aligns with the instructions provided in the design document. This review process encompasses checking labeling and the installation of fiber optic cables.

Use Case 2: Village Environment. Within this use case, users are placed in a village setting, standing in front of a house with a table (refer to Fig. 3.14b). Users must inspect a floating text and then retrieve a physical measuring tape placed on the same virtual table as in Use Case 1. Following this, they are instructed to measure the distance between the entrance and the door (as depicted in Fig. 3.14b).

Use Case 3: Urban Environment. In this use case, users are elevated to the height of fiber optic pole connections (refer to Fig. 3.14c). Their objective is to review the information presented on a tablet, which aids them in their task. Users interact with the physical tablet using their hands and the response is rendered and streamed to the VE using the casting app. Additionally, a hand tracking algorithm is implemented, allowing users to activate virtual buttons by placing their real hands on top. The task entails reviewing information on the tablet pertaining to the labeling and installation of the pole cable. Users conclude the session using the buttons once they have reviewed all the tasks outlined on the tablet and the informative text.

Use Case 4: Field Environment. The final use case encompasses a broader range of

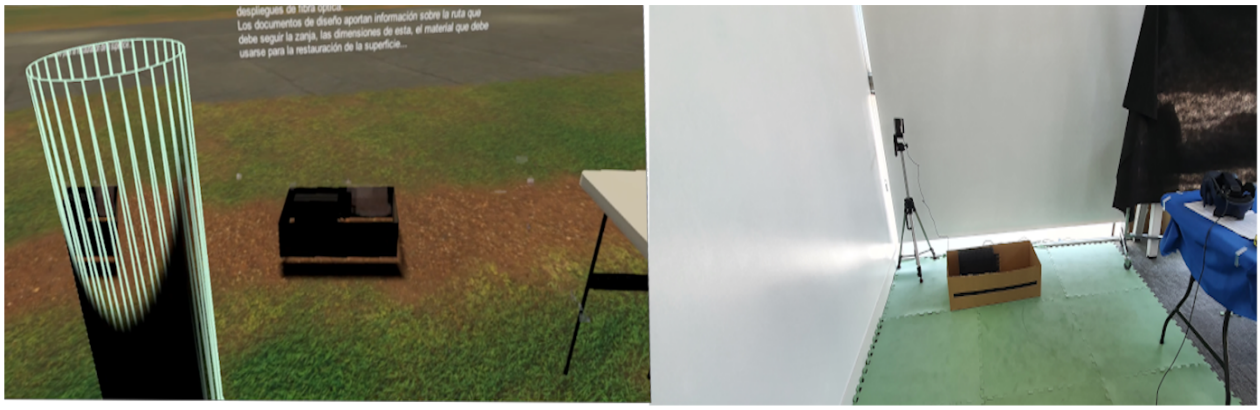


Figure 3.15: XR environment of fiber network construction area including a table with a design document and a fiber network handhole.

actions and elements, placing users in a field environment (see Fig. 3.14d). In this scenario, users follow a floating text to pick up a real tablet from a physical table. These tangible objects have virtual counterparts within the virtual environment, and users use segmented hands to interact with them. After acquiring the tablet and a GPS stick from the table (in the physical environment, users grab the controller), they must use the tablet to perform measurements with the GPS stick using an application installed on the tablet [47].

Experiment Setup

The experiment takes place across two levels: virtual and physical. In the virtual environment, we implement the use cases described in Stimuli subsection. In the physical environment, we have elements that support the experiment and replicate the training scenarios. Some elements bridge the gap between the physical and virtual worlds, such as the computer, table, HMD, and green plates, while others exist primarily to enable the virtual scenarios. To maintain consistency, we employ HTC Vive Pro traces for calibration, ensuring users always start from the same physical position. This calibration aligns physical and virtual spaces precisely; for instance, in Scenario 1, a user measuring a trench can trust that its physical size matches the virtual representation. In Fig. 3.15 we can see this trench in the virtual environment that has a physical representation in the form of a cardboard box.

Methodology

Participants Test Session Questionnaire

After the virtual experience, participants completed a 19-item questionnaire to assess the QoE. This questionnaire covered aspects like their sense of presence, visual quality, susceptibility to simulator sickness, and overall QoE. To measure presence, we used a sub-sampled version of Witmer's Presence Questionnaire [41], containing seven questions that aligned with three factors: Involvement, Haptic Sensation, and Adaption. Participants rated these questions on a 7-point scale (Q1-7). Visual quality was evaluated by having participants individually assess the quality of various elements within the virtual scene, as detailed in Table 5.6. Simulator sickness susceptibility was gauged using a 5-point scale, following ITU-T Rec. P.919

No.	Factor	Question
1	Involv.	How natural did your interactions with the environment seem?
2	Involv.	How compelling was your sense of objects moving through space?
3	Involv.	How much did your experiences in the virtual environment seem consistent with your real world ones?
4	Sens.Haptic	How well could you move or manipulate objects in the virtual environment?
5	Adapt.	How quickly did you adjust to the virtual environment experience?
6	Adapt.	How proficient in moving and interacting with the virtual environment did you feel at the end?
7	Adapt.	How well could you concentrate on the assigned tasks rather than on the mechanisms used to perform them?
8-16	Visual Quality	Please rate the perceived quality of the different scene elements
17	Simulator Sickness	Did you feel any sickness or discomfort during the experience? Please rate it
18	Global QoE	How would you rate the quality of the experience globally?
19	Usefulness	How useful this experience would be for training supervisors?

Table 3.4: Questionnaire used in the experiment.

guidelines [43]. Additionally, the questionnaire included two global QoE items: an overall experience evaluation using an ACR scale and an assessment of the method's usefulness for training construction supervisors on a 5-point scale. Finally, participants had the opportunity to provide open-ended feedback about their experience using the system.

Results

This section presents the outcomes of the QoE study, encompassing an evaluation of presence, visual quality, simulator sickness, overall quality, and recommendation.

Presence

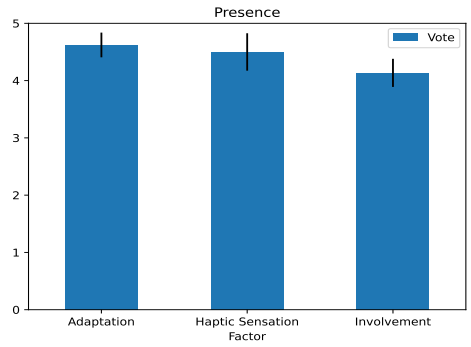
Presence results are illustrated in Figure 3.16a. Notably, the three presence factors examined in the study consistently received high scores, all surpassing 3.5 . These scores are comparable to those reported in [44], where a substantial sense of presence was also observed. Moreover, on average, our results exhibited an improvement, indicating that the introduction of new interaction methods not only expanded the scope of training use cases but also enhanced the sense of presence [44]. Visual quality findings, aggregated by groups of object representations, are depicted in Figure 3.16b. After establishing the normality of the data, an analysis of variance revealed a significant impact of representation methods on Visual Quality ratings ($p = 0.003, \eta^2 = 0.072$). Subsequent Tukey posthoc analysis confirmed significant distinctions between realistic representations of real objects and both types of virtual object representations. The results suggest that while interactions scored favorably, Real Objects with Realistic Representations had the lowest visual quality, possibly due to limitations in egocentric capture resolution.

Global Quality and Usefulness

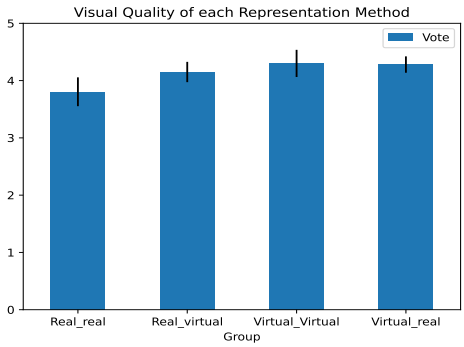
Figures 3.16c and 3.16d display scores for Global QoE and Usefulness. Most users reported favorable QoE scores ($QoE \geq 4$), with none rating it as poor or worse ($QoE \leq 2$). Similarly, a majority of users found the training method useful for construction supervisor training, aligning with the system's primary purpose. These results suggest that the technology effectively serves this training role.

User Feedback

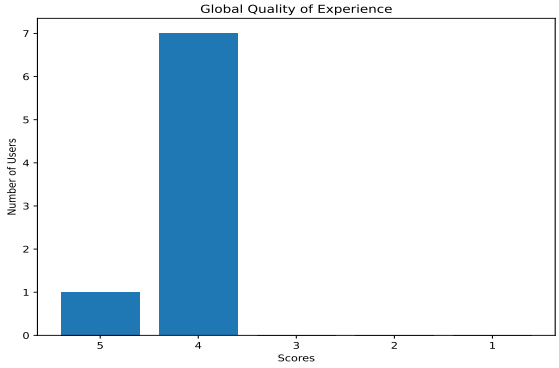
Users' experiences with the virtual training tool for supervisors yielded promising results,



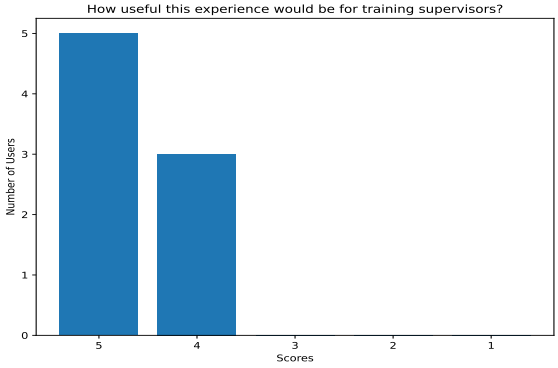
(a) Presence results averaging the presence factors. Mean and 95% confidence interval are represented.



(b) Average scores of each group of object and representation method. Mean and 95% confidence interval are represented.



(c) Bar chart of the global QoE scores.



(d) Scores of the usefulness question.

Figure 3.16: Graphs of the results of the study.

albeit some adjustments may be necessary to optimize its utility as a comprehensive training solution. Participants appreciated the visual fidelity of the training, enabling them to learn within a VE that closely mirrors real-life scenarios. While the tool excels in teaching simpler tasks, more complex scenarios may benefit from additional training methods. Nevertheless, it represents a valuable addition to supervisor training programs, offering a cost-effective, realistic training experience.

Simulator Sickness
Among the 8 participants, 7 reported no simulator sickness symptoms, while 1 reported experiencing mild effects.

3.5.2 Discussion

In this section, we have delved into the suitability of various interaction methods within XR. Specifically, these methods are grounded in the idea of facilitating natural interaction with both physical and virtual environments. Our classification scheme categorizes these interaction methods based on their visual representation and the presence or absence of corresponding physical objects.

Firstly, we examined interaction methods involving physical objects with photorealistic representations. It was evident that these methods excel in facilitating accurate interaction with physical elements, enabling manipulation of instruments like meters and tablets. This enhancement significantly enriches training and interaction experiences involving physical objects that are challenging to replicate faithfully in virtual reality. However, it's worth noting that the current limitations in egocentric camera technology, particularly concerning resolution, need to be addressed to ensure optimal visualization. This technological challenge is expected to be resolved with time.

Next, we evaluated interaction methods involving physical objects but with virtual representations within the virtual environment. This approach enables the augmentation of physical objects to resemble different entities in the virtual realm. For instance, a controller can represent a GPS antenna, offering a cost-effective alternative to using real, more expensive counterparts. The results indicate that, in general, these methods yield favorable outcomes in terms of quality and presence. This augmentation of physical objects with virtual counterparts not only enhances presence but also offers practical cost-saving benefits.

Moving forward, we explored interaction methods centered on virtual elements with virtual representations—purely virtual objects. The findings reveal that these methods generally perform well, underscoring their effectiveness in providing immersive experiences.

Lastly, we investigated interaction methods involving virtual objects with realistic representations, such as photorealistic environments. Users' feedback indicated that these methods notably contribute to immersion—an overarching goal of immersive technologies.

However, it's important to acknowledge certain limitations in these developments. Firstly, the validation of these techniques has been conducted with a relatively small number of users, potentially limiting their generalizability to larger and more diverse user populations. Secondly, the evaluation of these methods has largely taken place in separate experiments,

rather than within integrated, real-world XR systems. This separation of evaluations may not fully capture the complex interplay between local and social interactions. Lastly, the performance of these developments in real network scenarios, where factors such as latency and bandwidth constraints come into play, remains uncertain. Further research and testing in these areas will be crucial to ensure the robustness and applicability of these techniques in practical Social XR implementations.

In summary, our discussion sheds light on the diverse spectrum of interaction methods in XR and their respective merits. These findings provide valuable insights for future developments in XR technology, offering potential advancements in enhancing training and immersive experiences across various domains.

3.6 Conclusions and Future work

The EPSILON project successfully explored the integration of NUIs within a XR environment designed for a specific real-world application: fiber optic construction review training. This project not only focused on technical development but also incorporated assessments of user experience through QoE studies.

A critical lesson learned during the development process pertains to the importance of fidelity when introducing real-world elements into a virtual space for realistic interaction. Specifically, these elements, such as virtual representations of tools and equipment, need to closely match their physical counterparts in terms of visual representation and user feedback mechanisms. This fidelity is essential to maintain user immersion and prevent a sense of alienation from the virtual environment. Conversely, for entirely virtual elements with no real-world counterparts, user expectations for fidelity may be lower.

The findings from the EPSILON project contribute to the ongoing development of immersive training tools utilizing NUIs. Future research can explore the optimal balance between fidelity and usability for various NUI implementations, particularly within the domain of construction training

In relation to a previously established gaps in Fig. 1.3, this work contributes in two key ways. Firstly, the methodology applied in our QoE studies is being proposed for use in an ongoing interlaboratory experiment focused on assessing the QoE of interactive Social XR communications. This demonstrates the potential of our approach for broader applications within the XR field. Secondly, the project has provided valuable insights into the limitations of current XR systems. Notably, the accuracy of the implemented NUI developments was limited by two technical factors: the resolution of the built-in cameras and the delay associated with local reality processing. Our research highlights that delay is a particularly critical factor when XR systems aim to achieve a social dimension, as in Social XR applications. Addressing these limitations will be crucial for the continued advancement of immersive and interactive XR experiences.

According to the interaction classification detailed in Section 1.2, this work focused on the Spatial and Self areas of our Social XR interaction classification. A natural progression for future work would be to extend this research by incorporating social interaction within the

EPSILON system. An appendix to this thesis proposes a detailed experiment to evaluate the impact of social interaction on user experience within the context of the EPSILON project. The findings from this proposed experiment would provide valuable insights into user expectations and requirements for Social XR training environments. This would further contribute to the development of a comprehensive framework for evaluating UX in XR systems.

Chapter 4

Effects of the Delay on the Interaction in Social XR

4.1 Introduction

During the development of the thesis we have studied and developed interaction methods under the XR paradigm. These developments, both 360° video and immersive training environments, showed how different technical factors of the systems affected QoE. These technical factors must be to ensure the correct mapping of interactions in the physical reality to the virtual one.

With the knowledge acquired in the developments of Chapter 3, we were able to determine that, when developing NUIs in XR, delay was the technical factor that conditioned all methods. In this regard, we found that the current Social XR systems have been tested mainly under ideal laboratory conditions. Therefore, only low and very stable delays have been considered [48]–[50]. Additionally, the power limitations of the headsets and the resource-intensive nature of algorithms relying on neural networks add to the complexity of delivering a Social XR experience based on NUIs. As a response to these constraints, the concept of offloading has gained traction, where rendering and information processing are handled in the cloud. While this approach aims to alleviate the burden on local hardware, it introduces even more latency to the systems.

The impact of delay on the user experience resulting from the shift from local to remote processing, and the various processes affected by this latency, remain relatively unexplored territory. Addressing these gaps in our understanding is crucial for the successful development and adoption of Social XR technologies. In this chapter, we present a comprehensive framework for Social XR communications, aiming to bridge the gap between theoretical promise and real-world applicability, while unraveling the intricate interplay between technology and human interaction in this area.

Here, the structure of the chapter is presented. Section 4.2 presents the related work of the Social XR systems and presents a common framework that summarizes all the processes involved from a delay perspective. Section 4.3 presents a classification of delays according to

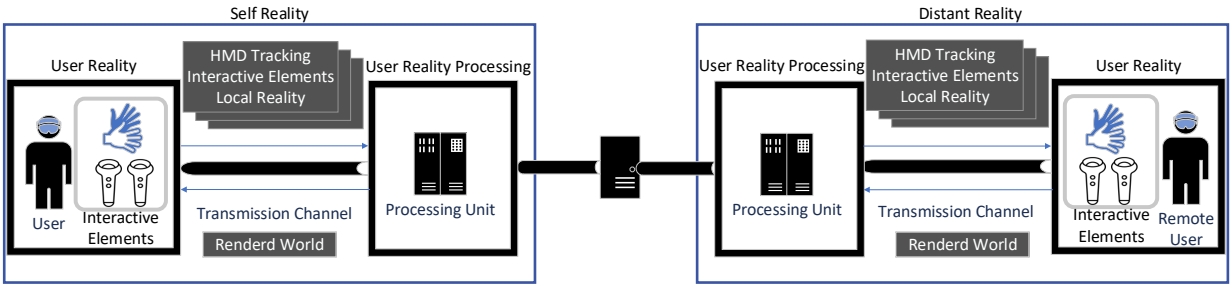


Figure 4.1: Social XR Framework.

their influence factors on perception, their location in the processing framework and its QoE perceptual implication. Section 4.4 presents a state-of-the-art analysis of the physiological implications of these delays on users. Besides, the section presents a new QoE prediction model for immersive use cases. Finally, Section 4.5 presents the conclusions and future work.

4.2 Related work

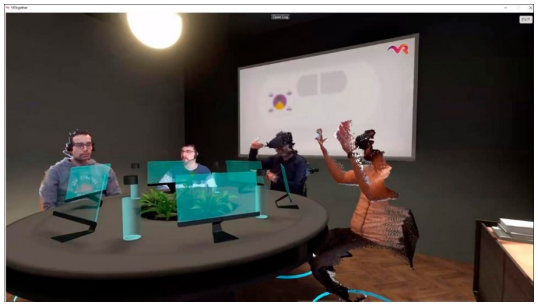
Social XR is inherently linked to social interaction, making it crucial to examine use cases that involve users being transported to remote environments for social engagements. Consequently, Social XR must incorporate a communication system built on immersive communication techniques. This requires the use of devices like HMDs along with the processing of an interactive shared world. To enable interaction within this shared world, audiovisual and haptic methods are essential. These functionalities are integrated through various processes to create the immersive environment. Research papers in the field of immersive communications systems frequently incorporate diagrams that illustrate these processes. For instance, in [48], a process diagram is featured, elucidating the 360° video-based immersive communication process. This diagram encompasses various stages, including camera capture, transcoding, and rendering. Likewise, [49] offers a process diagram that centers on volumetric video streaming. This particular system leverages a combination of depth and color cameras for the generation, transcoding, and rendering of volumetric video. Moreover, [51] provides a process diagram pertaining to free-viewpoint video communications, which relies on camera arrays. Figure 4.1 depicts the proposed framework. It encompasses two distinct levels for processing and transmitting information across users' realities. The first level is referred to as Self Reality, which encompasses all the necessary processes for interaction with one's own self and the surrounding physical environment, including controllers and nearby objects. Furthermore, the data collected within the self reality is disseminated to different users' realities to construct the Social XR.

The second level of processing involves the capture and processing of elements within the Distant Reality. Additionally, our framework accounts for the interconnection between these two realities, specifically the transmission channel linking each client User Reality Processing unit.

The User Reality Processing unit bears the responsibility of handling the information originating from the user's reality, such as the user's head and controller positions, as well as any

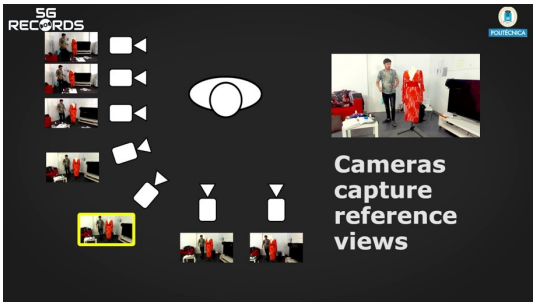
(a) 360° camera (Owl)

(b) 3D Avatar (Mozilla Hubs)



(c) Pointcloud user capture (VR2Gather)

(d) Webcam Integration



(e) Free-view point video

Figure 4.2: 3D models to reproduce during the task.

available data from other realities. This processing unit combines these inputs to compose the shared virtual environment in which users will become immersed.

Therefore, our proposed framework offers a valuable tool for immersive communication developers seeking to comprehend the constituent processes driving latency within their systems, pointing potential bottlenecks and enhancing their understanding of latency impact on user experience.

4.3 Delay in the Social XR

The generation of the Social XR involves a multitude of processes that can potentially lead to latency and subsequently influence the QoE for users. In particular, users are likely to detect latency in their actions once the virtual environment appears in the HMD. This phenomenon

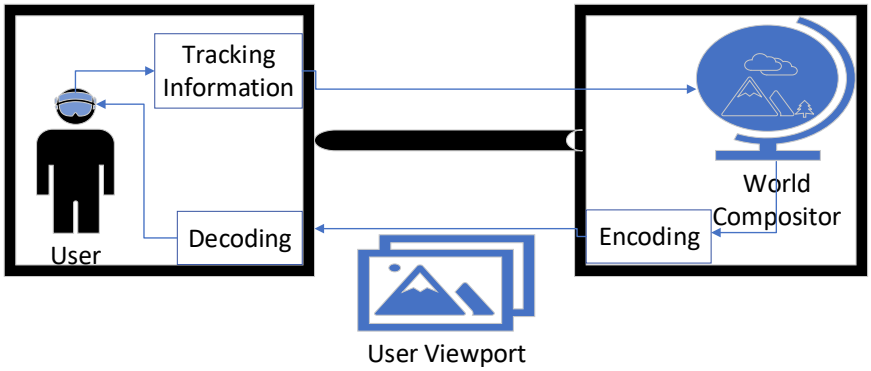


Figure 4.3: Viewport Rendering Process.

is widely recognized in the literature as motion-to-photon (M2P) latency.

However, not all M2P delays have the same impact on QoE. These range from physiological impact in the form of simulator sickness to psychological impact, such as immersion disruption. To clarify the difference between the impact of different delays in interactive immersive systems, we studied the factors that affect the perception of delay. Based on all this information, we have classified the delays according to their influence factors on perception, their location in the processing pipeline and its QoE perceptual implication: **Viewport Rendering**, **Local Interaction**, and **Distant Reality**. The following sections analyze the processes involved in each delay composition in a Social XR environment. Moreover, a selection of technologies and use cases that are sensitive to these delays is presented along with an analysis of their implication on the QoE.

4.3.1 Viewport Rendering Delay

In recent years, the most popular immersive systems are HMDs, such as the Meta Quest and the HTC Vive. These devices isolate our visual system by means of a nearby screen. The function of the HMD is to follow the position and orientation of the head to reproduce those movements in a virtual world. Consequently, HMDs always include some type of head tracking system. With this information, the graphics engine renders a portion of the virtual world based on our position and orientation. This portion of the virtual world is called Rendered Viewport. In this process there are several different steps or sub-processes necessary to achieve the viewport. In this subsection we analyze the processes involved in the viewport generation as well as the different configurations and use cases sensitive to it.

In Fig. 4.3, we can observe the different sub-processes involved in the viewport rendering process. The HMD should provide the tracking information to the World Compositor unit. Then, the World Compositor renders the user environment viewport according to its head position. Subsequently, the viewport has to be encoded, transmitted and decoded. Finally, the HMD shows the rendered viewport to the user as feedback on his/her head's movement. Therefore, viewport rendering delay extends from the acquisition of the user's position until the user perceives the correct feedback in the form of a rendered viewport. In general, the processes that contribute to the viewport rendering delay are:

1. The transmission of the tracking position.
2. The computational cost of computing the viewport.
3. The encoding of the viewport.
4. The transmission of the encoded viewport through the communication channel.
5. The decoding and displaying the viewport on the HMD screen.

From a communication perspective, the processes most affected by the network use the transmission channel. In our framework, those processes are the transmission of the tracking information to the processing unit and the transmission of the resulting viewport to the HMD. Essentially, The transmission delay depends on how fast the communication channel can transmit the information between the HMD and the processing unit. In addition, this delay depends on the transmission technology and distance between the User Reality Processing unit and the User Reality. While tethered (HTC Vive) or embedded (Meta Quest) solutions provide the fastest responses, technologies that process the reality in a distant server using remote rendering are highly network dependent. Furthermore, remote rendering solutions where the server is nearby offers acceptable results in terms of latency [52]. However, they involve tying immersive technology to specific places and limiting the free movement of users. Now that technologies like 5G and Beyond 5G are improving network features, remote rendering implementations like Cloud VR or untethered rendering are gaining interest in the Social XR area. Thus, it is crucial to consider the implications of viewport rendering delay in the QoE when designing the networks for these services. Since this delay is especially crucial because of the implications for QoE, there are several techniques that seek to mitigate it.

Prefetching, for example, consists of transmitting possible viewports to the HMD, these possible viewports are generated by pose prediction. If the user's pose finally matches one of them, the transmission delay is minimal. This type of technique is useful when the user is not moving (3DOF) and/or with a fixed background environment. Another technique to alleviate this delay is Image Warping. This technique consists in generating a new viewport by displacing the previous one according to the new pose offered by the head tracker while the next one is being received. For this technique to be effective, the previous viewport must cover more area than the user is able to see.

4.3.2 Local Interaction delay

In a shared immersive environment, interaction is an element that contributes to the immersion and realism of the experience. Interaction with the local environment is an enhancing feature in immersive communications compared with classic communication systems. Specifically, most immersive technology allows interaction using controllers or even your hands. Besides, this enriches communication for interactive use cases such as industrial training, education, or cloud gaming.

According to the Social XR communications use case, the processes involved in the interactive process are shown in Fig. 4.4. Following the framework, the interactive elements' information is obtained in the user's reality. Then, it is transmitted to the Processing Unit through the

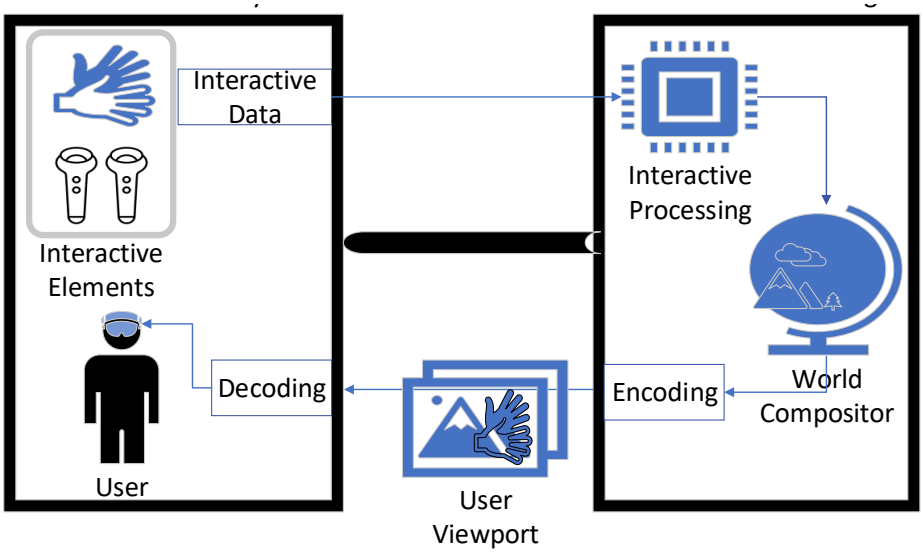


Figure 4.4: Interaction in Self Reality.

communication channel. After that, the information is processed to adjust the raw information of the interactive elements in the immersive environment. We call this process Interactive Elements Processing. Then, World Compositor processes the information to generate the viewport with interactive elements. Finally, the coded viewport is sent to the HMD through the communication channel.

The processes that contribute to the interactive delay follow these steps:

1. Interactive element data is captured and sent to the user's reality processing unit through the communication channel.
2. The interactive element processing unit utilizes adjustment algorithms to adapt the information to the virtual world's format.
3. The world compositor generates the user's viewport of the virtual environment.
4. The transmission of the viewport follows the same steps as the Viewport Rendering process: coding, transmission, decoding, and display.

The Interactive Processing module is responsible for processing and adapting the information captured by different sensors in the local environment. Also, this module adjusts this information so they fit into the virtual environment. Consequently, the magnitude of the delay is related to the complexity of the adjusting algorithms. Some examples are hand tracking, object pose estimation, and body segmentation. Currently, most of these algorithms use neural networks [53].

Although the algorithms based on deep learning are becoming more efficient, the computational cost of the state-of-the-art neural networks still too complex to build them in an HMD. Moreover, these complex algorithms made use of more powerful processing units needing from more hardware resources. However, the HMDs tend to be lighter as the weight is a milestone on the user's comfort. All these reasons promote the use of remote computing or

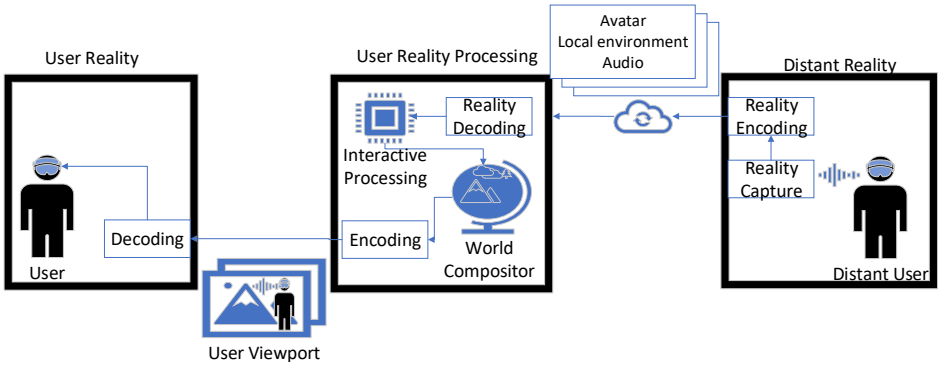


Figure 4.5: Distant Reality.

offloading. In these techniques, the input information is transmitted to a server that processes the information and sends back the result. As it happened for the **Viewport Rendering**, the magnitude of the delay introduced by this process depends mainly on the distance between the local user and the processing server.

4.3.3 Distant Reality Information

Interaction with distant elements involves numerous processes seen in the previous sections. Some of these processes are related to capturing the user's position and the interactive elements. In addition, information from distant realities must be captured, coded, and transmitted to the User World Processing unit. Then, this information is decoded, processed, and introduced into the virtual world of the local user through the world compositor. Finally, the viewport of the updated virtual world is shown in the user's HMD as in the previously described processes.

From the user's interaction point of view, the distant reality delay is defined as the time between an event triggered by the local user and the time the local user sees the distant world response. Although there is a large set of use cases that fit in this definition, in this work we focus on two Social XR use cases: teleoperation and teleconferencing.

To some extent, we summarize the transmitted information for both use cases in: audiovisual representation of users or avatars (teleconferencing). Audiovisual representation of the shared environment (teleoperation).

The representation of users or avatars are essential for visual communication in environments such as the Social XR. Somehow, we need to place distant users within our environment. To do this, we need a visual representation of the distant user. In the case of teleconferencing, the transmission of the distant user both visually (avatar) and aurally (voice) is particularly relevant. There are different techniques for avatar generation, ranging from photo-realistic to 3D static avatars. Some examples of static 3D avatars are Mozilla Hubs, Second Life, and Horizon Worlds. As the 3D avatar model is stored locally, this method only requires transmitting the remote user position. Although static 3D avatar model techniques require little transmission of information, they tend to be unrealistic and can cause a disruption of the immersion [54]. To solve this, other techniques as the volumetric capture generate realistic avatars [49], [51]. As they offer a photo-realistic version of the other user, volumetric avatar

capture can be considered the evolution of the classic video teleconferences.

Another use case in the Social XR is teleoperation: the ability of the user of the immersive environment to interact with a physical element of the distant reality. The representation of remote environments is essential for teleoperation in the Social XR. For this type of use case, we will need to transmit local user actions that will have an audiovisual response in the remote environment. In addition, this response must be transmitted back to the user. As in the case of teleconferencing, we can find methods that estimate the position of the remote elements and transmit their positions to update locally stored 3D models in self-reality. Moreover, we can find other techniques that capture the remote reality using cameras. Such cameras must be able to capture the entire environment. Consequently, some solutions use 360° cameras, either mobile or integrated into robots [48] [55].

In the case of environment representation for teleoperation cases, audio, and its source location become relevant as the audio sources of the remote reality are not attached to only one position as in the teleconference use case.

4.4 Studies Results and QoE Model

In the previous sections, we have identified three delays that are treated separately in Social XR communications: Viewport Rendering Delay, Local Interaction Delay, and Distant Reality Delay. Besides, we have analyzed the processes involved in each delay within a common framework. In this section, we present an analysis of the limits of perception and acceptance for the different delays and use cases based on QoE studies. We group these use cases according to the classification proposed in the previous sections (Viewport, Interaction, and Distant delays). In addition, we have also classified the different perceptual implications of delays according to the expectations of response in [56].

4.4.1 Viewport Rendering Delay

Nowadays, the current HMD devices tend to be lighter and untethered. Therefore, the main effort in the HMD design area is to embed more efficient hardware into the device. However, this affects the rendering performance. Because of this, solutions based on remote rendering are becoming popular. Remote rendering techniques consist in generating the virtual world in a processing server. Consequently, the viewport rendering delay may increase according to the physical and logical distance between the rendering server and the immersive device.

According to the literature, a low reality rendering delay is crucial for maintaining the QoE [57], [58]. When the viewport rendering is not synchronized well with the movements of your head, it generates a conflict in your vestibule system that drives the user to suffer from simulator sickness [57]. Numerous studies addressed the impact of this delay on the QoE. For example, in the study presented in [59] users had to complete a searching task with different viewport delay values. The results of this study suggest that the delay acceptance limit should be far below 58 ms. However, according to [60], limits of the viewport delay in the QoE may depend on the user behavior and the use case within the immersive environment. Regarding the user's behavior, we have selected a study that evaluate the worst-case scenario, a user that

rotates 180º at once [60], [61]. In this sense, we ensure that the most demanding use cases are taken into account. According to this study, there is a perception threshold of 7 ms and an acceptable threshold of 20 ms for Viewport delay. However, recent studies show that the semantics of reality is not so crucial for the vestibule system. Moreover, it is preferable to have less quality or outdated content rather than waiting for the user's viewport to update after a movement [16], [62], [63]. Thus, the different delays associated with reality content generation as interactive object position and other users' avatars are treated separately because their physiological implications are not so critical.

4.4.2 Interaction Delay

The impact of the interaction delay on QoE means that the user may not be able to interact correctly with their local reality. However, the direct effect of these delays seems to have no impact on causing simulator sickness [16]. Nevertheless, prolonged use of the unsatisfactory device may lead to the abandonment of the technology. From a psychological perspective, the influence of local interaction delay on the QoE affects factors such as the feeling of being immersed, the perception of usability, and the general opinion of the system. According to the classification of [56], the QoE implications of the interaction delay fit as Simultaneous Perceived Stimulus. Therefore, the user expects the visual feedback to be synchronized with their other senses. In the case of hand manipulation, the sync between haptic perception and visual feedback.

We have identified two use cases in the Social XR whose local interaction delay generates a different impact on the QoE: self-view delay and environment update delay.

The self-view delay implies that we act with our hands or controllers in the local reality, but that action takes some time to be shown to us in the rendered viewport. This delay is especially relevant when processing our body as part of the immersive environment relies on computationally expensive techniques. This situation could require such processing to occur away from the local environment.

To explore the implications of the delay on QoE, we addressed a study on how the self-view delay may affect the QoE in interactive Social XR environments [63]. The study made use of egocentric image segmentation to visualize hands and nearby objects realistically. In the experiments, subjects had to replicate a Lego-style Fig. using their own hands and real blocks in an XR environment. During the experiment, we applied a set of selected delay values in the visual feedback of the hands and blocks. The conclusions of this study point out that the global quality perception and immersion may be harmed because of the mismatch between the local reality feedback and the delayed visual response. However, the limits of the adaptation to these types of delays are less sensitive. In this study, the threshold of perception of self-view latency was estimated at 300 ms while the delay threshold for the acceptance was estimated at 450 ms. This study is part of the contributions of this thesis and is presented in Section 5.3.

Another use case related to the local interaction in the Social XR is the delay of environment updating. This means that when we move with our head, we receive correct visual feedback from our head movement, but, the information we see may take time to be correctly shown. In

the Social XR, this happens when the user processing unity is divided. Part of the rendering that gives consistency to the world happens nearby the users' reality while the rest of the processing happens in another server (typically further). This technique is known as Split Rendering. This solution avoids simulator sickness reducing the viewport delay despite the increased bandwidth cost. However, the impact of this delay in the QoE in terms of presence and global perceived quality in immersive environments hasn't been widely explored.

For this case, we also conducted a subjective quality study in which the subjects had to view 360° videos [62]. The experiment simulated the delay between the user's head movement and the presentation of the correct updated content in the video. The results of the study confirmed the importance of having an un-updated or degraded environment rather than having a viewport delay. Moreover, we evaluated different delay values to find the limits of the environment updating delay in terms of QoE. Regarding the delay thresholds, the results show a perception threshold of 150 ms and an acceptance threshold of 300 ms. This experiment is described in details in Section 5.2.

4.4.3 Distant Reality Delay

Finally, we analyze the implication on QoE of delays associated with distant realities. In the case of this work, we analyze two use cases, teleoperation, and teleconferencing. In both, the perception of latency implies waiting for a response after a user-triggered event. In teleconferencing, we expect feedback from the other user. However, in teleoperation, we expect it from a remote agent that responds to our movements, such as a crane, a surgical robot or a car [55], [64], [65].

According to the classification of [56], the QoE implications of the delay in classic teleconferencing and teleoperation systems fit into Asynchronous Perceived Stimulus. Hence, the user does not expect immediate feedback from the system. Because of this, the limits of the delay tend to be larger. Indeed, the delay in teleconference systems has been widely addressed for 2D systems [6]. Consequently, in this work we will use the recommendation as a reference to define the thresholds for the 2D teleconference use case [6]. In the studies from which the recommendation draws the results are [66], [67]. In [66], an experiment was conducted with subjects to understand how the delay limit for audio-only conversations changed when video was added. During the experiment, participants engaged in seven conversations on specific topics. After each conversation, they were asked to rate the conversation using various attributes. Additionally, varying degrees of delay were introduced into each of the seven conversations. The result was that the limit of acceptance for audiovisual conversation is 500 ms. Furthermore, work cited in the same study stands 200 ms to be the threshold for the perception of conversational delay in video conferences [68]. Therefore, we consider the perception threshold to be 200 ms and the acceptance threshold to be 500 ms. With respect to immersive telecommunications, Section 5.4 of this thesis presents a study of QoE in volumetric XR social video conferencing systems from which we can extract the perceptual and delay values, respectively 600 and 900 ms [69].

The latency values and scenarios proposed for the teleoperation case are presented below. Due to the variability of scenarios that fit the teleoperation use case we have decided to focus

on three scenarios that we understand as potential uses of immersive communications: remote manipulation, telesurgery and remote driving. In the context of the remote manipulation, there is a conducted a study in which users were tasked with remotely completing a log insertion operation using a crane [55]. In that experiment, users operated a crane with a joystick and viewed the environment with a 360° camera. The range of added delays spans from 0 to 30 ms for display updates and from 0 to 800 ms for the hand controller. Notably, the study reveals a pronounced influence of latency on display update and a noteworthy adverse effect arising from an 800 ms delay on the hand controller. Regarding delay thresholds, results show a perception threshold of 400 ms and an acceptance threshold of 800 ms.

Regarding the telesurgery scenario, [64] conducted a user study with sixteen medical students who performed energy dissection and needle-driving exercises on a robotic simulator. They performed the tasks with random delays between 0 and 1,000 ms (in 100-ms intervals). The study establishes that noticeable performance degradation becomes apparent after 300 ms, with delays surpassing 500 ms proving particularly challenging for intricate tasks. Hence, we consider the perception delay as 300 ms and the acceptance delay as 500 ms.

In the context of remote driving, there are no latency studies in the field of immersive communications. However, we can find works on the values of perception and acceptance latencies. In [65], the suitability of the 5G network to meet the 5G Automotive Association (5GAA) [70] latency acceptance requirement for remote driving is discussed. From the study [65] it is determined that the perception delay is 30 ms while the acceptance delay marked by 5GAA is 120 ms.

Table 1 summarizes the different latency scenarios with the effect on the QoE and technology use cases. Moreover, through the various QoE studies outlined in the previous sections, we have identified the thresholds of perception and acceptance for the use cases.

4.4.4 QoE Model

Some QoE prediction models regarding latency in interactive communications have been proposed. In [11], a model for mobile gaming using a linear approach is proposed. This model was also validated by subjective experimentation using different game genres, coding qualities, and network conditions. In the ITU E-Model [7], the delay impact is characterized by an algebraic equation used to predict subjective effects of transmission impairments. Moreover, Rec. ITU G.1072 [12] relies on a logistic decay equation as the standard predictive model for assessing gaming QoE in cloud gaming services.

These models have been established after numerous subjective QoE studies. In general, the values in predicting the quality for the same use cases vary significantly, meaning that the models show a high variance in the quality predicted. Although they may differ in values, all models share three zones. A first one in which the delay is imperceptible. The second in which the QoE begins to decline rapidly as the delay progresses, and a final one in which the QoE declines with a less pressing slope. To replicate such behavior, we have selected Rec. ITU G.1072 curve as the basis for our model. We have decided to use this model because the gaming paradigm is the closest one to Social XR. However, the Rec. ITU G.1072 model only accepts two use cases: scenarios with highly QoE-sensitive delays and those that allow

Table 4.1: Summary of perceptual implications, use cases and the perception and acceptance threshold for each delay.

Delays in Framework	Perceptual implications	Use cases	Perception Threshold	Acceptance Threshold	Ref.
Viewport Delay	Vestibular Motion Sickness	Unthetered rendering, CloudVR	7 ms	20 ms	[60], [61]
Local Interaction Delay	Simultaneous Perceived Stimulus	Self-View Delay	300 ms	450 ms	[63]
		Enviroment Updating	150 ms	300 ms	[62]
Remote Interaction Delay	Asyncronous Perceived Stimulus	Conversational (2D)	200 ms	500 ms	[66], [68]
		Remote Manipulation	400 ms	800 ms	[71]
		Telesurgery	300 ms	500 ms	[64]
		Remote Driving	30 ms	120 ms	[70]
		Conversational (Social XR)	600 ms	900 ms	[69]

longer delays. In our case, we have adapted the model so that it fits the curves according to the results of QoE studies depending on the use case. Therefore, the model parameters become relative to the conditions and results of the QoE study. These are, the maximum and minimum values of the quality scale, the quality values consider as thresholds of perception and acceptability and their corresponding latency values. Therefore, this model can be used, on the one hand, as a reference if the use cases fit those already listed in Table 4.1, and, on the other hand, to obtain a predictive model based on QoE results for other use cases.

$$QoE = Q_{max} - \frac{Q_{max} - Q_{min}}{1 + e^{f_a - f_b Ta}} \tag{4.1}$$

where:

$$\begin{aligned} f_b &= \frac{\ln\left(\frac{(Q_{max}-Q_{Tm})(Q_{Ts}-Q_{min})}{(Q_{max}-Q_{Ts})(Q_{Tm}-Q_{min})}\right)}{Tm - Ts} \\ f_a &= f_b Ts + \ln\left(\frac{Q_{Ts} - Q_{min}}{Q_{max} - Q_{Ts}}\right) \end{aligned} \tag{4.2}$$

Ts is the delay value for the perception threshold, Tm is the acceptation threshold and Ta is the M2P delay of the system. Q_{Ts} represents the quality value for the perception threshold and Q_{Tm} represents the acceptance quality value. In addition, Q_{min} and Q_{max}, represent the minimum and maximum quality values on the scale.

The following is an example of how to use the model with reference to the values in Table 4.1. First, the values of the use cases that appear in Table 4.1 use the Likert scale. That is, the range of values varies from 5 to 1, with 5 being the best score and 1 being the lowest. Then, we consider 3 as the acceptance threshold and 4.5 as the perception value. That is, the values of the equation take: $Q_{max} = 5, Q_{min} = 1, Q_{Ts} = 4.5, Q_{Tm} = 3$. Finally, using these values, the parameters become dependent on the value of the sensing delay (Ts) and the acceptance (Tm).

$$\begin{aligned} QoE &= 5 - \frac{4}{1 + e^{f_2 - f_3 Ta}} \\ f_b &= \frac{\ln 3.5}{Tm - Ts} \\ f_a &= f_b Ts + \ln 7 \end{aligned} \tag{4.3}$$

The plot shown in Fig. 4.6 is obtained by using the equation 1 for each use case with their

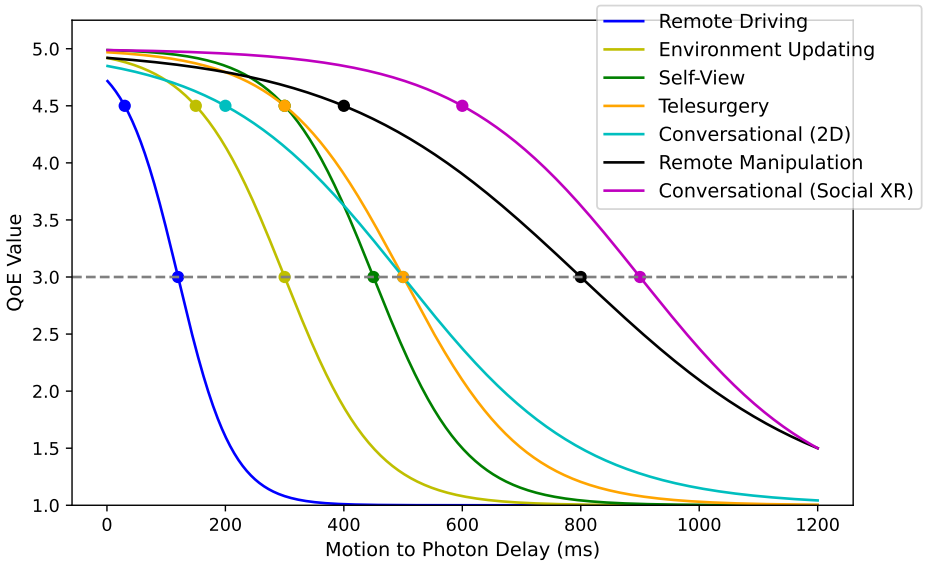


Figure 4.6: Evaluation of the model using Table 4.1.

respective Ts and Tm values. In the case of the Viewport Delay delay, it has not been included in the plot since both Tm and Ts are of an order of magnitude less than the rest, not being able to be appreciated in the plot correctly together with the rest of the use cases. We can observe in the Table 4.1 and in the Fig. 4.6 that interaction-related latency values allow for higher delays. Within this group, we find differences between local and distant interaction delays. These interaction values are reasonable considering the differences in expectations with respect to visual feedback. In local interaction perception, a faster response is expected, whereas in teleconferencing or remote manipulation, higher delay is accepted because an immediate response is not expected. In contrast, we can observe that those teleoperation scenarios that simulate more critical use cases such as remote driving or telesurgery are affected by tighter acceptable delays.

4.5 Conclusions

In this chapter, we have presented a common framework for immersive remote communications that lists the processes required to enable different use cases in the Social XR. Furthermore, we have identified different sub-processes (viewport, local and distant interaction delays) in the main framework according to their effect on the user experience. In addition, we have analyzed how the new technologies for placing the processing on remote servers may harm the immersion and QoE of the users because of the delay increasing.

Hence, to maintain the QoE in the future Social XR communications, it will be necessary to keep the latency of the different sub-processes below certain thresholds that guarantees the QoE. In this work we present a summary of different delay thresholds based on QoE studies with proper assessment methodology. Additionally, we present a QoE model to estimate the QoE according to the delay magnitude. This model was conform by using the acceptance and perception threshold of the QoE studies and adapting the ITU QoE delay impairment model

for gaming systems. Additionally, the model's capacity to adapt to new use cases using QoE latency results is a noteworthy contribution.

Chapter 5

Influence of Delay on the QoE in Video-based Social XR

5.1 Introduction

During the development of the thesis, video-based solutions have been proposed to achieve communications under the Social XR paradigm. Following the scheme presented in Fig. 1.2 we have proposed: 360° video for the virtual environment, interactions with the physical environment based on image segmentation and, finally, representation in the shared space by means of volumetric avatars. In parallel with the developments and collaborations to generate this system, we came to the conclusion that the major stumbling block in bringing these systems to reality was the delay.

This chapter delves into the core of our research, where we present the outcomes of three carefully conducted studies. The objective of these studies is to assess the impact of the delays discussed in Chapter 4, namely the viewport updating delay, self-view delay, and conversational/videoconferencing delay. Each study is meticulously designed to adhere to the recommendations set forth by the ITU. By systematically investigating the effects of these delays on various aspects of user experience, we aim to contribute valuable insights to the field and provide a better understanding of how these factors influence communication in immersive environments.

Here, the structure of the chapter is presented. Section 5.2 presents the study related with the environment updating delay. The study includes a description of the methodology, the setup used and the results and conclusions. Section 5.3 describes the self-view delay study. The section explains the setup to achieve a minimum delay environment, how we adapted an interactive task, the study design and its results. Finally, Section 5.4 describes the study on volumetric videoconferencing delay.

5.2 Environment Updating Delay Study

Current XR applications use omnidirectional video to boost users' immersion and sense of presence. As contents from distant video sources cannot be instantaneously delivered, the end-to-end delay becomes a key problem when user actions cannot be simultaneously matched by system reactions. Thus, we have designed and executed an experiment to assess its influence on the QoE, the sense of presence, and the sickness caused. To do it, we have developed a viewport adaptive simulator to render simultaneously two layers of immersive video to allow different adaptation schemes and delay values. Twenty observers have assessed 180 test videos from 9 sources. Our analysis shows a clear influence of the delay condition and the adaptation scheme on the perceived quality. Moreover, it also shows that the adaptation schemes and delay conditions have a small influence on the sense of presence and little effect on observers' sickness.

5.2.1 Real-time Video Based Environment

As shown in Fig. 1.2, the Social XR requires the representation of a shared environment. In this context, immersive video, offering a photorealistic view, can be used for enhancing the sense of presence. Furthermore, real-time video streaming is required in use cases discussed in Chapter 4 such as video surveillance and remote operation.

To acquire the physical environment, telepresence systems usually have Pan Tilt Zoom (PTZ) cameras taking advantage of their capability to move their angle of vision [72] [73]. Nevertheless, their use for immersive video in Social XR environments remains unexplored, as only a fraction of the 360° environment is delivered and, thus, the rest of the 360° scene is not being updated. An alternative is provided by Tile-Based encoding, where only a portion of the video is sent at a high quality while the rest of the 360° scene is encoded at minor quality. When the user moves to another position, the high quality region is updated [74]. All these schemes for immersive video streaming are classified as viewport adaptive schemes and they all share the same critical problem: the delay.

Delay is an important parameter for QoE in immersive environments [75], as increased end-to-end latency can lead to unpleasant experiences [31]. Thus, this section analizes the impact of delay on different schemes of adaptive viewport inmmersive video streaming in terms of QoE.

Several recent works have considered subjective assessments to measure the QoE and the immersion of the users [3], [76], [77]. Subjective factors such as sense of presence, simulator sickness [28], and Mean Opinion Score (MOS) are claimed to be a good method for measuring the QoE of immersive video [3], [30]. Other recent works have measured the impact of network delay in the QoE in immersive video delivery [74], [75], [78]. However, real-time streaming imposes an additional delay and the mechanical constrains of PTZ cameras add 200 ms [78].

We measured the QoE in four types of schemes based on the viewport adaptation schemes of tile-based videos and inmmersive PTZ video delivery: a hybrid 360° and PTZ camera system called foveated imaging [79], tile-based 360° video, PTZ video delivery, PTZ video delivery over pre-recorded background. To assess video quality, Degradation Category Rating (DCR) has been used. To assess presence and sickness factors, The Simulator sickness [28] and the

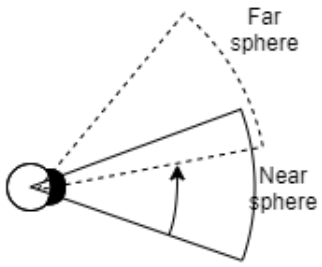


Figure 5.1: Top-view diagram of user movement.

mini-MEC [80] questionnaires have been used. In accordance with the methodology presented in chapter two, we have also included the short question on motion sickness in the study. The scenarios were simulated using a specific tool developed for this experiment, where the user initially watches a high quality 360° video and then watches different schemes of adaptation with a variety of delay scenarios. The study objectives were:

- Present a novel scheme for adaptive video schemes
- Assess the QoE over different transmission schemes involving PTZ immersive and tile-based video varying the delay.

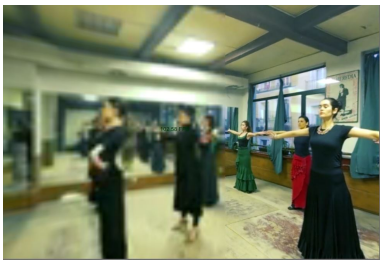
5.2.2 View-Port Adaptive Simulator

For this experiment, we developed a subjective assessment tool, using the Unity engine, which supersedes a previous version [81], by allowing the simultaneous play of two overlaid 360° videos to simulate different types of viewport adaptive schemes with various delay conditions. Thus, different delay values can be used to simulate the end-to-end latency between a user's movement (an action) and the instant when they sees again high quality video (a reaction).

Tool Design

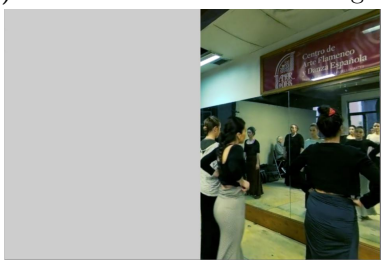
The main requisite for the design of the tool was the ability to render video in two different virtual spheres as can be seen in Fig. 5.1, which shows an horizontal cross-section of the virtual spheres viewed from above. While the distant sphere is fully rendered, the nearby sphere only renders a part corresponding to the Field Of View (FOV). The effect of the distance between the two spheres is imperceptible to the user.

The experiment considers two display conditions. In the initial condition, the user watches a non-delayed video within the FOV (near sphere). When the user turns, they begins to see the furthest sphere during the selected delay condition. Depending on the selected adaptive scheme, the effects of the delay will be different. Fig. 5.2 represents the effects associated with the different schemes assessed in our work. If the schemes include a 360° camera, the quality will worsen as Fig. 5.2a shows. This refers to tile-based scheme (the regions/tiles are pre-established), and to foveated imaging (they depend on the PTZ camera position). However, for an only PTZ camera scheme, a still image will be displayed during the delay. Thus, the main difference between only PTZ camera scenarios is whether a pre-recorded scene is available, and so the displayed still image will correspond to the scene acquired some time



(a) Tile-based and foveated imaging.

(b) PTZ over pre-recorded background.



(c) PTZ.

Figure 5.2: Effects of delay in the different viewport adaptive schemes.

Table 5.1: Summary of the different delay components.

Delay Component	Meaning
τ_{pos}	New head position transmission time.
τ_{adj}	Time from the reception of the new head position until the camera reaches the final position.
τ_{cod}	Coding time.
τ_{net}	Network delay.
τ_{dec}	Decoding time.

before the user's movement. Otherwise, a grey image is displayed, as shown in Figures 5.2b and 5.2c.

Delay Definition

The considered viewport adaptive streaming schemes are composed by a capture device, a video engine and the network. All the involved delays can be grouped into an overall value representing the latency between a movement of the user and the instant when high quality video can be seen again (Eq. 5.1). Firstly, the video engine sends the new position to the camera adding τ_{pos}, which is similar to the network delay τ_{net}. Then, if the camera is a PTZ, it will require a certain time to move to the new position τ_{adj}. This delay does not exist for 360° video cameras. The video is acquired by the camera and τ_{cod} includes the capturing, the local processing, and the encoding of the video. Finally, the video is decoded and rendered to the viewing device adding τ_{dec}. These delays are summarized in Table5.1.

$$\tau_{e2e} = \tau_{pos} + \tau_{adj} + \tau_{cod} + \tau_{net} + \tau_{dec} \approx \tau_{adj} + \tau_{cod} + 2 * \tau_{net} + \tau_{dec} \tag{5.1}$$

Our viewport adaptive simulator considers the overall delay value as a design parameter. Different simulations will involve several sets of delay alternatives (configurable in our tool).

5.2.3 Experimental Design

Subjective Questionnaires

Quality, presence, and simulator sickness questionnaires were used to assess QoE. As users always start watching the video in the highest quality, DCR [5] was used to measure video quality. As DCR is designed for evaluating the degradation between a pair of scenes, in this case, we requested users to assess how unpleasant was the effect of the delay in each sequence. Following the ITU-T Rec. P.910 [5], differential MOS (DMOS) are computed and, to do so, a hidden reference (i.e., absence of delay) was presented randomly to the subjects.

To measure presence, the mini-MEC test [9] was used, which is a subset of the MEC questionnaire [82]. Only certain questions about attention allocation, spatial stimulation, self-location, possible actions, cognitive involvement and suspension disbelief were asked.

Also, to measure simulator sickness, two different questionnaires were used at different stages of the experiment. On the one hand, the tool presented a question about sickness at the end of each each sequence [30]: "How is the level of dizziness or nausea?". On the other hand, before starting the experiment and after each part of the session, the SSQ [28] was filled by the users, to analyze the temporal evolution.

Equipment, Selected Delays and Videos

The overall end-to-end latency groups all kind of delays as described in the previous subsection. Related work in the evaluation of the impact of streaming delay in immersive video have used delays in the range of 100ms to 1s [78][83]. Thus, we considered two groups of delays in our experiment: short (150ms and 300ms) and long (500ms and 1s) delays. Both delay groups included 0ms as hidden reference.

Three high-quality equirectangular (4096x1980) video sources of 60 seconds were chosen for the test, covering different properties: 1) moving camera (video of a drone flying over a seaport), 2) fixed camera (fixed drone recording a medical assistance), and 3) exploratory content (a flamenco classroom, shown in Fig. 5.2). A total of 9 source clips were used, by dividing these videos into three clips of 20 seconds. Taking this into account, and given the 4 adaptive schemes and the 5 considered delays, a set of 180 test videos of 20 seconds was generated. It is worth noting, that for foveated imaging and tile-based schemes, the low quality versions were generated reducing the original resolution to 420x320. The HMD Lenovo Mirage Solo was used in the experiment as the player device.

Observers and Workflow

Twenty participants (ages between 22 and 40, average of 26.5, 3 females and 17 males) took part in the experiment.

Firstly, the four adaptation schemes were shown to them in a initial training session, continuously with the highest delay. After each scheme, the user scored the DCR. Therefore, the user could get used to the delay effects and the voting method.

Before the test, each user answered the SSQ to gather information about their initial sickness

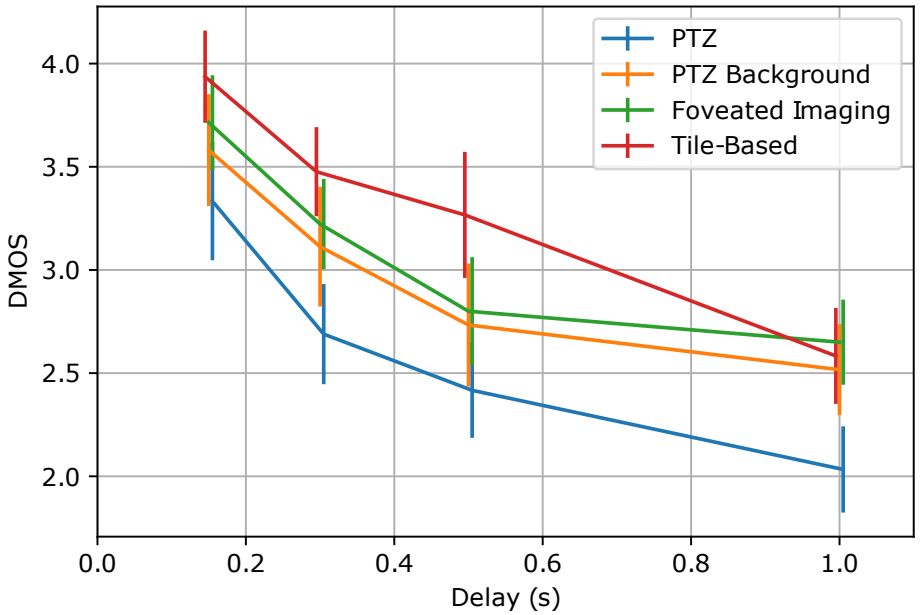


Figure 5.3: DMOS score for each scheme.

state. The test was divided into two sessions, one for each delay group (short or long). Then, the 3 clips of each source were shown for each adaptive scheme with the different delays from the selected group (randomized). After each test clip, DCR was scored. After each 3 clips, the observers filled the question about sickness as well as mini-MEC questionnaire, while the SSQ was filled at the end of each test session. Each session lasted around 30 minutes, with a break of 10 minutes in-between.

5.2.4 Results

Video Quality

Figure 5.3 shows the DMOS and 95% confidence interval of the the DCR scores for the different adaptation schemes and delays. In all cases, there is a strong dependency between the delay and the DMOS. Even a relatively low delay of 150ms is clearly perceptible. The slope of the curves decreases faster between 150ms and 300ms with the increase of the delay. However, the scores decrease slower between the 300ms and 1s. This situation shows that the end-to-end delay is specially relevant in the first hundreds of milliseconds of delay. These results also show that the PTZ scheme is significantly worse than the rest of schemes. Particularly, comparing PTZ with and without pre-recorded background it is shown that adding 360° information to the PTZ video (e.g., a static frame or a low-resolution video), significantly enhances video quality. Moreover, PTZ over pre-recorded background scheme shows similar results to 360° camera-based schemes.

A two-way ANOVA was done to check the dependency of the DMOS results on the adaptation scheme (AS) and the source sequence (SRC). As previously seen, a significant effect on the adaptation scheme was found (p = 6.4e-5, $p < 0.001$), but no significant dependency was

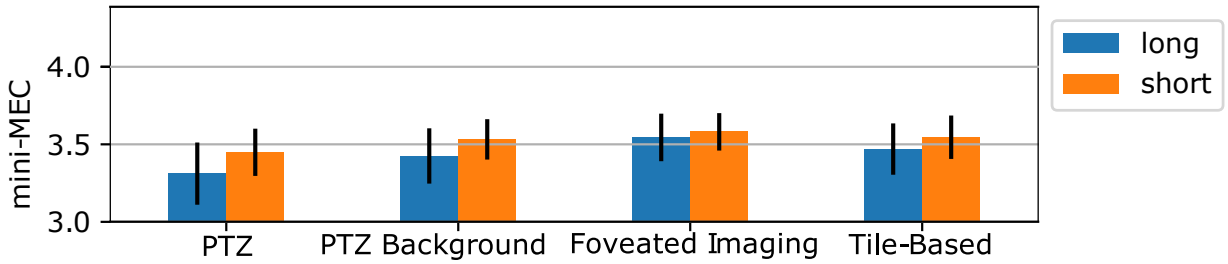


Figure 5.4: Mini-MEC average score for scheme and delay.

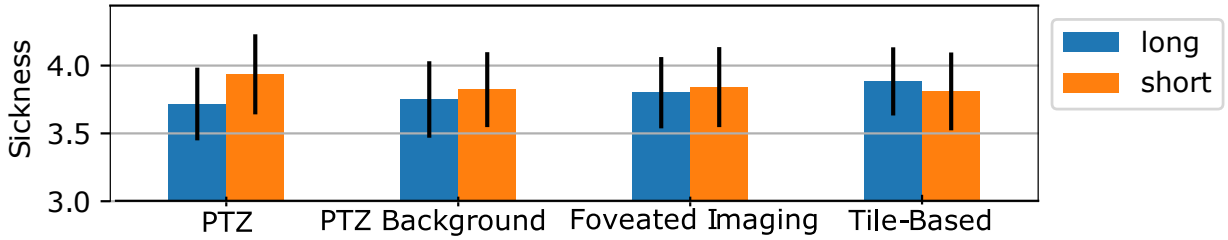


Figure 5.5: Sickness question average for scheme and delay.

Table 5.2: Simulator Sickness Questionnaire average results.

	O	**N**	**D**	σ	**Total Score**
Starting	15.17	15.92	7.33	7.33	11.02 ±8.10
Resting	37.11	33.43	15.56	24.77	23.86 ±10.85
Ending	50.66	43.51	20.68	32.43	31.73 ±14.21

found on the individual source sequences ($p = 0.08$) or the interaction of both conditions ($p = 0.35$). This supports the aggregation of scores from different sources in the analysis of the results.

Sense of Presence

Figure 5.4 shows the mini-MEC presence score for each adaptive scheme and delay group. Unlike video quality scores, there is no significant difference between adaptation schemes or delay groups. Under the conditions of our experiment (visible effect of the adaptation with respect to an homogeneous no-delay scenario), there is no influence of the specific delay or adaptation scheme on the sense of presence.

Simulator Sickness Figure 5.5 illustrates the mean value of the sickness question for scheme and delay group. As in the sense of presence, there is no significant difference among the schemes or delays. Table 5.2 shows the average results of the SSQ Nausea (N), Disorientation (D) and Oculomotor (O) factors, and the total SSQ score, with the confidence interval for 95% confidence [28]. An increase of sickness over time is observed, in line with similar tests watching 360° videos with HMDs [84].

5.2.5 Conclusions

In this experiment, subjective video quality, sense of presence and sickness were studied with adaptive schemes using different camera systems: PTZ and/or 360° cameras. The results of video quality show that the PTZ scheme is significantly worse than PTZ over pre-recorded background, foveated imaging and tile-based schemes. Also, adding 360° information to a PTZ adaptive scheme enhances the perceived QoE. Regarding sense of presence and simulator sickness, the results show that the level of delay or the selection of a specific adaptation scheme (even large delays and strong foveation schemes, such as PTZ) have no significant influence. This supports the possibility of using PTZ cameras within VR setups. Further research can extend these results using higher-quality background videos or lower delays.

5.3 Self-View Delay

In our Social XR proposal illustrated in Fig. 1.2, users must be able to interact with the environment in which they are physically located. For this, it is necessary to mix the information coming from the physical and virtual realities. In this context, the XR paradigm allows user interaction by blending the physical and virtual realities through a self-representation of the user. However, how this blending is done can affect realism and, ultimately, break the user's immersion. Some examples of interfaces that can affect the way the body is introduced into the XR are: controllers, haptic gloves or grounded haptics [8]. Other methods introduce the user's body without intermediate elements, for example through image processing algorithms [35], [85]. While these methods maintain the user's immersion, they may add a significant self-view delay to the user embodiment. The self-view delay in an XR environment is defined as the difference between the time of a user's movement and the time when the user sees their move into the XR.

As the delay can affect the realism and immersion in the XR environment, it is very important to know its limits for interactive tasks. Consequently, self-view delay has been studied extensively in both non-immersive and immersive scenarios [55], [86]–[89]. However, all studies to date emphasize non-immersive interaction or rely on very specific tasks.

In this study, we addressed the impact of different levels of self-latency on an standardized interactive task while keeping the level of self-representation at a good quality. The rest of the section is structured as follows. Section 5.3.1 describes the concept of the self-view delay in XR environments. Section 5.3.2 provides the details of the self-view delay experiment, while Section 5.3.2 presents the obtained results. Finally, Section 5.3.3 expose the conclusions of the self-view delay experiment.

5.3.1 Artificial Self-view delay XR environment

Our study addressed the impact of delay on QoE and user performance in XR environments. This delay is mainly caused by the techniques for mixing local and virtual reality. Therefore, delay is a key factor in developing extended reality interaction methods. In addition, delay can harm both QoE and user performance. Specifically, the delay related to the self representation in XR is specially interesting as seeing yourself in the virtual reality is important to preserve

Table 5.3: Summary of the different delay components.

Delay Component	Meaning
τ_{acq}	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_{fps}	Time between frames.

the user's immersion [90], [91]. Delay in immersive environments has been addressed by other studies. However, such studies make use of intrusive devices or are based on highly artificial interactions. Moreover, these results of these studies are limited in terms of subjective QoE due to the selected delay values [55], [87].

To address the limits of the self-view delay in XR we developed an XR environment with minimal self-view delay along with a subjective and objective quality experiment. In addition, we adapted a standardized interaction assessment task to maximize the generalization of our results. The following subsections explain the definition of the self-view delay and the system default delay measurements.

System Self-view Delay

In our camera-based XR solution there are two processes that contribute to the self-view delay: capturing and rendering. Table 5.3 the different components of the self-view delay.

Capturing stands for the elapsed time between the instant of the user's movement and the instant the computer has available the information from the camera. This latency is the sum of the time between frames of the camera (τ_{fps}) plus the acquisition time (τ_{acq}).

Rendering represents the elapsed time between the moment when the computer receives the frames from the camera and the moment when the computer displays the processed information to the user. This latency is the sum of the image processing (τ_{proc}) plus the display rendering delay (τ_{disp}). Thus, the overall self-view delay is defined as:

$$self\text{-}view\ delay = \tau_{fps} + \tau_{acq} + \tau_{proc} + \tau_{disp} \tag{5.2}$$

System Default Self-view Delay Measurement

The scope of this experiment is to measure the impact of the self-view delay in interactive XR environments. In this context, it is necessary to know the intrinsic delay of the system before adding the artificial ones. We used a method inspired in the "numerical latency measurement" [86] to measure the intrinsic delay. In contrast, our implementation uses a 2D display instead of a 7-segment one for the reference clock as is illustrated in the Fig. 5.6. The difference between the reference clock and the captured one in the high framerate video gives us an estimation of the intrinsic self-view delay of the XR environment. After 70 measurements, the mean estimated intrinsic delay was $190ms \pm 9ms$. This result is in line with other HTC Vive Pro pass-though delay measurements [1] [2]. We also developed a tunable artificial latency

[1] https://stereolabs.com/blog/vive-pro-ar-zed-mini/

[2] https://softserveinc.com/en-us/blog/passthrough-ar-headset-comparison

Table 5.4: Questionnaire used in the experiment.

Factor	Question
Global QoE	How would you rate the quality of the experience globally?
Sens.Haptic	How well could you move or manipulate objects in the virtual environment?
Sickness	Did you feel any sickness or discomfort during the experience? Please rate it
Involv.	How natural did your interactions with the environment seem?
Adapt.	How quickly did you adjust to the virtual environment experience?
Adapt.	How well could you concentrate on the assigned tasks rather than on the mechanisms used to perform them?
Involv.	How compelling was your sense of objects moving through space?
Involv.	How much did your experiences in the virtual environment seem consistent with your real world ones?

adder by buffering the camera frames before displaying them into the virtual environment. Finally, we used the same numerical latency method together with software to measure the added delay for each buffered frame. After several iterations, we concluded that each frame adds 37 ms.

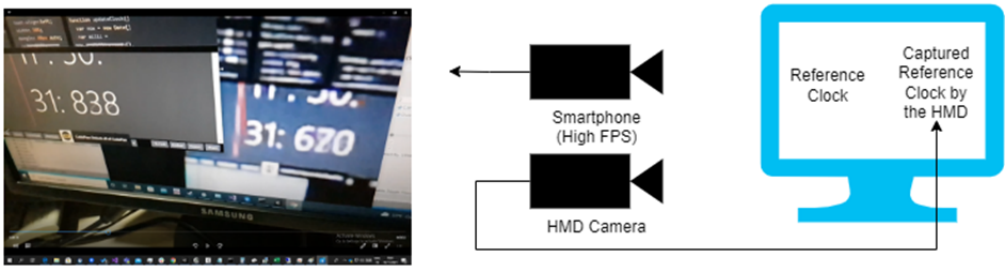


Figure 5.6: Example and diagram of the offset latency measurement system.

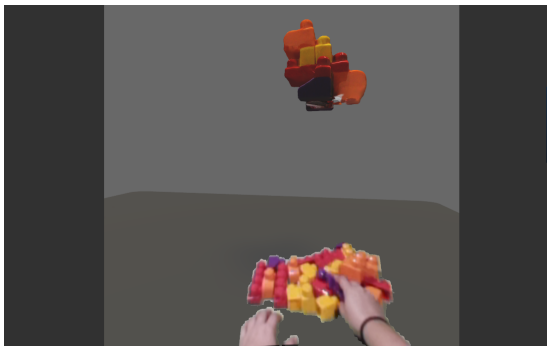
5.3.2 Experimental Design

To measure the impact on the QoE of the self-view delay in interactive XR environments, we designed a new task inspired in a standardized one for measuring the impact of audiovisual degradation in interactive communications [92]. The original task consisted of building a model using Lego-style blocks with the help of another user through a regular teleconference system. An example of implementation of this task in its original form is [93]. However, as we wanted to measure the impact of self-view delay, we developed a single user experience including a realistic 3D model of each complete shape. This model was inserted within the virtual environment so the user could see the 3D model. During the experience, users are able to see their hands and a set of realistically shaped blocks that are integrated by egocentric capture (with the same system used in chapter 3). Thus, they are able to replicate the figure shown to them in the virtual environment. The Fig. 5.7 shows a user building a model while watching the reference 3D model using their own hands and real blocks.

XR Environment Setup & Apparatus

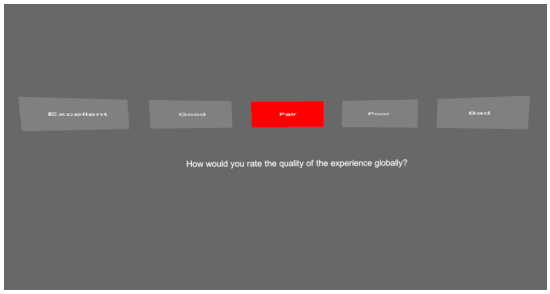
XR needs coordinated information from the real world embedded in the virtual one. The design of the virtual scenario has been made with the intention of not distracting the user. Thus, virtual scene is composed by a grey room with a simple gray table. Figure 5.7(a) shows a distant view of the scene.

(a) Virtual environment



(b) Segmentation result

(c) Physical reality setup



(d) Votation application

Figure 5.7: XR environment setup.

Building the XR environment requires the integration of the physical reality. In our implementation, physical reality is captured from the HTC Vive Pro cameras and segmented by the XR engine shaders. System's segmentation uses a chroma-key algorithm like in [9]. These methods for integrating the physical reality keep the delay at the minimum while preserving the user's immersion [9], [85]. Figure 5.7 (b) shows the result of the chroma segmentation during the experiment task. In order to add latency to the baseline, I developed a frame buffering system, so that by increasing the frame buffer, the time from when the users perform a movement until they see it reflected on the XR can be increased.

Methodology

Before starting the experiment, the user had to sit in front of a table, adjust the headset, and adjust their position according to the table. (see Figure 5.7(c)). The experiment consisted of two regular sessions and a training one. During the training session, the subject got used to the XR environment and the procedure of the task. In addition, the training session included the best and the worst conditions (190 and 597 ms), so the user knew the delay range in advance. This is a common practise in QoE assessment [4], [5]. The 3D model used during the training session was not included in the regular ones. During the regular sessions, each subject had to indicate the beginning of the experience before each construction. After the

confirmation, the experimenter started the session. Then, the subject had to complete the building task reproducing the 3D model with a fixed self-view delay value. After that, they had to vote within the virtual environment using the MIRO360 app [26]. Then, the process started again with another model and delay value. After four iterations, the user rested for 10 minutes before starting the second session [94].

Questionnaire

After finishing each condition (model plus delay), users had to fill a questionnaire evaluating important aspects of the QoE interaction in XR: global quality, involvement, adaption, haptic sensation and simulator sickness. The questionnaire included eight questions (see Table 5.6). This questionnaire is a subsampling of Presence Questionnaire of Witmer and Singer, which was validated in [95] for interactive immersive environments. In addition, during the EPSILON system evaluations (Chapter 2), this questionnaire was also used successfully. All items were evaluated in a Likert-like 5-level scale.

Stimuli

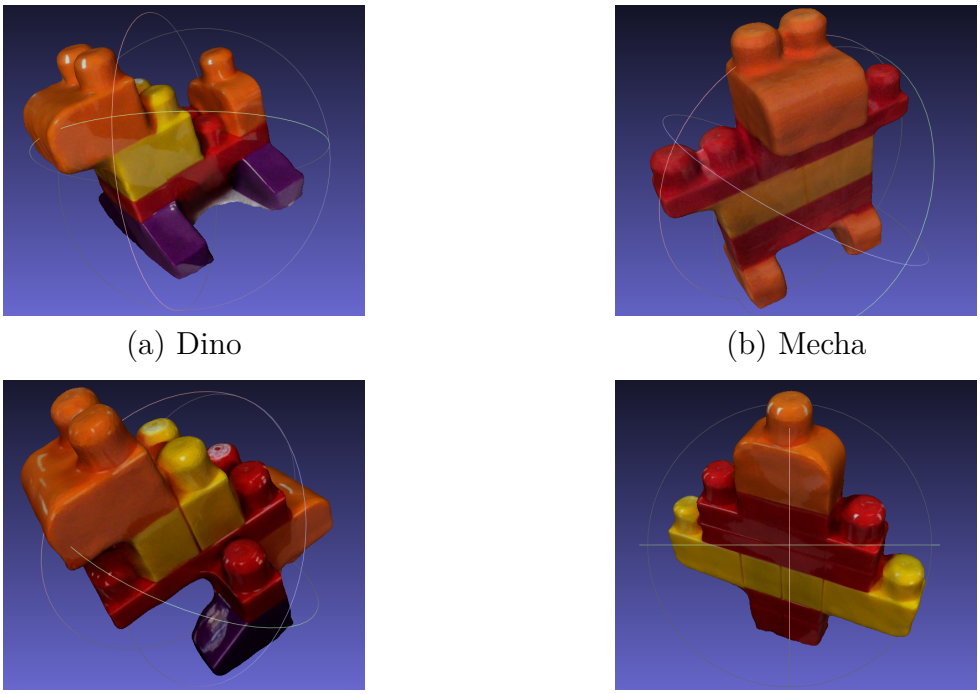
The users had to perform the task eight times with different delay values in two separate sessions. For selecting the delays, we performed a pre-test where 6 users tried from 0 to 12 buffered frames, which means a delay range from 190 to 634 ms. During the pre-test, each user played for approximately 5 minutes in total. For each delay, they had to vote verbally from 1 (bad) to 5 (excellent) how well they felt about interacting with the blocks. We found that the opinion scores fell from 3 to 2 between 375-486 ms. Moreover, these results are in line with the literature about delay impact on the self perception and task-based delay adaption [55], [87], [88], [96]. Consequently, we decided to use more delay stimulus around those values. The selected delays were [190, 264, 338, 375, 412, 449, 523, 597] ms. These eight values were separated in two sets for each regular session. Set A contained [190, 338, 412, 523] ms while set B contained [264, 375, 449, 597] ms. In addition, we balanced the starting set (12 users started with A, and 11 started with B), and the order of each delay value in each set was randomized during the experience. The available set of individual blocks for building each model was the same for all models. During regular sessions, the subjects had to build four different models (see Figure D.1). The model shapes were also randomized, except for model Rocket which was assigned to delay values 264 ms and 523 ms to have anchor values in each set.

Subjects

We conducted a lab trial with 23 subjects (7 female and 16 male; ages between 21 and 34). Each one had to complete the task eight times mixing four models and eight delay values separated in two sessions.

Results

During the experiment, we collected the scores for each QoE factor and the task elapsed time. Before aggregating all the scores, we analyzed the influence of the model shape on



(a) Dino (b) Mecha

(c) T-Rex (b) Rocket

Figure 5.8: 3D models to reproduce during the task.

the subjective scores. Firstly, we ensured that the scores for each factor conformed to a normal distribution ($kurt < |2|, skewness < |2|$ for all factors votations)[97]. The result of the ANOVA allowed us to discard the influence of content (model shape) on voting for all the QoE factors ($\rho > 0.05$). The scores for sickness have not been included in the analysis as the participants did not felt sick at all for any delay value. Considering the static nature of the XR environment, the absence of simulator sickness is in line with the results of previous studies[89], [94]. Taking this into account, Fig. 5.9 shows the Mean Opinion Scores (MOSs) and 95% confidence intervals obtained for the considered QoE factors.

Global QoE

From the results of the globalquality factor (GQOE), we can observe no statistical difference from the reference delay (190 ms) until the 375 ms score. After that, the GQOE maintains at an acceptable level (above 3) until 449 ms. For the 523 and the 597 ms delay the GQOE score decreases until the range of (2, 2.5) what denotes a strong QoE disruption.

Involvement

Involvement stands for the average score of the three involvement questions in the Table 5.6. Here, we can observe a similar behavior to the global quality results. However, for the 523 and 597 ms, we can observe that the mean values are even worse for this factor even though the starting score (190 ms) is around 0.5 points lower than the GQOE.

Adaption

Adaption factor is constructed averaging all the scores of the two adaption questions. The scores follow the same trend that the previous factors. However, for longer delays, adaptation

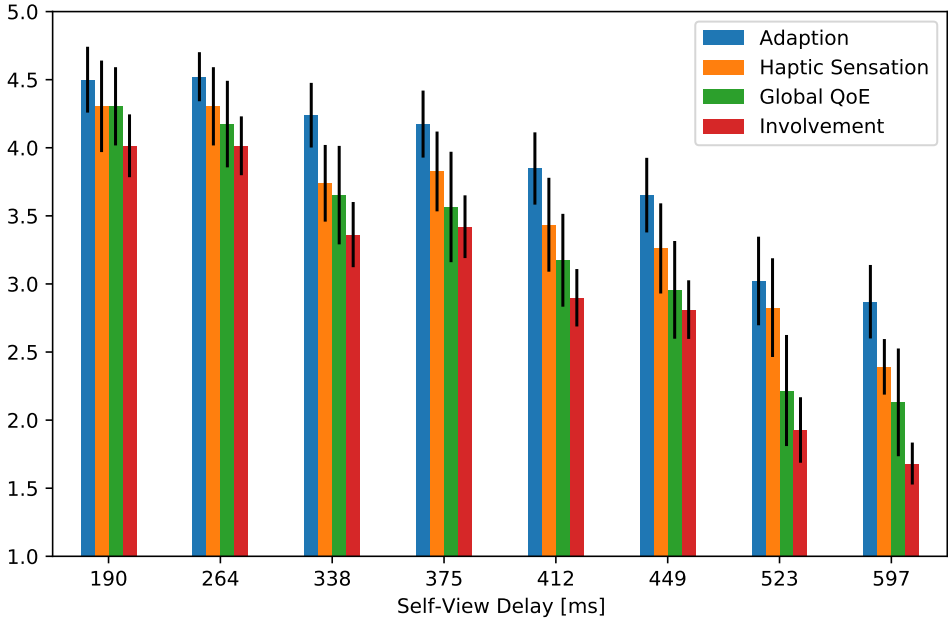


Figure 5.9: Mean scores of the different QoE factors per delay.

scores remained at acceptable levels (around 3). This idea that the people adapts somehow to the delays is supported by some previous studies [55], [96].

Haptic Sensation
The results for haptic sensation follow a similar trend. However, the values reached for the larger delay values are close to acceptable levels (3) as is the case for adaptation.

Time to accomplish
We can observe in Fig. 5.10 that there is a clear upward relationship between experiment performance and added delay. After an ANOVA analysis, we observed a significant influence of the delay on the time to build each model ($\rho < 0.05$). However, Tukey post-hoc analysis indicated that we could only find significant differences for the extreme values (190 vs 597) ms. Moreover, we can observe in Figures 5.9 and 5.10 that the impact of delay on involvement or global quality is much more pronounced than for execution time. This is in line with the results of the adaptation factor. That is, people felt less immersed as delay increased, but, in contrast, immersion disruption did not have such effect on their ability to perform and adapt to the task.

5.3.3 Conclusions

This section presents a study on how the self-view delay affects the QoE and the performance in interactive immersive XR environments. The task selected for the study is inspired by a widely validated ITU-T interactive task [92]. The results show that there is a threshold around 450 ms for the QoE factors (global QoE, involvement, and haptic sensation) where the QoE falls to levels of non-acceptance. However, the time to accomplish and the adaption factor show that the users can adapt to these delay scenarios.

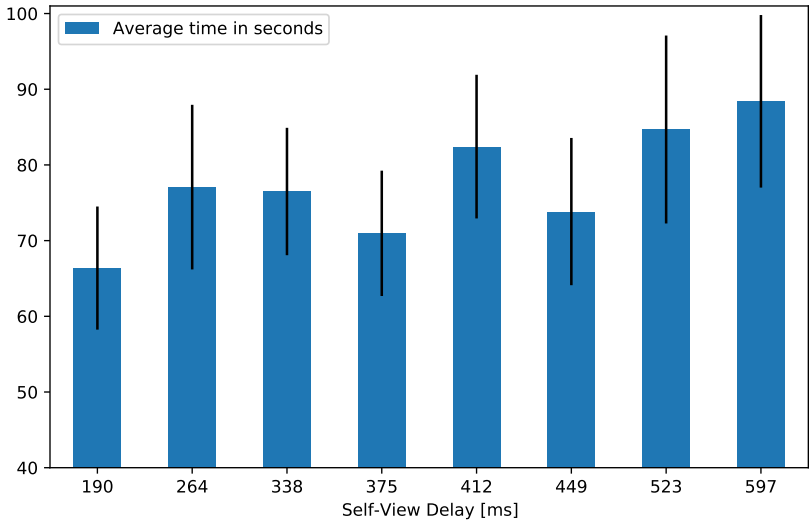


Figure 5.10: Average time of accomplishment per delay.

5.4 Volumetric Videoconferencing Delay

During this chapter we have presented the studies related to the environment and self-representation delay present according to the Social XR diagram presented in Fig. 1.2 .

In this section we present the third and last delay, the delay in volumetric audiovisual communication. The use of immersive technologies has aroused interest in several telecommunications-based applications, such as industrial training [98], [99], telecare [100], and telemeetings [54]. However, 2D videoconferencing is still the most widely used technology for teleconferences, although it presents certain drawbacks that affect the user experience. According to [54], 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 two-dimensional visual representation and the lack of user free movement.

To overcome the limitations of 2D videoconferencing, Social XR has emerged as a promising solution by offering a more natural and immersive communication alternative. 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 [2], [101], [102]. 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 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[2], [54], [102].

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 [103]–[106]. 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 [103]. 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 [3], [55], [93], [104], [107], [108]. Together with objective metrics, these experiments have identified acceptable delay thresholds for videoconferencing [6], [7], [109]. The recommended delay threshold for avoiding user annoyance is below 600 ms [6], but recent studies have suggested higher values, exceeding 900 ms [93], [104]. 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 study addressed the challenge of determining 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 [110] 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 [93], [104]. Therefore, this study contributes to:

- 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.

5.4.1 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 [2].

The main difference between Social XR and 2D videoconferencing is the Degrees of Freedom (DOF) for user exploration and interaction [54]. DOF signifies how freely a user can view different angles of media content. The level of DOF in Social XR systems ranges from 3DOF, which involves head movements (pitch, yaw, and roll), to 6DOF, 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, [48] presents a virtual environment where users can interact with a distant environment in 3DOF using a 360° camera. Another example is [111], which presents an environment with purely virtual avatars where users interact with 3DOF using their voice and controllers. However, this 3DOF environment does not use video for user representation.

Finally, [49] presents a 3DOF 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 realism and interactivity to XR experiences, making them feel even more like face-to-face interactions [2], [54]. With respect to volumetric video, we can find two representation techniques. On the one hand, we have the 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 [112]–[114]. 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 [115].

Point cloud is another approach to represent volumetric video. Point cloud is 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 [115]. 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 [50], Holoportation in Microsoft Mesh [116], and VR2Gather [49].

In this work, the VR2Gather Social XR system [49] 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 (see Fig. 5.11). Another decisive factor was that it is open source [49], 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.

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. Therefore, the system must capture aspects of two physical realities, namely where the two remote users are located, and position all that information in a Social XR environment. As an illustration, Fig. 5.11 shows two users placed in two different physical rooms (bottom) each wearing a Head Mounted Display (HMD), and corresponding snapshots

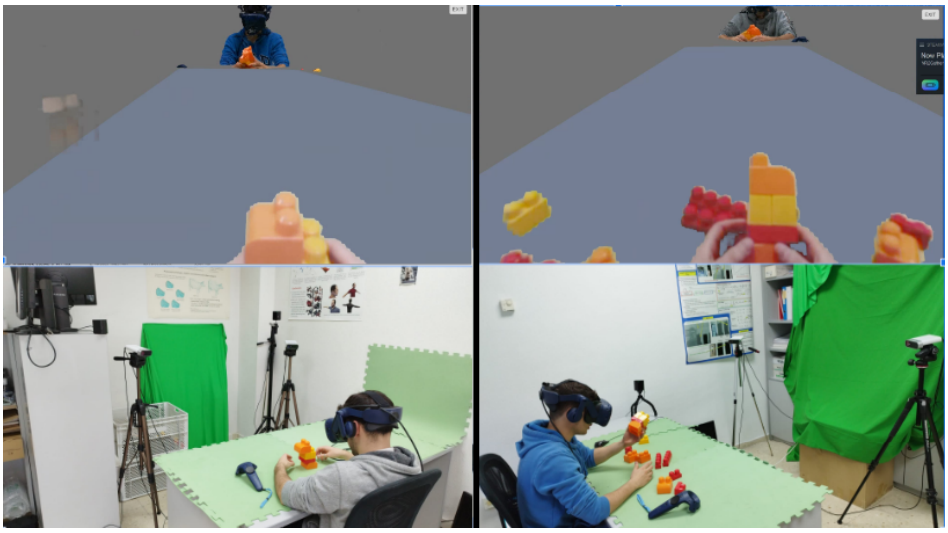


Figure 5.11: Two users sitting in two different physical rooms and meeting in the same Social XR environment during the experience.

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. Also, the volumetric representation of the remote user is visible at the end of the virtual table.

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 Fig. 5.12. 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. That is, the physical room where each user is located. In this study, we use a room with a table (see Fig. 5.11). In each black frame of Fig. 5.12, 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 above, 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

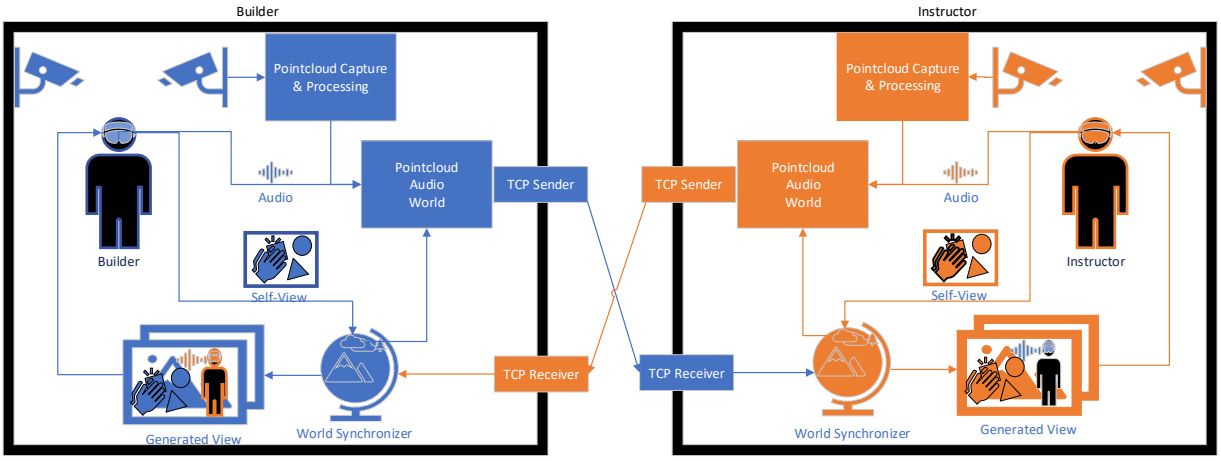


Figure 5.12: Diagram of volumetric XR communications.

Figure 5.13: Local environment self-view without distant user.

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 (see Fig. 5.13)

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 [110]. 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 (see Fig. 5.14). An analysis of the different processes that contribute to the end-to-end delay is presented in the next subsection.

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 5.5 summarizes the different components that consist of delays related to capturing, processing, display, transmission, and synchronization.

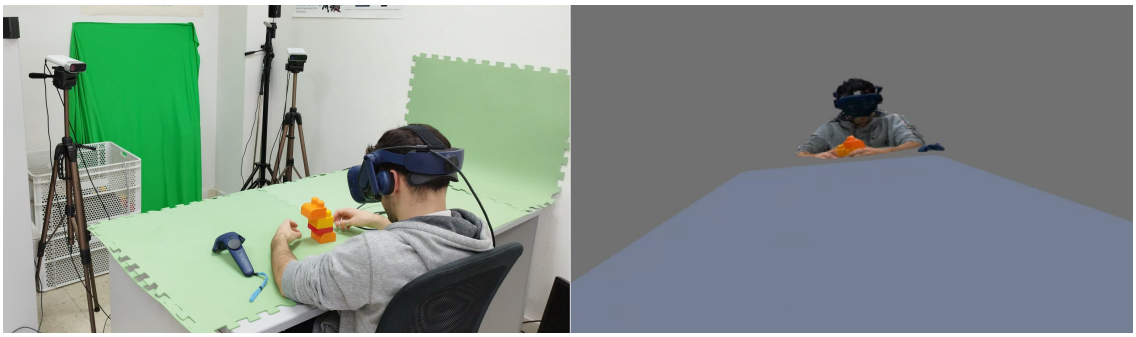


Figure 5.14: Physical environment of the instructor and the generated viewport of the builder in the Social XR environment.

Table 5.5: 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.

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 Fig.5.13). After that, the result is rendered in the virtual world and displayed in the HMD. In Fig. 5.14 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:

$$self\text{-}view\ delay = \tau_{cap} + \tau_{proc} + \tau_{disp} \tag{5.3}$$

In Equation (5.3), 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. Firstly, 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:

$$XR\ delay = \tau_{cap} + \tau_{proc} + \tau_{tx} + \tau_{sync} + \tau_{disp} \tag{5.4}$$

In Equation (5.4), 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 syncronization, i.e., audio and video syncronization 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.

5.4.2 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 Rec. P.920 [109]. 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, while multicamera-based volumetric capture was used to represent distant users. These adaptations are illustrated in Fig. 5.11.

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 [63], using an identical system configuration. To eliminate the effect of additional parametets, 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 Fig. 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 study.

This section introduces the methodology employed in the current study. The research involved the adaptation of the standardized ITU-T Rec. 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.

Hardware

The experimental hardware utilized in this study encompassed a range of functionalities, namely physical reality capture, point cloud capture and transmission, synchronization, and Social XR environment display, allocated per user. Physical reality capture and environment display were achieved through the use of the HMD HTC Vive Pro, while point cloud capture and generation were facilitated by using the CWIPC system [110], utilizing the Kinect Azure color and depth cameras. The synchronization of social worlds was managed by VR2Gather software [110], installed on Windows 10 PCs with an Intel Core i7-4790 with a clock speed of 3.6 GHz, boasting 8 cores, alongside an NVIDIA TITAN Xp GPU.

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 [117] 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 [81] application was used to conduct the questionnaires within the virtual environment.

Objective Data

During the experiment, objective data were captured to analyse the impact of delay on user performance. On the one hand, 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.

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 5.6 presents the subjective factors evaluated in conjunction with their respective questions. Subjective factors analysis included global quality,

Table 5.6: Questionnaire used in the experiment.

Category	Factor	Question	Reference
Subjective performance	Global QoE	How would you rate the quality of the experience globally?	[109]
	System Annoyance	How easy did you find it to communicate using the system?	[6]
	Delay perception	Did you perceive any reduction in your ability to interact during the conversation due to delay?	[6]
	Interruptions	How would you judge the effort needed to interrupt the other party	[6]
Presence	Involvement	How much did your experiences in the virtual environment seem consistent with your real world ones?	[63], [95]
	Adaption	How well could you concentrate on the assigned tasks rather than on the mechanisms used to perform them?	[63], [95]
	Accomplishment	I am confident that we completed the task correctly	[63], [95]
Social Factors	Social Presence	I felt connected with my partner	[118], [119]
	Social Annoyance	I was able to understand partner's message	[118], [119]
	Social Adaptation	My partner and I worked together well	[118], [119]
	Collaboration	Information from partner was helpful	[118], [119]

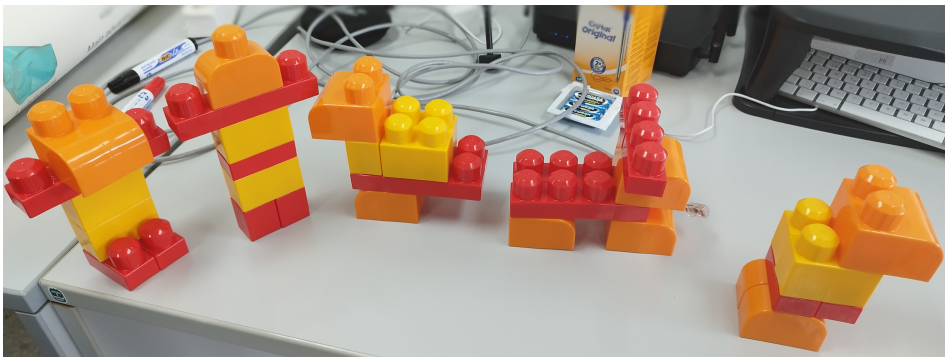


Figure 5.15: Selected block based figures, from right to left: *Mazinger*, *Rocket*, *Bird*, *Dog*, and *TRex*.

system annoyance, delay perception, and interruption perception, derived from international standards and specifically aimed at assessing the impact of delay on system acceptance [6], [109], [120]. 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 [95]. This questionnaire was also used for the self-view delay experiment [63]. To further examine the impact on subjective social quality, questions from [119] used in an experiment with a similar task [118] were included to assess subjective social factors.

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 1000 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 1000 was increased by reducing the number of blocks for each figure. The following delay values were selected: 300 ms (minimum), 600 ms, 900 ms, 1200 ms, and 1500 ms. In addition, the selected block-based figures are shown in Fig.5.15. Each Fig. is composed of 7 blocks.

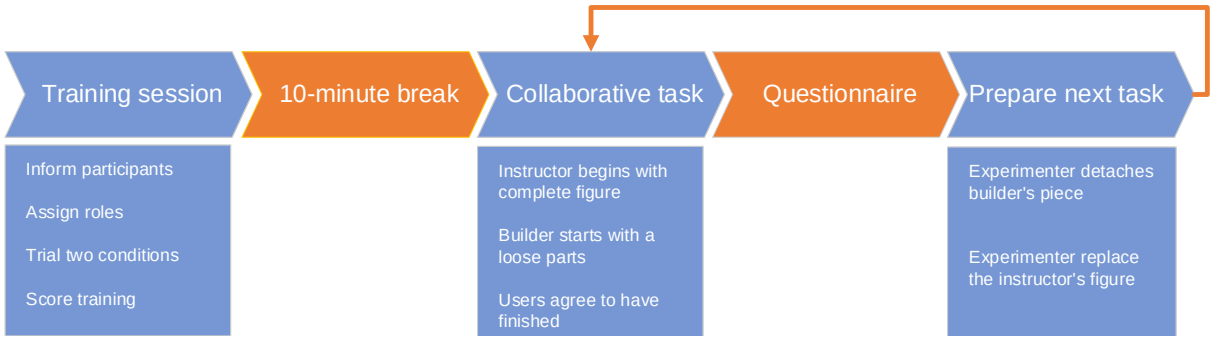


Figure 5.16: Experiment workflow diagram.

An essential consideration when establishing experimental conditions is randomization and balancing [4]. To ensure that conditions were balanced, the Graeco-Latin distribution was used to organize the delay and Fig. conditions [121]. 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.

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 (1500 ms) delay conditions. This methodology is in line with the conventional practices in subjective experiments [4], [5]. 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. Fig. 5.16 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, while 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 5.6. 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.

Participants

We conducted an experiment with 60 subjects (29 female and 31 male; ages between 20 and 33, mean: 22.8, standard deviation: 2.1). None of them were experts in the use of VR. 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.

Table 5.7: Subjective Performance Analysis.

Factor	Variable		ANOVA F	ANOVA p	ANOVA η^2	Significantly different
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	

5.4.3 Results

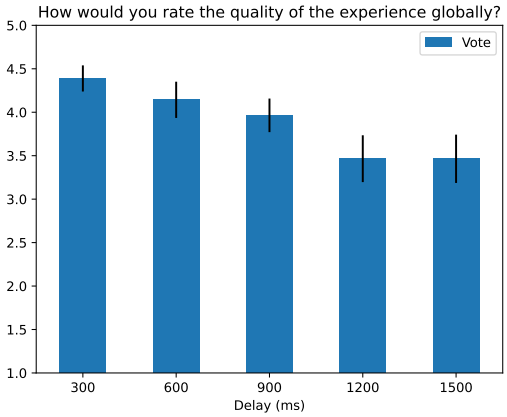
This section presents the results of the various factors assessed in the experiment. Each subsection comprises a normality test to assess the distribution of scores, an ANOVA to examine the impact of delay, figure, and role on voting outcomes, and a bar graph of the average score for each role and delay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

Subjective performance factors

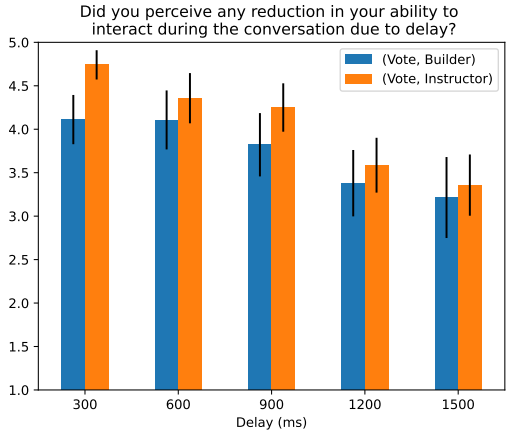
Initially, normality was confirmed for each of the factors either by a Kolmogorov-Smirnov normality test or by checking that both skew and kurtosis were in the range (-2, 2) as established by [97]. Table 5.7 shows the statistical results for each factor of the subjective performance of the system. This 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 factor. In addition, for variables showing significance ($p < 0.05$), Tukey's HSD (Honestly-significant-difference) 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.

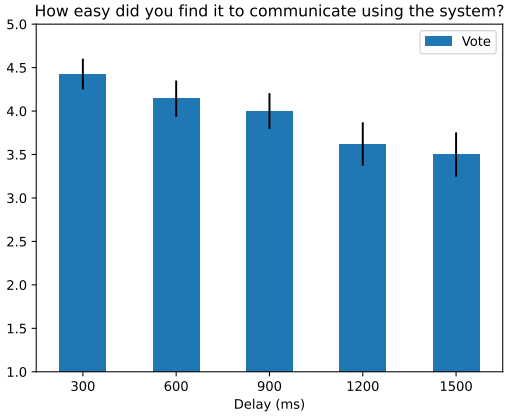
Figures 5.17a, 5.17c, 5.17b, 5.17d show the average scores for each factor and delay with their 95% confidence intervals. It can be observed that for the factors of perceived delay and interruptions, we can find differences between roles, the builders being more sensitive to delay (i.e., they notice it earlier). Moreover, we can find significant differences from 600 ms of delay



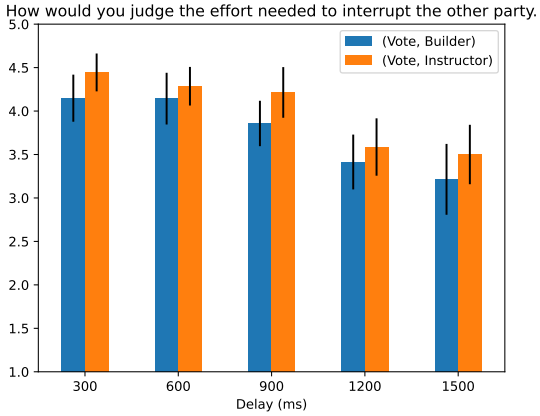
(a) Mean score values of the Global QoE with 95% confidence intervals.



(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.

Figure 5.17: Subjective Performance Results.

Table 5.8: 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	-

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 (1200 ms and 1500 ms) reaching levels on average of 3.5. At the 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 [6], however, it is in line with later studies [122] and [104].

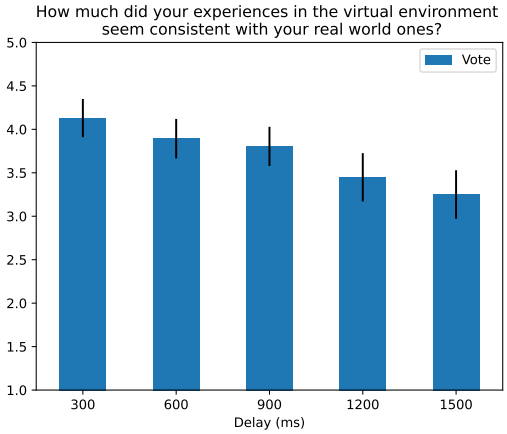
Presence

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 5.8, which includes the role, delay, and Fig. 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 (1200 and 1500 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.

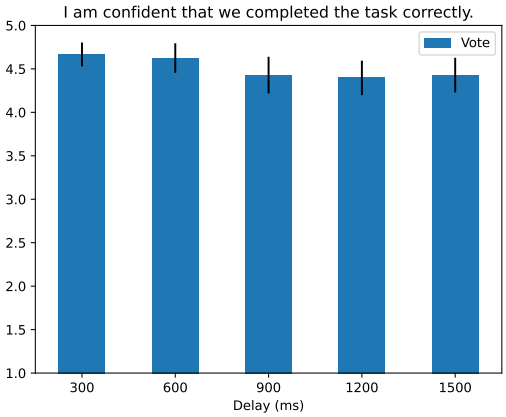
According to the average results in Figs. 5.18a, 5.18c, 5.18b, we only found differences between the roles in adaptation factor. Here, we can observe that the builders suffered more from the delay than the instructors. This is in line with the idea that builders notice the delay earlier and that it is 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.

Social factors

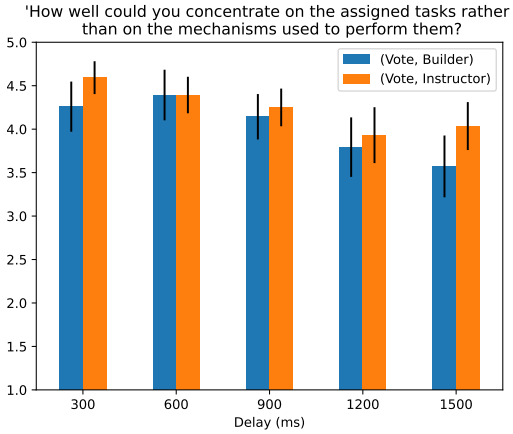
The present study examined some social factors. First, we verified the normality of the



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



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

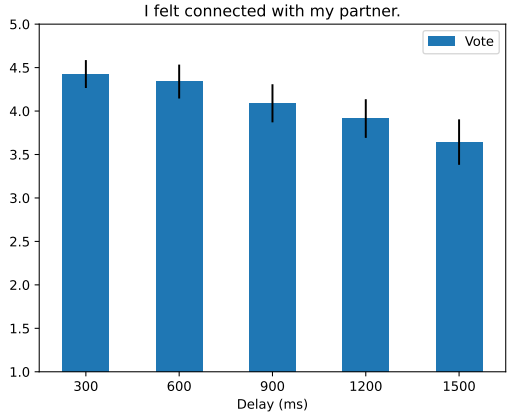


(c) Mean score values of the Adaptation with 95% confidence intervals.

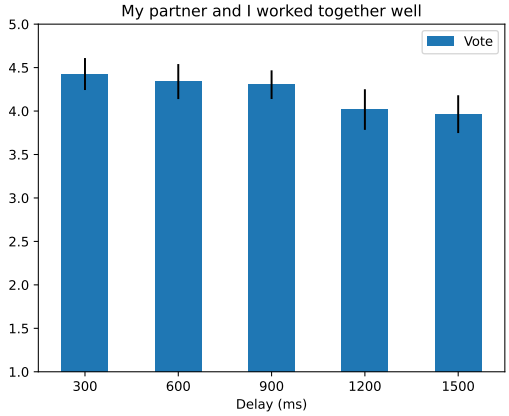
Figure 5.18: Presence Factors Results.

Table 5.9: Social Factors Analysis.

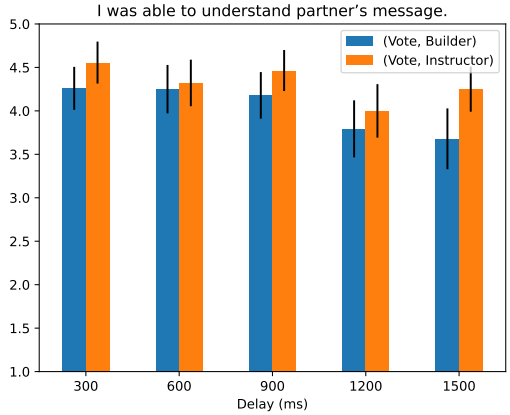
Factor	Variable		ANOVA F	ANOVA p	ANOVA η^2	Significantly different
Social Presence	Role		$F_{1,230}$ =3.549	0.061	0.0117	-
	Delay		$F_{4,230}$ =7.761	**<0.001**	0.102	(≤600) vs (≥1200) (900) vs (1500)
	Figure		$F_{4,230}$ =2.440	**0.048**	0.032	(Bird) vs (Rocket)
Social Annoyance	Role		$F_{1,230}$ =5.714	**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	-



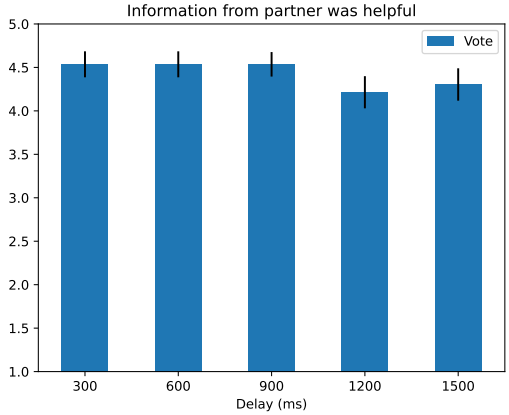
(a) Mean score values of the Social Presence with 95% confidence intervals.



(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.

Figure 5.19: Social Factor results.

skew and kurtosis ratings, which were found to have absolute values less than 2. Utilizing an ANOVA, it was determined that, for most of the social factors, only the delay factor had a significant impact on the ratings ($p < 0.05$), while the role and Fig. 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 an effect of the Fig. 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 1200 ms and 1500 ms.

According to the average results from the Figs. 5.19a, 5.19c, 5.19b, 5.19d, 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

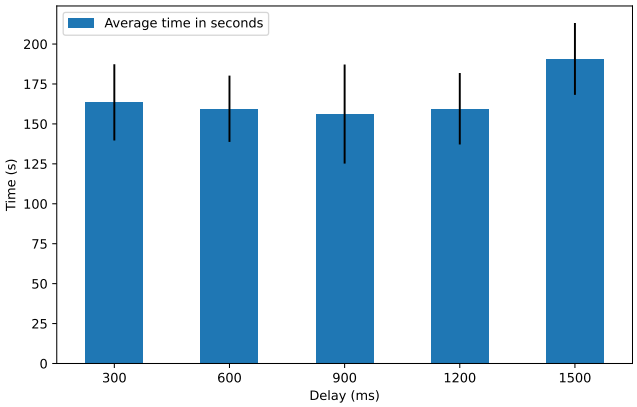


Figure 5.20: Mean score values of the task duration in seconds with 95% confidence intervals.

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 while the instructors kept their averages relatively stable.

Duration

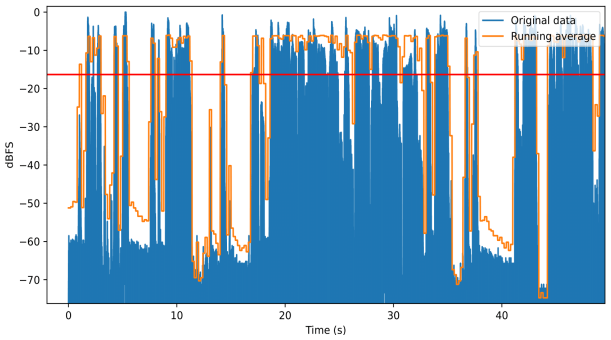
This section presents an analysis of the impact of completion time for each experimental condition, namely delay, and figure. First, a normality test was conducted to determine the distribution of the data, which indicated a nonnormal distribution with kurtosis that exceeded an absolute value of 2. Subsequently, a more detailed examination of the results was performed, revealing a significant 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 Fig. 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 Fig. 5.20, 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 ms to 1200 ms, while for the worst condition, an average of 190 was obtained, representing $\sim 19\%$ increase.

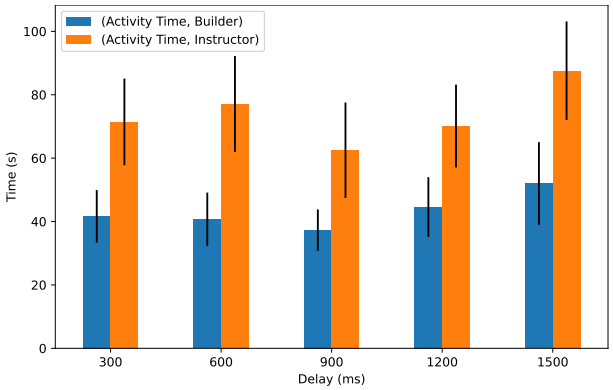
Audio

During the experimental sessions, the conversations of the participants for each condition (delay and figure) were captured using OBS software [117], 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.

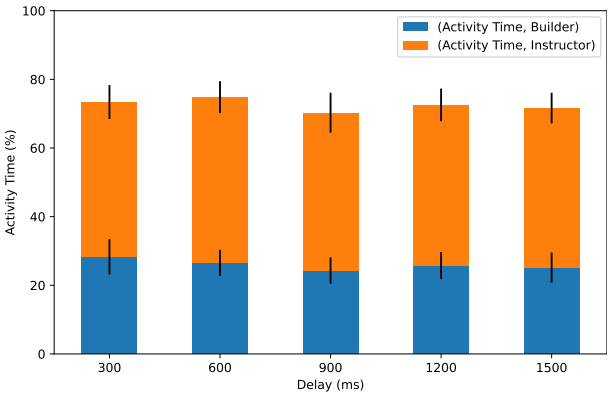
To ensure uniformity and standardization of the audio signals, the audio files were normalized



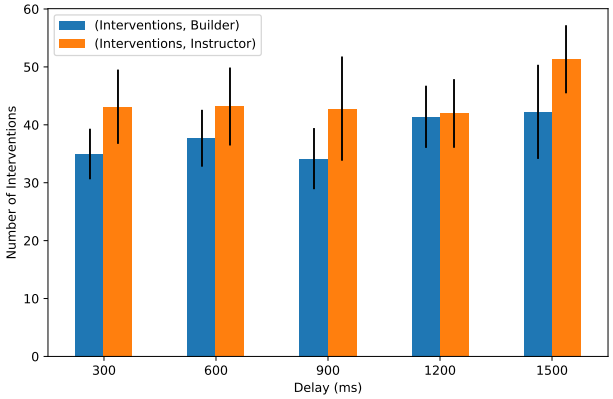
(a) Audio segment in dB with the signal in blue, running squared average of 200 ms in orange and threshold for activity in red.



(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.

Figure 5.21: Audio results.

to -26 dBov according to ITU-T Rec. P.56 [123]. 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 Fig. 5.21a 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 Fig. 5.21b the average time taken to finish the different figures for each role and delay. According to this graph, we can see that the average values increase by 1500 ms for the instructors and from 1200 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 Fig. 5.21d 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 each intervention as the time between two silences of more than 200 ms following the ITU-T P. 1305 [6]. The results of the number of interventions show similar results to those of the activity time per role. Together with the results shown in Fig. 5.21c, 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 1200 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.

5.4.4 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 [103], [108] 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 [104], [122].

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 [63], [124]. 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 ($\sim 45\%$) while the builder spoke for $\sim 25\%$ of the time (see Fig. 5.21b). 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 above, 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 [106] (prior experience, task complexity, and expectations), we can find a great influence of the type of task [125]. In particular, the block-building task represents the most common form of interactive collaboration in videoconferencing. That is, a conversation between two users that collaborate to perform a task [126]. 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 in order 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 think 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).

5.4.5 Conclusions

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 has 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.

Chapter 6

Contributions, Conclusions and Future Work

6.1 Contributions

The objectives of the doctoral thesis are framed within the evaluation of Quality of Experience (QoE) in Social eXtended Reality (XR) environments. Firstly, we aimed to generate methodology for the evaluation of QoE. Secondly, the study of Natural User Interfaces (NUIs)s as a form of interaction and, finally, the delay in Social XR as the most influential system parameter in QoE. This was done by contributing in two ways, the analysis of latency in Social XR systems and by providing QoE studies analyzing the impact of delay in three ways of interacting in Social XR.

Initially, it was decided to address the development of a standard methodology for the evaluation of QoE in immersive technologies. Among all the possible use cases, we contributed to the adaptation of the methodology for QoE evaluation in 2D video to 360° video. In this context, an inter-laboratory study was conducted with more than 300 participants to validate the use of ACR and DCR, as well as the proposed minimum hardware and stimuli included in the ITU recommendations for 2D video. In addition, this study also evaluated the influence of sequence duration, coding degradation and HMD device. Finally, we also contributed providing tools to the community in the form of datasets and an application for VR environments to allow the completion of questionnaires without leaving the immersive experience.

This study led to the publication of recommendation ITU P.919. This marked an important milestone in the development of the doctoral thesis. On the one hand, an important contribution was made in the area. On the other hand, by performing this study at the beginning of the dissertation, it made the following studies meet appropriate methodological standards.

In contributing to XR state of the art in interaction, we first contributed on developing image-based interactive environments based on NUIs. Specifically, the manipulation of physical objects through egocentric segmentation. After several proofs of concept, we got involved in a project to develop an industrial training tool using this interaction technique. For the

evaluation of these interaction paradigms, we used the methodology previously developed for 360° videos but adjusted to interaction scenarios. Therefore, we contributed in adapting and testing standardized methodology for the evaluation of interaction in XR.

After a first contribution positively analyzing the training environment, we made a second contribution analyzing even more interaction techniques that we developed during the project. In addition, with the second study we also contributed by adapting and testing the methodology for immersive technologies in an XR interaction context.

While developing different methods of interaction, we realized that one of those factors that must be aligned with reality as closely as possible, and which is fundamental to creating realistic experiences, is latency. Aligned with the scope of the thesis, we have contributed on the study of latency from the QoE point of view. When studying latency as a SIF in Social XR, the first thing we identified was that there were different latencies that could affect the user at the same time. During the development of the thesis, the delays studied are referred to from a user's point of view. That is, from the time the user interacts until he/she can watch it. To isolate each of the processes that introduced different delays we contributed to the state of the art by presenting a common framework in which the different processes that contribute to latency are dissected. In this regard, we differentiate the following delay paradigms: viewport delay, local interaction delay and remote interaction delay. The first major separation we made in the delays were those that affect us from a physiological point of view, and those that only cause us disaffection. Specifically, the viewport delay is closely related to a conflict in the vestibular system causing cybersickness. However, local and remote interaction delays, by preserving the coherence between our movements and the update of our environment, cause a decrease in the overall quality and the feeling of self and social presence.

Once in the realm of QoE, the impact of these delays on the user's perception depends on previous experience, expectations and the specific use case. In this sense, use cases involving immersive technologies are where there are more diverse interactions. Furthermore, there is a greater diversity of use cases in immersive technologies. This makes that depending on the use cases the tolerances to delays are different under the same interaction paradigm. During the state of the art study of different QoE studies evaluating latency we could observe a pattern. Up to a certain point of latency the quality was not affected at all, once the delay was noticeable (perception value), the quality dropped to a point where it is considered unacceptable (acceptance value). These values were already taken into account in previous ITU recommendations for QoE prediction for interactive multimedia transmission. However, in 2D transmission, use cases used to be restricted to very low latency scenarios, such as remote gaming or more tolerable latencies, such as videoconferencing. In XR, however, the number of use cases increases, in principle, because local interaction is mediated by devices, and because of the free movement possibilities it allows. By studying the state of the art with respect to latency values for different use cases, we contribute by adapting a model that, instead of having parameters according to low/high latency scenarios, accepts perception and acceptance values measured in subjective studies. In addition, thanks to the knowledge acquired during the thesis, we contributed in the drafting of ITU-T Rec. P.QMX which will be the next ITU-T Rec. P.1320. This ITU Recommendation is focused on the assessment of XR meetings.

However, not all delays were studied, during the development of the thesis we contributed to the state of the art by conducting three QoE studies addressing the different delays related to our video-based Social XR scheme. The first study addressed the latency allowed when viewing environments captured using 360° video in different configurations. The second study addressed the study of self-latency. To carry out this study, an ITU-T task for interaction evaluation was adapted. In the second study we addressed the self-view latency based on photorealistic egocentric segmentation. This study contributes to presenting the adaptation of a standardized 2D task to immersive environments and to establishing acceptable delays for self-perception. Lastly, the third and last study is the first to evaluate the impact of latency with volumetric video-based representation in Social XR videoconference. Furthermore, this provides the first study of its kind in the state of the art. This third study, in addition to being the first study of latencies for volumetric video in Social XR, provides guidelines on how to evaluate interactive XR environments both in the development of the tools to carry out the study as well as in a methodology for analyzing conversations in Social XR.

6.2 Conclusions

In the work carried out for the standardization of 360° video methodologies we established some guidelines to undertake QoE studies for immersive technology. Furthermore, during the development of natural interaction and Social XR studies, we have also successfully followed those guidelines for evaluating QoE (as far as they were adaptable). As a result we conclude that when conducting a QoE experiment with interactive video elements it is vital to follow the following guidelines. 1) The visual quality must be in accordance with the viewing device and the user's expectations. 2) The protocol of the P.919 recommendation must be followed, including breaks and immersive voting. 3) It is highly recommended to perform a previous pilot due to the novelty of the technology and the numerous possibilities of failure of the experiment setup.

Thanks to the developments of the EPSILON project, we have explored the integration of NUIs under real requirements. Along with these developments we have conducted QoE studies to evaluate the impact of these developments on the user experience. As a technical lesson learned in this area, it should be noted that when creating experiences that attempt to introduce physical space into virtual space to create realistic interactions, it is vital that the introduction of these elements be as close to reality as possible in terms of perception. In other words, it is important to meet the expectations of feedback and visual fidelity so that users do not feel alienated from the immersion. This may not be as true when we are talking about non-existent or virtual elements, as users do not have as high expectations in this regard.

Delay was the technical factor selected as the core of our research. First, we conclude that there are a multitude of new use cases where video-based solutions are being explored (telesurgery, teleoperation, teledriving) while others are being adapted from well established use cases (videoconferencing). Besides, the impact of delay can vary from a subjective discomfort to a situation of physical discomfort caused by the delay value and the consequence on the user's visual feedback. In this sense, viewport rendering is the most important and strict. Other

delays studied during the thesis such as local interaction and remote interaction , which cause mismatches and unupdated elements in the virtual world, affect mainly subjectively to the user. In addition, we have discovered that latency models for video based on boundary situations (very low or relatively high requirement) does not fit the variety of use cases postulated in Social XR. In this sense, the difference between the usual use cases for video and those postulated for Social XR lies in the fact that local interaction is computer-mediated. For example, the visual feedback from your hands, or controlling an object, in the physical reality occur with some physical delay, and in Social XR with a technology-mediated delay.

In light of our results, we propose that addressing the effect of delay on new cases of Social XR requires a QoE study. Specifically, the limits of acceptance and perceived latencies for new use cases of Social XR should be determined. For this purpose we provide a model adapted from the ITU along with our thesis results. In addition, along with our video-based Social XR proposal, we have studied three delays in different Social XR configurations. Thanks to these studies we have been able to give acceptability values to the entire proposed Social XR video-based pipeline.

Regarding how to perform latency-centric QoE experiments, we conclude that is essential to correctly measure M2P latency in video systems. We deduce that it is necessary to develop a setup that allows to modify the latency, starting from a minimum latency (it would be necessary to connect an Ethernet cable between the systems) so that we can have a complete picture of how the QoE value advances while the delay increases. To select latency values for the study, it is necessary to conduct a pilot study with few questions focused on system performance and many delay conditions with the purpose of identifying acceptance and perception values. Another conclusion regarding task-based QoE experiments is that it is important to set similar conditions in the tasks. Specifically, it is necessary to set the same number of subtasks for all conditions to avoid the learning effect. These guidelines are given in purpose to measure objective factors such as the time to complete a task with the lowest possible noise. However, in the particular case of delay, we observed that subjectively unacceptable values of delay do not necessarily have an impact on the execution time or performance of the task (see self-view and videoconferencing delay results). We attribute these results to our ability to adapt to delays. This makes it even more critical to correctly measure the sensing and delay values.

6.3 Future Work

Our work on NUI-based training systems shows great promise for creating realistic training experiences. Further, incorporating social interaction holds the potential to not only boost these capabilities but also enable researchers to establish new methodologies for QoE assessment specifically focused on NUIs within this training context. Exploring these distinctions and developing targeted QoE assessment methodologies for NUI-based construction training will be crucial for ensuring the effectiveness and user adoption of this promising XR application.

Regarding QoE methodology for video-based communications systems, the next future lines of research in immersive methology standardization must take into account the new hardware possibilities offered by immersive technologies. Among them, the possibility of having

untethered 6DOF experiences and the higher refresh rates and resolution they offer. In this sense, it is likely that the exploration patterns of users, as well as the type of content shown will have to be analyzed in detail to understand the possible methodological changes implied by these new technologies.

In this regard, during the final stages of the thesis, a QoE study was performed by visualizing pre-recorded dynamic volumetric avatars to analyze the validity of ACR and how different compression errors affected QoE.

Another line of future research should address how the new volumetric video solutions affect QoE measurement, taking into account not only bitrate and resolution measurements, but also volumetric capture positional faults as well as different forms of coding based on the user's position. In the area of volumetric video Social XR there are still many lines of research to be pursued. On the one hand, existing volumetric Social XR systems make use of rather cumbersome systems that require expert personnel to configure. On the other hand, offline avatar capture solutions offer unsatisfactory results for users (see Horizon Worlds case). In line with this, QoE evaluation methodologies must continue to be updated with the different developments that immersive technology offers.

References

[1] A. P. et al., "QUALINET white paper on definitions of immersive media experience (imex)", *CoRR*, vol. abs/2007.07032, 2020. arXiv: 2007.07032.

[2] P. Pérez, E. Gonzalez-Sosa, J. Gutiérrez, and N. García, "Emerging immersive communication systems: Overview, taxonomy, and good practices for qoe assessment", *Frontiers in Signal Processing*, vol. 2, 2022. DOI: 10.3389/frsip.2022.917684.

[3] K. Brunnström *et al.*, *Qualinet White Paper on Definitions of Quality of Experience*. Mar. 2013.

[4] Rec. ITU-R BT.500-15, "Methodology for the subjective assessment of the quality of television pictures", 2023.

[5] Recommendation ITU-T P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications", 2008.

[6] Rec. ITU-T P.1305, "Effect of delays on telemeeting quality", 2016.

[7] Rec. ITU-T G.107, "The E-model: a computational model for use in transmission planning", 2015.

[8] U. Radhakrishnan, K. Koumaditis, and F. Chinello, "A systematic review of immersive virtual reality for industrial skills training", *Behaviour and Information Technology*, vol. 40, no. 12, pp. 1310–1339, 2021.

[9] A. Villegas, P. Perez, R. Kachach, F. Pereira, and E. Gonzalez-Sosa, "Realistic training in VR using physical manipulation", *Proceedings - 2020 IEEE Conference on Virtual Reality and 3D User Interfaces, VRW 2020*, pp. 109–118, 2020.

[10] ITU-R, *Methodology for the subjective assessment of the quality of television pictures*, Recommendation BT.500-14, Oct. 2019.

[11] S. Wang *et al.*, "Modeling and characterizing user experience in a cloud server based mobile gaming approach", in *GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference*, 2009, pp. 1–7.

[12] Rec. ITU-T G.1072, "Opinion model predicting gaming quality of experience for cloud gaming services", 2021.

[13] ITU-T, *Subjective test methodologies for 360º video on head-mounted displays*, Recommendation P.919, Oct. 2020.

[14] ITU-T, *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment*, Recommendation P.913, Mar. 2016.

[15] ITU-T, *Subjective assessment methods for 3D video quality*, Recommendation P.915, Mar. 2016.

[16] ITU-T, *Influencing factors on quality of experience (QoE) for virtual reality (VR) services*, Recommendation G.1035, May 2020.

[17] H. Jun, M. R. Miller, F. Herrera, B. Reeves, and J. N. Bailenson, "Stimulus sampling with 360-Videos: Examining head movements, arousal, presence, simulator sickness, and preference on a large sample of participants and videos", *IEEE Transactions on Affective Computing*, Jun. 2020, early access.

[18] J. Song, F. Yang, W. Zhang, W. Zou, Y. Fan, and P. Di, "A fast FoV-switching DASH system based on tiling mechanism for practical omnidirectional video services", *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2366–2381, Sep. 2020.

[19] C. Cortés, P. Pérez, J. Gutiérrez, and N. García, "Influence of video delay on quality, presence, and sickness in viewport adaptive immersive streaming", in *International Conference on Quality of Multimedia Experience*, Jun. 2020, pp. 56–59.

[20] N. Staelens, S. Moens, W. Van den Broeck, *et al.*, "Assessing quality of experience of IPTV and video on demand services in real-life environments", *IEEE Transactions on Broadcasting*, vol. 56, no. 4, pp. 458–466, Dec. 2010.

[21] P. Frohlich, S. Egger, R. Schatz, M. Muhlegger, K. Masuch, and B. Gardlo, "QoE in 10 seconds: Are short video clip lengths sufficient for Quality of Experience assessment?", in *International Workshop on Quality of Multimedia Experience*, Jul. 2012.

[22] S. Tavakoli, K. Brunnström, J. Gutiérrez, and N. García, "Quality of experience of adaptive video streaming: Investigation in service parameters and subjective quality assessment methodology", *Signal Processing: Image Communication*, vol. 39, pp. 432–443, Nov. 2015.

[23] S. Knorr, C. Ozcinar, C. O. Fearghail, and A. Smolic, "Director's Cut - A combined dataset for visual attention analysis in cinematic VR content", in *European Conf. on Visual Media Production*, Dec. 2018.

[24] F. De Simone, J. Gutiérrez, and P. Le Callet, "Complexity measurement and characterization of 360-degree content", in *Human Vision and Electronic Imaging*, Jan. 2019, pp. 216-1–216-7.

[25] Y. Wang, Z. Chen, and S. Liu, "Equirectangular projection oriented intra prediction for 360-degree video coding", in *IEEE Int. Conference on Visual Communications and Image Processing*, Dec. 2020.

[26] C. Cortes, P. Perez, and N. Garcia, "Unity3D-based app for 360VR subjective quality assessment with customizable questionnaires", in *2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin)*, IEEE, Sep. 2019, pp. 281–282. DOI: 10.1109/ICCE-Berlin47944.2019.8966170.

[27] A. Singla, W. Robitza, and A. Raake, "Comparison of subjective quality test Methods for omnidirectional video quality evaluation", in *International Workshop on Multimedia Signal Processing*, Sep. 2019.

[28] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness", *The International Journal of Aviation Psychology*, vol. 3, no. 3, pp. 203–220, Jul. 1993. DOI: 10.1207/s15327108ijap0303_3.

[29] P. Pérez, N. Oyaga, J. J. Ruiz, and A. Villegas, "Towards systematic analysis of cybersickness in high motion omnidirectional video", in *Int. Conference on Quality of Multimedia Experience*, May 2018.

[30] H. T. Tran, N. P. Ngoc, C. T. Pham, Y. J. Jung, and T. C. Thang, "A subjective study on QoE of 360 video for VR communication", in *2017 IEEE 19th International Workshop on Multimedia Signal Processing, MMSP 2017*, vol. 2017-Janua, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 1–6. DOI: 10.1109/MMSP.2017.8122249.

[31] J. Kim, W. Kim, S. Ahn, J. Kim, and S. Lee, "Virtual Reality Sickness Predictor : Analysis of visual-vestibular conflict and VR contents", in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, May 2018, pp. 1–6. DOI: 10.1109/QoMEX.2018.8463413.

[32] W. B. Stone III, "Psychometric evaluation of the Simulator Sickness Questionnaire as a measure of cybersickness", Ph.D. dissertation, Iowa State University, 2017.

[33] S. Doolani, C. Wessels, V. Kanal, *et al.*, "A review of extended reality (xr) technologies for manufacturing training", *Technologies*, vol. 8, no. 4, 2020.

[34] K. Shankhwar, T.-J. Chuang, Y.-Y. Tsai, and S. Smith, "A visuo-haptic extended reality–based training system for hands-on manual metal arc welding training", *The International Journal of Advanced Manufacturing Technology*, vol. 121, no. 1, pp. 249–265, Jul. 2022.

[35] E. González-Sosa, P. Perez-Garcia, D. Gonzalez-Morin, and A. Villegas, "Subjective Evaluation of Egocentric Human Segmentation for Mixed Reality", in *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, IEEE, 2021.

[36] A. A. Akanmu, J. Olayiwola, O. Ogunseiju, and D. McFeeters, "Cyber-physical postural training system for construction workers", *Automation in Construction*, vol. 117, no. May, p. 103 272, 2020.

[37] O. Almousa, J. Prates, N. Yeslam, *et al.*, "Virtual Reality Simulation Technology for Cardiopulmonary Resuscitation Training: An Innovative Hybrid System With Haptic Feedback", *Simulation and Gaming*, vol. 50, no. 1, pp. 6–22, 2019.

[38] B. Spittle, M. Frutos-Pascual, C. Creed, and I. Williams, "A review of interaction techniques for immersive environments", *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2022.

[39] M. Orduna, P. Perez, and et al., "Methodology to assess quality, presence, empathy, attitude, and attention in 360-degree videos for immersive communications", *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.

[40] G. López, L. Quesada, and L. A. Guerrero, "Alexa vs. siri vs. cortana vs. google assistant: A comparison of speech-based natural user interfaces", in *Advances in Human Factors and Systems Interaction: Proceedings of the AHFE 2017 International Conference on Human Factors and Systems Interaction, July 17- 21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*, Springer, 2018, pp. 241–250.

[41] B. G. Witmer, M. J. Singer, and B. G. Witmer, "Measuring Presence in Virtual Environments: A Presence Questionnaire", *Presence: Teleoperators and Virtual Environments*, vol. 7, no. 3, pp. 225–240, Jun. 1998.

[42] P. Pérez, E. González-Sosa, R. Kachach, F. Pereira, and Á. Villegas, "Ecological validity through gamification: An experiment with a mixed reality escape room", in *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, IEEE, 2021, pp. 179–183.

[43] J. Gutierrez, P. Perez, M. Orduna, *et al.*, "Subjective evaluation of visual quality and simulator sickness of short 360 videos: Itu-t rec. p. 919", *IEEE Trans. on Multimedia*, 2021.

[44] C. Cortés, M. Rubio, P. Pérez, B. Sánchez, and N. García, "Qoe study of natural interaction in extended reality environment for immersive training", in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2022, pp. 363–368.

[45] M. Borges, A. Symington, B. Coltin, T. Smith, and R. Ventura, "Htc vive: Analysis and accuracy improvement", in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 2610–2615.

[46] S. Garrido-Jurado, R. Muñoz-Salinas, F. Madrid-Cuevas, and R. Medina-Carnicer, "Generation of fiducial marker dictionaries using mixed integer linear programming", *Pattern Recognition*, vol. 51, pp. 481–491, 2016.

[47] Dmytro Kryvoruchko, *Screenstream*, Feb. 19, 2024.

[48] R. Kachach *et al.*, "The Owl: Immersive telepresence communication for hybrid conferences", in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2021, pp. 451–452.

[49] I. Viola, J. Jansen, S. Subramanyam, I. Reimat, and P. Cesar, "VR2Gather: A collaborative social VR system for adaptive multi-party real-time communication", *IEEE MultiMedia*, pp. 1–13, 2023. DOI: 10.1109/MMUL.2023.3263943.

[50] P. Carballeira, C. Carmona, C. Díaz, *et al.*, "Fvv live: A real-time free-viewpoint video system with consumer electronics hardware", *IEEE Transactions on Multimedia*, vol. 24, pp. 2378–2391, 2022. DOI: 10.1109/TMM.2021.3079711.

[51] P. Carballeira *et al.*, "FVV Live: A real-time free-viewpoint video system with consumer electronics hardware", *IEEE Transactions on Multimedia*, pp. 2378–2391, 2022.

[52] V. Kelkkanen *et al.*, "Synchronous remote rendering for vr", *Int. J. Comput. Games Technol.*, Jan. 2021.

[53] D. G. Morín, P. Pérez, and A. G. Armada, "Toward the distributed implementation of immersive augmented reality architectures on 5g networks", *IEEE Communications Magazine*, no. 2, pp. 46–52, 2022.

[54] J. Skowronek *et al.*, "Quality of experience in telemeetings and videoconferencing: A comprehensive survey", *IEEE Access*, pp. 63 885–63 931, 2022.

[55] K. Brunnström and et al., "Latency impact on Quality of Experience in a virtual reality simulator for remote control of machines", *Signal Processing: Image Communication*, vol. 89, p. 116 005, 2020. DOI: 10.1016/j.image.2020.116005.

[56] C. Attig *et al.*, "System latency guidelines then and now – is zero latency really considered necessary?", in *Engineering Psychology and Cognitive Ergonomics: Cognition and Design*, Springer International Publishing, 2017, pp. 3–14.

[57] S. Palmisano, R. S. Allison, and J. Kim, "Cybersickness in head-mounted displays is caused by differences in the user's virtual and physical head pose", *Frontiers in Virtual Reality*, vol. 1, 2020. DOI: 10.3389/frvir.2020.587698.

[58] P. Pérez, "Exploring the realverse: Building, deploying, and managing qoe in xr communications", in *2022 ITU Kaleidoscope- Extended reality – How to boost quality of experience and interoperability*, 2022, pp. 1–11.

[59] P. Caserman, M. Martinussen, and S. Göbel, "Effects of end-to-end latency on user experience and performance in immersive virtual reality applications", in *Entertainment Computing and Serious Games*, Cham: Springer International Publishing, 2019, pp. 57–69.

[60] J.-P. Stauffert, F. Niebling, and M. Latoschik, "Latency and cybersickness: Impact, causes, and measures. a review", *Frontiers in Virtual Reality*, vol. 1, Nov. 2020. DOI: 10.3389/frvir.2020.582204.

[61] J. Kim, W. Luu, and S. Palmisano, "Multisensory integration and the experience of scene instability, presence and cybersickness in virtual environments", *Computers in Human Behavior*, p. 106 484, 2020.

[62] C. Cortés, P. Pérez, J. Gutiérrez, and N. García, "Influence of video delay on quality, presence, and sickness in viewport adaptive immersive streaming", in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–4. DOI: 10.1109/QoMEX48832.2020.9123114.

[63] C. Cortés, J. Gutiérrez, P. Pérez, I. Viola, P. César, and N. Garćıa, "Impact of self-view latency on quality of experience: Analysis of natural interaction in xr environments", in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 3131–3135. DOI: 10.1109/ICIP46576.2022.9897983.

[64] S. Xu *et al.*, "Determination of the latency effects on surgical performance and the acceptable latency levels in telesurgery using the dv-trainer® simulator", *Surgical Endoscopy*, vol. 28, no. 9, pp. 2569–2576, 2014.

[65] P. Pérez, "Exploring the realverse: Building, deploying, and managing qoe in xr communications", in *2022 ITU Kaleidoscope- Extended reality – How to boost quality of experience and interoperability*, 2022, pp. 1–11.

[66] J. Tam *et al.*, "Video increases the perception of naturalness during remote interactions with latency", in *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '12, Austin, Texas, USA: Association for Computing Machinery, 2012, pp. 2045–2050.

[67] J. C. Tang, "Why do users like video? studies of multimedia-supported collaboration", USA, Tech. Rep., 1992.

[68] Polycom, "Supporting real-time traffic: Preparing your IP network for video conferencing. Tech. rep.", 2006.

[69] C. Cortés, I. Viola, J. Gutiérrez, *et al.*, "Delay threshold for social interaction in volumetric extended reality communication", *ACM Trans. Multimedia Comput. Commun. Appl.*, Mar. 2024, Just Accepted. DOI: 10.1145/3651164.

[70] 5G Automotive Association, "5GAA (2020) Tele-operated Driving (ToD) use cases and technical requirements Tech. rep.", 2020.

[71] K. Brunnström *et al.*, "Latency impact on quality of experience in a virtual reality simulator for remote control of machines", *Signal Processing: Image Communication*, vol. 89, p. 116 005, 2020.

[72] Y. Ren and H. Fuchs, "Faster feedback for remote scene viewing with pan-tilt stereo camera", in *2016 IEEE Virtual Reality (VR)*, 2016, pp. 273–274. DOI: 10.1109/VR.2016.7504759.

[73] M. Finžgar and P. Podržaj, "Machine-vision-based human-oriented mobile robots: A review", *Strojniški vestnik - Journal of Mechanical Engineering*, vol. 63, pp. 331–348, May 2017. DOI: 10.5545/sv-jme.2017.4324.

[74] A. Singla, S. Göring, A. Raake, B. Meixner, R. Koenen, and T. Buchholz, "Subjective quality evaluation of tile-based streaming for omnidirectional videos", in *Proceedings of the 10th ACM Multimedia Systems Conference, MMSys 2019*, Association for Computing Machinery, Inc, Jun. 2019, pp. 232–242. DOI: 10.1145/3304109.3306218.

[75] A. Grzelka, A. Dziembowski, D. Mieloch, O. Stankiewicz, J. Stankowski, and M. Domanski, "Impact of Video Streaming Delay on User Experience with Head-Mounted Displays", in *2019 Picture Coding Symposium (PCS)*, IEEE, Nov. 2019, pp. 1–5. DOI: 10.1109/PCS48520.2019.8954527.

[76] C. Youngblut, "Experience of Presence in Virtual Environments", *Virgina. Institute for Defense Analyses*, no. September, p. 158, 2003.

[77] B. Lee, B. Bach, T. Dwyer, and K. Marriott, *Immersive Analytics*, 2019. DOI: 10.1109/MCG.2019.2906513.

[78] T. Aykut, J. Xu, and E. Steinbach, "Realtime 3D 360-Degree Telepresence with Deep-Learning-Based Head-Motion Prediction", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 231–244, Mar. 2019. DOI: 10.1109/JETCAS.2019.2897220.

[79] M. F. Syawaludin, C. Kim, and J. Hwana, "Hybrid Camera System for Telepresence with Foveated Imaging", in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2019, pp. 1173–1174. DOI: 10.1109/VR.2019.8798011.

[80] P. Perez and J. Escobar, "MIRO360: A Tool for Subjective Assessment of 360 Degree Video for ITU-T P.360-VR", in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2019, pp. 1–3. DOI: 10.1109/qomex.2019.8743216.

[81] C. Cortés, P. Pérez, and N. García, "Unity3d-based app for 360vr subjective quality assessment with customizable questionnaires", in *2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin)*, 2019, pp. 281–282. DOI: 10.1109/ICCE-Berlin47944.2019.8966170.

[82] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire", *Presence: Teleoperators and Virtual Environments*, vol. 7, no. 3, pp. 225–240, 1998. DOI: 10.1162/105474698565686.

[83] Y. S. De La Fuente, G. S. Bhullar, R. Skupin, C. Hellge, and T. Schierl, "Delay Impact on MPEG OMAF's Tile-Based Viewport-Dependent 360° Video Streaming", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 18–28, Mar. 2019. DOI: 10.1109/JETCAS.2019.2899516.

[84] A. Singla, S. Fremerey, W. Robitza, and A. Raake, "Measuring and comparing qoe and simulator sickness of omnidirectional videos in different head mounted displays", May 2017. DOI: 10.1109/QoMEX.2017.7965658.

[85] M. McGill and et al., "A dose of reality: Overcoming usability challenges in vr head-mounted displays", in *Proc. 33rd Annual ACM Conf. Human Factors in Computing Systems, CHI '15*, 2015, pp. 2143–2152. DOI: 10.1145/2702123.2702382.

[86] R. Gruen and et al., "Measuring system visual latency through cognitive latency on video see-through ar devices", in *2020 IEEE Conf. on Virtual Reality and 3D User Interfaces (VR)*, 2020, pp. 791–799. DOI: 10.1109/VR46266.2020.00103.

[87] P. Caserman and et al., "Effects of End-to-end Latency on User Experience and Performance in Immersive Virtual Reality Applications", *Lecture Notes in Computer Science*, vol. 11863, pp. 57–69, 2019. DOI: 10.1007/978-3-030-34644-7{_}5.

[88] V. Dam and et al., "Effects of prolonged exposure to feedback delay on the qualitative subjective experience of virtual reality", *PLoS ONE*, vol. 13, no. 10, 2018. DOI: 10.1371/journal.pone.0205145.

[89] C. Cortes and et al., "Influence of Video Delay on Quality, Presence, and Sickness in Viewport Adaptive Immersive Streaming", in *2020 12th Int. Conf. on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–4. DOI: 10.1109/QOMEX48832.2020.9123114.

[90] A. Doumanoglou and et al., "Subjective quality assessment of textured human full-body 3D-reconstructions", in *2018 Tenth Int. Conf. on Quality of Multimedia Experience (QoMEX)*, 2018, pp. 1–6. DOI: 10.1109/QoMEX.2018.8463385.

[91] S. Jung and et al., "Over My Hand: Using a Personalized Hand in VR to Improve Object Size Estimation, Body Ownership, and Presence Sungchul", *Proc. Symposium on Spatial User Interaction - SUI '18*, pp. 60–68, 2018. DOI: 10.1145/3267782.3267920.

[92] Rec. ITU-T P.920, "Interactive test methods for audiovisual communications", 2000.

[93] M. Schmitt and et al., "Towards Individual QoE for Multiparty Videoconferencing", *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1781–1795, 2018. DOI: 10.1109/TMM.2017.2777466.

[94] J. Gutiérrez, P. Pérez, M. Orduna, A. Singla, C. Cortés, *et al.*, "Subjective evaluation of visual quality and simulator sickness of short 360° videos: ITU-T Rec. P.919", *IEEE Transactions on Multimedia*, vol. 24, pp. 3087–3100, 2022. DOI: 10.1109/TMM.2021.3093717.

[95] P. Pérez and et al., "Ecological validity through gamification: An experiment with a mixed reality escape room", in *2021 IEEE Int. Conf. Artificial Intelligence and Virtual Reality (AIVR)*, 2021, pp. 179–183. DOI: 10.1109/AIVR52153.2021.00040.

[96] A. Farshchiansadegh and et al., "Adaptation to visual feedback delay in a redundant motor task", *Journal of Neurophysiology*, vol. 113, no. 2, pp. 426–433, 2015. DOI: 10.1152/jn.00249.2014.

[97] D. George, *SPSS for windows step by step: A simple study guide and reference, 17.0 update, 10/e*. Pearson Education India, 2011.

[98] Y. Wang and et al., "Research on the Application of Medical Teaching Based on XR and VR", *Proc. 2nd Int. Conf. on Big Data and Informatization Education, ICBDIE 2021*, pp. 688–691, 2021. DOI: 10.1109/ICBDIE52740.2021.00162.

[99] J. Sánchez-Margallo and et al., "Application of Mixed Reality in Medical Training and Surgical Planning Focused on Minimally Invasive Surgery", *Frontiers in Virtual Reality*, vol. 2, 2021. DOI: 10.3389/FRVIR.2021.692641.

[100] F. Cassola and et al., "A novel tool for immersive authoring of experiential learning in virtual reality", *Proc. IEEE Conf. Virtual Reality and 3D User Interfaces, VRW 2020*, pp. 44–49, 2021. DOI: 10.1109/VRW52623.2021.00014.

[101] T. Hennig-Thurau, D. N. Aliman, A. M. Herting, G. P. Cziehso, M. Linder, and R. V. Kübler, "Social interactions in the metaverse: Framework, initial evidence, and research roadmap", *Journal of the Academy of Marketing Science*, Dec. 2022. DOI: 10.1007/s11747-022-00908-0.

[102] J. Li and P. Cesar, "Chapter 22 - social virtual reality (vr) applications and user experiences", in *Immersive Video Technologies*, G. Valenzise, M. Alain, E. Zerman, and C. Ozcinar, Eds., Academic Press, 2023, pp. 609–648. DOI: https://doi.org/10.1016/B978-0-32-391755-1.00028-6.

[103] K. Schoenenberg, A. Raake, and J. Koeppe, "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*, vol. 72, no. 5, pp. 477–487, 2014. DOI: https://doi.org/10.1016/j.ijhcs.2014.02.004.

[104] G. Berndtsson, M. Folkesson, and V. Kulyk, "Subjective quality assessment of video conferences and telemeetings", in *2012 19th International Packet Video Workshop (PV)*, 2012, pp. 25–30. DOI: 10.1109/PV.2012.6229740.

[105] G. W. Cermak, "Multimedia quality as a function of bandwidth, packet loss, and latency", *International Journal of Speech Technology*, vol. 8, no. 3, pp. 259–270, Sep. 2005. DOI: 10.1007/s10772-006-6368-3.

[106] C. Attig, N. Rauh, T. Franke, and J. F. Krems, "System latency guidelines then and now – is zero latency really considered necessary?", in *Engineering Psychology and Cognitive Ergonomics: Cognition and Design*, D. Harris, Ed., Cham: Springer International Publishing, 2017, pp. 3–14.

[107] J. Tam, E. Carter, S. Kiesler, and J. Hodgins, "Video increases the perception of naturalness during remote interactions with latency", in *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '12, Austin, Texas, USA: Association for Computing Machinery, 2012, pp. 2045–2050. DOI: 10.1145/2212776.2223750.

[108] K. Schoenenberg, A. Raake, and P. Lebreton, "Conversational quality and visual interaction of video-telephony under synchronous and asynchronous transmission delay", in *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2014, pp. 31–36. DOI: 10.1109/QoMEX.2014.6982282.

[109] Rec. ITU-T P.920, "Interactive test methods for audiovisual communications", 2000.

[110] J. Li, S. Subramanyam, J. Jansen, *et al.*, "Evaluating the user experience of a photorealistic social vr movie", in *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2021, pp. 284–293. DOI: 10.1109/ISMAR52148.2021.00044.

[111] A. Becher, J. Angerer, and T. Grauschopf, "Negative effects of network latencies in immersive collaborative virtual environments", *Virtual Reality*, vol. 24, no. 3, pp. 369–383, Sep. 2020. DOI: 10.1007/s10055-019-00395-9.

[112] R. Mekuria, M. Sanna, S. Asioli, E. Izquierdo, D. C. A. Bulterman, and P. Cesar, "A 3D tele-immersion system based on live captured mesh geometry", in *Proceedings of the 4th ACM Multimedia Systems Conference*, ser. MMSys '13, Oslo, Norway: Association for Computing Machinery, 2013, pp. 24–35. DOI: 10.1145/2483977.2483980.

[113] S. Beck, A. Kunert, A. Kulik, and B. Froehlich, "Immersive group-to-group telepresence", *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 4, pp. 616–625, 2013. DOI: 10.1109/TVCG.2013.33.

[114] N. Zioulis, D. Alexiadis, A. Doumanoglou, *et al.*, "3D tele-immersion platform for interactive immersive experiences between remote users", in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 365–369. DOI: 10.1109/ICIP.2016.7532380.

[115] I. Viola and P. Cesar, "Chapter 15 - volumetric video streaming: Current approaches and implementations", in *Immersive Video Technologies*, G. Valenzise, M. Alain, E. Zerman, and C. Ozcinar, Eds., Academic Press, 2023, pp. 425–443. DOI: https://doi.org/10.1016/B978-0-32-391755-1.00021-3.

[116] S. Orts-Escolano *et al.*, "Holoportation: Virtual 3D teleportation in real-time", in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, ser. UIST '16, Tokyo, Japan: Association for Computing Machinery, 2016, pp. 741–754. DOI: 10.1145/2984511.2984517.

[117] OBS, *Obs studio*, Feb. 19, 2023.

[118] P. Wang, X. Bai, M. Billinghurst, *et al.*, "Using a Head Pointer or Eye Gaze: The Effect of Gaze on Spatial AR Remote Collaboration for Physical Tasks", *Interacting with Computers*, vol. 32, no. 2, pp. 153–169, Jul. 2020. DOI: 10.1093/iwcomp/iwaa012. eprint: https://academic.oup.com/iwc/article-pdf/32/2/153/38856822/iwc_32_2_153.pdf.

[119] K. Gupta, G. A. Lee, and M. Billinghurst, "Do you see what i see? the effect of gaze tracking on task space remote collaboration", *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 11, pp. 2413–2422, 2016. DOI: 10.1109/TVCG.2016.2593778.

[120] Rec. ITU-T G.1035, "Influencing factors on quality of experience for virtual reality services", 2021.

[121] J. R. Lewis, "Pairs of latin squares to counterbalance sequential effects and pairing of conditions and stimuli", *Proceedings of the Human Factors Society Annual Meeting*, vol. 33, no. 18, pp. 1223–1227, 1989. DOI: 10.1177/154193128903301812. eprint: https://doi.org/10.1177/154193128903301812.

[122] M. Schmitt, S. Gunkel, P. Cesar, and D. Bulterman, "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*, ser. SAM '14, Orlando, Florida, USA: Association for Computing Machinery, 2014, pp. 13–16. DOI: 10.1145/2661126.2661135.

[123] Rec. ITU-T P.56, "Objective Measurement of Active Speech Level", 1993.

[124] S. Garg, A. Srivastava, M. Glencross, and O. Sharma, "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*, ser. CHI EA '22, New Orleans, LA, USA: Association for Computing Machinery, 2022. DOI: 10.1145/3491101.3519678.

[125] L. Battle, R. J. Crouser, A. Nakeshimana, A. Montoly, R. Chang, and M. Stonebraker, "The role of latency and task complexity in predicting visual search behavior", *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1246–1255, 2020. DOI: 10.1109/TVCG.2019.2934556.

[126] Lifesize, *2019 impact of video conferencing report*, Industrial Report, 2019.

[127] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences", *Behavior research methods*, vol. 39, no. 2, pp. 175–191, May 2007.

[128] K. Brunnström and M. Barkowsky, "Statistical quality of experience analysis: On planning the sample size and statistical significance testing", *Journal of Electronic Imaging*, vol. 27, no. 5, pp. 053013-1–11, Sep. 2018.

[129] ISO/IEC JTC1/SC29/WG11, *Common test conditions for point cloud compression*, Moving Picture Experts Group Meeting, Output doc. N18474, Geneva, Switzerland, Mar. 2019.

[130] E. d'Eon, T. M. Bob Harrison, and P. A. Chou, *8i Voxelized Full Bodies - A voxelized point cloud dataset*, SO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) Input document WG11M40059/WG1M74006, Geneva, Switzerland, Jan. 2017.

[131] MPEG, *MPEG-PCC-TMC2*, https://github.com/MPEGGroup/mpeg-pcc-tmc2, 2022.

[132] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness", *The International Journal of Aviation Psychology*, vol. 3, no. 3, pp. 203–220, Nov. 1993.

[133] I. Viola, S. Subramanyam, J. Li, and P. Cesar, "On the impact of vr assessment on the quality of experience of highly realistic digital humans: A volumetric video case study", *Quality and User Experience*, vol. 7, no. 1, p. 3, Dec. 2022.

[134] S. Subramanyam, I. Viola, J. Jansen, E. Alexiou, A. Hanjalic, and P. Cesar, "Subjective qoe evaluation of user-centered adaptive streaming of dynamic point clouds", in *International Conference on Quality of Multimedia Experience*, Sep. 2022.

[135] E. Zerman, C. Ozcinar, P. Gao, and A. Smolic, "Textured mesh vs coloured point cloud: A subjective study for volumetric video compression", in *Int. Conf. on Quality of Multimedia Experience*, May 2020.

[136] S. Rossi, I. Viola, and P. Cesar, "Behavioural analysis in a 6-DoF VR system: Influence of content, quality and user disposition", in *Workshop on Interactive EXtended Reality*, Oct. 2022.

[137] M. McGill, D. Boland, R. Murray-Smith, and S. Brewster, "A dose of reality: Overcoming usability challenges in vr head-mounted displays", in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15, Seoul, Republic of Korea: Association for Computing Machinery, 2015, pp. 2143–2152. DOI: 10.1145/2702123.2702382.

[138] Y. Chillcoat and S. DeWine, "Teleconferencing and interpersonal communication perception", *Journal of Applied Communication Research*, vol. 13, no. 1, pp. 14–32, 1985.

[139] A. Steed and R. Schroeder, "Collaboration in immersive and non-immersive virtual environments", in *Immersed in Media: Telepresence Theory, Measurement & Technology*, M. Lombard, F. Biocca, J. Freeman, W. IJsselsteijn, and R. J. Schaevitz, Eds. Cham: Springer International Publishing, 2015, pp. 263–282.

[140] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik, "The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response", *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1643–1652, 2018.

[141] P. Carballeira, C. Carmona, C. Diaz, *et al.*, "Fvv live: A real-time free-viewpoint video system with consumer electronics hardware", *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[142] J. Jansen, S. Subramanyam, R. Bouqueau, *et al.*, "A pipeline for multiparty volumetric video conferencing: Transmission of point clouds over low latency dash", in *Proceedings of the 11th ACM Multimedia Systems Conference*, ser. MMSys '20, Istanbul, Turkey: Association for Computing Machinery, 2020, pp. 341–344.

[143] M. Orduna, P. Pérez, J. Gutiérrez, and N. García, "Methodology to assess quality, presence, empathy, attitude, and attention in 360-degree videos for immersive communications", *IEEE Transactions on Affective Computing*, Feb. 2022, Early Access.

[144] C. George, M. Spitzer, and H. Hussmann, "Training in ivr: Investigating the effect of instructor design on social presence and performance of the vr user", in *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '18, Tokyo, Japan: Association for Computing Machinery, 2018.

[145] A. J. Fairchild, S. P. Campion, A. S. García, R. Wolff, T. Fernando, and D. J. Roberts, "A mixed reality telepresence system for collaborative space operation", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 814–827, 2017.

[146] C. Cortés, M. Rubio, P. Pérez, B. Sánchez, and N. García, "Qoe study of natural interaction in extended reality environment for immersive training", in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2022.

Appendices

Appendix A

Scientific Contributions

JOURNALS

- **(2023)** C. Cortés et al., "Delay threshold for social interaction in volumetric eXtended Reality communication". ACM Transactions on Multimedia Computing, Communications, and Applications, accepted.
- **(2023)** C. Cortés, P. Pérez N. García, 'Understanding latency and QoE in Social XR". IEEE Consumer Electronics Magazine.
- **(2021)** J. Gutiérrez, P. Pérez, M. Orduna, A. Singla, C. Cortés et al., "Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360° Videos: ITU-T Rec. P.919", IEEE Trans. on Multimedia.

CONFERENCES

- **(2023)** C. Cortés, M. Rubio, J. Gutierrez, B. Sánchez, P. Pérez, N. García, "Evaluation of interaction methods in an Extended Reality training environment". MMVE, ACM MMSys, Vancouver, Canada.
- **(2022)** D. González, M. Orduna, C. Cortés, M. J. López Morales. "The Owl: An Accessible Immersive Telepresence System for the Future of Human Communication", IEEE Globecom, Rio de Janeiro, Brasil.
  First prize at IEEE Communications Society Student Competition.
- **(2022)** C. Cortés, J. Gutiérrez, P. Pérez, I. Viola, P. César, N. García, "Impact of Self-View Latency on Quality of Experience: Analysis of Natural Interaction in XR Environments". IEEE ICIP, Bordeaux, France.
- **(2022)** C. Cortés, M. Rubio, P. Pérez, B. Sánchez, N. García, "Qoe study of natural interaction in extended reality environment for immersive training". IEEE VR(W), Christchurch, New Zealand.
- **(2022)** C. Cortés, M. Orduna, P. Pérez, N. García. "Natural Collaborative interfaces for XR immersive learning", ACM IMX, Aveiro, Portugal.
- **(2021)** C. Cortés, P. Pérez, J. Gutiérrez, N. García, "Influence of video delay on quality,

presence, and sickness in viewport adaptive immersive streaming", QoMeX, Berlin, Germany.

- **(2020)** C. Cortés, P. Pérez, N. García, "Unity3D-based app for 360VR subjective quality assessment with customizable questionnaires", IEEE ICCE-Berlin, Berlín, Germany.

CONTRIBUTIONS TO STANDARDS

- **(2022)** G. Berndtsson, A. Raake, O. Rummukainen, P. Usai, J. Skowronek, A. Toet, I. Viola, D. Lindero, M. Fiedler, H.J. Zepernick, P. Pérez, C. Cortés, "QoE Assessment of eXtended Reality (XR) Meetings (for consent)", ITU-T Study Group 12 meeting, contribution SG12-C0054, Geneva, Switzerland, June 2022.
- **(2021)** C. Cortés, P. Pérez, N.García, "Evaluating the impact of delay on QoE in immersive interactive environments", London, United Kingdom, June 2021.
- **(2020)** F. Adeyemi-Ejeye, F. Battisti, K. Brunnström, M. Carli, P. César, Z. Chen, N. Cieplińska, C. Cortés, C. Díaz, S. Fremerey, N. García, J. Gutiérrez, O. Hamsis, J. Hedlund, F. Hofmeyer, Y. Hu, L. Janowski, D. Juszka, P. Lambert, M. Leszczuk, P. Mazumdar, M. Orduna, P. Pérez, A. Raake, A. Singla, G. Van Wallendael, I. Viola, "IMG Test Phase 1 - Short Sequences: Results and Outcomes". VQEG contribution.
- **(2020)** F. Adeyemi-Ejeye, F. Battisti, K. Brunnström, M. Carli, P. César, Z. Chen, N. Cieplińska, C. Cortés, C. Díaz, S. Fremerey, N. García, J. Gutiérrez, J. Hedlund, F. Hofmeyer, Y. Hu, L. Janowski, D. Juszka, P. Lambert, M. Leszczuk, P. Mazumdar, M. Orduna, P. Pérez, A. Raake, A. Singla, G. Van Wallendael, I. Viola, "VQEG Test Plan for Quality Assessment of 360-degree Video. Phase 1: Short sequences". MPEG 131th meeting contribution.
- **(2019)** M. Orduna, C. Cortés, P. Pérez, N. García, "IMG Work plan: Pre-test discussion - UPM tests". VQEG contribution.

PUBLIC AVAILABLE DATA

- P.919 Dataset with the video sources, processed, sequences and the study results: https://www.gti.ssr.upm.es/ ccs/VQEG360Dataset

PUBLIC AVAILABLE SOFTWARE

- Supplemental material of Evaluating the Influence of the HMD, Usability, and Fatigue in 360VR Video Quality Assessments: https://git.gti.ssr.upm.es/pub/Miro360

Appendix B

Results of the experiments of section 2.2 from laboratories external to the UPM

B.1 Influence of methodology

In principle, the two methodologies employed in the test, namely ACR and DCR, were not directly compared by any of the laboratories involved. Nonetheless, as the same conditions were employed in different labs to test the influence of the sequence length for ACR and DCR, it is possible to perform an inter-lab analysis to understand the influence of the selected methodology on the final scores. In particular, we compare the results obtained in Test A (ACR: 10s vs 20s) and Test C (DCR: 10s vs 20s), as well as the ones obtained in Test B (ACR: 20s vs 30s) and Test D (DCR: 20s vs 30s). In our analysis, we exclude any sequence that was not present in both the test sessions under exam, to ensure a fair comparison.

Results of the Mann-Whitney's U test show a significant effect of test methodology for Test A with respect to Test C ($z = -6.6370,\ p < 0.001,\ r = 0.1024$), as well as for Test C with respect to Test D ($z = -3.2416,\ p = 0.0012,\ r = 0.0560$), albeit with a smaller effect size. To further understand whether the sequence length might affect the differences among methodologies, we compare the two methodologies separately per sequence length. To do so, we aggregate the results obtained in Test A, B, C, D, and E, while considering only the lowest common group of contents and distortions. Mann-Whitney's U test shows a significant effect of methodology for sequence length of 10s (Test A, Test C, and Test E: $z = -8.1081,\ p < 0.001,\ r = 0.1700$) and 20s (Test A, B, C, and D: $z = -4.9043,\ p < 0.001,\ r = 0.0870$), whereas no significant effect of methodology was observed for sequence length of 30s (Test B, Test D, and Test E: $z = -1.6306,\ p = 0.1030,\ r = 0.0329$). Results indicate that the choice of methodology might have an impact on the distribution of the scores, especially for certain sequence lengths, as MOS values are on average 0.24 higher when using the DCR methodology as opposed to the ACR methodology (for Tests A, B, C, D, and E, and all sequence lengths: $z = -8.5471,\ p < 0.001,\ r = 0.0962$). However, the effect sizes we obtain in our comparisons imply that the effect, if existing, is quite small. In addition, the patterns of the results

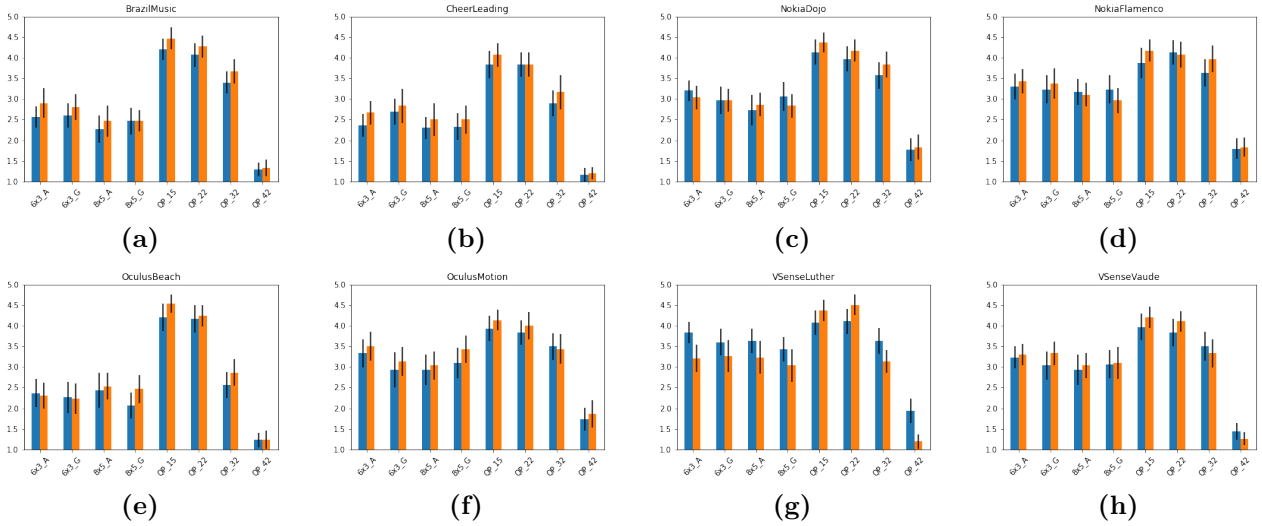


Figure B.1: Results of MOSs from Test A (Wuhan) using ACR with videos of 10s (blue) and 20s (orange). Uniform encoding schemes are indicated with the QP, non-uniform ones are named by the tiling division and transition (A: Abrupt, G: Gradual).

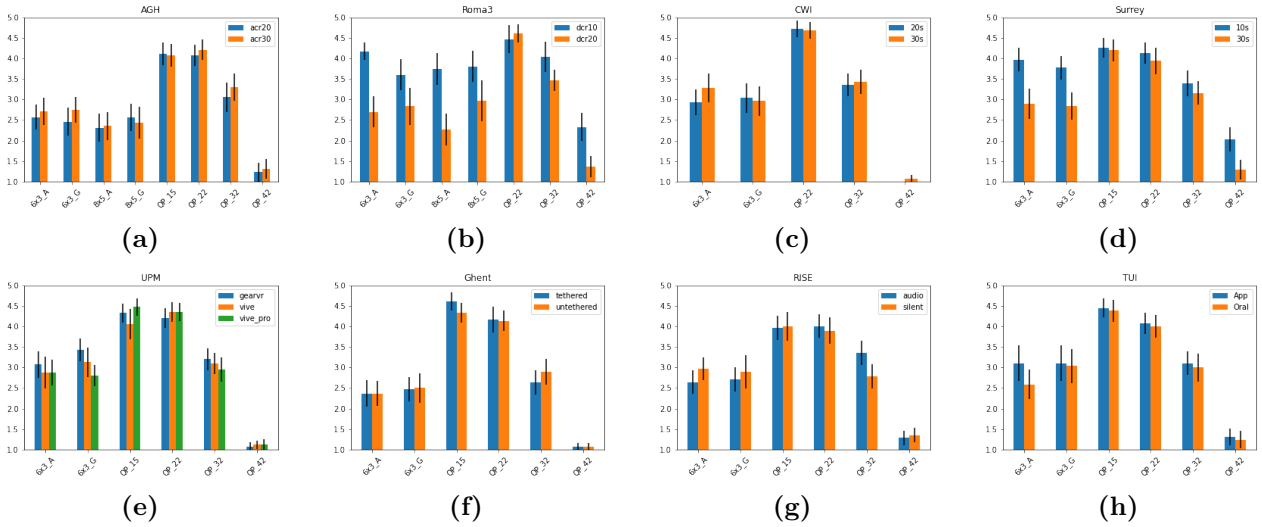


Figure B.2: Results of MOS from all laboratories (considering the tested conditions) for VSense-Luther. Charts for the rest of SRCs can be found in the supplemental material.

obtained in the involved labs (i.e., expected decreasing quality when increasing uniform QPs and no big differences among the non-uniform configurations considered in the tests, as shown in Fig. B.1 and Fig. B.2 validate the use of ACR and DCR methodologies for subjective assessment of coding quality for 360° video. Thus, these two methodologies were included in the ITU-T Rec. P.919 [13].

Table B.1: p-values for a mixed model and different test conditions. For conditions involving sequence duration also p-value without VSenseLuther sequence is presented.

ID	Lab	Test Condition	p-value with VSense-Luther	p-value without VSense-Luther
A	Wuhan	ACR, 10s vs. 20s	0.005	6.4e-06
B	AGH	ACR, 20s vs. 30s	0.326	0.754
C	Roma3	DCR, 10s vs. 20s	9.4e-09	0.089
D	CWI	DCR, 20s vs. 30s	0.014	0.001
E	Surrey	ACR: 10s vs. 30s	9.03e-06	0.035
F	UPM	GearVR vs. Vive	0.1087	N/A
	&	GearVR vs. Vive Pro	0.2230	
	Nokia	Vive vs. Vive Pro	0.0014	
G	Ghent	Tethered vs. untethered HMD	0.562	N/A
H	RISE	With vs. w/o audio	0.006	N/A
I	TUI	Scoring app vs. voice	0.046	N/A

B.2 Influence of sequence duration

Regarding the influence of sequence duration, we present in Table B.1 the p-values obtained for the different tests considering these conditions. As we can see, all compared conditions, except ACR 20s vs. 30s, are statistically significant but with different significance level. Since the obtained results are aggregated over different conditions and SRCs, the results' visual investigation is necessary. Further, inspection shows that one of the sequences (VSenseLuther) showed unexpected results compared to the other videos. For Wuhan (Test A), the 20-second sequences have, most often, higher MOS (see Fig. B.1), except for the sequence VSenseLuther. For Roma3 (Test C), again for VSenseLuther, we obtain a significant decrease in the quality, as observed in Fig. B.2 (c) (see also all the results from this lab in the supplemental material). That might have been caused by the new scene in this particular sequence, that is not displayed for the first 10 seconds. Thus, in addition we present results obtained without

the VSenseLuther sequence (see Table B.1). The new results showed the higher statistical significance of Wuhan (Test A) and CWI (Test D), while for Roma3 (Test C) the results stopped being significant. For Surrey (Test E) the significance was reduced, and for AGH (Test B) the results were still not statistically significant. For Wuhan, ACR 20s comes with higher scores. It is not an effect distinctly visible for a single scene. However, the mixed model's analysis allows us to see all the sequences together, also normalizing each sequence quality's influence. Since in Wuhan (Test A) general differences between MOS for 10s and 20s can be observed (see Fig. B.1), the overall result shows the statistical significance, and it was shadowed by VSenseLuther reverse influence. After removing this sequence, we conclude that 20-second sequences obtained higher MOS by 0.12 than 10s ($\chi^2(1) = 20.3,\ p = 6.4e-06$). Also for CWI (Test D), the effect without the VSenseLuther sequence is more substantial, and again more extended sequences obtain higher MOS by 0.14 ($\chi^2(1) = 10.2,\ p = 0.001$). The effect observed for Roma3 (Test C) is mainly, or even only, caused by the extreme difference obtained for the VSenseLuther sequence. Thus, after removing it, the effect is not observed anymore ($\chi^2(1) = 2.90,\ p = 0.089$). Again removing this sequence is necessary since it is not consistent for the first and last 10 seconds. It should be noted, that apart from Roma3 (Test C), AGH (Test B) also did not gather significantly different results, which, in this case, this could be caused by the subjects inconsistency. Since there are two contradicting subject removal algorithms described in ITU-R BT.500 [10] and ITU-T P.913 [14], we decided to not use any of them and leave for the further research this particular condition.

To go one step further and analyze for which test stimuli there were significant differences, Wilcoxon Signed-Rank tests (non-parametric tests for related samples) were computed, after checking the non-normality of the gathered scores, and applying Bonferroni corrections for multiple comparisons. Only significantly different pairs were identified with VSenseLuther: one pair (QP42, $p = 0.0002$) among 64 for Test A (Wuhan), 3 pairs (6x3-abrupt with $p = 8.6e-06$, 8x5-abrupt with $p = 8.4e-05$, QP42 with $p = 0.0003$) among 35 for Test C (Roma 3), and 2 pairs (6x3 gradual with $p = 0.0007$ and abrupt with $p = 0.0002$) among 48 for Test E (Surrey). No significantly different pairs were found for Test B (AGH) among 40 pairs and Test D (CWI) among 25 pairs.

These results evidence that no systematic effects of the sequence duration on the quality ratings are generally observed, while, as expected, differences can be obtained when using characteristic videos with changing properties during time (e.g., VSenseLuther). Thus, subjective tests of coding degradations with 360° videos can be done with sequences of 10 seconds, taking into account these effects, as reported in the ITU-T Rec. P.919 [13].

B.3 Influence of audio

To check the influence on quality assessment of watching the 360° videos with or without audio, the results from the Test H, carried out by RISE, were analyzed. The mixed model analysis shows that silent sequences obtained MOSs higher by 0.075 ($\chi^2(1) = 7.51,\ p = 0.006$). The measured difference is statistically significant but minimal, and visible only by analyzing all sequences. Analyzing the differences between all the pairs with Wilcoxon Signed-Rank tests (with Bonferroni corrections), no significant different pairs are detected among the 48

possible comparisons.

These results support that it is possible to use test stimuli either with or without audio to evaluate visual quality, as included in the ITU-T Rec. P.919 [13]. Nevertheless, it should be noted that no spatial audio was used in these tests, so it should be considered that, especially when dealing with non-uniform degradations, off-screen sound may influence audiovisual quality ratings.

B.4 Influence of method to collect ratings

To check the influence of the two tested methods to collect the observers' ratings (i.e., through the application and verbally), the results from the Test I, carried out by TU Ilmenau, were analyzed. The mixed model shows the border case with ($\chi^2(1) = 3.975396,\ p = 0.046$), which is theoretically statistically significant, but indicating a very similar performance of both methods. In fact, the post-hoc Wilcoxon Signed-Rank tests showed no significantly different pairs among the 48 test videos compared. Therefore, both voting interfaces or verbal voting are recommended in the ITU-T P.919 [13] for evaluations performed with 360° videos.

B.5 Minimum number of observers

To compute the minimum number of observers required per laboratory, we base our analysis on the desired statistical power 0.8. Given the within-subject design and the assumed non-normality of the data, we consider the case of a one-tailed Wilcoxon signed-rank statistical test aiming to determine whether one distortion leads to higher MOS scores concerning another. Assuming a type I error probability $\alpha = 0.05$, and an effect size of $r = 0.5$ (in our test, the observed range was $r = [0.46, 0.62]$), we use the free software G*Power [127] to obtain a minimum sample size of $N = 28$. This is in line with an estimation as outlined in Brunnström, and Barkowsky [128], using VQEGNumSubjTool[1]. For this, we considered a within-subject design with the same statistical power of 0.8, a standard deviation of 0.9 (which is a bit higher than we can expect in regular 2D video quality test), and a MOS difference of 1. Considering that the number of PVSs in each sub-experiment is about 50, and that we are looking at all possible comparisons (i.e., $50 \cdot 49/2 = 1225$), the result was also $N = 28$. This calculation is based on the t-test, which is more efficient as it relies on parametric statistics and would give a lower number, but considers multiple comparisons with an overall $\alpha = 0.05$ for each experiment. These results supported the recommendation, included in ITU-T Rec. P.919 [13], to have at least 28 participants in similar subjective tests with 360° videos.

[1] https://slhck.shinyapps.io/number-of-subjects/

Appendix C

Volumetric avatar assessment for Social XR

We have presented in the previous subsection a large study that led to the publication of an international recommendation. The objective of that study was to propose and validate an assessment methodology for immersive 360° video. Although 360° video fulfills the function of generating the shared environment depicted in Section 1.2, the representation of the user, also called avatar, is necessary to achieve Social XR.

However, as with 360° video, there are still a lack of specific methodologies for the evaluation of video-based avatars. Thus, this subsection presents a study that explores the validation of the methodology for QoE assessment focused on another form of immersive video, volumetric video. Specifically, volumetric video to generate avatars. Among the different ways of generating volumetric video, point clouds are one of the most suitable techniques for scenarios that require real time, therefore, they are ideal for Social XR [49]. Firstly, results on the impact of compression artifacts on the perceived QoE of the users are reported, showing the validity of Absolute Category Rating, although more than 20 observers may be needed to obtain robust conclusions. Results on users' exploration behavior show no significant differences when visualizing point clouds with different qualities, no changes in the behavior during the test session, and no correlation between exploration activity and quality assessments. Further research will be conducted to help identify appropriate methodologies for the subjective assessment of point clouds and for understanding users' exploration behavior.

C.1 Subjective experiment

According to the Social XR diagram presented in this thesis (see Fig. 1.2), Social XR involves the interaction of users within a shared environment and the representation of remote users in it. During this appendix, the term avatar will be used to refer to the visual representations of the users. Their main function is to locate and present the remote user. Therefore, avatars in social XR are audiovisual elements. Regarding techniques for visual representations, avatar generation range from simple 3D avatars to photorealistic representations in real time through

the use of volumetric video capture [49]. According to [54], volumetric video is the most realistic technique to generate avatars in Social XR.

However, as with 360° video, we are devoting efforts to validating methodology for QoE evaluation and user behavior analysis for volumetric avatars. During the last stages of the thesis, an exploratory experiment was carried out to validate methodologies in the area of volumetric avatars. Specifically, the objectives of this study were:

- To explore the validity of a simple and well-established methodology, originally designed for 2D content:
  - Check the validation of ACR for volumetric video based on pointclouds.
  - Check whether the proposed test induce simulator sickness.
- To analyze users' exploration behavior in this context and its possible influence on the subjective assessment
  - Check if there are differences in how users explore avatars according to their appearance.
  - Check if there are differences in the way people explore point clouds with different qualities.
  - Check the temporal evolution of users exploring behavior.
  - Categorize users according to their browsing patterns and check if they vote the same way.

C.2 Stimuli

Four dynamic point clouds representing humans, depicted in Fig. C.1, were used as source (SRC) contents in this experiment, specified in the MPEG Common Test Conditions [129] and published in [130]. Human point clouds were selected given our future interests on studying QoE in social XR scenarios, which include photo-realistic representations of the users. All of them contain 300 frames (at 30 fps) and 1024x1024x1024 (RGB) points. To generate the test stimuli, these point clouds were encoded using MPEG V-PCC (TMC2v15) [131] with five rate points (defining the quality of the texture, the geometry, and the precision of the occupancy map), as shown in Table C.1, and the provided configurations for all-intra encoding described in [129].

C.3 Equipment and Environment

The tests were performed at the *Universidad Politécnica de Madrid* (Spain), in a test room where the observers could move comfortably. Point clouds were visualized using Pico Neo 3, which is an untethered device. An application was developed with Unity to reproduce dynamic point clouds in a virtual environment based on an empty room with medium gray walls. The point clouds were displayed approximately in real (human) size and they were

(a) Loot **(b)** Red&Black **(c)** Soldier **(d)** Longdress

Figure C.1: Screenshots of the SRC point clouds.

Table C.1: V-PCC Rate settings for the test stimuli.

Rate	Geometry QP	Texture QP	Occupancy Map Precision
R01	32	42	4
R02	28	37	4
R03	24	32	4
R04	20	27	4
R05	16	22	2

placed at a distance of 1 meter from the starting position of the observer. The rendering shape of the points was a circle of 0.05 units, so discontinuities were not noticeable in the shapes of the point clouds from the initial position. Also, the application displayed an interface to rate the quality of the displayed point clouds. In addition to these ratings, the head position and rotation data were stored for each participant while visualizing each PC.

C.4 Methodology

The test protocol followed the general guidelines of ITU recommendations for subjective quality assessment experiments [4], [13]. In particular, ACR was used to evaluate the quality of the test point clouds, while the Simulator Sickness Questionnaire (SSQ) was used to measure cybersickness [132]. The point clouds were shown to the participants for 10 seconds and they could examine them by freely moving around them. Then, participants rated the perceptual quality within the virtual reality environment and started to visualize the following PC after pressing a button to continue. The sequence of point clouds shown to each participant was randomized. Concerning the SSQ, the participants were asked to fill it in three different moments during the tests (details in subsection C.5) to assess the evolution of the symptoms along the experiment.

C.5 Test Session

The structure of the whole test session performed with each participant was divided into seven parts, as depicted in Fig. C.2.

Figure C.2: Test session structure.

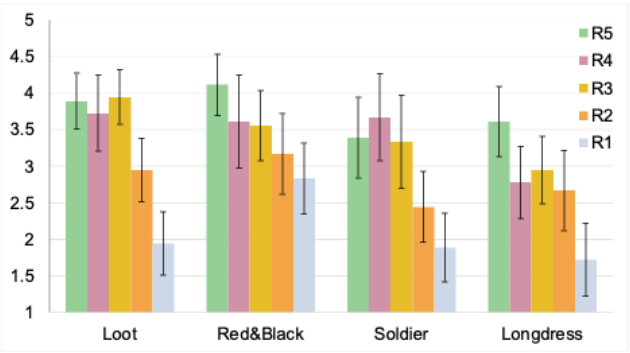


Figure C.3: Quality results.

First of all, the conditions and procedures of the experiment were explained to the participants. The welcome session also involved a vision test and the signing of the informed consent by the participant for processing his/her data according to the GDPR of the European Union. Afterward, the form with demographic data and SSQ were filled. Subsequently, a training session was conducted to make the participants familiar with the equipment, the interaction area, the rating methodology, etc., and to provide examples of the test stimuli using two dynamic point clouds with the lowest and highest quality levels. Then, a first test session, which lasted approximately 10 minutes, was conducted by visualizing and evaluating a first set of dynamic point clouds. Once it finished, there was a small break of 5 minutes for the participants to rest and fill again the SSQ. After this break, the rest of the test stimuli were displayed and evaluated. In the end, the observers filled out the last SSQ and provided their feedback about the tests. Finally, they were remunerated for their participation in this study.

C.6 Observers

Twenty participants (10 women and 10 men), aged 19-29 years (mean of 22.7 and standard deviation of 2.6), took part in the tests. Among them, 47% of the participants were international students. All observers were assessed on (corrected-to-)normal vision. Also, participants were requested to fill out a questionnaire about their experience in using VR headsets. According to the results, 74% of the participant were using it for the first time, 10% of them had used it less than 5 times, and 16% had used it more than 20 times. After all tests, one participant's data were discarded due to hardware problems during the session, and the quality ratings from another one were not considered due to errors in data collection.

C.7 Results

Quality

The MOSs obtained from the quality assessments provided by the participants for the test

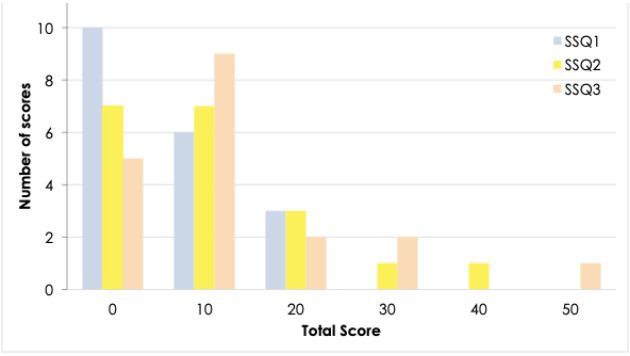


Figure C.4: SSQ results.

point clouds is shown in Fig. C.3, together with the 95% confidence intervals. In general, the expected trend of obtaining worse MOSs for more severe compression rates is shown, although similar results were obtained in various cases for R5, R4, and R3. In order to check statistically significant differences among the tested conditions, a Kolmogorov-Smirnov test was first performed, which showed normality of the data. So, paired t-tests were performed with Bonferroni corrections for multiple comparisons. The results from these tests are in accordance with the statistical significance shown by the confidence intervals in Fig. C.3. Thus, statistical significance can be assumed for those conditions where these intervals do not overlap. Firstly, it is worth noting that even the best compression rate do not provide MOSs higher than 4, which can be due to the inexperience of the participants in watching this type of content that presents holes and discontinuities that are more visible when getting too close to the point clouds. Secondly, compression artifacts have a different impact depending on the SRC point cloud, as shown by low MOSs obtained for R4 and R3 in Longdress, which is a more dynamic PC than, for example, Red&Black. It is worth mentioning that there were a few undesirable freezes with Loot, but they do not seem to have impacted the main results. These results show that, as hypothesized, ACR can be a suitable methodology for the quality assessment of dynamic point clouds with compression artifacts, although more test participants may be required to obtain more robust and significant results.

A similar trend of the results can be observed in [133], [134]. Also, in comparison with [135], the MOSs obtained for the corresponding point clouds in both tests present a high Pearson correlation (0.818), even though in that test the point clouds were visualized in a 2D screen.

Simulator Sickness

As aforementioned, the SSQ was used to evaluate the simulator sickness Our hypothesis was that, given the structure of the test session (see Fig. C.2 and the limited time in which the participants were using the HMD, the simulator sickness symptoms would be mild. Figure C.4 shows the histogram distribution of the Total Score (obtained from the ratings of the individual symptoms, according to [132]) for the three times that the participants answered the questionnaire along the whole session (i.e., at the beginning of the test, during the break between the two test sessions, and at the end of the test). Although the results show that simulator sickness may increase along the session, the obtained scores are low enough (in comparison with other validated experiments [94]) to guarantee that the procedure followed in this experiment is appropriate in terms of participants' physiological discomfort.

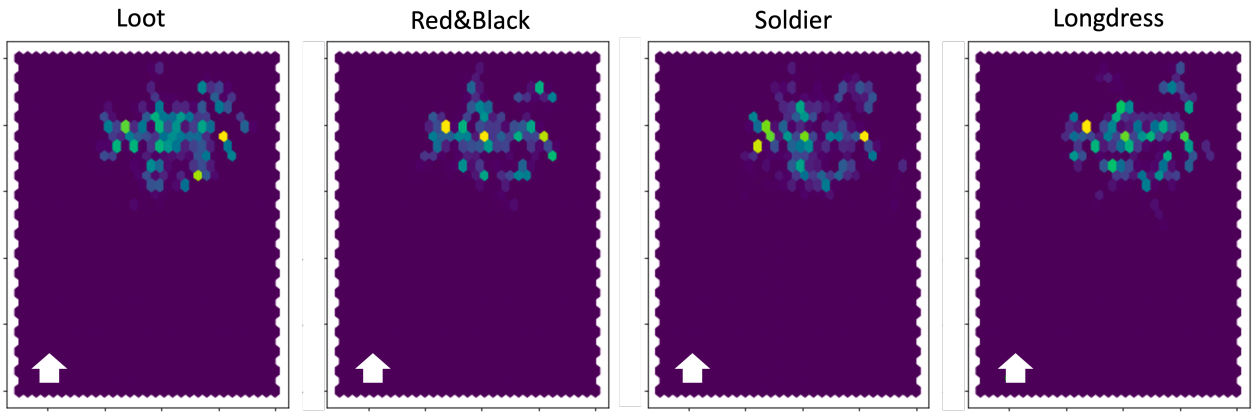


Figure C.5: Heat maps (aggregated per SRC) of the distribution of the observers' position while exploring the point clouds (white arrow with the PC's orientation).

Exploration behavior

Fig. C.5 shows the heat maps of the most visited locations (on the floor) by the observers for each SRC point cloud (aggregated for all compression rates). As it can be observed, the participants mainly explored the point clouds from a position that allowed them to not only see the front part but also around them. We hypothesized that the exploration behavior would be similar for the four considered point clouds since they are all human representations. The results support this hypothesis since no significant differences can be observed in Fig. C.5 on the way people explore the different point clouds. Possibly, a slightly higher exploration activity can be observed with Longdress, which would be in line with the findings in [136]. In this study, more dispersion in exploratory movements was found for more dynamic point clouds, which is the case of Longdress since it moves forwards and does not stay around a fixed point like the other point clouds. In addition, Fig. C.6 shows the distribution of the viewing directions in elevation for each SRC point cloud. We focus on the elevation (i.e., pitch) since we observed that the participants mainly looked straight ahead to the point clouds with minimal rotation in the yaw axis, which was also observed in [136]. As can be noticed, there are no significant differences among the different point clouds, which supports our previous statements. It can be observed that the participants had the tendency to look slightly down, which may be because they probably tend to direct their heads a bit downwards to fit the whole PC in the viewport and be able to spot and notice imperfections in any part of the PC. In general, participants (average height to the HMD of 1.6 meters) watched the point clouds at a distance of 4.3 meters, so, looking straight to the point clouds at this distance they do not fit in the visible viewport.

Similarly, Fig. C.7 and Fig. C.8 do not show differences among the exploration behaviors for different compression rates. These results contradict our hypothesis, since we expected more activity with point clouds in the high-quality range, where artifacts may be less noticeable, so the observers may search more actively to identify them for their assessment. Nevertheless, similar conclusions were obtained in [136].

To check whether the exploration behavior of the participants changed throughout the

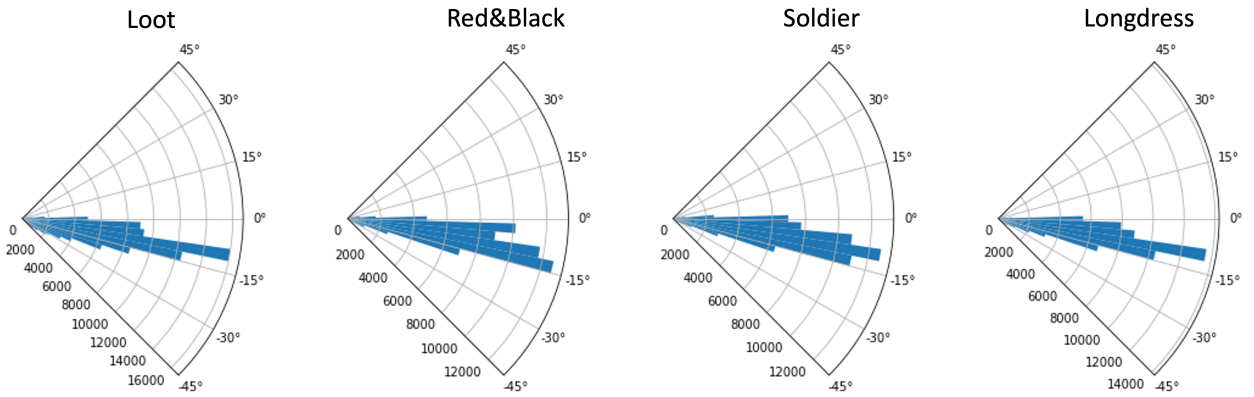


Figure C.6: Distribution of the viewing direction in elevation of the observers while exploring the PC's (aggregated per SRC).

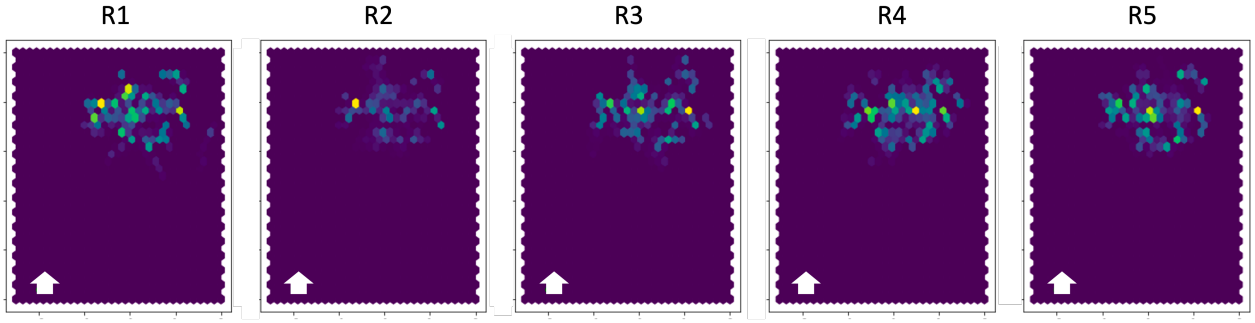


Figure C.7: Heat maps (aggregated per rates) of the distribution of the observers' position while exploring the point clouds (white arrow with the PC's orientation).

whole session, we analyze the distribution of their positions while watching the point clouds (aggregating for all point clouds) in the first and the second test session (i.e., before and after the break). In this sense, we hypothesized that, since the observers watched each SRC point cloud several times during the test, they would explore less after visualizing them the first time. The results shown in Fig. C.9c(a) contradict this hypothesis, since users seem to move more in the second session. This behavior could be explained by the inexperience of most of the participants in visualizing this type of content and in using HMDs, so in the first session users tend to be more cautious in exploring and moving, while in the second session, they start to get used to it and try to experience more. This is also supported by the average distance traveled by the participants in both test sessions, which resulted in 0.66 meters for session 1 and 0.85 meters for session 2. In addition, Fig. C.9c(b) shows the distribution of the viewing angles in elevation for both sessions. As can be seen, in the first session users looked higher and lower, extending their viewing angles between 0 and -20º, while in the second session, they focused on a narrow range between -5º and -15º. Probably, during the first session, users learned that the best way to evaluate the quality of point clouds is to look within a range of viewing angles, so that, as aforementioned, the whole PC falls within the viewport.

Finally, to analyze if there are different types of users in terms of their exploration behavior, we

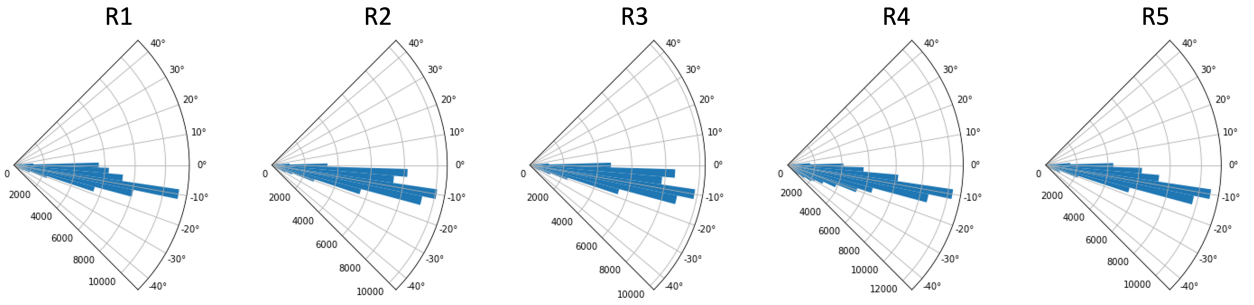
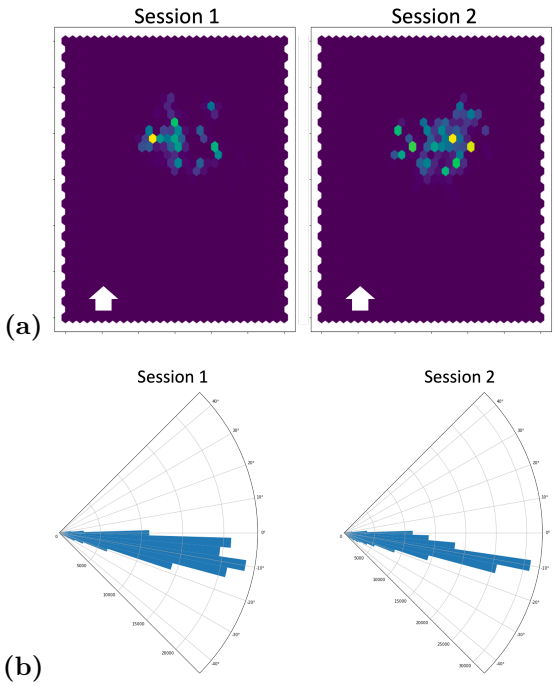


Figure C.8: Distribution of the viewing direction in elevation of the observers while exploring the PC's (aggregated per rates).

analyzed their activity by computing the average distance traveled by each participant in both test sessions. The results are shown in Fig. C.10. As we hypothesized, some observers tend to move more (e.g., participants 1, 8, etc.), while others stay almost static while observing the point clouds (e.g., participants 6, 9, etc.). To investigate if there is any relationship between the way that the participants assessed the quality of the point clouds and their exploration activity, Fig. C.11 depicts the voting patterns of each observer for all the test point clouds. While it can be seen that some participants were more positive (e.g., user 8) or negative (e.g., user 1) with their scores, no clear relationship can be found in this sense between users that explored more and those that moved less. It is worth noting that no outlier removal was applied since given the novelty of quality assessment of immersive media, traditional methods (e.g., recommended in ITU-T BT.500 [4] and ITU-T P.913 [14]) are not suitable and further research is required [94].

C.8 Conclusion

This study explored the subjective quality assessment of dynamic point clouds with compression artifacts using ACR methodology and analyzed the exploration behavior of users while visualizing them with an HMD. The results showed that ACR can be a valid methodology, but more than 20 observers may be needed for significant results. The analysis of the exploration behavior of the users did not show significant differences in exploration activity between point clouds of different qualities, changes in behavior over the test session, or correlation between exploration activity and quality assessments. Future work will focus on validating the methodology for the evaluation of transmission errors and on investigating eye-tracking data to further understand how users watch point clouds. Also, the resulting datasets and tools will be made publicly available to support the research on this topic.



(c) Distributions of the positions and viewing directions of the observers while exploring the point clouds in the two test sessions.

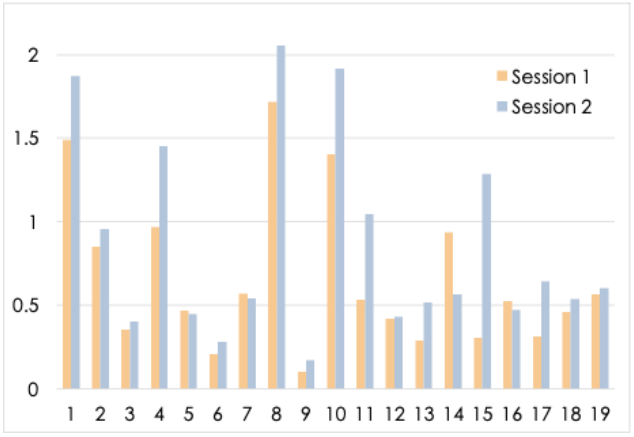


Figure C.10: Average distance in meters traveled by each user while exploring the point clouds in the two sessions.

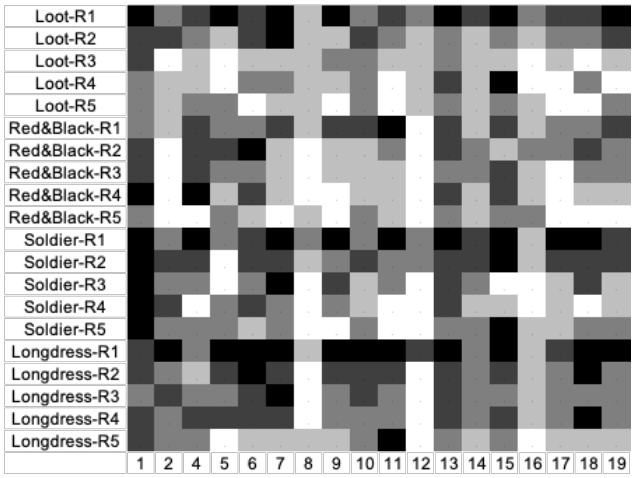


Figure C.11: Diagram of the quality scores provided by each user (black: 1, white: 5).

Appendix D

Social XR training system

Stemming from the single-user platform developed within EPSILON, our goal is the evaluation of interaction in the context of Social XR technologies. Therefore, this subsection shows the design of an experiment in the presence of an instructor. In the actual version of the tool, the role of the instructor does not exist. Instead, an updateable text guides the user through the training. This fact clashes with the primary goal of keeping realism and ultimately of the immersive technology itself. In addition, this may affect the plausibility of the representation, which is key to the feeling of immersion in interactive environments [137]. Thus, we present an study of how natural collaborative interfaces can be introduced within the XR environment.

We aim to introduce an external user into the XR environment presented in the Fig 3.1. The requirements of the XR environment make it necessary to carry out an analysis of the needs of each user to capture and integrate technology in terms of physical reality and virtual reality. For enabling communication in the XR environment it is essential to introduce verbal communication (audio) and, to boost realism, non-verbal communication [138].

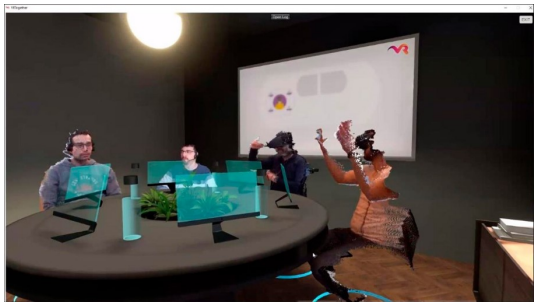
In terms of physical reality, to introduce the instructor it will be necessary to specify whether the users will share the real space or not. On the one hand, the fact that the users share a virtual space might harm the instructor immersion. On the other hand the logistics are simpler as only one room is needed. However, our intention is to propose a solution that allows remote collaboration. So, our decision is that users will not share the local space. Consequently, we will focus on solutions where the user can at least communicate verbally with the learner. Moreover, most of the analysed methods also show a visual representation of the outer user (instructor's avatar). Under the umbrella of the instructor's audiovisual capture, we classify the different solutions according to:

- The instructor's avatar can be observed from any point of view, i.e., the instructor's presentation is 3D or 2D.
- The instructor's avatar is updated in real time.
- The instructor's avatar is realistic.

After reviewing the SoA techniques [139]–[141] for avatar representation, we found five methods that fit the use case requirements. Firstly, a simple method is to introduce the instructor's

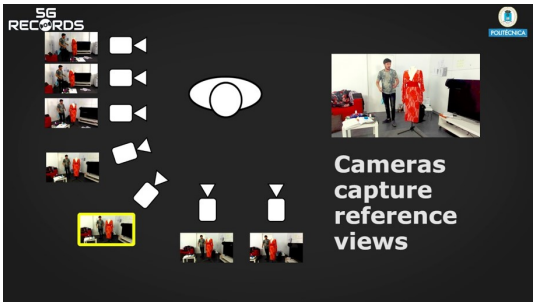
(a) 360° camera (Owl)

(b) 3D Avatar (Mozilla Hubs)



(c) Pointcloud user capture (VRTogether)

(d) Webcam Integration



(e) Free-view point video

Figure D.1: 3D models to reproduce during the task.

voice into the XR environment while a view of the XR environment is transmitted to the instructor. This solution does not need immersive hardware for the instructor. Then, if you add a simple avatar representation, we found 3D custom avatar solutions like Mozilla Hubs. In this solution, the user has a 3D avatar that can move around the XR environment using controllers. However, this is an unrealistic avatar representation and a non-real-time visual solution. Beyond, some methods capture the user and create updated realistic avatars. These methods are Freeview point video [141], pointcloud [142] capture, and simple or 360° camera capture[143]. In Table D.1, there are some specific implementations of these avatar methods. In addition, Fig. D.1 shows visual examples of these technologies within immersive environments.

	Spatial		Visual Update		Visual	
	3D	2D	Real-Time	Offline	Realistic	Virtual
Mozilla Hubs	x			x		x
PointClouds[142]	x		x		x	
Webcam[144]/ OWL[143]		x	x		x	
FVV[145]	x		x		x	
None (voice)						

Table D.1: Classification of the avatar methods

D.1 Construction use case proposal

For a truly immersive experience, realism is a key factor. In particular, virtual reality devices aim to isolate users from the outside world to transport users to another place. However, when it is necessary to interact with part of your physical reality, immersive environments use invasive elements that can worsen realism and, to some extent, immersion [9].

In a previous work [146], we developed a XR environment for immersive learning in civil work training. The interaction with tools within the environment was designed using natural interfaces. Specifically, we used the information of the HMD's frontal cameras for integrating the body of the user and the necessary tools though color segmentation. This way, we achieved a good performance in terms of satisfaction and QoE. However, the interaction with the teacher/instructor role was translated to a floating text (see Fig 3.1). In that specific element we found that the interaction was deficient. This means, the element of the floating text for giving instructions was a non-natural interface method.

For resolving this issue, we propose the integration of an instructor within the XR environment. In terms of XR design, this means to add another user to the experience. This is, building an XR collaborative learning environment. In the introduction we analyzed different methods for representing outer users in XR.

In our use case, the users are able to see their bodies into the XR in a real time and realistic way using the the front camera of the HMD. Consequently, we consider that the best options for representing the outer user while maintaining the coherence between both user representations are: Pointcloud registration, Webcam, and Free viewpoint video. Before the experiment, we will perform a pre-pilot of these three methods to have insights about the opinion of the users before assessing the QoE factors and learning in depth.

D.2 Methodology

The evaluation of the collaborative interface will be addressed through a subjective assessment in which participants will carry out tasks of the described use case. Participants will be expert participants (real instructors). So, the subjects will focus on the task and will understand it easily. The experiment considers two tasks. The first one, called "Rural task", will be to check a fiber installation on a rural environment, like Fig. 3.1 shows. The second one, called "Urban Task" will be to check a fiber installation on a telephone pole in an urban environment like

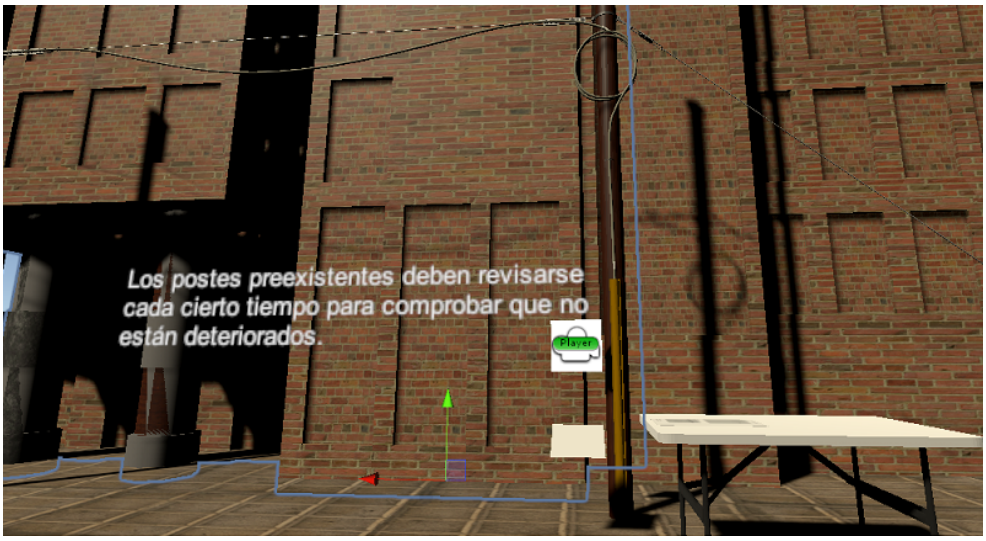


Figure D.2: Urban task environment.

	Condition 1 (with guidance)	Condition 2 (without guidance)
Test A	Rural task with text	Rural task
Test B	Rural task with avatar	Rural task
Test C	Urban task with text	Urban task
Test D	Urban task with avatar	Urban task

Table D.2: Experiment conditions

Fig. D.2. Both tasks are similar in difficulty and complexity. In addition, the two tasks use the same tools (a tape measure and a screwdriver) to check if the installation conforms to the requirements. Also, the role of the instruction is quite similar in both use cases. Additionally, the experiment considers two experimental conditions without guidance and with guidance which can be through text, as presented in Figure 3.1, or with an avatar. All participants will start the test with the guided condition because the tasks requires from a previous instruction to be performed. After this guided task (Condition 1), the subjects will have to perform the same task without any help (Condition 2). Table D.2 presents the tests considered in the experiment: A, B, C, and D. Each participant will perform two tests: A and D or B and C in a randomized order. This design will allow us to analyze the effect of collaboration on learning and retention. Additionally, we will perform a statistical analysis to evaluate the influence of the test order on the results.

The evaluation will address two types of analysis. First, the objective analysis will consider the time spent in each test and the need for assistance in the unguided condition. Besides, we will assess the QoE using subjective questionnaires after each task for evaluating the following factors:

- Overall quality
- Visual quality
- Simulator Sickness

Factor	Question
Involv.	How natural did your interactions with the environment seem?
Involv.	How compelling was your sense of objects moving through space?
Involv.	How much did your experiences in the virtual environment seem consistent with your real world ones?
Sens.Haptic	How well could you move or manipulate objects in the virtual environment?
Adapt.	How quickly did you adjust to the virtual environment experience?
Adapt.	How proficient in moving and interacting with the virtual environment did you feel at the end?
Adapt.	How well could you concentrate on the assigned tasks rather than on the mechanisms used to perform them?
Social Presence	I felt that people were talking to me
Social Presence	I felt that I was listening to the others in the video
Social Presence	I felt I was present with the other people in the video
Social Presence	I felt like the people in the video could see me
Social Presence	I felt I was actually interacting with other people
Visual Quality	Please rate the perceived quality of the instructor's avatar.
Sickness	Did you feel any sickness or discomfort during the experience? Please rate it
Global QoE	How would you rate the quality of the experience globally?
Usefulness	How useful this experience would be for training

Table D.3: Questionnaire used in the experiment.

- Spatial Presence (Involvement, Adaption and Haptic sensation)
- Social Presence

For assessing the QoE in the first four factors we will use a questionnaire that has been validated for the same environments and for others with the same kind of natural interfaces [9]. This questionnaire was selected from a subsampling of Presence Questionnaire of Witmer and Singer validated in [95] for interactive immersive environments. In addition, the social presence factor wasn't present in the previous experiments. Consequently we decided to address the social presence using questions from [143] validated for interactive conferences. Table D.3 shows the complete questionnaire.

D.3 Conclusions and future work

In this subsection we have analysed different methods for enabling collaboration in XR training environments. Also, we presented a methodology for assessing the impact on the learning and QoE of the new feature. In future work, a experiment following the described methodology should be carried out. In addition, it would be necessary to perform a pre-pilot experiment comparing the different avatar representation techniques. Finally, a further analysis should be performed on whether different tasks have different avatar representation requirements.

D.4 Conclusions and future work

In conclusion, our research has successfully validated the efficacy of realistic natural interaction methods grounded in image-based techniques for local interaction for Social XR.

These findings not only contribute to advancing the field of XR technology but also provide valuable insights for developers and designers aiming to create immersive and seamless XR experiences that resonate with users on a profound level. In addition, our research indicates that the techniques we have proposed for Social XR maintain a high-quality experience. This

suggests their suitability for future Social XR systems that require realistic interactions at both the local and social levels. As the demand for immersive Social XR experiences grows, our findings support the development of innovative platforms that can create meaningful connections among users while delivering realistic interactions and user satisfaction. This represents a significant advancement in harnessing the potential of Social XR across various applications.