October 27–31, 2025
Dublin, Ireland

Association for Computing Machinery

Advancing Computing as a Science & Profession

IXR '25

Proceedings of the 3rd International Workshop on

Interactive eXtended Reality

Sponsored by:

ACM SIGMM

General Chairs:

Irene Viola (Centrum Wiskunde & Informatica, Netherlands)
Silvia Rossi (Centrum Wiskunde & Informatica, Netherlands)
Marta Orduna (Nokia, Spain)
Maria Torres Vega (KU Leuven, Belgium)

Co-Located with ACM multimedia

Dublin, Ireland 27-31.10.2025

Association for
Computing Machinery

Advancing Computing as a Science & Profession

The Association for Computing Machinery
1601 Broadway, 10th Floor
New York, NY 10019-7434

ISBN: 979-8-4007-2051-2

Additional copies may be ordered prepaid from:

ACM Order Department
PO Box 30777
New York, NY 10087-0777, USA

Phone: 1-800-342-6626 (USA and Canada)
+1-212-626-0500 (Global)
Fax: +1-212-944-1318
E-mail: acmhelp@acm.org
Hours of Operation: 8:30 am – 4:30 pm ET

Cover photo obtained from bigstockphoto.com

IXR 2025 Chairs' Welcome

It is our great pleasure to welcome you to the third workshop on Interactive eXtended Reality (IXR'25). After the success of the inaugural edition of this workshop in 2022 and its second edition in 2023, our purpose is to make this the premier forum for the presentation of research results and experience reports on the topic of interactive extended reality (XR), including contributions in terms of use cases, applications, novel protocols, and algorithms. Moreover, IXR'25 gives researchers and practitioners a unique opportunity to share their perspectives with others interested in the various aspects of XR and how to make it ready for the future.

The call for papers attracted submissions from Austria, Italy, the Netherlands, Poland, Spain, the United Kingdom, and the United States. The program committee received 7 full technical paper submissions. Each paper was reviewed by our Technical Program Committee, and a total of 6 submissions were accepted.

We encourage attendees to attend the keynote presentation. This valuable and insightful talk can and will guide us to a better understanding of the future:

- *Towards Responsible Data-Driven Interactive and Immersive Media Experience understanding*, Niall Murray (who is currently at the Technical University of the Shannon)

Putting together *IXR'25* was a team effort. We first thank the authors for providing the content of the program. We are grateful to the program committee, who worked very hard in reviewing papers and providing feedback for authors. Finally, we thank the hosting organization MM'25, and our sponsor, ACM SIGMM. We hope that you will find this program interesting and thought-provoking and that the symposium will provide you with a valuable opportunity to share ideas with other researchers and practitioners from institutions around the world.

Irene Viola
IXR'25 General Chair
CWI, The Netherlands

Silvia Rossi
IXR'25 General Chair
CWI, The Netherlands

Marta Orduna
IXR'25 General Chair
Nokia, Spain

Maria Torres Vega
IXR'25 General Chair
KU Leuven, Belgium

Table of Contents

IXR 2025 Workshop Organization

General Chairs: Irene Viola *(CWI, The Netherlands)*
Silvia Rossi (*CWI, The Netherlands)*
Marta Orduna (*Nokia, Spain)*
Maria Torres Vega *(KU Leuven, Belgium)*

Program Committee: Roberto Azevedo (*Disney Research, Switzerland*)
Sara Baldoni (*University of Padua, Italy*)
Carlos Cortés (*Universidad Politécnica de Madrid, Spain*)
Sam Van Damme (*Ghent University, Belgium*)
Anna Ferrarotti (*Università degli Studi Roma TRE, Italy*)
Alan Guedes (*University of Reading*)
Jesús Gutiérrez (*Universidad Politécnica de Madrid, Spain*)
Tim Wauters (*Ghent University, Belgium*)
Xuemei Zhou (*CWI, The Netherlands*)

Sponsor: 

Towards Responsible Data-Driven Interactive and Immersive Media Experience Understanding

Niall Murray
nmurray@research.ait.ie
Technological University of the Shannon: Midlands Midwest
Athlone, Westmeath, Ireland

Abstract

This talk will present an evolution of interactive and immersive media experiences, from the perspective of how the user experience is and has been evaluated. Such media experiences have gained significant action in recent years across a range of application domains, with a particular examples to be discussed within the creative and cultural, education, and health sectors. To understand the pragmatic and hedonic qualities of such technologies, the user Quality of experience (QoE) has become a crucial field of study. QoE aims to target how various types of system, context and user factors influence the quality of user experience. In particular, the talk will outline the use of various physiological, behavioural, interaction and performance based metrics towards a continuous data-driven approach of implicit based user experience understanding. The talk will highlight some of the challenges with existing approaches, and with an eye to the next steps within this research field, it will propose how we should, in a more responsible manner, use data captured from users to understand their experience.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**;

Keywords: Quality of Experience, Interactive Media, Extended Reality

ACM Reference Format:
Niall Murray. 2025. Towards Responsible Data-Driven Interactive and Immersive Media Experience Understanding. In *Proceedings of the 3rd International Workshop on Interactive eXtended Reality (IXR '25), October 27–28, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3746269.3760423

IXR '25, Dublin, Ireland

ACM ISBN 979-8-4007-2051-2/2025/10.
https://doi.org/10.1145/3746269.3760423

1 Biography

Niall Murray (Member, ACM, IEEE) is a senior lecturer in the Department of Computer and Software Engineering in the Faculty of Engineering and Technology in the Technological University of the Shannon: Midlands Midwest (TUS). His teaches on programs related to computer networks, multimedia distribution and computing systems. He co-leads the HCIDMS research centre in TUS and his main research areas are Quality of Experience (QoE) evaluation of Immersive and Interactive multimedia experiences, multisensory multimedia and Human Centric AI.

He is a Research Ireland funded investigator in the Adapt Centre (AI Driven Digital Content Technology) and the VistaMilk Centre (Agri-Food and information communications technology (ICT)). He is a co-applicant on the recently created ARC Hub for ICT, a translational national research centre in Ireland. He is currently part of the Horizon Europe AMPLIFY project and is the coordinator of the Horizon Europe TRANSMIXR project (https://transmixr.eu/).

2 Acknowledgments

The author acknowledges the support of the Research Ireland Adapt Center (13/RC/2106 P2) and the Horizon Europe TRANSMIXR (under the Grant Agreement 101070109) and AMPLIFY (under the Grant Agreement 101177413) projects.

Integration of 3D FLS Displays with 3D Authoring Tools

Nima Yazdani
University of Southern California
Los Angeles, USA
nimayazd@usc.edu

Shahram Ghandeharizadeh
University of Southern California
Los Angeles, USA
shahram@usc.edu

Abstract

Authoring tools typically convert 3D animation data into 2D rasterized formats for display, a process that imposes computational burdens, degrades depth fidelity, and complicates the recovery of original 3D information. This paper presents the first known direct integration of Blender, a widely used 3D animation tool, with Dronevision (DV), a 3D multimedia display composed of synchronized Flying Light Specks (FLSs). In addition to raster-based pipelines, this integration enables real-time, high-fidelity illumination of 3D animations directly in physical space. Our system leverages Blender's piecewise animation model—comprising manually placed keyframes and mathematically defined interpolation functions—to generate spatiotemporal trajectories that are mapped to the motion of autonomous FLS drones. We introduce the concept of a display clock that is maintained by each FLS and synchronized across all FLSs using a broadcast based protocol. Our techniques have a fixed complexity independent of the number of FLSs. They enable a DV to scale to millions and billions of FLSs.

CCS Concepts

• **Human-centered computing** → **Visualization design and evaluation methods**; **User studies**; • **Computing methodologies** → **Tracking**; **Motion path planning**; • **Computer systems organization** → **Embedded and cyber-physical systems**; • **Information systems** → **Multimedia content creation**.

Keywords

Authoring Tools, Flying Light Speck, Dronevision, Blender, Keyframes

ACM Reference Format:
Nima Yazdani and Shahram Ghandeharizadeh. 2025. Integration of 3D FLS Displays with 3D Authoring Tools. In *Proceedings of the 3rd International Workshop on Interactive eXtended Reality (IXR '25), October 27–28, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746269.3760418

1 Introduction

The Flying Light Speck (FLS) project (https://www.flslab.org) is developing a new class of eXtended Reality (XR) systems that integrates virtual artifacts into the physical world [17, 19]. An FLS is a small drone equipped with an onboard processor, wireless networking capabilities, and one or more light sources with adjustable color and brightness [18]. Utilizing decentralized localization algorithms [2–4], a swarm of FLSs can autonomously self-assemble and de-assemble to create dynamic, illuminated objects. This process is analogous to rapid 3D printing. The resulting artifacts are visible to the naked eye and facilitate haptic interaction with bare hands [12, 13, 17]. Figure 1 shows a user interacting with an FLS illuminated Jenga brick and a swarm of FLSs pushing back against the user's finger tip.

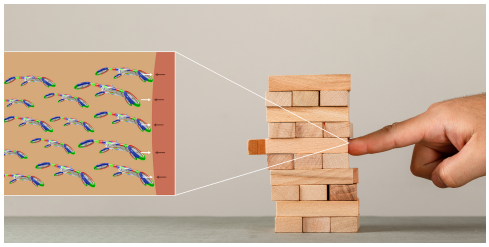


Figure 1: A user pushing an FLS illuminated Jenga brick.

3D authoring tools such as Blender [11] will play a major role in developing the artifacts illuminated by FLSs. Today's tools typically convert native 3D data into 2D rasterized formats for display on standard computer screens, see Figure 2. This rasterization process sacrifices depth quality and is not appropriate for 3D illumination. We address this by presenting an integration of Blender, a 3D authoring tool, directly with Dronevision (DV) [1, 33, 34], a true 3D multimedia display. We envision DV as a 3D display for desktops, enabling next-generation multimedia applications using FLSs [17, 32]. See Figure 3. A key feature of this integration is its ability to provide an animator with real-time visual feedback of their changes using Blender by avoiding the noticeable delay[1] attributed to the rasterization process. The DV will illuminate the latest changes to the 3D content immediately and on-demand.

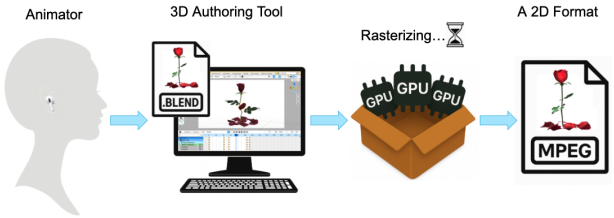


Figure 2: Today's state of the art; 2D rasterization.

[1]Rasterization is computationally expensive. To illustrate, consider rendering of a highly detailed 3D animation, such as a complex particle system simulating falling snow or rain. Each particle's trajectory must be individually rasterized frame-by-frame into a sequence of high-resolution images. This process demands significant computational resources, including intensive CPU and GPU utilization, often leading to extended rendering times that can span from several minutes to hours for short sequences.

IXR '25, Dublin, Ireland

ACM ISBN 979-8-4007-2051-2/2025/10
https://doi.org/10.1145/3746269.3760418

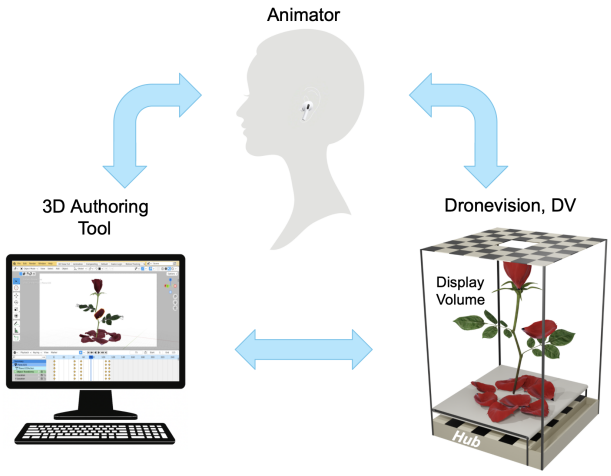


Figure 3: Animator using a DV with a 3D authoring tool.

Rasterization and 3D illumination are complementary techniques for different use cases. While rasterization is well-suited for 2D displays, integrating Blender with Dronevision enables the real-time illumination of 3D shapes and animations. This approach eliminates computational delays and preserves all spatial and depth information, avoiding the data degradation common in 2D screens and rasterized formats. Our prior work [18] introduced a technique for computing FLS flight paths to render a sequence of mesh files illuminated at a prespecified rate, e.g., 24 files per second. This paper builds on that foundation by showing how FLSs implement the concepts of keyframes and interpolation functions directly.

The **contributions** of this paper are as follows:

- An integration of Blender animations with DV, a 3D multimedia display. We envision this integration as an alternative to today's rasterization on 2D screens and file formats. The DV illuminations complement these rasterization. (Section 2.)
- A simple broadcast-based technique customized for DV to maintain a synchronized display clock for each FLS. Its overhead is in the form of memory (a few megabytes) required from each FLS. It has a fixed overhead, enabling the DV to scale to millions and billions[2] of FLSs. (Section 2.3.)
- An implementation using Blender's keyframes and interpolation functions. (Section 3.)
- An evaluation of clock drift and its impact on the FLS illuminations and how our technique addresses it effectively. (Section 4.)

In addition to the above, we present related work in Section 5 and brief conclusions and future research directions in Section 6.

2 An Integrated Approach

This section presents the fundamentals of a 3D authoring tool, a DV, and their integration in turn. We illustrate the integration using an example and describe the resolution of a DV. We state the clock synchronization problem formally. Subsequently, we present a broadcast based solution to the clock synchronization challenge.

[2]Achieving an immersive, room-sized DV may require billions of FLSs, resembling the holodeck from Star Trek [19, 20]. It may consume a substantial amount of energy.

3D Authoring Tools: In modern authoring tools used for animation, the motion of objects is often controlled by curves, commonly known as F-curves, that represent the variation of animated parameters such as position, rotation, and scale over time [16]. Instead of modeling these changes with one continuous, complex function, authoring tools such as Blender use a piecewise approach. This approach builds the overall curve from two components: keyframes and interpolation functions.

Keyframes are specific moments in time where the animator manually sets the desired value of a parameter [29]. They serve as anchor points that define critical states in the animation, such as the start and end of a movement or a sudden change in direction, e.g., abrupt end to a scene. Interpolation functions are the mathematical methods such as Bézier, linear, or constant functions that automatically compute the values between keyframes [10, 24]. These functions ensure smooth transitions by filling in the gaps between the manually defined keyframes.

While both components are integral to creating realistic and controllable animations, they serve distinct purposes and do not substitute for one another. Keyframes provide the precise, user-defined targets that capture the creative intent, while interpolation functions determine how the animation evolves between those targets. Without keyframes, there would be no specific moments to guide the motion; without interpolation functions, the animation would lack the smooth transitions needed to connect these discrete points effectively.

Dronevision, a 3D Display: Figure 3 illustrates the architecture of the Dronevision (DV), a 3D display system [1]. At the base of the DV, the FLSs are housed and prepared for deployment. The DV architecture contains a hub, a small computer. The hub hosts a process named *Orchestrator* responsible for coordinating the swarm [18]. The Orchestrator has its own physical clock that is used as the global system clock. It also serves as the *display* clock of FLSs illuminating an animation. The Orchestrator continuously broadcasts its time to all FLSs; those that are housed and charging on their launch pads and those in the display volume illuminating an animation. Each FLS listens to this broadcast and employs linear regression to correlate the Orchestrator's clock with its own local physical clock, see Figure 5 and Section 2.3. This display clock ensures tight temporal synchronization amongst the FLSs.

The volume of a DV is partitioned into **display cells**. A display cell may be occupied by at most one FLS. It corresponds to the FLS and its downwash [7, 9, 15, 27, 31], a region of instability caused by its flight. Assuming an FLS is a quadrotor, a display cell may be an ellipsoid [7, 9, 27] or a cylinders [15, 31] that results in a larger separation along the height dimension. Each display cell has a unique Length, Height, and Depth (L,H,D) coordinate [18]. We use the L, H, D coordinate system instead of X, Y, Z to identify a cell because there is no consensus on one definition of the Y and Z axes. While the picture industry uses the Z axis as the depth, mathematicians use the Y axis for the depth. It is trivial to map our L, H, D coordinate system to either definition without ambiguity.

The light emitted from an FLS fills an **illumination cell**, see Figure 4. Without loss of generality and to simplify the discussion, we assume a display cell and an illumination cell are cubes. We denote the ratio of an illumination cell to a display cell along a dimension as Q. Assuming the same Q value across L, H, and D,

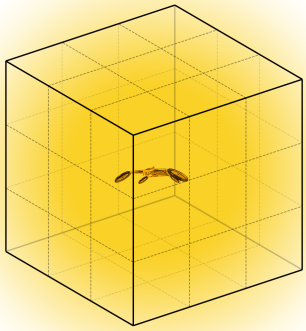


Figure 4: An illumination cell with ratio Q=3 along each dimension, consisting of 27 display cells.

the number of display cells contained in an illumination cell is Q^3; 3^3=27 in Figure 4. A maximum of Q^3 FLSs may occupy different cells of an illumination cell with one FLS illuminating the cell [6, 18]. We assume the remaining Q^3-1 FLSs remains invisible. This assumption is reasonable because these FLSs may render the same light and color as the illuminating FLS with lower intensity depending on the location of their occupied display cell.

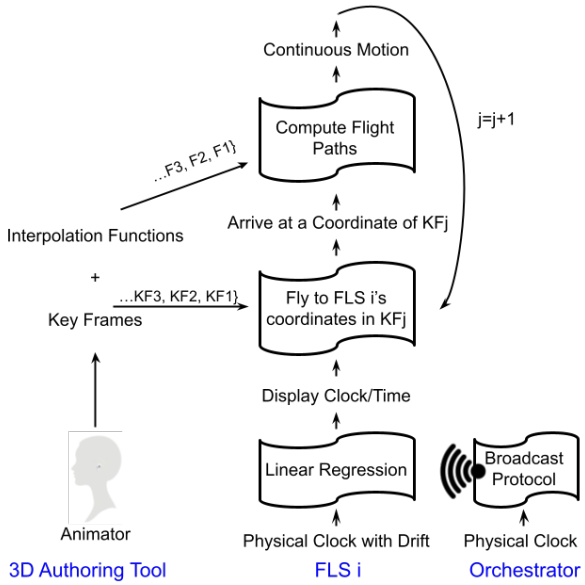


Figure 5: FLS integration with a 3D Authoring Tool.

An Integrated Approach: Figure 5 shows the integrated architecture of a 3D authoring tool, such as Blender, with a DV. The DV may consist of swarms of potentially millions of FLSs to illuminate animations. Figure 5 illustrates the software stack of a single FLS for clarity, though variants of this stack occur concurrently across the different FLSs that constitute a swarm.

The animator authors keyframes (KF1, KF2, KF3, etc.) specifying the exact position each FLS should occupy at particular moments in time. Additionally, the animator define interpolation functions (F1, F2, F3, etc.) that mathematically describe smooth, continuous movements between these discrete keyframes. Each FLS is assigned coordinates corresponding to the start of each keyframe, and the interpolation functions determine the FLS's precise flight path between these positions.

FLSs continuously listen for a broadcast message from the Orchestrator. Each FLS implements a *display* clock using these broadcast messages and its physical clock. Section 2.3 details this display clock. The broadcast protocol is specific to a DV and takes advantage of the known physical characteristics of the DV. Its overhead is fixed in the form of (a) memory in the order of megabytes from each FLS, and (b) periodic messages transmitted by the Orchestrator. It scales to support millions and billions of FLSs.

Once the animator presses the "Play" button, the Orchastrator processes the keyframes and assigns the coordinates of the keyframes to individual FLSs. The Orchastrator broadcasts the interpolation functions to all FLSs and informs each FLS when to start processing the interpolation functions, T_{start}. This is the start time of the animation. The Orchestrator computes it by computing the maximum time T_{last} required for the furthest FLS to arrive its initial coordinate from deployment; T_{start}=T_{last}+1. The Orchestrator may perform this computation and provide T_{start} to the FLSs while they are in the hangar and prior to being dispatched. Each FLS flies to its assigned starting coordinate defined by the first keyframe, KF1. At T_{start} of its display clock, FLSs initiate the execution of the first interpolation function. Each FLS independently executes its stored interpolation functions, smoothly transitioning from one keyframe to the next as specified by the animator. After processing one interpolation function by flying along its specified path, each FLS proceeds directly to the starting coordinate of the next keyframe. This iterative process continues until the entire animation sequence concludes and the FLSs arrive at their assigned coordinate in the last[3] keyframe.

An Example: To illustrate these concepts, consider an animation depicting a falling rose petal (Figure 6). The animation comprises 112 frames, defined by five keyframes occurring at frames 1, 42, 55, 105, and 112. Within each keyframe there is a specified starting coordinate for each FLS. If each coordinate component is represented by a floating-point value of 4 to 8 bytes, then each full 3D coordinate occupies up to 24 bytes. Thus, storing all starting coordinates for the five keyframes requires only about 120 bytes per FLS. The motion between keyframes is governed by nine Bézier interpolation functions per keyframe interval (Figure 8). These Bézier functions, which precisely control the petal's translation, rotation, and scale, are exceptionally compact, requiring just tens of bytes for representation. Together, this compact data representation facilitates efficient storage and rapid computation.

To initiate an animation, swarms of the participating FLSs are dispatched to their assigned initial display cell specified by the first keyframe (KF_j). Several algorithms for such dispatching is presented in [18]. Once the FLSs arrive their initial position in

[3]With Blender, it is possible to not have a final keyframe by specifying a timeline range. This special case is trivial to implement by using extrapolation mode or the coordinate of the FLS in the last keyframe.

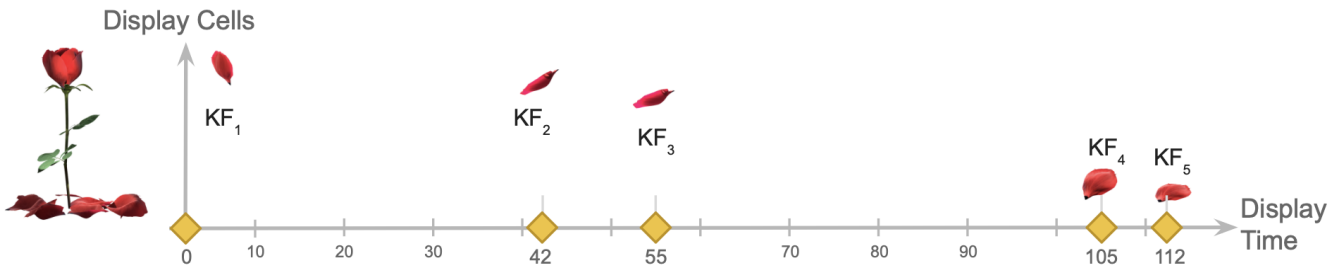


Figure 6: The rose petal animation consists of five keyframes spread across the 112 frames of the full animation.

(KF_1), they proceed to autonomously compute their flight paths by executing the interpolation functions provided by the authoring tool. See the Bézier interpolation functions shown in Figure 8. This novel integration drastically reduces computational overhead, enabling near-instantaneous trajectory calculation. Additionally, the linear regression synchronization technique itself represents another novel advancement in this work, providing high-precision clock synchronization across potentially millions of FLS nodes.

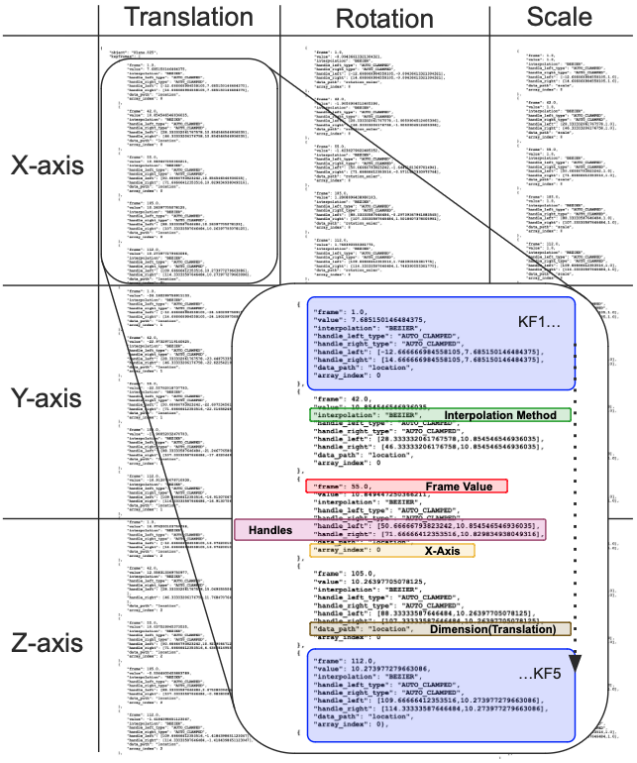


Figure 7: JSON file containing the keyframing and interpolation data.

Figure 6 shows the five keyframes of the rose petal animation relative to the entire rose structure, highlighting the space-time spectrum of the animation. In this figure, the Y-axis represents the 3D space within the DV display, i.e., its display cells. The X-axis represents the display time, which is produced by the display clock shown in Figure 5. The animator specified keyframes are not necessarily evenly spaced along the X or the Y axes.

2.1 Display Resolution

The parameter Q directly determines the resolution of a Dronevision (DV) display by defining the number of display cells contained within each illumination cell. Given a fixed DV display volume and illumination cell size, a larger Q value implies smaller individual display cells, significantly increasing the display's granularity. One practical way to achieve higher Q values is by reducing the physical dimensions of the FLSs and their associated downwash regions.

This relationship is visualized in Figure 9, where different keyframes of the rose petal animation are shown across varying Q values. As Q increases, the density and number of display cells representing each keyframe grow, thereby enhancing the resolution and detail visible within the animation.

Moreover, Q plays a crucial role in maintaining visual detail when viewers zoom into or out of the animation. At a distance (zoomed out), groups of closely situated display cells within a single illumination cell appear indistinguishable to the human eye, allowing many of these cells to remain inactive or uniformly colored without compromising perceived image quality. As viewers zoom closer, previously indistinguishable cells become visually distinct, activating and displaying unique colors to preserve detailed animation features. Thus, higher Q values enable deeper zoom levels while maintaining sharp visual fidelity, revealing finer animation details as they become perceptually discernible.

For accurate alignment, the global coordinate system of Blender animations maps directly onto Dronevision by aligning Blender's origin—the center of its XYZ coordinate frame—with the center of Dronevision's base within the display volume. Portions of the animation extending beyond Dronevision's physical display boundaries are clipped during rendering, further highlighting the importance of Q in managing visual transitions and preserving animation detail near the display edges.

2.2 Problem Statement

Consider a swarm of F FLSs, $\{f_1, f_2, ..., f_F\}$. Each FLS f_i is provided with a sequence of K coordinates $\{k_1, k_2, ..., k_K\}$ corresponding to K key frames. An interpolation function I_j specifies the flight path of FLS f_j for two consecutive coordinates. It may consist of $K-1$ interpolation functions, $I_j = \{I_{j,1}, I_{j,1}, ..., I_{j,K-1}\}$. Function $I_{j,1}$ defines how the FLS f_j travels from the coordinate k_1 to k_2. It is possible for I_j to consist of fewer than $K-1$ functions where a function applies to more than two consecutive coordinates.

A function may be shared by multiple FLSs. At one end of the spectrum, all FLSs may share one function I. At the other end of the spectrum, F interpolation functions may exist with one function for each FLS.

$$
\begin{aligned}
\text{Location:} \quad & B_X(t) = (1-u)^3(7.68515) + 3(1-u)^2u(7.68515) + 3(1-u)u^2(10.85455) + u^3(10.85455) \\
& B_Y(t) = (1-u)^3(-24.1803) + 3(1-u)^2u(-24.1803) + 3(1-u)u^2(-23.44875) + u^3(-22.9733) \\
& B_Z(t) = (1-u)^3(16.9762) + 3(1-u)^2u(16.9762) + 3(1-u)u^2(15.04936) + u^3(12.55831) \\
\text{Rotation Euler:} \quad & R_X(t) = (1-u)^3(-5.40678) + 3(1-u)^2u(-5.40678) + 3(1-u)u^2(-109.15364) + u^3(-109.15364) \\
& R_Y(t) = (1-u)^3(2.19041) + 3(1-u)^2u(2.19041) + 3(1-u)u^2(43.41660) + u^3(55.86539) \\
& R_Z(t) = (1-u)^3(-1.35955) + 3(1-u)^2u(-1.35955) + 3(1-u)u^2(-68.46226) + u^3(-68.46226) \\
\text{Scale:} \quad & S_X(t) = S_Y(t) = S_Z(t) = 1.0 \\
& \text{where } u = \frac{t-1}{41}, \quad t \in [1, 42]
\end{aligned}
$$

Figure 8: Bézier equations defining Location, Rotation, and Scale between Keyframe 1*(frame 1)* and Keyframe 2*(frame 42)*. Each dimension is controlled independently using separate Bézier functions.

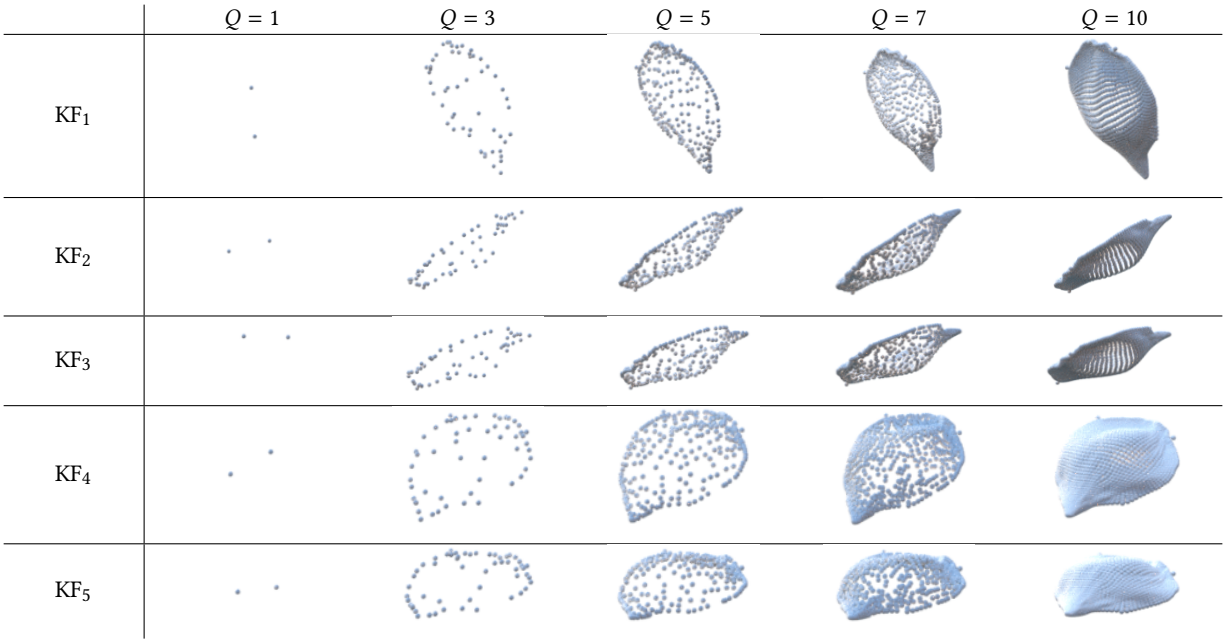


Figure 9: Keyframe snapshots (KF_i) across varying values of Q, which dictates the resolution of a 3D multimedia display (DV).

The problem solved in this paper is how to synchronize F FLSs to start at time T_{start}, execute their function to compute a path, and move along the path such that they are at their assigned coordinate k_i within a margin of error Δ_i time units. This Δ_i may be specified by the requirements of an animation application [25, 26, 28] and dependent on the characteristics of the interpolation function. It may be different for each of the K coordinates. The ideal value of Δ_i is zero. This means all F FLSs are synchronized as a function of time.

2.3 Synchronized Clocks

We solve the problem of Section 2.2 by implementing a synchronized display clock across all FLSs. This clock is synchronized with the Orchestrator's clock. See bottom right hand corner of Figure 5. The maximum difference between the two clocks is either equal to or lower than Δ_i. This ensures an FLS will reach each keyframe precisely enough to render the intended animation without noticeable temporal discrepancy. Our clock synchronization technique is:

(1) **Scalable:** The overhead of the synchronization technique is fixed and independent of the number of FLSs, F. Hence, the technique may support millions and billions of FLSs.
(2) **Robust:** Each FLS computes its display time independent of the other FLSs. This avoids a single point of failure. Even if certain FLSs lose communication or fail [5, 6], others maintain a synchronized display clock.

We assume each FLS i has a physical clock that ticks at a fixed rate. The display clock of FLS i is logical and consists of two parameters: Current time, $T_C(i)$, and the rate $\lambda(i)$ at which it advances for every physical clock tick. The display clock is synchronized when $T_C(i)$ matches the Orchestrator's current time and the combination of the FLS i's physical clock tick and $\lambda(i)$ match the rate at which the Orchestrator's clock ticks. Hence, the two parameters that an FLS i must fine-tune are:

- $T_C(i)$, its current time.
- $\lambda(i)$, the rate at which FLS i advances its current time as a function of its physical clock tick.

We implement this fine-tuning by requiring the Orchestrator to broadcast its local clock time to FLSs periodically while the FLS is in the hangar or in the display volume illuminating an animation. The receiving FLSs set their $T_C(i)$ to the Orchestrator's clock time after adjusting for the message delivery time. They maintain recent samples to build a simple regression model to compute $\lambda(i)$.

2.3.1 Message Delivery Time. FLSs make use of the physical characteristics of a DV to compute the delay incurred in receiving a broadcasted message from the Orchestrator. Prior research [8] has established that the one-way broadcast delay primarily depends on distance. We assume a table of known locations in the DV and their distance from the Orchestrator. It may be constructed by a human when designing the DV. FLSs use a localization technique to identify their location in this table to look up the delay that they should use for the broadcasted messages from the Orchestrator.

For those FLSs in a hangar, the position of their pad and its relative distance from the transmitter are well known. Similarly, we partition the display volume of a DV into fixed quadrants. Once again, the distance between a quadrant and the Orchestrator is well known.

We use techniques similar to [21] to compute the expected signal travel time using the known relative positions. The resulting lookup table is compact. For example, if it consists of one million known locations then it is expected to be in the order of a few megabytes. A hash table facilitate look up of the delay given a location. Each receiving FLS knows its location, looks up the transmission delay in the table, to compute its current time $T_(i)$.

Given the tabletop scale of a DV, the transmission delays are typically sub-microseconds in duration [21]. If the granularity of $T_C(i)$ is milliseconds, then the delays will be insignificant. Hence, in almost all cases, the offset between the FLS display clocks and the Orchestrator time will remain comfortably within the tolerated margin of error Δ_i (Section 2.2). However, if the measured offset between an FLS's local clock and the global clock exceeds Δ_i, an offset correction is applied.

2.3.2 Skew Correction, $\lambda(i)$. Even with precise offset alignment, hardware clocks in FLSs may drift at varying rates relative to the Orchestrator's clock. For instance, an FLS physical clock may increment by 2 counts per Orchestrator tick, whereas another FLS increments by 4 counts for the same Orchestrator tick. To systematically handle these variations, an FLS employs linear regression over multiple pairs of *(physical clock time, Orchestrator clock time)*.

Concretely, each FLS logs broadcast timestamps from the Orchestrator at fixed intervals, pairing them with its own physical clock readings. By fitting a line through these *(physical, Orchestrator)* time pairs, the FLS estimates how its clock rate differs from the Orchestrator's clock rate. If the linear regression reveals that the local clock increments only half as quickly, then a skew correction factor of 2 is applied to the FLS's physical clock increments. Conversely, if the physical clock is found to be running too quickly, a factor less than 1 will be used. Each FLS implements this technique independently to realize a display clock. By synchronizing with the Orchestrator's clock, the display clock of all FLSs is synchronized.

3 An Implementation

We implemented the concepts presented in this paper using Blender [11]. A Python script exports each FLS's keyframe coordinates and interpolation functions from an animated sequence (e.g., a falling rose petal). This data is stored in JSON files and used as direct inputs to a DV simulator, allowing each FLS to interpolate its local trajectory independently using the discussion of Figure 5.

To illustrate, with an animation such as the rose with a falling petal, each segment between consecutive keyframes uses a cubic Bézier function such as the following: $B(t) = (1-t)^3 P_0 + 3(1-t)^2 t\, P_1 + 3(1-t)t^2\, P_2 + t^3 P_3$, where $t \in [0,1]$ is the normalized parameter for that segment, and P_0, P_1, P_2, P_3 are extracted from the Blender-generated JSON. Figure 7 shows the structure of the JSON file that represents the keyframes and their interpolation functions. Once this is instantiated in memory, it will be represented as Figure 8.

The animation timeline is divided into segments defined by keyframes. Each FLS identifies its active segment based on its' display clock, then evaluates the corresponding cubic Bézier curves for position, rotation, and scale. These values form a transformation matrix applied every $\frac{1}{24}$ second, i.e., a Blender frame.

FLS path length and velocity are derived by numerically integrating the position derivative between two consecutive timestamps:

$$\text{frame\_length} = \int_{t_1}^{t_2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\, dt, \qquad (1)$$

and multiplying by the animation frame rate (e.g., 24). This ensures precise velocity tracking for each FLS.

Each FLS executes its path using its synchronized display clock, see Section 2.3. While in theory two timestamp pairs suffice to solve for $T_C(i)$ and $\lambda(i)$, we implemented a queue with an adjustable length to handle real-world noise. This allows for stable clock correction even under measurement uncertainties. In our implementation we used a queue of ten samples.

4 An Evaluation

We used the rose with the falling petal for our evaluation purposes. It consists of 65,321 FLSs. However, only 2008 of these FLSs are of interest because they constitute the falling petal. The five keyframes

(a) Error model 1. **(b) Error model 2.** **(c) With clock synchronization.**

Figure 10: Location of FLSs with two different error models for the physical clock when the first FLS reaches its final destination in KF5, see (a) and (b). (c) shows the results with the same two error models using the clock synchronization technique of Section 2.3. This is the desired location of FLSs as shown by the last column and the last row of Figure 9.

that they constitute are shown in Figure 9 under the column Q=10. The remaining 63,313 FLSs are stationary.

To evaluate the robustness of our clock synchronization technique under challenging conditions, we implemented two heterogeneous physical clock rate models that introduce significant drift. With the first model, FLSs with odd-numbered IDs operate with physical clocks running at two times (2x) the rate of the Orchestrator clock, while FLSs with even-numbered IDs operate at one-half (0.5x) the Orchestrator clock rate. We have observed a similar variation in practice when purchasing hundreds of servers. With the second error model, the speed of the physical clock of each FLS is selected randomly within ±10% error of the Orchestrator clock.

Figure 10 shows the rose and the location of the 2008 FLSs as soon as the first FLS arrives at its final coordinate in KF5. Figure 10a and 10b show the location of FLSs with each error models and no clock synchronization. Figure 10c shows the location of FLSs with clock synchronization. The results is the same with both error models. Hence, we show only one illumination.

The first error model causes the FLSs to be partitioned into two sets, 1004 fast and 1004 slow FLSs. The slow FLSs are flying along a path specified by the interpolation function between KF1 and KF2 when the fast FLSs arrive at their final coordinate, see Figure 10a. This provides the illusion of two petals, one moving very fast while the other moves very slow.

The second error model results in a significant distortion of the petal, see Figure 10b. While 1585 FLSs are flying along a path specified by the interpolation function between KF3 and KF4, 3 are at KF4, 379 FLSs are flying along a path specified by the interpolation function between KF4 and KF5, and 41 fast FLSs are at their final coordinate in KF5. These FLSs are not synchronized because of the random 10% skew in their physical clocks.

With both error models, our technique provides each FLS with a display clock that is synchronized with one clock, namely the Orchestrator's local clock. This provides the correct illumination shown in Figure 10c.

5 Related Work

This work builds upon our prior research [1–6, 17, 18, 33, 34], specifically our technique for computing FLS flight paths for 3D animations presented in [18] and conceptual modeling of 3D animations using FLSs [32]. We describe these in turn. In [18], we present a highly parallel offline technique to compute flight paths given a sequence of mesh files. It requires exporting and processing the entire sequence of mesh files from a 3D authoring tool. For example, creating the "rose petal" animation involved processing 115 individual mesh files, a process that is both time-consuming and resource-intensive. The key difference in the current paper is its efficiency. Instead of using mesh files, our new method only exports the animation's keyframes and interpolation functions from the 3D authoring tool. This significantly reduces the data required; the same "rose petal" animation now only needs 5 keyframes and 9 functions (see Figures 8 and 9). Techniques of [18] are still applicable when working exclusively with mesh files.

This paper extends our previous work in [32], which proposed a conceptual model for intelligent multimedia data through content-based queries and annotations. The prior work focused on an interactive animation studio that allowed animators to manipulate 3D illuminations directly, effectively using the FLS display as both a viewing and an authoring tool. Here, we focus on viewing and present the framework of Figure 5 and a novel technique to synchronize the FLS clocks, which are key additions to our system.

Clock drift is a well studied problem. There are numerous causes for this drift ranging from imperfect quartz crystal oscillators sensitive to temperature and power conditions to power supply variations and system load. The requirement for periodic resynchronization [22] is well known. The Network Time Protocol (NTP) is widely adopted in traditional computing systems [23]. NTP employs a hierarchical synchronization structure, which faces scalability challenges in extremely large swarms consisting of tens of thousands of FLSs. As swarm size increases, the bandwidth and computational overhead required for frequent synchronization escalate, introducing significant latency and potentially reducing overall synchronization accuracy.

The clock synchronization technique of Section 2.3 is specific to a DV. This simple broadcast technique uses the physical characteristics of a DV to provide the FLSs with a display clock synchronized with the Orchestrator. It is similar to other broadcast-based methods, such as Reference Broadcast Synchronization (RBS) [14], designed for sensor networks and wireless ad hoc networks. RBS is similar to our technique in that it offers a high relative synchronization precision by eliminating uncertainty associated with sender-side delays. Our technique is different from RBS in that it does not require the FLSs to compare their clocks to each other based on when they receive a reference broadcast message. This works for our setting because the DV characteristics are known in advance.

There are drone specific algorithms to address clock skew. For example, adaptive clock skew compensation allows an ocean bottom drone to predict and correct its clock offset relative to its neighbors [30]. These techniques are specific to an application use case in a specific environment. They inspired the development of our clock synchronization technique that is specific to FLSs in a DV.

6 Conclusion and Future Research

This paper presents an integration of a 3D authoring tool with a 3D display using FLSs, i.e., a DV. The system enables the export of keyframes and interpolation functions from the authoring tool, which are then used by the FLSs to render the animation within the DV. A key requirement is for each FLS to maintain a display clock that is synchronized with the DV's Orchestrator clock. To address this, we presented a simple broadcast-based synchronization technique using linear regression.

We are implementing a prototype of the integrated approach using Blender, an open source 3D modeling tool [11]. We have an implementation of the Blender interfaces as reported in Section 3. We also have FLS prototypes with processing and networking capabilities to implement localization techniques such as Swarical [3] and the software stack of Figure 5 to transform interpolation functions into flight paths. However, the prototype faces a variety of challenges ranging from the reliability [6] of the FLSs to the limited flight time of an FLS on a fully charged battery [18], approximately 7 minutes. We are addressing these challenges in the near future.

The proposed integration will empower animators to refine animations interactively within the DV platform, with changes seamlessly propagating back into Blender. This bidirectional workflow supports efficient version control and accelerates iterative design. Building system support for this kind of interactive workflow represents an exciting and ambitious long-term goal.

Acknowledgments

We are grateful to the anonymous reviewers of the IXR workshop for their insightful and constructive feedback, which has enhanced the quality of this paper. This research was supported in part by the NSF grants IIS-2232382 and CMMI-2425754.

References

[1] Hamed Alimohammadzadeh, Rohit Bernard, Yang Chen, Trung Phan, Prashant Singh, Shuqin Zhu, Heather Culbertson, and Shahram Ghandeharizadeh. 2023. Dronevision: An Experimental 3D Testbed for Flying Light Specks. In *The First International Conference on Holodecks* (Los Angeles, California) *(Holodecks '23)*. Mitra LLC, Los Angeles, CA, USA, 1–9. doi:10.61981/ZFSH2301

[2] Hamed Alimohammadzadeh and Shahram Ghandeharizadeh. 2023. SwarMer: A Decentralized Localization Framework for Flying Light Specks. In *The First International Conference on Holodecks* (Los Angeles, California) *(Holodecks '23)*. Mitra LLC, Los Angeles, CA, USA, 10–22. doi:10.61981/ZFSH2302

[3] Hamed Alimohammadzadeh and Shahram Ghandeharizadeh. 2024. Swarical: An Integrated Hierarchical Approach to Localizing Flying Light Specks. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 6153–6161. doi:10.1145/3664647.3681080

[4] Hamed Alimohammadzadeh and Shahram Ghandeharizadeh. 2024. Swazure: Swarm Measurement of Pose for Flying Light Specks. In *The Second International Conference on Holodecks* (Los Angeles, California) *(Holodecks '24)*. Mitra LLC, Los Angeles, CA, USA, 17–25. doi:10.61981/ZFSH2403

[5] Hamed Alimohammadzadeh, Daryon Mehraban, and Shahram Ghandeharizadeh. 2023. Modeling Illumination Data with Flying Light Specks. In *ACM Multimedia Systems* (Vancouver, Canada) *(MMSys '23)*. Association for Computing Machinery, New York, NY, USA, 363–368. doi:10.1145/3587819.3592544

[6] Hamed Alimohammadzadeh, Shuqin Zhu, and Shahram Ghandeharizadeh. 2025. Techniques to Conceal Dark Standby Flying Light Specks. *ACM Trans. Multimedia Comput. Commun. Appl.* (April 2025). doi:10.1145/3724399 Just Accepted.

[7] Senthil Hariharan Arul and D. Manocha. 2020. DCAD: Decentralized Collision Avoidance With Dynamics Constraints for Agile Quadrotor Swarms. *IEEE Robotics and Automation Letters* 5 (2020), 1191–1198. doi:10.1109/LRA.2020.2967281

[8] Francois Baccelli, Bartlomiej Blaszczyszyn, and Paul Muhlethaler. 2009. Bounds on Information Propagation Delay in Interference-Limited Wireless Networks. In *IEEE International Workshop on Spatial Stochastic Models for Wireless Networks (SpaSWiN)*. IEEE, 86–90. doi:10.1109/SPAWC.2009.5161805

[9] Daman Bareiss and Joran van den Berg. 2013. Reciprocal Collision Avoidance for Robots with Linear Dynamics using LQR-Obstacles. In *Proceedings - IEEE International Conference on Robotics and Automation*. 3847–3853. doi:10.1109/ICRA.2013.6631118

[10] Senay Baydas and Bulent Karakas. 2019. Defining a Curve as a Bezier Curve. *Journal of Taibah University for Science* 13, 1 (2019), 522–528.

[11] Blender Online Community. 2024. Blender - A 3D Modelling and Rendering Package. https://www.blender.org. Version 4.0.

[12] Yang Chen, Hamed Alimohammadzadeh, Shahram Ghandeharizadeh, and Heather Culbertson. 2023. Towards Enabling Complex Touch-based Human-Drone Interaction.. In *IROS Workshop on Human Multi-Robot Interaction* (Detroit, USA).

[13] Yang Chen, Hamed Alimohammadzadeh, Shahram Ghandeharizadeh, and Heather Culbertson. 2024. Force-Feedback Through Touch-based Interactions With A Nanocopter. In *2024 IEEE Haptics Symposium (HAPTICS)*. 271–277. doi:10.1109/HAPTICS59260.2024.10520851

[14] Jeremy Elson, Lewis Girod, and Deborah Estrin. 2002. Fine-Grained Network Time Synchronization using Reference Broadcasts. In *Proceedings of the 5th symposium on Operating systems design and implementation (OSDI)*. USENIX Association, 147–163. doi:10.1145/844128.844143

[15] Eduardo Ferrera, Alfonso Alcántara, J. Capitán, Á. R. Castaño, P. Marrón, and A. Ollero. 2018. Decentralized 3D Collision Avoidance for Multiple UAVs in Outdoor Environments. *Sensors (Basel, Switzerland)* 18 (2018).

[16] Gordon Fisher. 2014. *Blender 3D Basics Beginner's Guide: A Quick and Easy-to-Use Guide to Create 3D Modeling and Animation using Blender 2.7*. Packt Publishing Ltd.

[17] Shahram Ghandeharizadeh. 2021. Holodeck: Immersive 3D Displays Using Swarms of Flying Light Specks. In *ACM Multimedia Asia* (Gold Coast, Australia). doi:10.1145/3469877.3493698

[18] Shahram Ghandeharizadeh. 2022. Display of 3D Illuminations using Flying Light Specks. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 2996–3005. doi:10.1145/3503161.3548250

[19] Shahram Ghandeharizadeh. 2025. Flying Light Specks: Dronevision, Holodecks and Spatial Computing. In *Proceedings of the 3rd International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice (McGE '25)*

(Dublin, Ireland) *(McGE '25)*. Association for Computing Machinery, New York, NY, USA, 2 pages. doi:10.1145/3746278.3759395

[20] Shahram Ghandeharizadeh and Vincent Oria. 2023. Virtual Reality, Augmented Reality, Mixed Reality, Holograms and Holodecks. In *The First International Conference on Holodecks* (Los Angeles, California) *(Holodecks '23)*. Mitra LLC, Los Angeles, CA, USA, 38–40. doi:10.61981/ZFSH2304

[21] André Günther and Christian Hoene. 2005. Measuring Round Trip Times to Determine the Distance Between WLAN Nodes. In *International conference on research in networking*. Springer, 768–779.

[22] Hedi Marouani and Michel R Dagenais. 2008. Internal Clock Drift Estimation in Computer Clusters. *Journal of Computer Systems, Networks, and Communications* 2008 (2008), 1–9.

[23] David L Mills. 1991. Internet Time Synchronization: The Network Time Protocol. *IEEE Transactions on Communications* 39, 10 (1991), 1482–1493.

[24] Donald E Myers. 1994. Spatial Interpolation: An Overview. *Geoderma* 62, 1-3 (1994), 17–28.

[25] Jakob Nielsen and Rolf Molich. 2005. A Systematic Comparison of Three Remote Usability Testing Methods: Concurrent Think-aloud, Concurrent Retrospective, and Delayed Retrospective. *International Journal of Human-Computer Interaction* 18, 1 (2005), 39–62.

[26] Filipe Pacheco, Fernando Rebelo, and Paulo Noriega. 2018. Towards Assessing Cognitive Load using Physiological and Interaction Measures in a Mobile task. In *Proceedings of the 9th Augmented Human International Conference (AH'18)*. ACM, Kaiserslautern, Germany, 1–8. doi:10.1145/3174910.3174919

[27] James Preiss, Wolfgang Honig, Gaurav Sukhatme, and Nora Ayanian. 2017. Crazyswarm: A Large Nano-Quadcopter Swarm. In *IEEE International Conference on Robotics and Automation (ICRA)*. 3299–3304. doi:10.1109/ICRA.2017.7989376

[28] Julie K Prigge and Amy K Webb. 2018. Animated Narratives in e-Learning: Effects on Recall and Engagement. *Journal of Educational Technology Systems* 46, 2 (2018), 273–296.

[29] Oliver Villar. 2021. *Learning Blender*. Addison-Wesley Professional.

[30] Bo Wang, John A Collins, and Anne Q Van Ufford. 2014. Clock Drift in Ocean Bottom Seismometers. *Geophysical Journal International* 196, 2 (2014), 1034–1048.

[31] Yang Xu, Shupeng Lai, Jiaxin Li, Delin Luo, and Yancheng You. 2019. Concurrent Optimal Trajectory Planning for Indoor Quadrotor Formation Switching. *Journal of Intelligent & Robotic Systems* 94 (05 2019). doi:10.1007/s10846-018-0813-9

[32] Nima Yazdani, Hamed Alimohammadzadeh, and Shahram Ghandeharizadeh. 2023. A Conceptual Model of Intelligent Multimedia Data Rendered using Flying Light Specks. In *The First International Conference on Holodecks* (Los Angeles, California) *(Holodecks '23)*. Mitra LLC, Los Angeles, CA, USA, 38–44. doi:10.61981/ZFSH2309

[33] Shuqin Zhu and Shahram Ghandeharizadeh. 2023. Flight Patterns for Swarms of Drones. In *The First International Conference on Holodecks* (Los Angeles, California) *(Holodecks '23)*. Mitra LLC, Los Angeles, CA, USA, 29–33. doi:10.61981/ZFSH2303

[34] Shuqin Zhu and Shahram Ghandeharizadeh. 2024. Circular Flight Patterns for Dronevision. In *The Second International Conference on Holodecks* (Los Angeles, California) *(Holodecks '24)*. Mitra LLC, Los Angeles, CA, USA, 1–11. doi:10.61981/ZFSH2404

Beyond Bounding Boxes: 2D Semantic Segmentation for Live Volumetric Video Streaming

Tamás Bukits
tbukits@vicomtech.org
Fundación Vicomtech, Basque Research and Technology Alliance (BRTA)
Donostia-San Sebastián, Spain

Ander Elosegi
aelosegi@vicomtech.org
Fundación Vicomtech, Basque Research and Technology Alliance (BRTA)
Donostia-San Sebastián, Spain

Ana Domínguez
adominguez@vicomtech.org
Fundación Vicomtech, Basque Research and Technology Alliance (BRTA)
Donostia-San Sebastián, Spain

Sergio Cabrero Barros
scabrero@vicomtech.org
Fundación Vicomtech, Basque Research and Technology Alliance (BRTA)
Donostia-San Sebastián, Spain

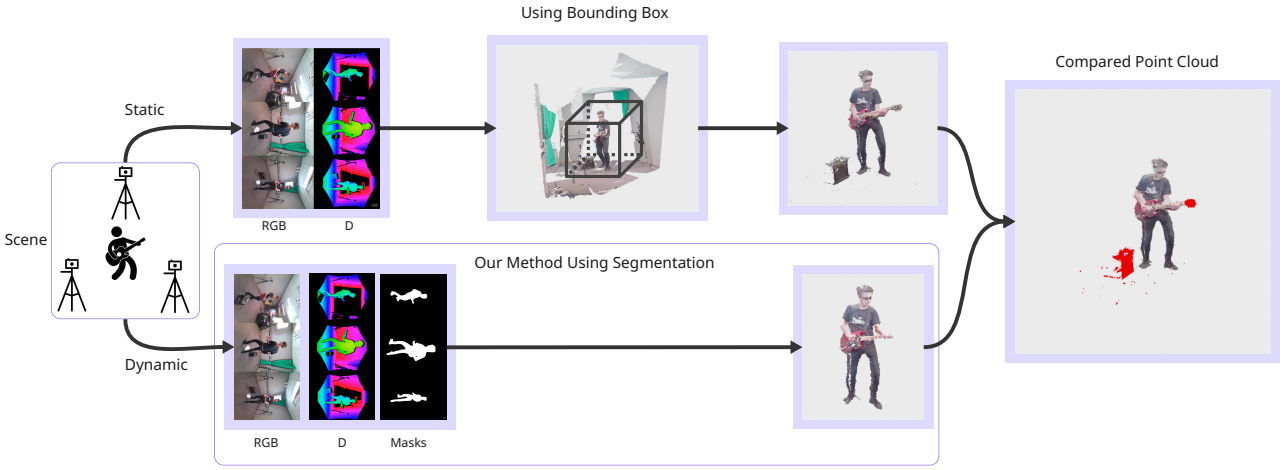


Figure 1: Our pipeline captures RGB-D video, applies 2D object segmentation, and projects masks onto depth images to generate 3D point clouds. Red points highlight differences between segmentation-based and bounding-box-based results.

Abstract

Live volumetric video streaming is an interesting technology for interactive Extended reality (XR) applications, mainly to represent users in telepresence applications as holograms, but also to digitize objects and spaces. Most capture systems rely on static methods to define what should be captured and what not, such as bounding boxes or chroma key segmentation, which is a limitation in many scenarios. To address it, we propose replacing static techniques with 2D semantic segmentation, allowing the system to capture only the important parts of the scene, but limiting its real-time performance as a consequence. In this paper, we describe the design and implementation of a system able to cope with the demands of multi-camera live volumetric video streaming. We evaluate our proposal in a reproducible testbed, using the CWIPC-SXR point cloud video dataset, to understand if this approach is feasible, scalable and accurate, and to find the limitations of its performance comparing it with using a static bounding box.

CCS Concepts

• **Computing methodologies** → **Image segmentation**; **Computer vision**; • **Information systems** → **Multimedia systems**; **Multimedia streaming**.

Keywords

Real-time volumetric video streaming, Multimedia telepresence, 2D semantic segmentation, Extended Reality (XR), Computer vision

IXR '25, Dublin, Ireland

ACM ISBN 979-8-4007-2051-2/2025/10
https://doi.org/10.1145/3746269.3760417

ACM Reference Format:
Tamás Bukits, Ander Elosegi, Ana Domínguez, and Sergio Cabrero Barros. 2025. Beyond Bounding Boxes: 2D Semantic Segmentation for Live Volumetric Video Streaming. In *Proceedings of the 3rd International Workshop on Interactive eXtended Reality (IXR '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746269.3760417

1 Introduction

In multimedia, volumetric video capture has emerged as a powerful medium for immersive applications in entertainment, telepresence, education, and more. While traditional systems often rely on static, controlled environments with predefined constraints—such as chroma key backgrounds or fixed bounding boxes—these approaches do not scale well to real-world, dynamic settings. In particular, real-time scene capture remains a challenge, as it demands systems that can both process and transmit data live, without the benefit of post-processing steps commonly used for offline recording. For example, it is not possible for current systems to include or exclude specific objects from a scene.

Deployments outside laboratory environments or dedicated capture studios face several limitations: either the capture misses relevant elements of the scene, or it includes unwanted background and noise. This is largely due to the static nature of current solutions, which lack the flexibility to adapt to changes in the scene. At the same time, recent advances in artificial intelligence—particularly in segmentation—have enabled accurate, real-time understanding of complex visual content, offering an opportunity to rethink volumetric capture systems as object-based multimedia systems.

In this work, we propose a novel system for live volumetric video streaming that leverages dynamic 2D semantic segmentation to capture and filter the most relevant parts of a scene. By replacing traditional static constraints with object-aware segmentation, our approach allows for more flexible, efficient, and accurate streaming of volumetric content, but it adds computation that can compromise real-time performance. To demonstrate the feasibility and performance of our method, we present an experimental setup in which pre-recorded scenes from the public CWIPC-SXR dataset [17] are streamed with H.264 codec and processed using a segmentation pipeline deployed on a virtual machine. The process is shown in Figure 1. The contributions of this paper are:

- **C1.** A live volumetric video streaming system design and implementation that replaces static constraints and bounding box approaches with dynamic semantic segmentation of the scene.
- **C2.** An empirical evaluation of this design over a reproducible testbed trying to answer the following questions around **feasibility**, **performance** and **accuracy**:
  - **RQ1.** Is the proposed system feasible?
  - **RQ2.** What is its performance and scalability? What are its limitations and where are the bottlenecks?
  - **RQ3.** Are resultant volumetric video reconstructions accurate?

In the remainder of the paper, after discussing Related Work next, we describe the general design of our system, along with its implementation details, in Section 3. We then present a testbed to evaluate our proposal in Section 4. We discuss performance results and present visual outcomes in Section 5, leading to the discussion in Section 6 and conclusion in Section 7.

2 Related Work

2.1 Volumetric Video

Volumetric video is increasingly recognized as a transformative medium across various domains, including entertainment, telepresence, and education [11]. This technology captures scenes by recording RGB-D information to generate a 3D representation and enables users to interact with them in immersive environments. However, processing such scenes remains a significant challenge due to performance limitations.

In our previous work, Cabrero et al. [3], we introduced a web-based platform prototype for transmitting and rendering volumetric videos, utilizing web technologies like WebSockets to ensure compatibility across a wide range of devices. The system's performance was evaluated by measuring the number of transmitted points while rendering recordings on eight different devices, including Oculus Quest and Pico headsets. A key issue identified was that encoding a large number of points using Google's Draco[1] compression negatively impacted performance—scenes containing 300,000 points could only be transmitted at 15 fps due to the computational overhead. This highlights an important aspect: not all points in a scene need to be processed—only those that are relevant.

Early research in this area primarily focused on static environments using controlled setups, often employing techniques such as chroma keying or fixed bounding boxes to isolate objects of interest. For instance, some professional capture studios define the recording area strictly to ensure optimal conditions. In [8], the authors built an $18' \times 18' \times 8'$ volumetric capture volume—a free-standing, fully enclosed green chroma-key structure that provided an uniform environment for background subtraction under controlled lighting. Other studies have explored virtual social environments by placing users within the same virtual experience. Gunkel et al. [7] developed a video conferencing system in which users' backgrounds were replaced with a fixed chroma colour before transmission. On the receiving end, alpha blending was applied to remove the background, producing a transparent image that showed only the user. Similarly, Fernández et al. [13] used this approach to integrate users into the same TV broadcast. Background subtraction was proposed as a solution for separating entities from the background in [6]; however, that work used stereo cameras instead of RGB-D cameras to reconstruct depth maps.

Alternatively, some researchers define static constraints, such as bounding areas, to focus on important parts of the scene in immersive telepresence. Reimat et al. [17] published the CWIPC-SXR dataset, which includes real-life scenarios such as watching TV, playing sports, and giving presentations. This dataset employs a cylindrical bounding area to separate objects from the background. In this setup, the cylinder's height ranges from 0.02 m to 2.38 m, with a radius of 1.2 m around the subject, thereby defining the processing region. Building on this approach to separate users from the background, Viola et al. [20] and Jansen et al. [9] works present end-to-end systems for transmitting humans in volumetric content in

[1] Draco: https://google.github.io/draco/

multiparty, real-time communication. Moreover, Alexiadis et al. [1] implemented live human reconstruction by placing humans within predefined bounding boxes to visualise and capture their motion. While such systems perform well in controlled environments, they are less adaptable to dynamic, real-world settings where objects move unpredictably.

To overcome the limitations of static constraints, the work of [15] proposed a foreground/background segmentation approach based on a Conditional Random Field (CRF) that fuses RGB-D data with pre-captured background models. While this enables binary segmentation of foreground objects in controlled environments, it lacks the semantic understanding of the scene.

2.2 Semantic Segmentation in 2D and 3D

With the evolution of Artificial Intelligence, semantic segmentation has become crucial for identifying and separating objects in a scene based on their categories. Current approaches can be broadly categorized into 3D point/voxel-based segmentation and 2D image-based segmentation.

In 3D semantic segmentation, existing methods primarily follow two paradigms. Point-based networks directly process raw point clouds, with PointNet [4] serving as a foundational architecture that operates directly on unordered point sets while lacking local feature aggregation. This limitation was addressed by PointNet++ [16] through hierarchical learning of local structures, while subsequent innovations like AdaptConv [22] further enhanced feature learning through adaptive graph convolutions. Alternatively, volumetric approaches operate on 3D grid representations, among which 3D U-Net [5] extends the U-Net architecture to volumetric data using encoder-decoder paths with skip connections to capture multi-scale context. Though effective for dense voxel grids in domains like medical imaging, 3D U-Net faces memory and computational challenges when scaling to large outdoor scenes due to its cubic complexity. Although 3D segmentation methods typically offer high semantic accuracy, their complex architectures make them unsuitable for real-time applications.

In contrast, 2D semantic segmentation methods are significantly faster and more efficient for real-time tasks. However, this speed comes at a cost—results from 2D views must be matched across multiple perspectives, which can introduce alignment issues and reduce overall accuracy. In the context of object detection and segmentation, Ultralytics[2] YOLO models have gained widespread popularity in recent years due to their high speed and efficiency in real-time applications [10]. Recent research has introduced the Segment Anything Model (SAM) [12], which performs semantic segmentation of arbitrary objects based on textual prompts. To improve processing speed, Fast Segment Anything (FastSAM) [21] divides the task into two stages: an all-instance segmentation stage using YOLOv8-seg, followed by a prompt-guided selection stage to isolate the target region.

In our work, we directly employ the eighth version of YOLO with a segmentation head, YOLOv8-seg, trained on the COCO dataset [14], to perform segmentation. Unlike FastSAM, which introduces additional processing stages, YOLOv8-seg enables us to achieve real-time segmentation with minimal latency while maintaining a strong balance between speed and accuracy.

3 2D Segmentation-based Live Volumetric Video Streaming

The system design is divided in three main components: Capture, 2D Segmentation and Rendering.

The **Capture** component uses colour and depth cameras—in our case Microsoft Azure Kinect cameras[3] connected via USB 3.0 to a processing machine. This setup provides synchronized RGB and depth (D) images that are encoded in two H.264 streams for each camera, one for colour and one for depth. The method to encode depth is similar to the one used by [19]. To simplify the system in the current version and avoid issues with synchronization, all cameras are merged into the same frame, as can be seen in the images shown in Figures 1 or 2.

The **Rendering** component on the other end takes the merged H.264 stream, decodes it and reconstructs the point cloud using the appropriate calibration parameters from the cameras. The resulting 3D output can be rendered and displayed in various end-user applications like web applications over head-mounted displays (HMDs), or desktop applications using OpenGL. During rendering or capture, the system allows the application of static bounding boxes. Capture and Rendering work together with or without the 2D Segmentation block.

The **2D Segmentation** component sits in the middle between Capture and Rendering. It implements a pipeline that takes the H.264 stream produced by Capture and a set of input labels describing the objects to segment. The output is a segmented version of the input H.264 stream that can be forwarded to the Rendering component. Inside the 2D Segmentation component, RGB images are processed using a segmentation model, which produces binary masks indicating the presence of objects of interest. These masks are then projected onto the corresponding colour and depth images, effectively isolating only the relevant colour and depth data.

Figure 2 presents a detailed diagram of the 2D segmentation pipeline, with each step described below. We use the YOLOv8-seg nano model, chosen for its high inference speed among the available model variants.

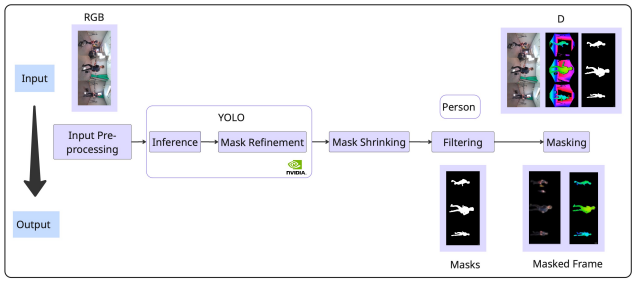


Figure 2: Pipeline for the 2D Segmentation Component.

[2]Ultralytics: https://docs.ultralytics.com

[3]Microsoft Azure Kinect SDK: https://azure.microsoft.com/es-es/products/kinect-dk

3.1 Input Pre-processing

The RGB images arriving from the Capture component are used in the inferencing step of the model. In order to have the correct format for what YOLO model expects of the input RGB images preprocessing needs to be done. The YOLO model accepts images with the size of 640×640 so the image is first resized. After this operation, colour conversion and normalisation are done in order to have the format for inferencing using YOLO. These operations are being run on the CPU using OpenCV[4] computer vision library.

3.2 Inference

Inference is used to predict segmentation masks from input images using a pre-trained model. While YOLO models are originally designed for object detection, newer versions have been extended with a segmentation head to support instance segmentation. To achieve optimal performance, we exported the model to a TorchScript file, enabling GPU inference using the Torch library[5] in C++. The pre-processed images are used as input to the exported model. These images are batched together, as batch processing is more efficient on the GPU due to its ability to perform parallel computations if there are more than one cameras in the capturing. The model outputs two sets of results: one for generating bounding boxes and another for producing segmentation masks.

3.3 Mask Refinement

This YOLO segmentation model is a derivative of the YOLACT architecture [2], which means that the result of the YOLO segmentation method returns with two tensors of output: Prediction Head Block with dimensions $(b, 116, 8400)$, and Protonet Block Output with dimensions $(b, 32, 160, 160)$, where b is the number of images (batch size). These outputs require further processing steps to obtain the desired masks as the model is missing to filter out irrelevant detections and keep only one detection for each object.

3.3.1 Prediction Head Block: Post-processing predictions. With a shape of $(b, 116, 8400)$, the first value represents the batch, followed by 8400 different predictions, each containing 116 values. The initial four values denote the x and y centre of the detection box, width, and height in pixels (ranging between 0 and 640 in our example). Subsequently, there are x-values based on the number of classes in the model (in this case, $x = 80$), representing the class-wise confidence scores for each prediction (ranging between 0 and 1). These scores are used during the Non-Maximum Suppression (NMS) step to filter out redundant or overlapping bounding boxes by retaining only the most confident detections per class. Finally, there are y-values representing coefficients to blend the prototypes to obtain object masks. With 32 prototypes, there are 32 corresponding values. Altogether, these account for the total of 116 values.

3.3.2 Protonet Block Output: Post-processing prototypes. With a shape of $(b, 32, 160, 160)$, comprises 32 feature maps of the image with mask size of 160×160. These maps should be blended together according to the coefficients from Prediction Head Block to generate masks for different objects. Now, the remaining task is to blend the prototypes together based on the detection coefficients. Following this, we will crop around the objects and upscale the masks to the original image size and binarize the masks to obtain the final masks. These operation also run on the GPU on tensors to have the fastest performance.

3.4 Mask Shrinking

After obtaining the segmentation masks we transfer these images to the CPU in order to process them in the following steps. These predicted masks are not always highly accurate, which can lead to outliers in the 3D scene. To address this issue, we developed a shrinking mechanism that refines the masks by applying the OpenCV *erode* function with a kernel size of 3×3. This operation reduces the size of the mask, helping to eliminate noisy or inaccurate edge pixels. As a result, fewer erroneous pixels are used during the masking process, improving the accuracy of the 3D reconstruction.

3.5 Filtering

After the masks are obtained for all detected objects in the image, object filtering is applied to specify which objects should be extracted in the subsequent processing steps. The remainder objects are ignored. In our setup, the model is capable of distinguishing between 80 different object categories, as it was trained on the COCO dataset [14].

3.6 Masking

With the RGB, depth (D), and mask information available, masking is performed by projecting the pixels with a value of 1 in the binary mask onto the corresponding depth image. This allows us to retain only the relevant pixels from both the RGB and depth images—those that should appear in the final scene—while removing unnecessary points from the reconstruction. This process is applied individually for each camera.

4 Empirical Study

4.1 Experiment Design

This empirical study aims at analysing the performance of the proposed 2D segmentation pipeline with an off-the-shelf YOLO segmentation model, and compare it with using bounding boxes. For the sake of targeting the performance evaluation of the 2D Segmentation component, we constructed the testbed in Figure 3. Each step of the experiments is explained in Section 4.2 to provide a detailed overview of our setup. For reproducibility, we used a publicly available dataset CWIPC-SXR [17], as their hardware and camera setup closely resembles ours. Our experiment was designed to evaluate the performance, modularity, and robustness of our distributed pipeline for 3D volumetric reconstruction. To show the effectiveness of the segmentation we selected scenes from the dataset where there were extra objects in the scene, not just the main object that we wanted to detect. For this reason, during our evaluation we focused on the following 10 sequences listed below:

(1) `cwi_electric_guitar`: segmenting the person
(2) `cwi_basketball`: segmenting the person
(3) `cwi_watching_football_r1_t1`: segmenting the person
(4) `cwi_watching_football_r2_t1`: segmenting the person

[4]OpenCV: https://opencv.org
[5]Torch: https://pytorch.org

(5) `cwi_acoustic_guitar`: segmenting the person
(6) `cwi_padel_training_r1_t1_person_racket`: segmenting the person and the paddle-racket
(7) `cwi_padel_training_r1_t1`: segmenting the person
(8) `cwi_padel_training_r2_t1_person_racket`: segmenting the person and the paddle-racket
(9) `cwi_padel_training_r2_t1_`: segmenting the person
(10) `cwi_book_presentation`: segmenting the book as remote (the book was detected as a remote by the model)

The goal of our experiment is to understand the feasibility, performance, scalability, accuracy and limitations of our system, answering the research questions listed in the introduction.

4.2 Testbed and Method

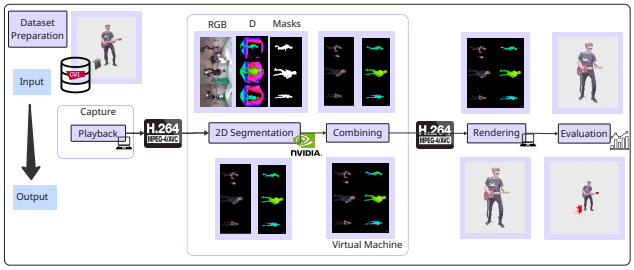


Figure 3: Testbed pipeline for evaluating our 2D segmentation method on the public CWIPC-SXR dataset for volumetric capture.

In the testbed there are 3 machines, one for each component of our system: Capture, 2D Segmentation and Rendering showing the steps on the Figure 3. The flow of the stream in the testbed is similar to that one described in Section 3.

To simulate the **Capture** component in a live system, we implemented an equivalent version using recorded data as input. We preprocessed the CWIPC-SXR dataset and used FFmpeg[6] to simulate streaming live from the real system. The dataset originally was recorded using Microsoft Azure Kinect cameras, providing synchronized RGB-D video in Matroska format (mkv). We kept the original resolution of the frames which is 2048×1536. Additionally, it includes comprehensive calibration files and bounding cylinder annotations, enabling both the reproduction of original results and accurate alignment of point clouds across multiple camera views. The Dataset Preparation step consisted on collecting these data and adapting to our architecture. From the collected data, we created a combined stream that contains the RGB and depth recordings from the Matroska file as many as many cameras are used during the experiment. Additionally, a session file is created in a JSON file where the numerical information is collected like image resolution, number of cameras, used bounding cylinder parameters (min height, max height, filter radius) and the calibration matrices which consists of the intrinsic and extrinsic parameters in order to align the multi view cameras. The RGB and depth images from each camera are stacked to form a single combined frame that includes all camera views. Each camera contributes an image of size 1536×2048 for both RGB and depth, resulting in a width of 2×2048. When combining n cameras vertically, the total frame resolution becomes $(n \times 1536) \times (2 \times 2048)$, where n is the number of cameras. After data preparation, the combined frame—containing both RGB and depth information—is encoded using the H.264 codec and streamed at the original frame rate of 30 FPS via the RTP/UDP protocol using FFmpeg. The stream is sent from the Playback block to the Rendering component through the virtual machine.

The **Rendering** component builds, renders the point cloud using the colour (RGB) and depth (D) information and the bounding cylinder parameters from the session file created during Dataset Preparation step. The output stream is rendered in an OpenGL window on the GPU running on a client machine in real time. The Capture and Rendering components together can operate either with or without the 2D Segmentation module. In both cases, the output streams are saved, allowing the Evaluation block to compare point clouds generated using only bounding constraints with those refined by segmentation as part of our evaluation. The Capture and Rendering machines hardware parameters are the following:

- **GPU:** NVIDIA GeForce RTX 3050 with 8GB VRAM.
- **CPU:** 16 cores with memory of 62 GB RAM.

Our **2D Segmentation** mechanism is deployed on the virtual machine with GPU to achieve the highest performance to the greatest extent possible. This component generates other meta data files for the Evaluation block which contains the execution time for each block in the developed code for the saved, encoded timestamps of the frames. As explained in the implementation section, this component reads the stream transferred from the Capture block, decodes it using the H.264 codec, and applies the segmentation algorithm to the streamed RGB frames to have the segmentation masks. After masking the RGB and depth frames for each camera, we combine them together into one frame in the Combining step. The combined image consists of two columns (RGB image, depth image) and n rows, where n corresponds to the number of cameras. The Combining function is processed on a different CPU thread in order to achieve the highest performance. The image is then encoded using the H.264 codec and streamed via the RTP/UDP protocol on a separate thread. The virtual machine hardware parameters are the following:

- **GPU:** NVIDIA Quadro RTX 4000 with 8GB VRAM.
- **CPU:** 20 cores with memory of 31 GB RAM.

4.3 Evaluation Metrics

We conduct both quantitative and qualitative assessments, focusing on performance metrics and visual point cloud comparisons.

4.3.1 Performance Metrics. We measure the execution time of key components in the implementation and the running Frame-per Second (FPS) is calculated from this execution time to assess system performance and identify potential bottlenecks of the system. The execution times in milliseconds of each block are collected in metadata files on the virtual machine after encoding each frame with the H.264 codec. The measured blocks are the following:

2D Segmentation pipeline: represents the overall performance of the system, including decoding, input pre-processing (Section 3.1),

[6]FFmpeg: https://ffmpeg.org

Table 1: FPS performance of system components for sequences 1–5 (CWIPC-SXR dataset). The colouring indicates: green (≥ 30, real-time FPS), yellow (15 ≤ framerate < 29, close to real-time), and red (< 15, underperforming).

	ID of the experiment														
	1			2			3			4			5		
Number of cameras	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
2D Segmentation pipeline	32	15	11	34	16	11	34	14	11	37	13	10	35	20	13
Input Pre-processing	253	147	96	273	137	97	271	134	96	288	137	98	272	135	100
Segmentation	72	42	29	73	41	29	73	41	30	75	41	29	73	43	30
Inference	195	162	167	191	164	179	183	138	179	197	146	183	194	177	166
Post-processing predictions	263	115	65	270	114	65	290	124	64	259	128	59	263	110	65
Post-processing prototypes	2424	4486	1522	2336	1598	2051	2346	1874	1910	2428	2097	1757	2383	2306	1541
Masking	101	37	35	138	52	33	119	39	34	137	30	23	122	103	56
Combining	165	32	15	161	28	20	158	36	23	116	24	23	177	34	24

Table 2: FPS performance of system components for sequences 6–10 (CWIPC-SXR dataset). The colouring indicates: green (≥ 30, real-time FPS), yellow (15 ≤ framerate < 29, close to real-time), and red (< 15, underperforming)

	ID of the experiment														
	6			7			8			9			10		
Number of cameras	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
2D Segmentation pipeline	33	15	9	30	17	11	36	12	9	35	15	9	38	21	17
Input Pre-processing	269	137	98	254	142	95	273	133	97	267	142	96	269	141	97
Segmentation	73	42	29	69	43	29	74	41	30	73	44	30	70	42	31
Inference	186	167	136	155	182	157	193	118	175	187	187	183	146	147	185
Post-processing predictions	281	112	76	368	107	67	269	145	64	276	105	62	390	121	64
Post-processing prototypes	2516	2957	1611	2131	2311	1598	2425	1850	1738	2362	2345	1768	1792	1971	1813
Masking	99	51	32	111	47	19	134	36	22	136	26	20	213	170	197
Combining	117	29	25	153	36	24	166	35	24	168	37	24	145	35	22

Figure 4: Qualitative results for three cameras on the CWIPC-SXR dataset show how segmentation can replace static bounding boxes by excluding points (red) more accurately. Sequence IDs (see Section 4.1): 3, 1, 9, 5, 2, 4, 7, 10 (left to right).

segmentation (see *Segmentation* below), masking (Section 3.6), combining (see *Combining* below), and encoding the output stream.

Input Pre-processing: measures the performance of image preprocessing before passing it to the YOLO model, implemented using OpenCV on the CPU. (See Section 3.1)

Segmentation: evaluates together the overall performance of the Inference (Section 3.2), Mask Refinement (Section 3.3), Shrinking (Section 3.4), and Filtering (Section 3.5).

Inference: measures the time taken by the YOLO model for inference on the GPU, excluding preprocessing and post-processing. (See Section 3.2)

Post-processing Predictions: evaluates the time required to process the raw model outputs, including bounding box decoding and non-maximum suppression. (See Section 3.3.1)

Input Pre-processing Prototypes: evaluates the time required to obtain the final masks from the predictions. (See Section 3.3.2)

Masking: measures the time taken to apply the segmentation masks to both the RGB and depth images on the CPU. (See Section 3.6)

Combining: measures the CPU time to merge frames from multiple cameras before encoding. (See 2D Segmentation block in Section 4.2)

4.3.2 Visual Point Cloud Comparison. To evaluate how accurate is our method in relation with using bounding boxes, we compare point clouds generated using both, at corresponding timestamps. These point clouds are generated by the Rendering component. To obtain qualitative results, we calculate point-wise distances between the two point clouds and label the differences as follows:

- **Red points**: present in the bounding box point cloud but not in the segmentation.
- **Green points**: present in the segmentation result but not in the bounding box.

5 Results

Tables 1 and 2 summarize the performance (in FPS) of key implementation blocks (see Section 4.3.1) of the 2D segmentation component, evaluated across different numbers of cameras and sequences defined in Section 4.1.

The results show that performance decreases as more cameras are added, increasing the system load. Among all components, the Combining block proved to be the most computationally expensive when using multiple cameras. For that reason our implementation carries out the combining step in different a separate CPU thread, which improved overall system performance compared to our initial experiments. The Masking block also shows a drop in performance as the number of cameras increase. While it performed well with a single camera, its efficiency declined with two or three cameras. However, in some cases, the difference between using two and three cameras was not significant. Failures in segmentation (i.e., not being able to detect objects) are shown in sequence 10 (`cwi_book_presentation`), where the algorithm frequently fails to detect the book in several frames. As a result, the masking step shows high FPS values in those frames, since no mask is returned from the segmentation and no actual computation is performed. In contrast, the YOLO Inference step maintained a stable performance across all configurations. Running the model on the GPU with batched input showed consistent results regardless of the number of cameras. It is important to note that the Post-processing prototypes block showed high performance in all cases, indicating that this operation was executed efficiently on the GPU. The remaining blocks also presented real-time performance; however, they still add additional computational load to the system, influencing the overall performance (2D Segmentation pipeline block).

Apart from performance metrics, we present initial findings on the visual accuracy of the results to understand if the proposed system can replace static methods to filter a scene, such as fixed bounding boxes. Figure 4 presents representative frames from the 10 sequences, demonstrating how segmentation can effectively focus on relevant regions of the scene without static constraints. Red points indicate areas that were included using static bounding boxes but are missing in the segmented point cloud—highlighting points unnecessarily added to the scene by the static approach.

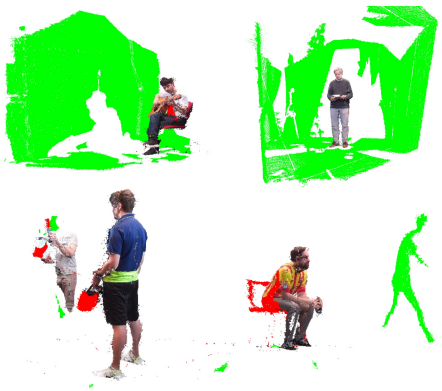


Figure 5: Qualitative results show segmentation failures with adding green points to the scene

Figure 5 shows examples of frames in which not objects were segmented by the model. Green points represent false positives, i.e. incorrectly added points due to segmentation failures. These occur when the segmentation algorithm fails to generate a valid mask and returns the entire RGB frame instead, which is the mechanism intentionally implemented to allow us a better visual inspection of such failures. This is illustrated in the first row of Figure 5, in the sequences `cwi_acoustic_guitar` and `cwi_book_presentation`. The second row shows examples where segmentation includes other objects, such as people outside the bounding box. Although technically a segmentation error, this is an expected behaviour of the model. In the sequence `cwi_padel_training_r1_t1`, parts of the recording person (e.g., a leg, half of the face) appear due to their presence in the space defined by the bounding box. Similarly, the full recording person is included in the point cloud in the `cwi_watching_football_r2_t1` sequence. Figure 6 illustrates how segmentation targets can significantly affect the reconstructed scene. In the upper part of the figure, only the person is segmented, and the racket appears as an extra object. In the lower part, both the person and the racket are segmented. This example highlights the power of semantic understanding of the scene and the flexibility of a segmentation-based approach in controlling what is included in the final reconstruction.

6 Discussion

In this section we review our initial research questions.

***RQ1**: Is the proposed system feasible?* We implemented an initial prototype that integrates 2D segmentation into a live volumetric video streaming system and evaluated it using a reproducible testbed with the publicly available CWIPC-SXR dataset, demonstrating the feasibility of the system. The prototype presented successfully processes synchronized RGB-D streams from multiple cameras in real-time or near-real-time, demonstrating that integrating 2D segmentation is practical within existing volumetric capture workflows. Constraints and limitations exist in the number of cameras to use and their resolution, but these can be partially solved in future iterations of this concept. In addition, 2D segmentation enables more efficient scene reconstruction by focusing on relevant regions and discarding unnecessary points, thus providing a more

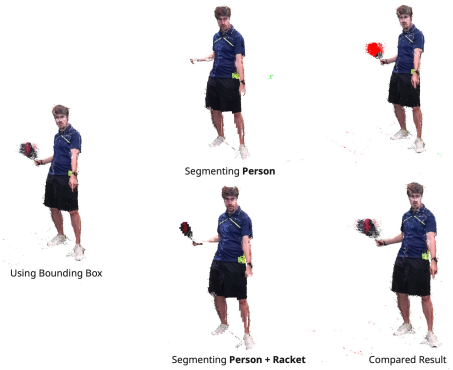


Figure 6: Example from `cwi_padel_training_r1_t1` showing how detecting only person (first row) vs. person and racket (second row) affects the point cloud at the same timestamp.

precise alternative to replace scene isolation with static bounding box.

***RQ2**: What is its performance and scalability? What are its limitations and where are the bottlenecks?*

The performance metrics showed that keeping the real time performance streamed the videos with 30 FPS is possible in the case of using one camera. The achieved frame rate is lower with more cameras, but it still offers acceptable values. The high frame resolution (2048×1536) of the dataset used in the evaluation is also a limiting factor. Several operations—such as RGB pre-processing, masking, and combining—are performed on the CPU using OpenCV, which slows down the processing. When the frame processing cannot keep up with the original stream speed at 30 FPS (Shown in Tables 1 and 2 as yellow or red coloured cells) the frames are dropped in order to not let waiting the processing for the next frame. For this reason not all of the frames are included in the output stream. In the metrics evaluation the YOLO-based segmentation inference remained stable and GPU-efficient because of using batched input. However, the masking and combining operations revealed bottlenecks when scaling to multiple cameras, with combining being the most computationally expensive step.

Additionally, masking performance in one case showed high FPS (10th sequence) due to segmentation failures (where no masks were returned, skipping masking computation). These represent limitations in the segmentation robustness. Another limitation occurs when the algorithm fails to differentiate between multiple objects, it may incorrectly include unwanted subjects—such as the person recording—instead of isolating only the main person in the bounding box.

***RQ3**: Are resultant volumetric video reconstructions accurate?*

Qualitative results show that our proposed 2D segmentation based reconstruction can isolate relevant scene components similar, and sometimes more precisely, than static bounding boxes. Figure 6 showed when focusing on small objects, like the paddle racket, can change the experience in the output. In many sequences, unnecessary background points were successfully excluded, enhancing scene clarity as shown in the Figure 4. However, accuracy depends heavily on segmentation quality. Failures to detect objects, or inclusion of unintended subjects (e.g., other person as camera operators), introduced noise or artifacts. While segmentation improves reconstruction accuracy overall, some errors—particularly false positives and missed detections—remain a challenge.

7 Conclusion and Future Work

This paper presented a novel method that integrates semantic 2D segmentation into live volumetric video pipelines, replacing static constraints like bounding boxes. This is a crucial enabler to create object-based volumetric video scenes. A system like this can enable new media capture workflows and novel methods to create immersive scenes using an object-based approach. To evaluate its feasibility and performance, we carried out an empirical study using the CWIPC-SXR dataset on a reproducible testbed. The results show that not only it is possible to perform 2D semantic segmentation, but it can also improve scene clarity by focusing on relevant regions and excluding unnecessary content. However, segmentation sometimes fails to detect the desired object or includes unwanted background elements, reducing accuracy. Our experiments also revealed that real-time performance could not be achieved in some scenarios, especially when using two or more high-resolution camera streams simultaneously.

Beyond the results presented in this paper, 2D Segmentation has been successfully integrated in our live volumetric video system, initially implemented in [3] and showcased in [18] with nice performance under the constraints showed here. This paper opens interesting areas of research and developments, for instance related to system stability and scalability. This architecture should be evaluated for different AI models, and in parallelized implementations, where synchronization of streams is paramount. CUDA-based optimizations[7] may enhance performance, although GPU memory transfer for high-resolution frames remains a bottleneck. A redesigned system architecture may be required. To address segmentation errors, incorporating a tracking algorithm that assigns unique IDs to objects could help filter out irrelevant detections. These improvements would bring us closer to understand the 3D scene and integrate physical entities into immersive environments.

Acknowledgments

This work has also been partially funded by the Ministry of Economic Affairs and Digital Transformation and European funds from the Recovery and Resilience Mechanism (RRM), through the UNICO I+D 5G-6G 2023 Call, under project INMERLIVE (TSI-064200-2023-4); and by the European Union (SPIRIT, 101070672) through the Open Call 1 project BAZKARIA. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The SPIRIT project has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

[7] https://docs.nvidia.com/cuda/cuda-c-programming-guide/

References

[1] Dimitrios S. Alexiadis, Anargyros Chatzitofis, Nikolaos Zioulis, Olga Zoidi, Georgios Louizis, Dimitrios Zarpalas, and Petros Daras. 2017. An Integrated Platform for Live 3D Human Reconstruction and Motion Capturing. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 4 (April 2017), 798–813. doi:10.1109/tcsvt.2016.2576922

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. 2019. YOLACT: Real-Time Instance Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9156–9165. doi:10.1109/ICCV.2019.00925

[3] Sergio Cabrero Barros, Ander Elosegi, Iñigo Tamayo, Ana Dominguez, and Mikel Joseba Zorrilla. 2024. Volumetric Video on the Web: a platform prototype and empirical study. In *Proceedings of the 29th International ACM Conference on 3D Web Technology* (Guimarães, Portugal) *(Web3D '24)*. Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. doi:10.1145/3665318.3677170

[4] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 77–85. doi:10.1109/CVPR.2017.16

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells (Eds.). Springer International Publishing, Cham, 424–432.

[6] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Trans. Graph.* 34, 4, Article 69 (July 2015), 13 pages. doi:10.1145/2766945

[7] Simon N. B. Gunkel, Rick Hindriks, Karim M. El Assal, Hans M. Stokking, Sylvie Dijkstra-Soudarissanane, Frank ter Haar, and Omar Niamut. 2021. VRComm: an end-to-end web system for real-time photorealistic social VR communication. In *Proceedings of the 12th ACM Multimedia Systems Conference* (Istanbul, Turkey) *(MMSys '21)*. Association for Computing Machinery, New York, NY, USA, 65–79. doi:10.1145/3458305.3459595

[8] Jonathan Heagerty, Sida Li, Eric Lee, Shuvra Bhattacharyya, Sujal Bista, Barbara Brawn, Brandon Y. Feng, Susmija Jabbireddy, Joseph JaJa, Hernisa Kacorri, David Li, Derek Yarnell, Matthias Zwicker, and Amitabh Varshney. 2024. HoloCamera: Advanced Volumetric Capture for Cinematic-Quality VR Applications. *IEEE Transactions on Visualization and Computer Graphics* 30, 5 (May 2024), 2767–2775. doi:10.1109/TVCG.2024.3372123

[9] Jack Jansen, Shishir Subramanyam, Romain Bouqueau, Gianluca Cernigliaro, Marc Martos Cabré, Fernando Pérez, and Pablo Cesar. 2020. A pipeline for multiparty volumetric video conferencing: transmission of point clouds over low latency DASH. In *Proceedings of the 11th ACM Multimedia Systems Conference* (Istanbul, Turkey) *(MMSys '20)*. Association for Computing Machinery, New York, NY, USA, 341–344. doi:10.1145/3339825.3393578

[10] Nidhal Jegham, Chan Young Koh, Marwan F. Abdelatti, and Abdeltawab M. Hendawi. 2024. YOLO Evolution: A Comprehensive Benchmark and Architectural Review of YOLOv12, YOLO11, and Their Previous Versions. https://api.semanticscholar.org/CorpusID:273798273

[11] Preetish Kakkar and Hariharan Ragothaman. 2024. The Evolution of Volumetric Video: A Survey of Smart Transcoding and Compression Approaches. *International Journal of Computer Graphics & Animation* 14, 1/2/3/4 (Oct. 2024), 01–11. doi:10.5121/ijcga.2024.14401

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4015–4026.

[13] Sergi Fernández Langa, Mario Montagud Climent, Gianluca Cernigliaro, and David Rincón Rivera. 2022. Toward Hyper-Realistic and Interactive Social VR Experiences in Live TV Scenarios. *IEEE Transactions on Broadcasting* 68, 1 (2022), 13–32. doi:10.1109/TBC.2021.3123499

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:14113767

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

[16] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. 2017. PointNet++: deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 5105–5114.

[17] Ignacio Reimat, Evangelos Alexiou, Jack Jansen, Irene Viola, Shishir Subramanyam, and Pablo Cesar. 2021. CWIPC-SXR: Point Cloud dynamic human dataset for Social XR. In *Proceedings of the 12th ACM Multimedia Systems Conference* (Istanbul, Turkey) *(MMSys '21)*. Association for Computing Machinery, New York, NY, USA, 300–306. doi:10.1145/3458305.3478452

[18] Andrés Santos-Torres, Patricia de Torres Coll, Tamás Bukits, Ramón Perisé, and Sergio Cabrero Barros. 2025. The XR Table: Envisioning The Future of Remote Dining Experiences Using Immersive Telepresence. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 1674–1690. doi:10.1145/3715336.3735730

[19] Tetsuri Sonoda and Anders Grunnet-Jepsen. 2021. Depth image compression by colorization for Intel RealSense depth cameras. *Intel Rev* 1 (2021).

[20] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, and Pablo Cesar. 2023. VR2Gather: A Collaborative, Social Virtual Reality System for Adaptive, Multiparty Real-Time Communication. *IEEE MultiMedia* 30, 2 (2023), 48–59. doi:10.1109/MMUL.2023.3263943

[21] Xu Zhao, Wen-Yan Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. 2023. Fast Segment Anything. *ArXiv* abs/2306.12156 (2023). https://api.semanticscholar.org/CorpusID:259212104

[22] Hao Zhou, Yidan Feng, Mingsheng Fang, Mingqiang Wei, Jing Qin, and Tong Lu. 2021. Adaptive Graph Convolution for Point Cloud Analysis. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 4945–4954. https://api.semanticscholar.org/CorpusID:237194767

Perceptual Quality Assessment of Spatial Videos on Apple Vision Pro

Afshin Gholami
University of Klagenfurt
Klagenfurt, Austria
afshingh@edu.aau.at

Sara Baldoni
University of Padova
Padova, Italy
sara.baldoni@unipd.it

Federica Battisti
University of Padova
Padova, Italy
federica.battisti@unipd.it

Wei Zhou
Cardiff University
Cardiff, UK
ZhouW26@cardiff.ac.uk

Christian Timmerer
University of Klagenfurt
Klagenfurt, Austria
christian.timmerer@aau.at

Hadi Amirpour
University of Klagenfurt
Klagenfurt, Austria
hadi.amirpour@aau.at

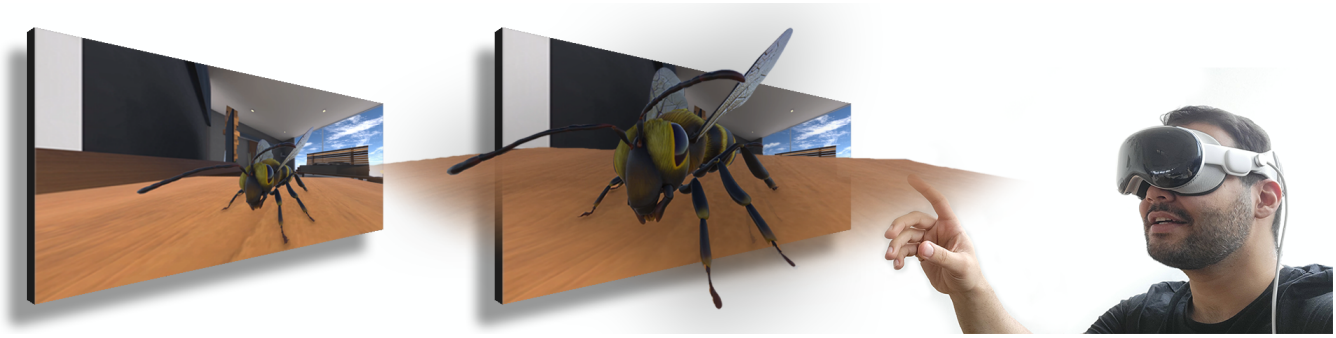


Figure 1: Quality evaluation of 2D and stereoscopic videos on the Apple Vision Pro (AVP).

Abstract

Immersive stereoscopic (3D) video experiences have entered a new era with the advent of smartphones capable of capturing stereoscopic videos, advanced video codecs optimized for multiview content, and Head Mounted Displays (HMDs) that natively support stereoscopic video playback. In particular, Apple's recent introduction of *spatial video* capture on the recent iPhone Pro series and immersive playback on the Apple Vision Pro (AVP) has accelerated the mainstream adoption of stereoscopic content. In this work, we evaluate the quality of spatial videos encoded using optimized x265 software implementations of Multiview HEVC (MV-HEVC) on the AVP and compare them with their corresponding 2D versions through a subjective test.

To support this study, we introduce *SV-QoE*, a novel dataset comprising video clips rendered with a twin-camera setup that replicates the human inter-pupillary distance. Our analysis reveals that spatial videos consistently deliver a superior Quality of Experience (QoE) when encoded at similar bitrates, with the benefits becoming more pronounced at higher bitrates. Additionally, renderings at closer distances exhibit significantly enhanced video quality and depth perception, highlighting the impact of spatial proximity on immersive viewing experiences.

We further analyze the impact of disparity on depth perception and examine the correlation between Mean Opinion Score (MOS) and established objective quality metrics such as PSNR, SSIM, MS-SSIM, VMAF, and AVQT. Additionally, we explore how video quality and depth perception together influence overall quality judgments. The complete dataset, including videos and subjective scores, is publicly available at https://github.com/cd-athena/SV-QoE.

CCS Concepts

• **Information systems** → **Multimedia streaming**.

Keywords

stereoscopic, MV-HEVC, QoE, spatial video, depth

ACM Reference Format:
Afshin Gholami, Sara Baldoni, Federica Battisti, Wei Zhou, Christian Timmerer, and Hadi Amirpour. 2025. Perceptual Quality Assessment of Spatial Videos on Apple Vision Pro. In *Proceedings of the 3rd International Workshop on Interactive eXtended Reality (IXR '25), October 27–28, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746269.3760422

IXR '25, Dublin, Ireland

ACM ISBN 979-8-4007-2051-2/2025/10
https://doi.org/10.1145/3746269.3760422

1 Introduction

Immersive media has transformed how users engage with digital content, extending beyond traditional viewing to provide highly interactive and engaging experiences [1, 2]. Advances in Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) have led

to the development of immersive environments that enhance user presence and interaction. As display hardware continues to evolve, with innovations such as high-resolution HMDs, eye-tracking, and spatial audio, immersive media is becoming increasingly realistic and accessible. Additionally, improvements in content creation tools are pushing the boundaries of digital storytelling, gaming, training simulations, and remote collaboration [3–5]. With the growing demand for lifelike and interactive experiences, immersive media is reshaping entertainment, education, and professional applications [6].

Among the various forms of immersive media, stereoscopic videos have regained popularity thanks to their ability to enhance realism through depth perception. Stereoscopic imaging works by capturing two slightly different perspectives of a scene (*i.e.*, one for the left eye and one for the right eye), mimicking the natural disparity of human vision [7]. This disparity allows the brain to interpret depth, resulting in a 3D perception of the content. Typically, stereoscopic videos are recorded using a dual-camera setup. During playback, these two images are displayed separately to each eye, either through passive polarization, active shutter glasses, or direct display on head-mounted devices [8].

Widespread adoption of this format has historically been limited by challenges across the entire multimedia pipeline, including content creation, efficient encoding, and display compatibility. Recent advancements have significantly lowered these barriers, making stereoscopic video a more accessible and scalable format for immersive media. The integration of stereoscopic video capture into consumer-grade devices, such as the iPhone Pro, has enabled effortless content creation without the need for specialized camera setups. On the encoding side, optimized HEVC-based [19] compression software, such as x265, ensures that high-quality stereoscopic video can be efficiently stored and streamed while maintaining perceptual quality. Furthermore, the emergence of HMDs with native support for stereoscopic video playback, such as the AVP and Meta Quest 3, has provided a dedicated ecosystem for consuming stereoscopic content. These developments have effectively bridged the gap between content creation, encoding, and rendering, thus enabling stereoscopic videos to become a viable and accessible format for next-generation multimedia applications.

While in Apple's definition, *spatial* videos refer to videos recorded with the iPhone Pro or AVP [18], encoded with their MV-HEVC [20] codec and displayed on the AVP, in this paper, we define *spatial* videos as stereoscopic content encoded using the MV-HEVC format and designed for seamless playback on the AVP.

Despite the advances in spatial video capture and display, there remains a significant gap in the research on spatial videos for immersive platforms, as most studies focus primarily on traditional 2D videos. While objective and subjective quality evaluation methodologies are well established for conventional video formats, their applicability to spatial videos, particularly on HMDs, remains less explored. Factors such as depth perception, binocular disparity, compression artifacts, and motion cues all influence the perceptual quality of spatial video in ways that differ from 2D content [21, 22]. Another key challenge in this domain is the lack of publicly available datasets for spatial video quality assessment. Existing datasets primarily focus on 2D content and do not adequately capture the perceptual nuances of spatial formats. Moreover, most of the stereoscopic datasets are limited to standard resolutions and frame rates, lacking high-resolution (*e.g.*, 4K) and high frame-rate (*e.g.*, 60fps) content, which are increasingly common in modern immersive video applications.

To address these limitations, (*i*) we introduce a novel dataset, *SV-QoE*, specifically designed for spatial video quality assessment, featuring diverse high-resolution 4K and high-frame-rate 60fps video sequences encoded at multiple quality levels (available at https://github.com/cd-athena/SV-QoE.) Furthermore, (*ii*) this paper bridges the research gap by conducting a comprehensive subjective evaluation of both 2D and spatial video quality on the AVP. We systematically assess viewer responses across three distinct quality levels, ensuring that all content is encoded using the same optimized open-source x265 codec at similar bitrates. Our experimental design involves controlled subjective tests, where participants experience a series of video sequences in both 2D and spatial formats. To comprehensively assess the viewing experience, participants evaluate three key aspects:

(1) video quality (Q1),
(2) depth perception (Q2), and
(3) overall quality (Q3).

(*iii*) We also evaluate the correlation between disparity and perceived depth, shedding light on how disparity cues affect depth perception and contribute to overall quality in stereoscopic content. (*iv*) Additionally, we examine the correlation between well-known objective quality metrics (*i.e.*, PSNR, SSIM, VMAF, and AVQT) and subjective video quality scores, offering a deeper understanding of the reliability and applicability of existing objective models in evaluating 2D and spatial video content on AVP. (*v*) Finally, we analyze the impact of video quality and depth perception on overall quality, providing insights into how these factors influence the immersive viewing experience.

2 Related Work

The stereoscopic video pipeline comprises several key modules: creation, encoding, delivery, rendering, and quality evaluation. The creation phase involves capturing left and right views using dual-camera systems, which require precise calibration and synchronization to ensure consistency. Existing stereoscopic video datasets have been summarized in Table 1. These datasets typically lack high-resolution and high frame-rate content, which limits their applicability to modern immersive viewing scenarios.

Recently, deep learning methods have also been applied to generate stereoscopic videos. For example, Zhang *et al.* [23] introduced a novel framework for converting 2D videos into stereo videos. Their approach employs depth-warping and blend-inpainting techniques, incorporating a mask-based hierarchical feature update refiner and a disparity expansion strategy to improve inpainting accuracy and reduce foreground bleeding. During the encoding stage, advanced compression methods are utilized to maintain high perceptual quality while reducing bitrate. In particular, inter-view redundancy between the left and right views is exploited to enhance compression efficiency, as explored in MV-HEVC [20] and its optimized variants [24, 25]. The delivery phase focuses on reliably transmitting this content over varying network conditions

Table 1: Overview of stereoscopic video datasets.

Dataset Name	Year	Resolution	Description
KITTI Stereo 2012 [9]	2012	1226×370	Outdoor driving scenes
KITTI Stereo 2015 [10]	2015	1242×375	Outdoor driving scenes with dynamic scenes with objects
SceneFlow [11]	2016	960×540	Synthetic stereo sequences
MPI-Sintel [12]	2012	1024×436 (24fps)	Synthetic scenes with complex motion and visual effects
RMIT3DV HD 3D Video [13]	2012	1920×1080 (25fps)	Diverse urban scenes
EPFL MMSPG 3DVQA [14]	2010	1920×1080 (25fp)	High-quality visual variations
Stereo Video Database [15]	2010	1920×1080 (25fps)	Stereo cinema post-production
NAMAD3D [16]	2012	1920×1080(25fps)	Natural 3D scenes with twin-lens camera
SVSR-Set [17]	2022	1920×1080 (30fps)	Indoor/outdoor with varied motion and lighting
SVD [18]	2025	1920×1080 (30fps)	Spatial videos taken by iPhone and Apple Vision Pro
SV-QoE (Ours)	2025	3840×2160 (60fps)	High resolution, high frame-rate synthetic sequences captured from two distances

by employing adaptive streaming protocols [26] and network optimization strategies, as explored by Chen *et al.* [27]. Finally, the quality evaluation module employs both subjective assessments and objective metrics to gauge the perceptual impact of distortions unique to stereoscopic content [21].

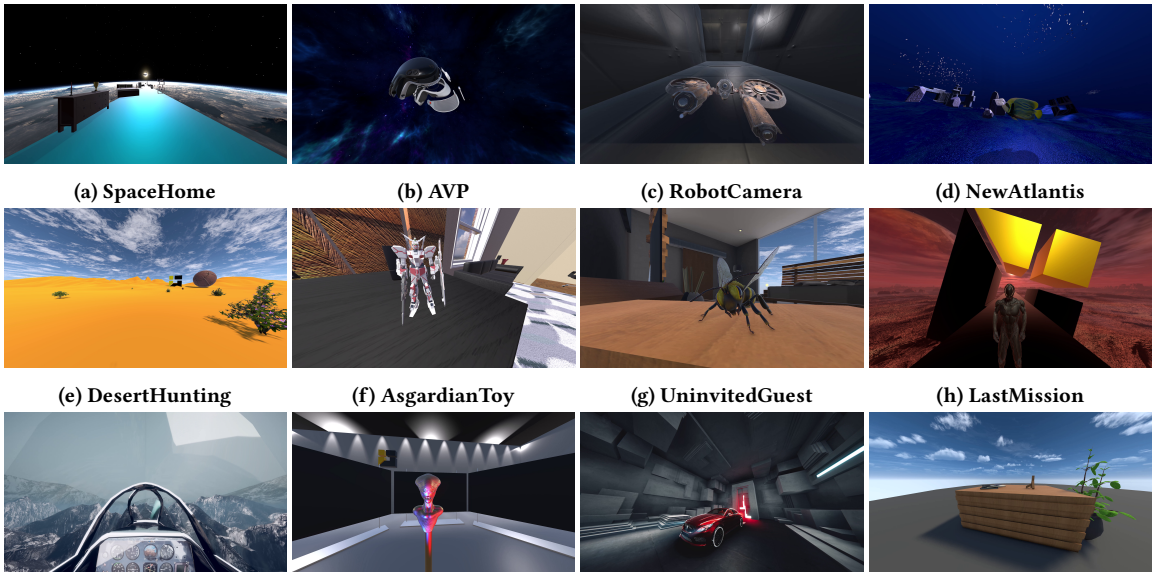
The quality evaluation of stereoscopic multimedia involves multiple parameters, including video quality, depth perception, and overall satisfaction. Goldmann *et al.* [28] investigated the impact of acquisition distortions, such as the baseline distance between left and right cameras, on the perceived quality of stereoscopic images and videos. Their findings indicate that as the camera baseline distance increases, perceived quality decreases, highlighting the critical role of acquisition setup in maintaining high-quality stereoscopic content. Zhou *et al.* [21] presented a comprehensive study on the visual quality assessment of 3D-HEVC compressed stereoscopic videos. They analyzed the impact of video compression and depth quality on the overall QoE. They develop a No Reference (NR) bitstream-level objective quality assessment model that extracts key features from 3D-HEVC bitstreams, such as quantization parameters and prediction residuals, to predict perceived video quality. Chen *et al.* [29] introduced a depth perception quality metric and extended it to an NR stereoscopic video quality assessment. Wan *et al.* [30] analyzed the impact of coding artifacts on depth perception in stereoscopic 3D videos, revealing that compression distortions introduced by the Advanced Video Coding (AVC) standard can significantly alter depth quality. Their subjective experiments showed that coding artifacts affect different spatial frequency components unequally, with high-pass and band-pass components being more crucial for depth perception than low-pass components. They also found that horizontal orientation structures play a dominant role in depth perception, and distortions in these components lead to more noticeable depth degradation.

Compared to 2D images, a 3D image consists of two 2D images – left and right views – introducing additional challenges in objective quality assessment. When the left and right views exhibit different types and levels of artifacts, asymmetric distortions occur, making it more complex to evaluate the quality of 3D images [31]. The simple average of the predicted quality scores from both views does not account for the binocular processing mechanisms of the Human Visual System (HVS) [32, 33]. To address these challenges, various 3D image quality assessment (IQA) methods have been developed, incorporating specific 3D characteristics. Notable approaches include the cyclopean model [34], weighted SSIM (W-SSIM), and weighted FSIM (W-FSIM) [35]. For 3D omnidirectional image quality assessment (OIQA), a multi-viewport-based model has been introduced [36], while the stereoscopic omnidirectional image quality evaluator (SOIQE) was designed based on predictive coding theory [37]. Galkandage *et al.* [38] introduced a novel HVS model inspired by physiological findings characterizing the motion-sensitive response of complex cells in the primary visual cortex. The proposed full-reference stereoscopic video quality assessment method leverages this model to enhance the prediction accuracy of perceived video quality.

3 Dataset Creation

In this paper, we present *SV-QoE*, a dataset that includes 12 scenes created using Unity engine 6 (Version: 6000.0.28f1) to showcase a variety of artistic styles, ensuring accurate depth representation and minimizing capturing distortions. Each scene was recorded as a 10-second clip. The Unity engine camera was configured with a perspective projection to ensure realistic depth perception, with a 137° horizontal Field of View (FoV) for a wide yet natural perspective, chosen to closely match the effective horizontal viewing range of the human visual system and modern VR headsets, thereby providing an immersive experience without introducing noticeable geometric distortion.

To capture both 2D and spatial video content, we employed a three-virtual camera setup. The central camera, located at the origin (0,0,0), was used to create the 2D content. Meanwhile, a stereo pair was placed 65 mm apart to replicate the human interocular distance, capturing the left and right views for immersive spatial video. All videos were rendered in 4K resolution at 60 fps. Sample frames from these videos are presented in Figure 2. To incorporate variations in object distance, five scenes (*AsgardianToy*, *AVP*, *CommaDotStudio*, *NewAtlantis*, and *UninvitedGuest*) were captured twice to represent both 'near' and 'far' perspectives. The 'near' perspective features objects positioned close to the camera, emphasizing fine spatial details and depth, while the 'far' perspective captures the same scenes with objects located farther away, offering a broader and more distant view of the environment. To maintain a consistent compression standard across formats, both 2D and spatial videos



(a) SpaceHome (b) AVP (c) RobotCamera (d) NewAtlantis

(e) DesertHunting (f) AsgardianToy (g) UninvitedGuest (h) LastMission

(i) Bumblebee (j) CommaDotStudio (k) Mercedes (l) TrainingScene

Figure 2: Sample frames of the created *SV-QoE* dataset.

were encoded using the open-source x265 encoder (version 4.1)[1]. The x265 encoder was chosen for its efficiency in high-quality video compression and its support for both standard 2D encoding and multiview coding, allowing MV-HEVC compatibility for spatial content. A Constant Rate Factor (CRF) quality control approach was employed during encoding. For spatial videos, three quality levels (*i.e.*, low, medium, high) were generated, while for 2D videos, the CRF values were selected to closely match the corresponding bitrates of these levels. High quality corresponds to a VMAF [39] score of 95, medium to 85, and low to approximately 75, with VMAF being the average of the left and right views.

4 Testing Procedure

We conducted a subjective test using the 5-point Absolute Category Rating (ACR) [40] methodology to evaluate the perceived quality of both 2D and spatial video content. In this test, participants viewed a series of 10-second video sequences and rated each on three distinct aspects: video quality, perceived depth, and overall quality using a five-point categorical scale. In this study, video quality refers to the fidelity of the videos (*e.g.*, compression artifacts), while overall quality captures the way the viewer experiences the content as a combination of video quality with perceived depth and immersion. Prior to the main test, participants underwent a training session designed to familiarize them with the rating procedure and the different video formats they would encounter. In addition, a Snellen visual acuity test was administered to ensure that all participants had normal vision. A total of 30 participants (11 females and 19 males; average age: 31 ± 6 years) took part in the study, and the entire testing session took, on average, 31 minutes to complete. Each participant rated each video six times: three times in spatial video encoded at three quality levels, and three times in 2D encoded at the corresponding bitrates of the spatial videos. To mitigate ordering effects, the presentation order of the videos was randomized for each participant.

[1]https://bitbucket.org/multicoreware/x265_git/src/Release_4.1/

5 Evaluation and Results

In this section, we analyze the results of the subjective test. First, we identify and remove any outliers from the collected data. Outlier detection was conducted using statistical methods, including the Z-score and Interquartile Range (IQR) [41], to identify extreme values that deviate significantly from the dataset's central distribution. As a result, three outlier responses were excluded from the analysis, and the subsequent results are based on data from 27 subjects.

5.1 Integrated perceptual quality analysis

Figure 3 shows the integrated results from the subjective test. In terms of video quality, both 2D and spatial formats were perceived similarly by the participants. However, depth perception showed a significant difference between the two modalities, with spatial content providing a notably enhanced depth experience. Additionally, overall quality ratings were higher for spatial videos compared to 2D videos. ANOVA [42] results further confirmed these findings, *i.e.*, video quality showed no significant difference ($p = 0.479$), while depth perception and overall quality exhibited significant differences ($p < 0.0001$ for both).

5.2 Perceptual quality analysis across different quality levels

Figure 4 shows that the perceived video quality at medium and high quality levels is similar for 2D and spatial videos encoded at the same bitrate. However, all other metrics exhibit statistically significant differences (see Table 2). This outcome may be attributed to the use of the default player in AVP, where the combination of

Table 2: ANOVA significance test results for different quality levels. A significance level of $\alpha = 0.05$ was used.

Quality level	Video quality	Depth perception	Overall quality
Low	Significant ($p = 1.90 \times 10^{-4}$)	Significant ($p = 1.23 \times 10^{-20}$)	Significant ($p = 1.10 \times 10^{-2}$)
Medium	Not Significant ($p = 1.60 \times 10^{-1}$)	Significant ($p = 3.69 \times 10^{-30}$)	Significant ($p = 5.12 \times 10^{-9}$)
High	Not Significant ($p = 2.45 \times 10^{-1}$)	Significant ($p = 1.27 \times 10^{-24}$)	Significant ($p = 7.06 \times 10^{-11}$)

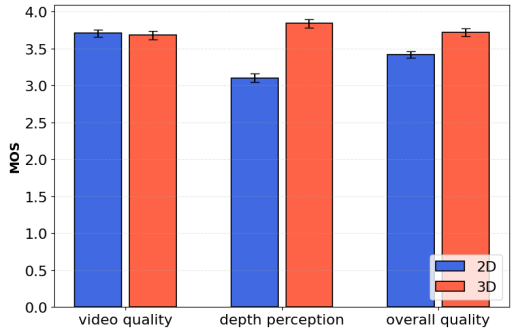


Figure 3: Integrated results: video quality is perceived similarly for both 2D and spatial formats, while depth perception is significantly enhanced in spatial content (3D), leading to higher overall quality ratings.

viewing distance [43] and player size might have made it difficult for users to notice artifacts at medium and high quality levels. This observation highlights the need for further investigation in future work.

Figure 5 shows sample frames from videos rendered at near and far object distances, while Figure 6 presents the average subjective ratings for five sequences, each captured with both near and far distances from the camera. The results consistently show that scenes featuring near object distances received significantly higher ratings in video quality, depth perception, and overall quality compared to their far-distance counterparts. When objects are positioned closer to the stereo cameras, the binocular disparity between the left and right views increases. This increased disparity enhances stereoscopic depth cues, making the 3D structure of the scene more salient and compelling for viewers. Consequently, the near-distance scenes produced a more immersive depth experience. These findings highlight the critical role of object proximity in shaping the viewing experience: as objects are rendered closer to the camera, spatial cues become more prominent, thereby enhancing perceived video quality, depth, and overall content appreciation. One potential explanation for the improved video quality is that closer objects occupy a larger portion of the screen and often reveal more visual detail, which may lead viewers to perceive the image as sharper or more vivid. However, this hypothesis warrants further investigation in future work.

5.3 Relationship between disparity and depth perception

To assess the strength of the relationship between stereoscopic disparity and depth perception (Q2), we first extracted disparity using the Stereo Semi-Global Block Matching (StereoSGBM) algorithm [44]. We then trained regression models on the full dataset and evaluated the Pearson Linear Correlation Coefficient (PLCC) and the coefficient of determination (R^2). These metrics respectively quantify the linear correlation and the proportion of variance in Q2 MOS explained by the average disparity. The results are presented in Table 3. All models exhibit a positive correlation, with performance steadily improving from simple linear models to more complex regressors. The highest accuracy is achieved by the Random Forest model, which yields a PLCC of 0.9184 and an R^2 of 0.8434, indicating a strong nonlinear relationship between disparity and depth perception.

Next, to evaluate out-of-sample predictive performance, we applied a leave-one-video-out cross-validation scheme and computed the Mean Absolute Error (MAE) for each model. MAE measures the average magnitude of prediction errors in the same units as Q2 MOS. Table 4 lists these MAE values, sorted from highest (worst) to lowest (best). The MAE results follow a consistent trend: as model complexity increases, prediction errors decrease. While linear models yield relatively high MAE values (*e.g.,* 0.3205 for linear regression), nonlinear models such as SVR (0.2411), KNN (0.2222), and Random Forest (0.2127) provide more accurate depth perception predictions.

5.4 Relationship between video quality and objective quality metrics

Similar to the previous section, we evaluate the strength of the relationship between widely used objective quality metrics—PSNR, SSIM [45], MS-SSIM [46], VMAF [39], and AVQT[2]—and subjective video quality (Q1). AVQT version 2 is specifically designed for spatial video content. For all other metrics, we computed the average scores between the left and right views.

We trained various regression models on the full dataset and computed PLCC and R^2. These metrics quantify the linear correlation and the proportion of variance in Q1 explained by each objective metric, respectively. The results are summarized in Table 5.

Among the metrics, AVQT exhibits the highest correlation with subjective video quality, particularly when modeled with nonlinear regressors. For instance, using Random Forest, AVQT achieves a PLCC of 0.9650 and an R^2 of 0.9231, outperforming all other metrics. VMAF also shows strong predictive power, especially with complex models like Random Forest (PLCC = 0.9575, R^2 = 0.9027). In contrast, traditional metrics such as PSNR, SSIM, and MS-SSIM show weaker correlations in simpler models, though their performance improves with increased model complexity. Notably, even with linear regression, AVQT (PLCC = 0.7302) and VMAF (PLCC = 0.6643) demonstrate stronger alignment with subjective quality than all other metrics.

[2]https://developer.apple.com/download/all/?q=avqt

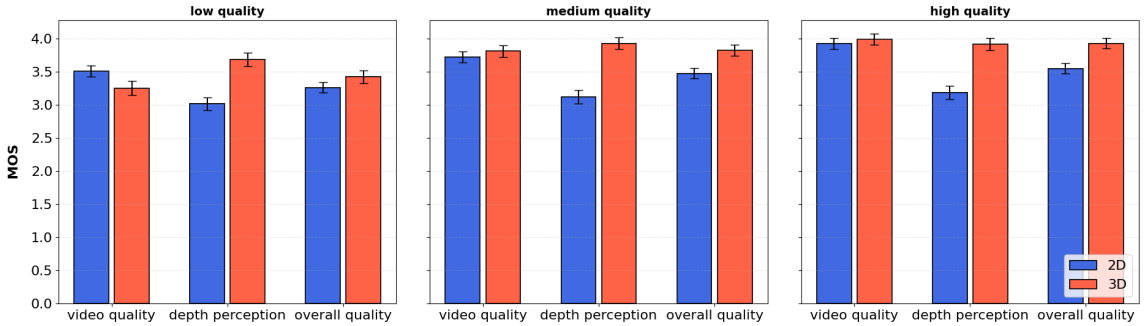
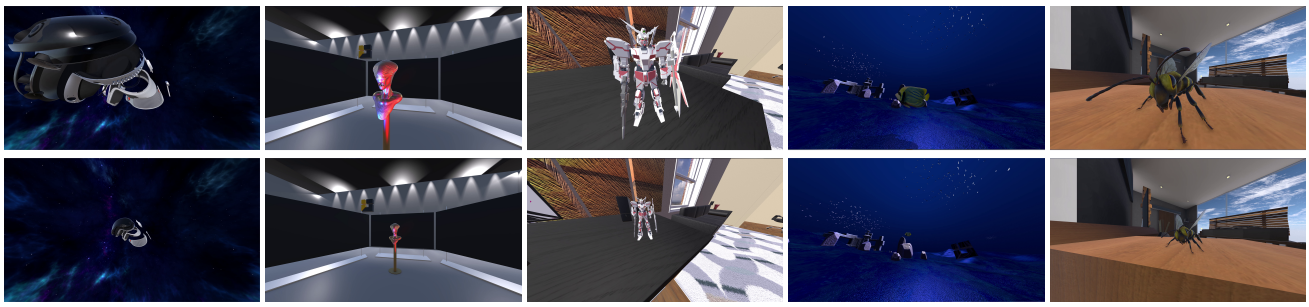


Figure 4: Results by quality level: at lower quality, 2D content exhibits higher video quality with similar overall quality, but immersive content provides enhanced depth perception. With increasing quality, video quality ratings converge, depth perception remains superior in immersive content, and overall QoE improves.



(a) AVP **(b) CommaDotStudio** **(c) AsgardianToy** **(d) NewAtlantis** **(e) UninvitedGuest**

Figure 5: Sample frames from videos rendered at near (upper row) and far (lower row) distances.

Table 3: PLCC and R^2 between disparity and depth perception.

Model	PLCC	R^2
Linear Regression	0.4788	0.2293
Polynomial Regression (Degree 2)	0.5533	0.3062
Polynomial Regression (Degree 3)	0.6921	0.4790
Support Vector Regressor (SVR)	0.7673	0.5810
K-Nearest Neighbors (KNN)	0.8648	0.7393
Random Forest	0.9184	0.8434

Table 4: Leave-One-Video-Out MAE Comparison.

Model	MAE
Linear Regression	0.3205
Polynomial Regression (Degree 2)	0.3121
Polynomial Regression (Degree 3)	0.2879
Support Vector Regressor (SVR)	0.2411
K-Nearest Neighbors (KNN)	0.2222
Random Forest	0.2127

Table 5: PLCC and R^2 between objective quality metrics and Q1 across regression models.

Model	**PSNR**		**SSIM**		**MS-SSIM**		**VMAF**		**AVQT**	
	PLCC	R^2	PLCC	R^2	PLCC	R^2	PLCC	R^2	PLCC	R^2
Linear Regression	0.4186	0.1753	0.4933	0.2434	0.3904	0.1524	0.6643	0.4413	**0.7302**	**0.5332**
Polynomial Regression (Degree 2)	0.4379	0.1917	0.5396	0.2912	0.4397	0.1934	0.7356	0.5412	**0.7968**	**0.6349**
Polynomial Regression (Degree 3)	0.4449	0.1980	0.5401	0.2917	0.4448	0.1979	0.7571	0.5732	**0.8034**	**0.6454**
Support Vector Regressor (SVR)	0.5778	0.3311	0.5634	0.3185	0.5067	0.2299	0.7587	0.5746	**0.8122**	**0.6492**
K-Nearest Neighbors (KNN)	0.5778	0.3311	0.5304	0.2793	0.5416	0.2891	0.7399	0.5445	**0.8284**	**0.6706**
Random Forest	0.9393	0.8573	0.9339	0.8278	0.9274	0.8263	0.9575	0.9027	**0.9650**	**0.9231**

We then evaluated generalization performance using leave-one-video-out cross-validation and reported the MAE for each model. As shown in Table 6, MAE provides a complementary view to PLCC and R^2, capturing prediction accuracy on unseen videos. The table highlights that AVQT consistently outperforms the other metrics across all regression models, achieving the lowest MAEs in every

Table 6: MAE between objective quality metrics and video quality (Q1) across regression models using leave-one-out cross-validation.

Model	PSNR	SSIM	MS-SSIM	VMAF	AVQT
Linear Regression	0.3475	0.3334	0.3521	0.3064	**0.2749**
Polynomial Regression (Degree 2)	0.3500	0.3249	0.3460	0.2921	**0.2462**
Polynomial Regression (Degree 3)	0.3591	0.3361	0.3493	0.2884	**0.2438**
Support Vector Regressor (SVR)	0.3484	0.3357	0.3528	0.3005	**0.2774**
K-Nearest Neighbors (KNN)	0.3743	0.3763	0.3773	0.3278	**0.2774**
Random Forest	0.4267	0.4385	0.3889	0.3626	**0.3042**

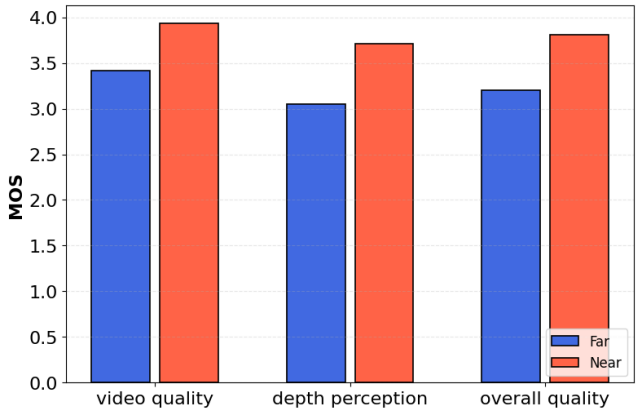


Figure 6: The average subjective results for five sequences captured from 'Near' and 'Far' distances.

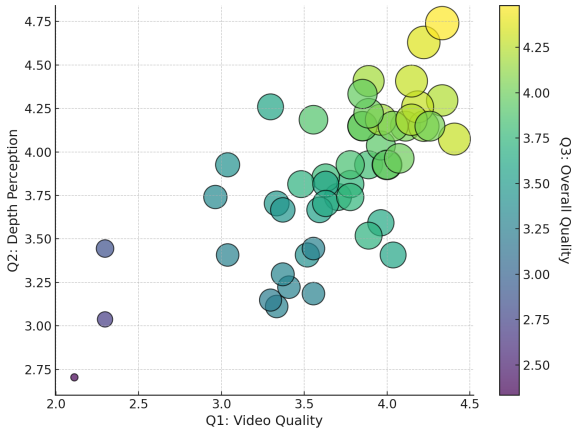


Figure 7: Scatter plot illustrating the relationship between video quality (Q1) and depth perception (Q2), with circle size and color representing overall quality (Q3). Larger and lighter circles indicate higher perceived overall quality.

case. For example, with Polynomial Regression (Degree 3), AVQT reaches an MAE of 0.2438, notably lower than VMAF (0.2884), MS-SSIM (0.3493), SSIM (0.3361), and PSNR (0.3591).

5.5 Relationship between video quality, depth perception, and overall quality

To explore how perceived video quality (Q1) and depth perception (Q2) jointly influence overall quality (Q3), we visualize their relationship in a scatter plot. As shown in Figure 7, Q1 is plotted on the x-axis, Q2 on the y-axis, and Q3 is encoded through both the color and size of each circle. The plot reveals a clear trend: higher values of Q1 and Q2 generally correspond to larger, lighter-colored circles, indicating higher overall quality (Q3). This visual pattern confirms that both video quality and depth perception contribute meaningfully to the overall viewing experience.

The relationship can be approximated using a linear model, expressed as:

$$Q3 = 0.4830 \cdot Q1 + 0.5234 \cdot Q2 - 0.0663. \quad (1)$$

This model explains 97.4% of the variance in Q3 ($R^2 = 0.974$), confirming that a linear combination of Q1 and Q2 effectively captures users' perception of overall quality. It also suggests that depth perception (Q2) has a slightly stronger influence on Q3 than video quality (Q1), though both factors are essential.

To further assess how well Q3 can be predicted using a combination of objective quality (AVQT) and depth-related cues (disparity), we trained several regression models and computed their PLCC and R^2. These results, summarized in Table 7, show that all models capture a meaningful relationship between these predictors and Q3. The Random Forest model yields the highest correlation (PLCC = 0.9822) and the greatest explained variance ($R^2 = 0.9532$), indicating its superior ability to model the nonlinear interactions between AVQT and disparity.

To evaluate how these models generalize to unseen content, we applied a leave-one-video-out cross-validation strategy and measured prediction accuracy using MAE. Table 8 presents these results. Again, Random Forest delivers the best performance with the lowest MAE of 0.2463, followed closely by KNN and SVR. These results confirm that leveraging both disparity and AVQT leads to robust predictions of overall quality across diverse video content.

Table 7: PLCC and R^2 for predicting Q3 using AVQT and disparity.

Model	PLCC	R^2
Linear Regression	0.6961	0.4845
Polynomial Regression (Degree 2)	0.7549	0.5699
Polynomial Regression (Degree 3)	0.8106	0.6571
Support Vector Regressor (SVR)	0.7792	0.5658
K-Nearest Neighbors (KNN)	0.8120	0.6479
Random Forest	0.9822	0.9532

Table 8: Leave-One-Video-Out MAE for predicting Q3 using AVQT and disparity.

Model	MAE
Linear Regression	0.2897
Polynomial Regression (Degree 2)	0.2732
Polynomial Regression (Degree 3)	0.3298
Support Vector Regressor (SVR)	0.2762
K-Nearest Neighbors (KNN)	0.2639
Random Forest	0.2463

6 Conclusion

This study provides a comprehensive evaluation of spatial video quality on HMDs, addressing critical gaps in immersive multimedia research. Our subjective assessment using the *SV-QoE* dataset –developed as part of this work– reveals that spatial videos consistently outperform their 2D counterparts in depth perception and overall quality, especially at higher bitrates The analysis further highlights the influence of viewing distance, with videos captured at closer ranges offering a notably enhanced QoE. Through a rigorous statistical analysis, we establish that both perceived video quality (Q1) and depth perception (Q2) are significant predictors of overall quality (Q3), with Q2 exerting a slightly stronger influence. These findings emphasize the importance of considering depth perception alongside traditional quality metrics when evaluating spatial content.

7 Acknowledgment

The financial support of the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, and the Christian Doppler Research Association is gratefully acknowledged. Christian Doppler Laboratory ATHENA: https://athena.itec.aau.at/.

References

[1] J. Van Der Hooft, H. Amirpour, M. T. Vega, Y. Sanchez, R. Schatz, T. Schierl, and C. Timmerer, "A Tutorial on Immersive Video Delivery: From Omnidirectional Video to Holography," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1336–1375, 2023.

[2] W. Zhou, H. Amirpour, C. Timmerer, G. Zhai, P. L. Callet, and A. C. Bovik, "Perceptual Visual Quality Assessment: Principles, Methods, and Future Directions," 2025. Version Number: 1.

[3] J. J. Cummings, M. Tsay-Vogel, T. J. Cahill, and L. Zhang, "Effects of immersive storytelling on affective, cognitive, and associative empathy: The mediating role of presence," *New Media & Society*, vol. 24, pp. 2003–2026, Sept. 2022.

[4] J. Psotka, "Immersive training systems: Virtual reality and education and training," *Instructional Science*, vol. 23, pp. 405–431, Nov. 1995.

[5] S. C. Mallam, S. Nazir, and S. K. Renganayagalu, "Rethinking Maritime Education, Training, and Operations in the Digital Era: Applications for Emerging Immersive Technologies," *Journal of Marine Science and Engineering*, vol. 7, p. 428, Dec. 2019. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[6] S. C. Bronack, "The Role of Immersive Media in Online Education," *The Journal of Continuing Higher Education*, vol. 59, pp. 113–117, May 2011. Publisher: Routledge _eprint: https://doi.org/10.1080/07377363.2011.583186.

[7] N. Dodgson, "Autostereoscopic 3D displays," *Computer*, vol. 38, pp. 31–36, Aug. 2005.

[8] N. S. Holliman, N. A. Dodgson, G. E. Favalora, and L. Pockett, "Three-Dimensional Displays: A Review and Applications Analysis," *IEEE Transactions on Broadcasting*, vol. 57, pp. 362–371, June 2011.

[9] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, June 2012. ISSN: 1063-6919.

[10] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3061–3070, June 2015. ISSN: 1063-6919.

[11] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation," pp. 4040–4048, IEEE Computer Society, June 2016. ISSN: 1063-6919.

[12] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A Naturalistic Open Source Movie for Optical Flow Evaluation," in *Computer Vision – ECCV 2012* (A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), (Berlin, Heidelberg), pp. 611–625, Springer, 2012.

[13] E. Cheng, P. Burton, J. Burton, A. Joseski, and I. Burnett, "RMIT3DV: Pre-announcement of a creative commons uncompressed HD 3D video database," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, pp. 212–217, July 2012.

[14] L. Goldmann, F. D. Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," in *Three-Dimensional Image Processing (3DIP) and Applications*, vol. 7526, pp. 242–252, SPIE, Feb. 2010.

[15] D. Corrigan, F. Pitié, V. Morris, A. Rankin, M. Linnane, G. Kearney, M. Gorzel, M. O'Dea, C. Lee, and A. Kokaram, "A Video Database for the Development of Stereo-3D Post-Production Algorithms," in *2010 Conference on Visual Media Production*, pp. 64–73, Nov. 2010.

[16] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutiérrez, and N. García, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, pp. 109–114, July 2012.

[17] H. Imani, M. B. Islam, and L.-K. Wong, "A New Dataset and Transformer for Stereoscopic Video Super-Resolution," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 705–714, June 2022. ISSN: 2160-7516.

[18] M. H. Izadimehr, M. Ghanbari, G. Chen, W. Zhou, X. Hao, M. Dasari, C. Timmerer, and H. Amirpour, "SVD: Spatial Video Dataset," 2025. Version Number: 1.

[19] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the multiview and 3d extensions of high efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 35–49, 2016.

[20] G. Tech, Y. Chen, K. Muller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, pp. 35–49, Jan. 2016.

[21] Wei Zhou, Ning Liao, Zhibo Chen, and Weiping Li, "3D-HEVC visual quality assessment: Database and bitstream model," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, (Lisbon, Portugal), pp. 1–6, IEEE, June 2016.

[22] R. Hussain, M. Chessa, and F. Solari, "Improving Depth Perception in Immersive Media Devices by Addressing Vergence-Accommodation Conflict," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, pp. 6334–6346, Sept. 2024. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[23] J. Zhang, Q. Jia, Y. Liu, W. Zhang, W. Wei, and X. Tian, "SpatialMe: Stereo Video Conversion Using Depth-Warping and Blend-Inpainting," 2024. Version Number: 1.

[24] T. Guionnet, K. Jerbi, T. Burnichon, and M. Raulet, "MV-HEVC: How to optimize compression of immersive 3D content," in *Proceedings of the 3rd Mile-High Video Conference on zzz*, (Denver CO USA), pp. 87–87, ACM, Feb. 2024.

[25] W. Liu, J. Li, and Y. B. Cho, "A novel architecture for parallel multi-view HEVC decoder on mobile device," *EURASIP Journal on Image and Video Processing*, vol. 2017, p. 24, Dec. 2017.

[26] C. Timmerer, H. Amirpour, F. Tashtarian, S. Afzal, A. Rizk, M. Zink, and H. Hellwagner, "HTTP Adaptive Streaming: A Review on Current Advances and Future Challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications*, p. 3736306, May 2025.

[27] G. Chen, S. Wang, J. Chakareski, D. Koutsonikolas, and M. Dasari, "Spatial video streaming on apple vision pro xr headset," 2025.

[28] L. Goldmann, F. De Simone, and T. Ebrahimi, "Impact of acquisition distortion on the quality of stereoscopic images," in *Proceedings of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2010.

[29] Z. Chen, W. Zhou, and W. Li, "Blind Stereoscopic Video Quality Assessment: From Depth Perception to Overall Experience," *IEEE Transactions on Image Processing*, vol. 27, pp. 721–734, Feb. 2018.

[30] W. Wan, D. Huang, B. Shang, S. Wei, H. R. Wu, J. Wu, and G. Shi, "Depth Perception Assessment of 3D Videos Based on Stereoscopic and Spatial Orientation Structural Features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, pp. 4588–4602, Sept. 2023.

[31] Y.-H. Lin and J.-L. Wu, "Quality Assessment of Stereoscopic 3D Image Compression by Binocular Integration Behaviors," *IEEE Transactions on Image Processing*, vol. 23, pp. 1527–1542, Apr. 2014.

[32] F. Qi, T. Jiang, S. Ma, and D. Zhao, "Quality of experience assessment for stereoscopic images," in *2012 IEEE International Symposium on Circuits and Systems*, (Seoul, Korea (South)), pp. 1712–1715, IEEE, May 2012.

[33] F. Battisti, M. Carli, P. Le Callet, and P. Paudyal, "Toward the assessment of quality of experience for asymmetric encoding in immersive media," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 392–406, 2018.

[34] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, pp. 1143–1155, Oct. 2013.

[35] J. Wang, A. Rehman, K. Zeng, S. Wang, and Z. Wang, "Quality Prediction of Asymmetrically Distorted Stereoscopic 3D Images," *IEEE Transactions on Image Processing*, vol. 24, pp. 3400–3414, Nov. 2015.

[36] J. Xu, Z. Luo, W. Zhou, W. Zhang, and Z. Chen, "Quality Assessment of Stereoscopic 360-degree Images from Multi-viewports," in *2019 Picture Coding Symposium (PCS)*, (Ningbo, China), pp. 1–5, IEEE, Nov. 2019.

[37] Z. Chen, J. Xu, C. Lin, and W. Zhou, "Stereoscopic Omnidirectional Image Quality Assessment Based on Predictive Coding Theory," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, pp. 103–117, Jan. 2020.

[38] C. Galkandage, J. Calic, S. Dogan, and J.-Y. Guillemaut, "Full-Reference Stereoscopic Video Quality Assessment Using a Motion Sensitive HVS Model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 452–466, Feb. 2021.

[39] "VMAF: The Journey Continues. by Zhi Li, Christos Bampis | by Netflix Technology Blog | Netflix TechBlog."

[40] International Telecommunication Union, "Recommendation ITU-T P.910: Subjective video quality assessment methods for multimedia applications," Tech. Rep. P.910, International Telecommunication Union, 2008.

[41] A. S. Yaro, F. Maly, P. Prazak, and K. Malý, "Outlier Detection Performance of a Modified Z-Score Method in Time-Series RSS Observation With Hybrid Scale Estimators," *IEEE Access*, vol. 12, pp. 12785–12796, 2024.

[42] F. Z. H., "Calculation and Interpretation of Analysis of Variance and Covariance: By George W. Snedecor. Ames, Iowa: Collegiate Press, Inc. 105 pages. 1934. $1," *Agronomy Journal*, vol. 26, pp. 255–256, Mar. 1934.

[43] H. Amirpour, R. Schatz, C. Timmerer, and M. Ghanbari, "On the Impact of Viewing Distance on Perceived Video Quality," in *VCIP 2021*, pp. 1–5, IEEE, Dec. 2021.

[44] H. Hirschmuller, "Stereo Processing by Semiglobal Matching and Mutual Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 328–341, Feb. 2008. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

[45] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, Apr. 2004. Conference Name: IEEE Transactions on Image Processing.

[46] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale Structural Similarity for Image Quality Assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, (Pacific Grove, CA, USA), pp. 1398–1402, IEEE, 2003.

Enhancing Access to 360-Degree Video Collections: A Novel Interface for Immersive Exploration in Virtual Reality

Bas van Eck
Department of Information and Computing Sciences,
Utrecht University
Utrecht, Netherlands
bas.vaneck@ziggo.nl

Wolfgang Hürst
Department of Information and Computing Sciences,
Utrecht University
Utrecht, Netherlands
huerst@uu.nl

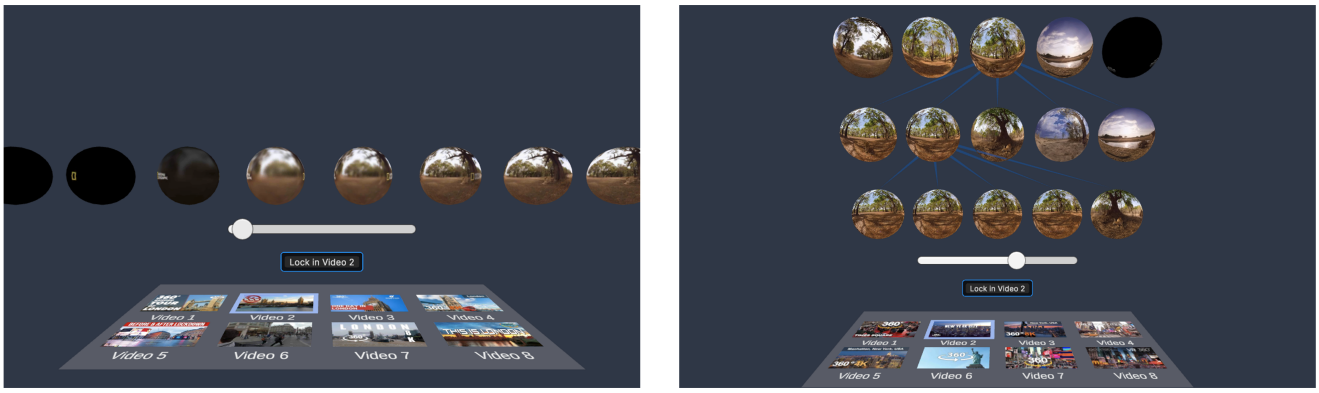


Figure 1: Interfaces to access and explore 360-degree videos in VR: Filmstrip design (left) and Hierarchical interface (right).

Abstract

Traditional methods for browsing 360° video libraries in virtual reality (VR) often use a two-dimensional grid with thumbnails, similar to how standard videos are displayed. However, this fails to take advantage of the immersive capabilities of VR. This study introduces two alternative interfaces that make better use of VR's three-dimensional space: a conventional filmtrip interface and a new hierarchical interface to explore videos. In a comparative study with 32 participants aged 18 to 30, users completed a timed search task to find a specific video using both interfaces in counterbalanced order. Performance was measured by search time and accuracy, while user experience was evaluated with the UEQ-S and post-task feedback. Results showed that participants found videos significantly faster with the hierarchical interface, though accuracy was similar across both interfaces. Participants also rated the hierarchical interface higher in terms of hedonic and overall quality, suggesting it offers a more enjoyable and efficient browsing experience for 360° video libraries in VR.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; *Empirical studies in interaction design.*

IXR '25, Dublin, Ireland

ACM ISBN 979-8-4007-2051-2/2025/10
https://doi.org/10.1145/3746269.3760420

Keywords

360° video, video browsing in VR, 360° video access.

ACM Reference Format:
Bas van Eck and Wolfgang Hürst. 2025. Enhancing Access to 360-Degree Video Collections: A Novel Interface for Immersive Exploration in Virtual Reality. In *Proceedings of the 3rd International Workshop on Interactive eXtended Reality (IXR '25), October 27–28, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746269.3760420

1 Introduction

As virtual reality (VR) technology continues to advance and head-mounted displays (HMDs) become increasingly affordable for the average consumer, a growing number of users are gaining access to the immersive experience of 360° video content. Typically, these videos are accessed through applications such as YouTube VR or Oculus Video, where the user interfaces present videos as thumbnails arranged in a two-dimensional (2D) grid pattern, like traditional video content is displayed on regular, non-immersive 2D screens. In VR environments, these interfaces are projected into three-dimensional (3D) space, yet a significant portion of the available screen space often remains underutilized. Moreover, 2D thumbnails are not well-suited for conveying the full context of 360° videos, often introducing image distortion, hindering the browsing experience.

In this user study, we address these challenges by designing an interface that combines the traditional 2D thumbnail grid with enhanced video visualizations, allowing for hierarchical exploration of video content. This approach seeks to balance browsing efficiency and user experience, optimizing the navigation of 360° video

libraries in VR. Our research specifically focuses on improving video preview visualizations, with key metrics including browsing efficiency and overall user experience.

2 Related Work

Research focusing specifically on 360° video collection browsing is sparse, and studies rarely combine the three key components – 360° videos, video browsing, and virtual reality (VR) – into a single investigation. This section reviews enhanced video browsing methods for traditional videos, 360° videos, and video browsing in VR.

2.1 Enhanced Video Browsing

Enhanced video browsers build on traditional browsing methods by incorporating additional features to improve navigation within a video. Many of these features focus on controllable playback, time-jumping, and timeline visualization. For example, [14] introduced navigation summaries that visualize extracted metadata in a temporal manner, using a timeline. Users can select summaries that range from low-level details (e.g., motion intensity, dominant colors) to high-level details (e.g., positions of commercial breaks, emotions of actors) based on their preferences. Other studies focused on improving interaction with video timelines. [9] and [10] addressed the issue of granularity in traditional timeline interfaces, which are challenging to navigate when zoomed out and limiting when zoomed in. Their solution was the ZoomSlider interface, which was later refined by [7] as the AV-ZoomSlider. This interface allows users to scroll horizontally through a timeline and adjust its scale vertically, with fast-forward browsing activated when reaching the window border.

While these advanced interfaces facilitate navigation within videos, browsing a video collection requires a different approach, one that provides a quick, clear overview of video content. This is often achieved through visualizations such as filmstrips that display consecutive frames from a video. Early studies, such as [4] and [3], employed this approach. [5] integrated filmstrip-based browsing into their interface, where filmstrips appear during video scrolling—a method now used by platforms like YouTube. More recent studies, such as [1], have explored hierarchical approaches, with [8] introducing a novel grid of temporally ordered thumbnails that users can zoom into for more detailed storyboards of selected video segments.

2.2 Video Browsing in Virtual Reality

A valuable source for novel video browsing interfaces in VR is the Video Browser Showdown (VBS) event [13], where teams compete to solve content-based video retrieval tasks. While the event focuses on video retrieval, it also highlights innovative video library visualization methods. Though most interfaces are designed for 2D screens, some teams have developed VR-based solutions, including Vitrivr VR [19] and EOLAS [20].

Vitrivr VR extends the Vitrivr content retrieval system to VR, allowing users to explore video libraries through a cylindrical, rotatable display. The media items' positions are determined by similarity scores, and users can inspect videos in detail via a media segment inspector, which displays frames in a depth-wise manner, resembling the filmstrip method.

EOLAS represents video keyframes embedded into a latent vector space and allows users to explore these keyframes based on spoken queries. Keyframes are clustered in a 3D space based on their encoded features, which users can navigate to explore the video content.

Both interfaces introduce innovative methods for exploring video libraries in VR, but they rely on conventional 2D thumbnails, making them less effective for browsing 360° videos.

2.3 Browsing 360° Videos

360° videos, also known as omnidirectional videos, allow users to dynamically change their viewing direction during playback. These videos can be viewed on various devices, including computers, smartphones, tablets, and VR headsets. Despite advancements in 3D space browsing, 360° videos are often represented by conventional 2D thumbnails. Typically, 360° frames are projected onto a 2D plane using equirectangular projection. [21] explored alternative shapes for representing 360° thumbnails, concluding that conventional equirectangular thumbnails were most effective for recognizing high-level concepts quickly. However, for identifying finer details, projecting frames onto a sphere yielded better accuracy without increasing task completion time. Furthermore, using spheres for thumbnail representation resulted in an improved user experience, particularly when focusing on low-level details.

Building on this, Knoop [11] conducted a comparative study on three VR video browser designs: an abstract movie-inspired design, a video-store-based design, and a record-store-based design. The designs were evaluated based on performance and user experience. The record-store design performed the worst across all metrics, while the video-store design scored higher on pragmatic quality measures. The abstract design, by contrast, scored better on hedonic quality measures, though perceived performance did not differ significantly between the designs. While this study demonstrated promising results for user experience, browsing speed remained inferior to traditional methods, as users were required to physically navigate the virtual space.

Building on these prior works, our research proposes a novel approach for browsing 360° video collections in VR, combining the traditional grid-based layout of thumbnails with enhanced video previews. This approach aims to improve both user experience and browsing efficiency, addressing the unique challenges posed by omnidirectional video content in a VR environment.

3 Approach

In a typical scenario of browsing 360° video collections, users filter content using search queries and select videos from the resulting list. These query results are conventionally displayed in a grid format, with videos represented by equirectangular or custom thumbnails, much like traditional video content. While 360° videos offer a more immersive experience through virtual reality (VR) environments than 2D screens, the potential advantages of VR—such as the ability to display content in 3D space—remain largely untapped in the context of 360° video browsing.

This research seeks to improve the 360° video browsing experience in VR by designing and evaluating two interfaces that aim to enhance both user efficiency and overall experience. Since this area remains underexplored, especially for 360° videos in VR, we draw from prior work in enhanced video browsers for traditional videos. Various studies have introduced methods for displaying additional details, such as neighboring frames, color-coded metadata, or complete overviews of video content upon selection. Previous research on traditional video browsing has tackled this challenge by providing visualizations of videos when selected. Following a similar approach, this study develops two VR interfaces for browsing 360° videos, leveraging the extended screen space offered by VR to visualize video content more effectively.

The first interface (Fig. 1, left) translates the state-of-the-art approach from traditional video browsing into the VR and 360° video context. It uses the familiar filmstrip method, where frames are presented in temporal order. Conventionally, filmstrips are displayed as a long horizontal row, but recent studies such as [19] and [11] have experimented with alternative designs, such as a box of frames that users can riffle through.

The second interface (Fig. 1, right) takes a different approach, allowing users to explore videos hierarchically. This method is inspired by studies such as [8], which visualize video content at varying levels of granularity. Detailed descriptions of both interfaces are provided in Section 4. A comparative analysis will be conducted to evaluate the effectiveness of the two interfaces. The following research questions guide this evaluation:

- RQ1: Which interface provides better user performance in terms of efficiency and accuracy?
- RQ2: How does the user experience of the filmstrip interface compare to that of the hierarchical interface?

To address these research questions, an experiment will be conducted in which participants evaluate both interfaces by performing a search task. This task requires users to locate a video containing a scene they have previously explored. Data collected during the experiment will be analyzed to compare the performance and user experience across the two interfaces. Performance will be measured by the time taken to inspect each video and the accuracy of video selection, while user experience will be assessed using a standardized questionnaire and qualitative feedback. Further details regarding the experimental design and methodology are outlined in Section 5.

4 Interface designs

Users may have different goals when browsing a collection of 360° videos. We have identified three primary use cases to guide the design of our interfaces:

(1) **High-level browsing**: When a user is searching for a video on a general topic (e.g., a video of someone skiing).
(2) **Detailed browsing**: When a user is looking for a video with specific details (e.g., a video of someone skiing in Austria using red skis).
(3) **Exploring video collections**: When a user is casually exploring a collection of videos without a specific target.

The traditional interface performs well for high-level browsing (use case 1), as users can often grasp the general concept of a video from a traditional thumbnail [21]. However, for detailed browsing and exploration (use cases 2 and 3), users may require a more detailed overview of the video content. In traditional video contexts, detailed visualizations typically involve displaying multiple frames either horizontally or in a grid format. Moreover, for 360° videos, [21] demonstrated that low-level details are more easily recognized when frames are represented as spheres rather than in equirectangular projection.

To improve the browsing experience across all use cases, we propose a novel interface that integrates additional details into the traditional grid of thumbnails, while performing a comparative analysis against an interface that represents the current state of the art.

4.1 Interfaces

We developed two interfaces for browsing 360° videos in VR: the filmstrip interface, which adapts traditional video browsing methods to the context of 360° videos and VR, and the hierarchical interface, which introduces a novel method for hierarchically exploring video content. While both interfaces initially present videos in a similar manner, they diverge in their approach to visualizing detailed video content.

In VR, the additional screen space allows for both high-level and low-level browsing without replacing the original grid of thumbnails. Instead, the grid is relocated to accommodate the detailed visualizations. Frames are represented as spheres, rather than traditional equirectangular thumbnails, as this has been shown to be more effective for recognizing details [21]. Drawing inspiration from holographic tables, the traditional grid of thumbnails is laid out as if on a virtual table. When a video is selected, a detailed visualization is projected above the table, differing between the two interfaces (Fig. 1).

In addition to the grid of thumbnails and the detailed visualizations, two more interface elements are included: a timeline and a "lock-in video" button. Users can interact with the timeline to navigate the video in a manner similar to traditional video players. The "lock-in video" button is specifically used to confirm video selection during the search task used in our user study, as illustrated in Section 5.

4.1.1 Filmstrip interface. In traditional video browsing, detailed visualizations often utilize a filmstrip that presents consecutive frames in a sequential manner. This approach provides a continuous temporal representation of the video content. The filmstrip interface retains this sequential presentation but projects the frames onto spheres instead of 2D planes. The result is a row of spheres in 3D space, with each sphere representing a frame from the video. Users can interact with the timeline or manually grab the spheres to scroll through the video. Additionally, the spheres can be rotated using the controller joystick, allowing users to view the 360° content from different angles. A screenshot of this interface is presented in Figure 1 (left).

4.1.2 Hierarchical interface. Several studies, including [8] and [7], have explored methods to give users control over the granularity of video browsing. In the context of 360° videos in VR, we address

this challenge by visualizing video content across different layers of granularity. The hierarchical interface presents a detailed representation of the video through a tree-like structure of spheres.

Initially, users are presented with one layer consisting of five spheres, each corresponding to a different segment of the video. For instance, if the video duration is five minutes, each sphere represents one minute. All spheres play their respective segments simultaneously, providing an overview of the entire video.

However, in longer or more dynamic videos, a single layer with five segments may not provide sufficient detail, as each segment could contain multiple scenes. To address this, users can select any sphere to generate a second layer (Layer 2), which further divides the selected one-minute clip into five equally long segments. This process can be repeated to create a third layer, allowing for increasingly granular exploration of the video. The result is a hierarchical tree of video segments, with each layer providing progressively finer details. A screenshot of the hierarchical interface with all layers activated is shown in Figure 1 (right).

5 Methodology

To address the research questions introduced in Section 3, we conducted an experiment where participants evaluated both interfaces (see Section 4). Participants completed a search task for each interface, followed by a questionnaire. During the search task, we measured both time and accuracy to evaluate performance, while the questionnaire gathered insights on the user experience. To minimize individual variability, we employed a within-subject design, counterbalancing the order in which participants used the interfaces.

5.1 Data collection

5.1.1 User performance. To answer the first research question, application data was collected to assess user performance during the search task. The two primary metrics were inspection time per video and the correctness of the selected video.

- **Inspection time per video** measures the time participants spent examining the detailed visualization of a video. By focusing on inspection time instead of total completion time, we account for the possibility that participants might inspect a different number of videos across interfaces. This metric more precisely measures the time participants require to determine whether they have identified the correct video.
- **Correctness of the selected video** evaluates whether the participant correctly identified the target video.

Additionally, the application recorded participants' actions in a log file with timestamps, providing further insights where necessary. The logged actions include opening a video, rotating spheres, adjusting the timeline, and selecting spheres. Selection of spheres is logged only for the hierarchical interface, as this functionality is not available in the filmstrip interface.

5.1.2 User experience. User experience (UX) is a subjective measure and can be challenging to quantify. UX is typically assessed using standardized questionnaires such as the System Usability Scale (SUS) [2], the Usability Metric for User Experience (UMUX) [6], or the User Experience Questionnaire (UEQ) [12]. [18] argue that for ranking two products by their UX quality, the choice of questionnaire has little impact, as all measure similar concepts. However, since UMUX and SUS primarily focus on pragmatic quality, while the UEQ assesses both pragmatic and hedonic qualities, we opted to use the short version of the UEQ (UEQ-S) [17], which balances these aspects.

The UEQ-S contains eight 7-point Likert scale questions that yield three scales: Overall quality, Hedonic quality, and Pragmatic quality. This shorter version allows us to gather meaningful feedback while keeping the questionnaire concise and manageable for participants.

5.2 Questionnaire

The experiment's questionnaire consisted of four parts:

(1) **Pre-experiment phase**: This phase included an information sheet, a consent form, and demographic questions. Demographics covered age, gender, occupation or degree, experience with VR, and experience with 360° videos. This part of the questionnaire was completed before the experiment began.
(2) **During evaluation**: Participants answered questions about each interface individually. First, they completed the UEQ-S for each interface. Then, they provided ratings and open-ended responses about the strengths and weaknesses of each interface.
(3) **Post-evaluation**: After completing the VR section of the experiment, participants answered final questions aimed at comparing the two interfaces.

5.3 Experiment design

5.3.1 Procedure. Participants began the experiment by completing the information sheet and consent form, followed by demographic questions. They were then assigned to one of two groups. Group A evaluated the Filmstrip interface first, followed by the Hierarchical interface, while Group B evaluated the Hierarchical interface first, followed by the Filmstrip interface.

For each interface, participants performed a search task. After reading the instructions, they completed a tutorial task to familiarize themselves with the interface. The tutorial task was similar to the actual search task but with only four videos, compared to the eight videos used in the actual task. The tutorial encouraged participants to explore the interface and formulate a browsing strategy, while the actual task required them to find the correct video as quickly as possible.

The search task consisted of two phases:

(1) **Frame exploration**: Participants explored a still frame from one of the videos in VR and were instructed to remember as much detail as possible.
(2) **Video identification**: Participants used the assigned interface to find the video containing the previously explored frame. Videos were presented in a random order for each participant, but the order remained consistent between interfaces. After inspecting the videos and using the detailed visualization, participants "locked in" their choice once they believed they had identified the correct video.

Participants completed the UEQ-S immediately after using each interface. This break between VR sessions also helped reduce the risk of motion sickness. After completing the final search task, participants filled out a questionnaire comparing the two interfaces.

5.3.2 Environment. The experiments were conducted in person, as the VR environment required a controlled, distraction-free setting. Participants used a non-rotating chair to ensure consistency in their seated position throughout the VR session.

5.4 Participants

We recruited 32 participants (56% male, 44% female) for this study through convenience sampling. The participants, aged between 18 and 30, were likely early adopters of new technologies. Specifically, 22 participants were aged 20-24 (68.75%), while 10 participants were aged 25-30 (31.25%).

Regarding VR experience, 10 participants (31.25%) had never used a VR headset before the study, while the majority (59.38%) had used a VR headset a few times. One participant occasionally used a VR headset, and two participants used a VR headset regularly. Nine participants with VR experience (28.13%) had previously used a VR headset to watch 360° videos. Additionally, 18 participants (56.25%) had watched 360° videos on a phone or tablet, and 14 participants (43.75%) had watched 360° videos on a computer or laptop. Seven participants (21.88%) had no prior experience watching 360° videos on any platform.

5.5 Materials and Technical details

The application, containing both interfaces and the VR viewer, was implemented in Unity version 2022.3.10f1. We used Unity's VR template, which includes the XR toolkit, to provide a framework for VR experiences. The application was developed for the Oculus Quest 3 headset.

5.5.1 Videos. The videos used in the study were selected to prevent participants from identifying the correct frame solely based on thumbnails, which would negate the purpose of detailed visualizations. Videos of specific cities were chosen for their comparable content and general thumbnails. Tutorials featured videos of Paris and Prague, while the actual search task used videos of New York and London.

The selection process involved filtering 360° videos on YouTube based on specific criteria: the videos had to be 15 minutes or less in duration, have a resolution of 4K, and include varied scenes throughout. A duration of under 15 minutes ensured that the videos were similar in length, and the 4K resolution maintained clarity when participants viewed frames in the VR viewer. The variety of scenes was crucial, as participants needed to identify a specific video based on a frame.

Ultimately, 24 videos were selected: 16 for the search tasks and 8 for the tutorials, all downloaded from YouTube. To conserve memory, the videos were downscaled to a resolution of 480x240, which was low enough to save memory while retaining sufficient detail. Videos not originally in equirectangular format were converted accordingly. FFMPEG (https://ffmpeg.org/) was used to perform video conversions and extract the still frames for the VR viewer.

	N	Mean	Std. Deviation	Minimum	Maximum
Filmstrip time per video	32	00:35.87	00:23.77	00:06.30	01:39.55
Hierarchical time per video	32	00:26.17	00:16.29	00:06.48	01:08.66

Table 1: Descriptive statistics for time spent per video in a timed search task, reported in mm:ss.ms

	Pragmatic Quality	Hedonic Quality	Overall Quality
Filmstrip interface	0.77	0.65	0.71
Hierarchical interface	1.11	1.92	1.52

Table 2: Average pragmatic, hedonic, and overall quality scores obtained from UEQ-S questionnaire analysis

6 Results

6.1 Performance

A timed search task was employed to assess participants' performance using each interface. Since the sequence in which participants inspected videos could not be controlled, we focused on time spent per video rather than the total time spent within the interface. This metric captures the time required for participants to determine whether a video contains the target frame. The results are shown in Table 1.

During the experiment, four participants encountered an issue with the hierarchical interface that prevented the first few videos' inspection times from being recorded. However, since our primary metric is the inspection time per video, their partial data remains usable for evaluating their performance.

Due to the relatively small sample size, we conducted a Shapiro-Wilk test, which revealed a significant deviation from normality for both the filmstrip interface ($W = .90$, $p = .008$) and the hierarchical interface ($W = .88$, $p = .002$). Consequently, the non-parametric Wilcoxon Signed-Ranks test was chosen for further analysis. The test indicated that participants spent significantly more time per video using the filmstrip interface than the hierarchical interface ($Z = -2.693$, $p = .007$).

In addition to time metrics, we assessed the correctness of participants' video selections during the search task. For the hierarchical interface, one participant selected the wrong video (96.88% accuracy), while two participants selected the wrong video using the filmstrip interface (93.75% accuracy). A McNemar's test determined that this difference in accuracy between the two interfaces was not statistically significant ($p = 1.00$).

6.2 User Experience

6.2.1 UEQ-S. The results of the UEQ-S questionnaire were analyzed using the tool provided by the authors [15]. This tool transforms the questionnaire responses from a 7-point Likert scale to values ranging from -3 (most negative) to +3 (most positive). The mean values for Pragmatic, Hedonic, and Overall quality are presented in Table 2.

We compared these results against a benchmark dataset derived from over 21,000 participants across 468 studies, which uses the full UEQ [16]. This benchmark categorizes results from "bad" (worst

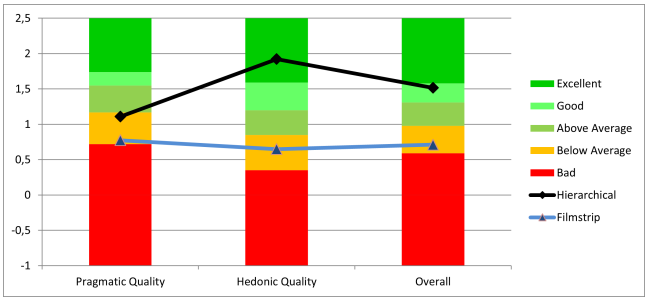


Figure 2: UEQ-S results for the hierarchical interface and the filmstrip interface compared to a benchmark dataset

25% of studies) to "excellent" (best 10%). Figure 2 presents the comparison of the UEQ-S results with this benchmark. The hierarchical interface scored just below average for Pragmatic quality but ranked "excellent" for Hedonic quality, resulting in a "good" overall quality rating. In contrast, the filmstrip interface scored below average across all scales.

The analysis tool also performed t-tests to assess the significance of differences between the two interfaces. No significant difference was found for Pragmatic quality ($p = .27$), but significant differences were observed for Hedonic quality ($p = .0001$) and Overall quality ($p = .002$).

In addition to the UEQ-S, participants rated each interface on a scale of 1 to 10. The filmstrip interface received an average rating of 6.31, while the hierarchical interface received an average of 7.63. A Wilcoxon Signed-Ranks test confirmed a significant difference between these average ratings ($p = .002$).

6.2.2 Qualitative data. Participants were also asked to indicate their preferred interface and discuss the advantages and disadvantages of each. Of the participants, 24 (75%) favored the hierarchical interface, while 8 (25%) preferred the filmstrip interface. The most commonly cited reasons for preferring the filmstrip interface were its familiarity, speed, and control over scrolling. Conversely, participants who favored the hierarchical interface praised its speed, simplicity, and the better overview it provided of video content. Table 3 summarizes the most notable advantages and disadvantages of each interface.

6.3 Analysis of search patterns

An analysis of the search patterns employed during the task provides additional insights into the strengths and weaknesses of each interface, highlighting potential areas for improvement.

For the filmstrip interface, search patterns were relatively straightforward, with limited variation. After opening the video visualization, participants either scrolled or skipped through the video to make their decision. In contrast, the hierarchical interface offered more flexibility in search strategies due to the multiple layers of granularity available.

To categorize search strategies, we examined the proportion of time participants spent at each layer of the hierarchical interface. Participants were classified into different strategy categories based on the relative time spent at each layer. For example, a participant who spent most of their time in Layer 1, a moderate amount of time in Layer 2, and the least time in Layer 3 was categorized as using the L1>L2>L3 strategy. We compared each participant's strategy with their time spent per video (see Section 6.1). Table 4 presents the distribution of search strategies and the corresponding average time spent per video.

Our analysis revealed that most participants spent the majority of their time in Layers 1 and 2. Additionally, participants who spent significant time in Layer 3 had noticeably higher average inspection times per video.

7 Discussion

This study set out to address two primary research questions related to improving the browsing experience for 360° video libraries in VR. The first research question focused on user performance, while the second explored user experience.

7.1 RQ1: Performance

To answer Research Question 1: *"Which interface results in better user performance, considering efficiency and accuracy?"*, we analyzed the results from the search task (Section 6.1).

In terms of **accuracy**, no statistically significant difference was found between the two interfaces. Nearly all participants correctly identified the target video, with only one incorrect selection in the hierarchical interface and two in the filmstrip interface.

Regarding **efficiency**, participants were statistically significantly faster at identifying the correct video using the hierarchical interface. This suggests that, despite its complexity, the hierarchical interface allows users to interpret video content more quickly than the filmstrip interface. A key factor contributing to this efficiency is the hierarchical interface's ability to display five video segments simultaneously, offering a more comprehensive overview.

The filmstrip interface, while relatively straightforward, had a common issue: participants with longer search times often skipped over the correct frame while scrolling, forcing them to revisit videos. On the other hand, Section 6.3 highlights that search strategies in the hierarchical interface influenced time efficiency. Most participants focused on Layers 1 and 2, which proved to be more effective than relying heavily on Layer 3. For this study, the three-layer structure likely provided more detail than necessary, as the videos were relatively short (under 15 minutes). Participants who focused on Layer 3, where each sphere represents shorter segments, were less effective because those spheres often showed the same scene.

Observationally, three strategies appeared to be most effective, all of which prioritized Layers 1 and 2. The first strategy exclusively used Layer 1, with participants scrolling through the timeline. In the second, participants systematically selected spheres in Layer 1 to utilize Layer 2. The third strategy involved a top-down approach, where participants only selected spheres in Layer 1 that visually resembled the target frame's characteristics (e.g., skipping aerial shots when searching for a city scene).

7.2 RQ2: User Experience

To answer Research Question 2: *"How do the user experience of the filmstrip interface and the hierarchical interface compare?"*, we examined the questionnaire results (Section 6.2).

Advantages and disadvantages mentioned by participants	Filmstrip Mentions	Hierarchical Mentions
This interface provides a **clear** overview of video content	3	9
This interface provides an **quick** overview of video content	7	8
This interface is **simple/easy** to navigate	7	5
Scrolling the timeline was very sensitive, causing spheres to move very fast	8	0
The interface was complex	0	5

Table 3: Notable advantages and disadvantages for both interfaces

Strategy	# Participants	Average time spent per video	Std. Deviation
L1>L2>L3	11	00:24.32	00:14.56
L2>L1>L3	9	00:23.23	00:14.44
L2>L3>L1	5	00:35.59	00:21.90
L3>L2>L1	3	00:43.46	00:12.27
L1>L2=L3	4	00:13.15	00:05.21

Table 4: Distribution of participants across different search strategies based on time spent in each layer of granularity

The UEQ-S results showed that the hierarchical interface scored higher on pragmatic quality, although this difference was not statistically significant. However, the hierarchical interface did show significantly higher scores for hedonic quality and overall quality. Additionally, participants rated the hierarchical interface higher on a scale of 1 to 10, with 75% of participants preferring it over the filmstrip interface.

Interestingly, participants cited similar reasons for preferring each interface—speed and simplicity. Despite the perceived complexity of the hierarchical interface, many users found it effective for quickly gaining an overview of video content. The hierarchical interface's advantage of providing a clear overview was mentioned three times more frequently than for the filmstrip interface, suggesting that the visual layout contributed to both its pragmatic and hedonic appeal.

Some participants noted the hierarchical interface's complexity as a drawback. However, despite these remarks, most of these participants performed better with the hierarchical interface, indicating that the complexity was manageable. On the other hand, a recurring complaint with the filmstrip interface was the sensitivity of the timeline, which caused spheres to move too quickly, especially for longer videos. This may have contributed to its lower hedonic quality rating.

7.3 Limitations & Future Work

This study has several limitations that should be considered when interpreting the results. First, there were some technical challenges during the experiment. Due to Unity's handling of video players, the Oculus Quest was unable to play 15 videos simultaneously, likely due to memory constraints. As a result, only the five most recent video segments in the hierarchical interface were played, while previous layers were paused. Additionally, three participants experienced crashes during the experiment, but they were able to continue after a quick restart, so the impact on the results was minimal.

Another limitation involved the streaming of participants' views to a phone for observational purposes. In some cases, the stream experienced lag, but this was not the primary source of data collection and did not significantly impact the results.

A final limitation was the relatively short time participants had to learn the interfaces. While this likely had minimal impact on the filmstrip interface due to its straightforward nature, the hierarchical interface may have benefited from additional learning time, as strategy played a significant role in performance (see section 6.3).

7.3.1 Filmstrip interface. Several participants noted difficulties with the sensitivity of the filmstrip interface's timeline. Potential improvements could include reducing the timeline's sensitivity by freezing the spheres in place and only changing their content. Alternatively, introducing adjustable granularity, similar to the hierarchical interface, could help mitigate this issue. Another possible improvement could involve increasing the visibility of spheres by arranging them in a cylindrical or S-shape, which could aid participants in identifying scene changes more effectively.

7.3.2 Hierarchical interface. The hierarchical interface proved effective, but alternative designs could further improve the experience. For instance, section 6.3 showed that prioritizing Layer 1 was a highly effective strategy. Therefore, creating an interface with just one layer, possibly with a larger number of spheres, could enhance performance. Another interesting variation to explore could involve incorporating a filmstrip as a second layer upon selecting a sphere. This would reduce the sensitivity issues associated with the filmstrip interface, as each segment in the second layer would cover a shorter duration.

8 Conclusion

This study set out to enhance the browsing experience of 360° video collections in virtual reality (VR) by designing and implementing two distinct interfaces: the filmstrip interface and the hierarchical interface. Both interfaces were evaluated in terms of efficiency and usability. To assess user experience, we employed the UEQ-S questionnaire alongside qualitative questions, while performance was measured through a timed search task in which participants identified a video containing a previously viewed frame.

The results revealed that participants completed the search task significantly faster using the hierarchical interface compared to

the filmstrip interface. In terms of accuracy, there were no significant differences between the interfaces, as nearly all participants correctly identified the target video. The UEQ-S results showed no significant difference in pragmatic quality between the two interfaces; however, the hierarchical interface demonstrated statistically significantly better hedonic and overall quality. Furthermore, qualitative feedback indicated a clear preference for the hierarchical interface among participants.

These findings suggest that the hierarchical interface offers superior performance and user experience, making it a more efficient and enjoyable method for browsing 360° video libraries in VR environments. While both interfaces exhibit strengths, the hierarchical interface stands out in terms of user satisfaction and browsing efficiency. Continued development and refinement of this approach will be crucial in advancing VR technology for 360° video browsing, offering users a more intuitive and effective experience.

Future work should focus on further refining the hierarchical interface and addressing the sensitivity issues observed in the filmstrip interface, potentially enhancing the usability and effectiveness of both methods.

References

[1] Abir Al Hajri, Matthew Fong, Gregor Miller, and Sidney Fels. 2014. Fast Forward with your VCR: Visualizing Single-Video Viewing Statistics for Navigation and Sharing. *Proceedings - Graphics Interface*.

[2] John Brooke. 1996. *SUS – a quick and dirty usability scale*. 189–194.

[3] Linjun Chang, Yichen Yang, and Xian-Sheng Hua. 2008. Smart video player. In *2008 IEEE International Conference on Multimedia and Expo*. 1605–1606. doi:10.1109/ICME.2008.4607760

[4] M.G. Christel, A.G. Hauptmann, A.S. Warmack, and S.A. Crosby. 1999. Adjustable filmstrips and skims as abstractions for a digital video library. In *Proceedings IEEE Forum on Research and Technology Advances in Digital Libraries*. 98–104. doi:10.1109/ADL.1999.777702

[5] Steven M. Drucker, Asta Glatzer, Steven De Mar, and Curtis Wong. 2002. SmartSkip: Consumer Level Browsing and Skipping of Digital Video Content. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI '02)*. Association for Computing Machinery, New York, NY, USA, 219–226. doi:10.1145/503376.503416

[6] Kraig Finstad. 2010. The Usability Metric for User Experience. *Interacting with Computers* 22, 5 (05 2010), 323–327. arXiv:https://academic.oup.com/iwc/article-pdf/22/5/323/1992916/iwc22-0323.pdf doi:10.1016/j.intcom.2010.04.004

[7] Wolfgang Hürst. 2006. Interactive Audio-Visual Video Browsing. In *Proceedings of the 14th ACM International Conference on Multimedia* (Santa Barbara, CA, USA) *(MM '06)*. Association for Computing Machinery, New York, NY, USA, 675–678. doi:10.1145/1180639.1180781

[8] Wolfgang Hürst and Dimitrios Darzentas. 2012. HiStory - A hierarchical storyboard interface design for video browsing on mobile devices. *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, MUM 2012*, 17:1–17:4. doi:10.1145/2406367.2406389

[9] W. Hürst, G. Götz, and T. Lauer. 2004. New methods for visual information seeking through video browsing. In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.* 450–455. doi:10.1109/IV.2004.1320183

[10] W. Hürst and P. Jarvers. 2005. Interactive, dynamic video browsing with the zoomslider interface. In *2005 IEEE International Conference on Multimedia and Expo*. 4 pp.–. doi:10.1109/ICME.2005.1521484

[11] Mike Knoop. 2023. Exploring Interactive Experiences to Browse Extensive 360° Video Libraries in Virtual Reality. Available at https://studenttheses.uu.nl/handle/20.500.12932/45252.

[12] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. *USAB 2008* 5298, 63–76. doi:10.1007/978-3-540-89350-9_6

[13] Klaus Schoeffmann. 2019. Video Browser Showdown 2012-2019: A Review. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. 1–4. doi:10.1109/CBMI.2019.8877397

[14] Klaus Schoeffmann and Laszlo Boeszoermenyi. 2009. Video Browsing Using Interactive Navigation Summaries. In *2009 Seventh International Workshop on Content-Based Multimedia Indexing*. 243–248. doi:10.1109/CBMI.2009.40

[15] Martin Schrepp. [n. d.]. Data analysis tool (ueq-s). https://www.ueq-online.org/Material/Short_UEQ_Data_Analysis_Tool.xlsx Accessed on 12-06-2024.

[16] Martin Schrepp. 2023. *User Experience Questionnaire Handbook*. https://www.ueq-online.org/Material/Handbook.pdf Version 11, Accessed on 12-06-2024.

[17] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4 (01 2017), 103. doi:10.9781/ijimai.2017.09.001

[18] Martin Schrepp, Jessica Kollmorgen, and Jörg Thomaschewski. 2023. A Comparison of SUS, UMUX-LITE, and UEQ-S. *J. User Exper.* 18, 2 (jun 2023), 86–104.

[19] Florian Spiess, Ralph Gasser, Silvan Heller, Mahnaz Parian-Scherb, Luca Rossetto, Loris Sauter, and Heiko Schuldt. 2022. Multi-Modal Video Retrieval Innbsp;Virtual Reality Withnbsp;vitrivr-VR. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part II* (Phu Quoc, Vietnam). Springer-Verlag, Berlin, Heidelberg, 499–504. doi:10.1007/978-3-030-98355-0_45

[20] Ly-Duyen Tran, Manh-Duy Nguyen, Thao-Nhu Nguyen, Graham Healy, Annalina Caputo, Binh T. Nguyen, and Cathal Gurrin. 2021. A VR Interface for Browsing Visual Spaces at VBS2021. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II* (Prague, Czech Republic). Springer-Verlag, Berlin, Heidelberg, 490–495. doi:10.1007/978-3-030-67835-7_50

[21] Alissa Vermast and Wolfgang Hürst. 2023. Introducing 3D thumbnails to access 360-degree videos in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023). doi:10.1109/tvcg.2023.3247462

Social Density and its Impact on Behaviour in Virtual Environments

Julie Williamson
julie.williamson@glasgow.ac.uk
University of Glasgow
Glasgow, UK

Silvia Rossi
s.rossi@cwi.nl
Centrum Wiskunde & Informatica (CWI)
Amsterdam, Netherlands

Ross Johnstone
ross.johnstone@glasgow.ac.uk
University of Glasgow
Glasgow, UK

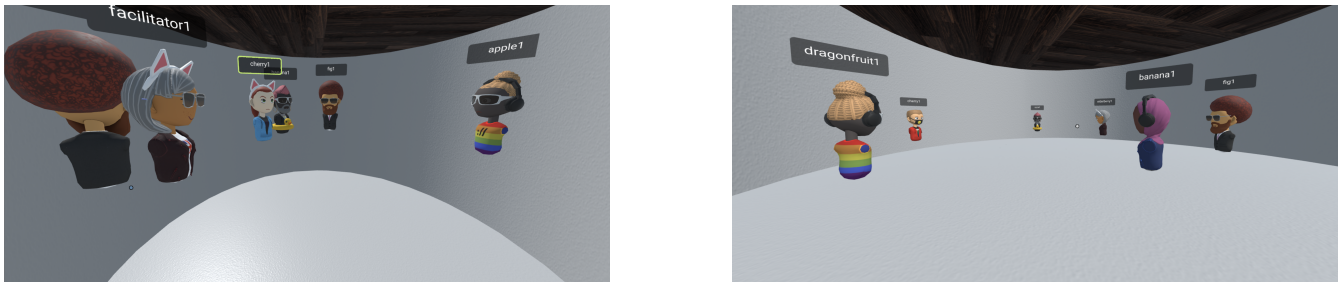
Irene Viola
irene.viola@cwi.nl
Centrum Wiskunde & Informatica (CWI)
Amsterdam, Netherlands

John Williamson
johnh.williamson@glasgow.ac.uk
University of Glasgow
Glasgow, UK

Thomas Röggla
t.roggla@cwi.nl
Centrum Wiskunde & Informatica (CWI)
Amsterdam, Netherlands

David A. Shamma
aymans@acm.org
Centrum Wiskunde & Informatica (CWI)
Amsterdam, Netherlands

Pablo Cesar
garcia@cwi.nl
Centrum Wiskunde & Informatica (CWI) and Delft University of Technology
Amsterdam, Delft, Netherlands



(a) The small room with high social density

(b) The large room with low social density

Figure 1: A large and a small virtual room created spaces with low or high social density.

Abstract

Virtual environments make it possible to connect and collaborate in social immersive realities, but there are still open questions about the influence of their design on the user experience. We conducted a study with 48 participants divided into groups of 6, completing conversational tasks in an instrumented virtual environment. Using a mix-methods approach, combining qualitative and quantitative research methods (interviews, questionnaires, conversation and movement analysis), we compared between two virtual environment designs. We found that the social density (or effective capacity) of the designed virtual environment influenced the quality of interaction between the participants.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Virtual reality**; **Empirical studies in collaborative and social computing**.

Keywords

Social VR, social density, social signal processing

IXR '25, Dublin, Ireland

ACM ISBN 979-8-4007-2051-2/2025/10
https://doi.org/10.1145/3746269.3760421

ACM Reference Format:
Julie Williamson, Silvia Rossi, Ross Johnstone, Irene Viola, John Williamson, Thomas Röggla, David A. Shamma, and Pablo Cesar. 2025. Social Density and its Impact on Behaviour in Virtual Environments. In *Proceedings of the 3rd International Workshop on Interactive eXtended Reality (IXR '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746269.3760421

1 Introduction

We are moving towards a future where immersive environments and virtual realities (VR) mediate our social and collaborative interactions. How our interpersonal interactions play out in VR is significantly influenced by how these experiences are designed, but we have limited models for understanding social phenomenon when mediated by immersive environments. *Digital proxemics* is an emerging model for understanding how we use space in virtual environments, focusing on social signals of body position relative to others in virtual space [4, 27, 28]. *Interpersonal social signals* like proximity, mutual gaze, and conversational fluency, which represent the signals created when two or more people interact, provide rich insights into what happens between individuals. Defining these interpersonal social signals and developing more sophisticated models of interpersonal interaction in immersive environments is needed to inform the design of more satisfying and successful communications and collaborations in social VR.

This paper explores the *quality* of interactions with different social densities [21, 26], that is the ratio of the number of people to the size of a space. Social density influences how crowded, cozy, or empty a space feels. Our hypothesis is that manipulating social density will influence how interaction unfolds. Our analysis focuses on the interpersonal social signals that we give off when interacting in immersive environments, focusing on how social density creates observable changes in these interpersonal social signals. We completed a lab-based user study (N=48) where eight groups of six participants completed conversational tasks while wearing head-mounted displays (HMDs) in an instrumented virtual environment. Results show different interaction and conversation patterns, depending on the social density of the virtual environment.

This paper makes the following contributions:

(1) A comparison of two social density conditions (low and high), resulting from a between-subjects evaluation (N=48) of conversational tasks in a virtual environment.
(2) An open source dataset capturing rich social signals including proximity, gesture, gaze, and conversation under low and high *social densities*.

2 Related Work

2.1 Proxemics

Research on proxemics in physical spaces has explored how people use space [12, 16], intended usage of spaces [22], and space design [10, 13]. Hall's foundational work on proxemics [12] provides the baseline for understanding how people use physical space. Factors like body visibility, voice loudness, body heat, odour, and personal preferences all contribute to how people use space, which is dynamic and contextually influenced by culture. Hall categorises four proxemic zones: intimate, personal, social, and public. The *intimate* zone (< 0.46 meters) entails more physical contact, close field of view, and potential awareness of breath or body heat. The *personal* zone (0.46–1.2 meters) allows for physical contact, but not consistently, and facial expressions and visual attention are more noticeable. Transitioning to the *social* zone (1.2–3.6 meters), one would expect less physical contact, if any, and the ability to observe the other person's entire body. Lastly, the *public* zone (> 3.6 meters) makes facial expressions and voice less impactful but enhances peripheral awareness of the immediate environment.

Beyond interpersonal distance, proxemics research also incorporates visual attention and body orientation when people interact in groups. Kendon introduces the concept of *F-formations*, emphasising the relative orientation of a group over time as a significant aspect of proximity [16]. F-formations emerge during social interactions among people in close proximity, who share a space with exclusive, direct, and equal access.

Applying physical proxemics to behaviour in virtual environments raises challenges given the absence of typical social and environmental constraints in the virtual world. Incorporating olfaction and haptics, which play vital roles in physical proxemics, presents difficulties within virtual settings. Additionally, the auditory experience through Head-Mounted Display (HMD) headphones diverges significantly from real-world auditory perceptions [24]. However, Virtual Environments (VEs) offer the intriguing possibility of intentionally manipulating, distorting or enhancing proxemic cues to shape the social dynamics of the virtual space.

2.2 Social Density

Social density refers to the number of individuals present in a given physical [3] or virtual space [21]. Social density is a crucial factor that influences social interactions, personal comfort, and behaviour patterns within space. Whyte's research asks what makes public spaces feel spacious or crowded, exploring the availability and layout of seating, street positioning, and exposure to elements [26]. His work integrates physical space design with human proxemics, examining effective capacity and perceived crowding. Effective capacity, which considers factors like proximity to others, comfort, and amenities, can significantly differ from physical capacity when spaces are well-designed. For instance, a vibrant city square with strategically placed benches and green spaces may create a sense of low social density despite a large number of people present.

Social density is particularly interesting when designing interactions in a VE. In the absence of physical constraints, the perception of social density can vary depending on factors such as the size and design of the VE, the number and size of avatars or virtual entities present, and the availability of interactive elements. Guitton et al. present an ethnographic study of merfolk in Second Life, highlighting the challenge of achieving social density in expansive virtual ocean environments [11]. Moore et al. also find that expansive environments in massively multiplayer online games make it challenging to achieve social density and create spaces where it is easy for people to interact [21]. There is limited quantitative work modelling how social density affects interaction in a VE, representing a current gap in how these spaces should be designed for communication and collaboration.

2.3 Focused/Unfocused Interactions

The focus (or lack of focus) during interaction also impacts how people interact and use space. Goffman categorises face-to-face interactions as either *unfocused* or *focused* [9].

Unfocused interactions are characterised by a lack of specific visual attention or shared interest among people. In such scenarios, individuals may coexist in the same space without actively

interacting with one another or directing their attention towards a common point of interest. In contrast, focused interactions involve individuals coming together with a shared point of visual attention or a common goal. In these situations, people gather to collaborate, discuss, or participate in activities that require collective focus. Focused interactions often have more defined social regulations, explicit or implicit boundaries for participation, and specific expectations or affordances for involvement. The atmosphere during focused interactions tends to be more structured and purpose-driven. [9].

Focused interactions can be further refined when considering if the rules for social engagement are "tight" or "loose" [7–9]. Social interactions with "tight" rules have highly structured and mutually understood rules for engagement. For example, a presentation to an audience is built on shared and mutually maintained rules around where the focus is and acceptable modes of participation and interruption. In contrast, social interactions with "loose" rules are more unstructured and there can be much flexibility in how the rules, if any, are enforced.

3 Understanding Behaviour in Social VR with High and Low Social Density

The key motivation of our evaluation was to understand how behaviour would be influenced when the virtual environment put constraints on social density within the virtual space. We recruited participants for conversational tasks in an instrumented virtual environment, including a *loosely* and *tightly* structured conversation that would elicit different kinds of behaviours. The virtual environments were sized to create *low* or *high* social density, which we hypothesised would influence how people use space and their comfort in maintaining personal space and turn-taking during conversation.

3.1 Experimental Protocol

For our lab study, we conducted a between-subjects evaluation that compared user experiences within two virtual spaces with different sizes, resulting in either low or high social density as described in Section 3.1.1. We recruited a total of 48 participants, organised into 8 groups of 6 participants each. These groups engaged in conversational tasks as described in Section 3.1.2, with the task order being counterbalanced to mitigate order-related effects. A facilitator was present in the virtual environment to guide participants through the experiment, as described in Section 3.1.3.

To begin, participants were gathered in the same physical space for on-boarding, reading information sheets, and consenting to participate. Participants were assigned unique identifiers: Apple, Banana, Cherry, Dragonfruit, Elderberry, and Fig. The assignment of these identifiers was randomised, each associated with a different physical room to which the facilitators directed each participant. These identifiers were beneficial for both facilitators and participants, aiding in study organisation and participant recognition throughout the evaluation.

As the first task, the Simulator Sickness Questionnaire (SSQ) [17] was administered to assess any effect of the experimental conditions on cybersickness. After completing the SSQ questionnaire, each participant was taken to an individual room, as described in Section 3.2.1. The facilitator gave a short tutorial to each participant, explaining how to fit and control the device, as well as how to enter and navigate the virtual environment; either with joystick movement or teleportation. Participants were given a period of time to accustom themselves to the experience before start the first of two conversation tasks, as described in Section 3.1.2. The order of the conversation tasks was counterbalanced across participant groups. A short questionnaire for Quality of Interaction (QoI), composed of 8 items, was administered after each conversation task, allowing for finer granularity in understanding the impact of each task on the QoI. At the end of the second conversation task, three longer questionnaires were given to the participants to fill out, namely the immersion questionnaire [14], the iGroup Presence Questionnaire (iPQ) [23], and the Social Presence questionnaire [18]. The post-study SSQ was also filled by participants.

To conclude, participants gathered in the same physical room and engaged in a focus group session, during which they shared their perspectives with each other. These encompassed their impressions of the virtual environment, comparative analyses with video calls and face-to-face interactions, their comfort levels, positional dynamics within the virtual space, and their overall experiences in the conversation tasks. Participants were also encouraged to provide feedback on their preferred and least preferred activities, alongside suggestions for potential enhancements. This concluded with thanking the participants for their time and allowing them any questions about the study.

3.1.1 Virtual Environment. Both rooms, as shown in Figure 1, were designed in a circular layout, with one room having a radius of 26.5 meters with low social density, and the other being smaller with a radius of 2.5 meters with high social density. Whyte describes the "effective capacity" of a space in terms of sitting space, where effective capacity is roughly .9 meters of sitting space per person. Within Hall's proxemic zones, this is within the *personal zone* [12]. Moore describes the low social density in one space as a small space where one must "rub elbows with other participants" where it is impossible to watch from a distance [21].

For the 6 participants plus a facilitator, our low social density room allows for 315 meters per person and the high social density room allows for just 2.8 meters per person. The high social density room was designed such that participants were crowded within the *personal and social* proxemic zones, where the low social density provided a large space where participants could move freely. The rooms were intentionally designed to be sparse, with a textured floor plane, smooth walls, and a textured ceiling plane to reduce the impact of environmental elements to influence behaviour. The rooms were circular to provide cornerless spaces after pilots indicated participants often positioned themselves within corners, which introduced arbitrary hot spots of activity.

3.1.2 Conversation Tasks. Goffman describes "tightness" and "looseness" in social settings as the degree to which individuals are expected to be present and attending to interaction, and how strictly social rules must be adhered to [9]. We designed two conversation tasks structured with either "tight" or "loose" social rules, with the hypothesis that different social rules would result in different experiences in a social VR.

Tight Story Task: "Yes, and" Scene in VR. For this task, six participants and one facilitator come together within the VR space to collaboratively construct a story consisting of short statements with tight social rules. The story begins when a participant makes the first statement, for example "I left my house today but I forgot my bicycle helmet." Each following statement must begin with "yes, and" and must be a continuance of the previous statement. Each statement should be short, but must continue the storyline established by the first statement. The facilitator monitored the passage of time and conclude the activity after five minutes. If the story naturally concludes, the facilitator instructed the participants to begin a new story until the overall task lasted five minutes. The story task is structured based on tight social rules that dictate the general content of conversation and how conversation proceeds, where turn-taking will be stilted and formal.

Loose Truths Task: Two Truths and a Lie in VR For this task, six participants and one facilitator come together within the VR space to tell two truths and a lie about themselves with loose social rules. Each participant takes a turn introducing themselves by sharing two truths and one lie about themselves. The group then tries to guess which statement is the lie. Once the lie is discovered, another participant introduces themselves. The facilitator monitored the time and helped move the task along as needed until the overall task lasted five minutes. The truths task is structured based on loose social rules, where individuals may speak informally, interruptions and questions are expected, and turn-taking will be regular and informal.

3.1.3 Facilitator. A facilitator was present in the virtual environment to guide the participants through the conversation tasks. The facilitator began by ensuring all participants were present and not experiencing any technical issues before beginning the conversation task. The facilitator would introduce the task and give participants an opportunity to ask any questions. Once the conversation task began, the facilitator monitored the task timing and facilitated the session as needed to keep the task running for five minutes. The facilitator would conclude each task and provide instructions for the participants to exit VR.

3.2 Experimental Setting and Hardware

3.2.1 Physical Space & Infrastructure. The participants initially gathered in a large room where the experiment began and ended. For the tasks in VR, participants were each taken to a private room that was physically and aurally separated from the other participants. In each private room, there was a desk, chair, and Quest 2 running a customised instance of Mozilla Hubs[1]. Each room was equipped with a 6GHz WiFi router to ensure maximum bandwidth availability and stability. The experiment was run using a modified version of the social VR platform Mozilla Hubs, deployed on a private cloud instance.

The participants used Meta Quest 2 headsets and the facilitator used a Meta Quest Pro headset. All VR users were seated for the duration of the experiment. Participants and facilitator used handheld controllers for hand movement capture during the experiment.

[1]https://hubs.mozilla.com/

3.2.2 Data Logging. During the experiment, participants' body, hand, and voice data was logged continuously. This was achieved through the use of a custom JavaScript-based subsystem, integrated into a modified version of the Mozilla Hubs client, described in [27]. The data logger collected metrics from the browser environment and directly from the Document Object Model (DOM) which specifies the virtual environment. We augmented the system to collect additional metrics for the purpose of this experiment. Technical details about the data collection is included in the appendices.

Key data streams used for this analysis include the avatar position, direction and rotation based on the headset sensors, position, direction and rotation of each hand based on the hand-held controller sensors, and the amplitude of the participants voice based on the directional microphone within the headset. All this data was updated and saved at a rate, constrained by the frame rate of the user's headset. In order to minimise the number of requests sent to the data collection server, the tick data is buffered and a POST request containing all the data as a JSON-formatted payload is only sent every 4000 ticks. The data collection server is responsible for validating and storing all session data and runs on a separate cloud machine.

3.3 Participants

A total of 48 individuals participated in the experiment. Participants were divided into groups of six for a total of eight sessions, each lasting approximately 1 hour. To ensure a diverse and balanced set of people (i.e., age, gender and nationality), participants were primarily recruited through a professional recruitment company but also from the organisation where the study took place. Since the experiment was conducted in English, all the participants were comfortable in reading, speaking, and writing in English. Additionally, none of them had uncorrected visual, hearing, or motor impairments, and they were instructed to attend the experiment with their glasses or lenses. All participants were aged between 22 and 68 years old with an average age equal to 37.48 and a standard deviation of 12.89. Among all of them, 50% identified as female, 46% as male and the remaining 4% as non-binary (female = 24, male = 22, non-binary = 2). In terms of ethnicity, most of the participants self-identified as being from Europe (40%), followed by Asia (19%), mixed origins (19%) and Africa (2%). 12 participants never had experience with VR before the experiment while 8 of them were experts (e.g., VR designers, or researchers); 22 were novices (e.g., have played less than 5 times) and only 6 were knowledgeable about this technology (e.g., have played 5+ times, or own VR headset). One participant self-identified as a person with a disability (i.e., dyslexia) and one preferred not to say. The majority of participants did not know each other before their involvement in the study. This experimental study was reviewed and approved by an institutional ethics committee.

4 Results

4.1 Quantitative Results

The quantitative results are based on the interaction logs generated by the instrumented virtual environment where we conducted this research. For our 48 participants (excluding the facilitator), our dataset includes 2.42 million logs of 131 minutes of interaction. This data was resampled to 30Hz for the analysis, resulting in 1.48

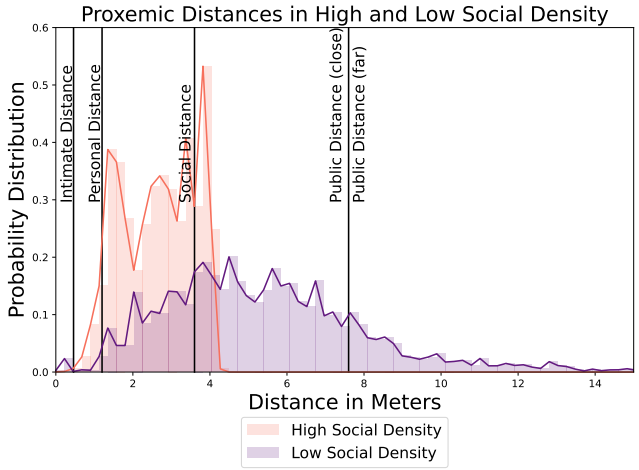
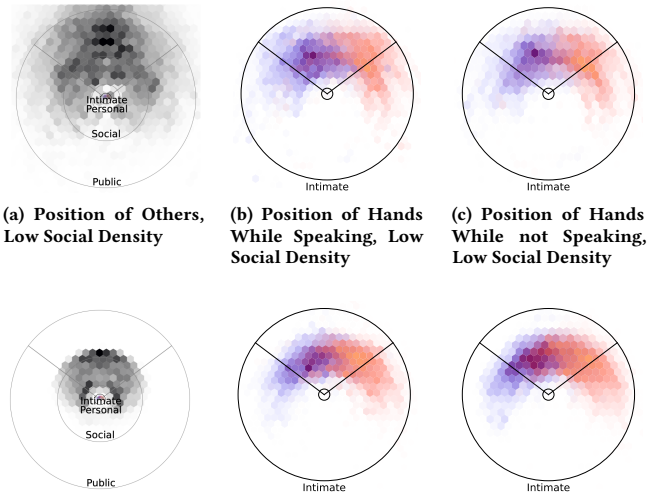


Figure 2: Proxemic analysis compares how people used personal space in the low and high social density conditions.



(a) Position of Others, Low Social Density **(b) Position of Hands While Speaking, Low Social Density** **(c) Position of Hands While not Speaking, Low Social Density**

(d) Position of Others, High Social Density **(e) Position of Hands While Speaking, High Social Density** **(f) Position of Hands While not Speaking, High Social Density**

Figure 3: Hexagonal bin plots with proxemic zones overlayed visualise the position of others relative to the field of view for all participants (facilitator excluded). The detailed insets from the Intimate zone visualise the position of the hands while speaking or not speaking. Others are visualised in grey, the left hand is visualised in blue, and the right hand is visualised in orange.

million logs of 131 minutes of interaction. All of the data, scripts, and visualisation techniques used in this analysis will be made publicly available as part of this paper.

Due to an issue with data logging on the facilitator's headset, the facilitator data is unavailable for six of the eight groups. However, quantitative analysis of the social metrics we have developed excludes the facilitator by design for some metrics and these metrics are not affected. Where the facilitator's data is required, we have performed the analysis on the subset of data where the facilitator is included. The subset of the data where the facilitator is present includes over 900,000 interaction logs, which was resampled to over 400,000 logs at 30 frames per second. Given the scale and detail of the data made possible by the instrumented virtual environment, the absence of the facilitator data from some groups has a limited impact on the quantitative analysis.

4.1.1 Social Density: Proximity and Personal Space. Previous works on digital proxemics have been concerned with how people use virtual space and how personal boundaries are maintained in virtual environments. In our design, we specifically put participants in virtual spaces with high and low social densities. In the high social density space, participants could not avoid collisions in the *personal zone* and would have limited space to maintain positions outside the personal zone throughout the task. In the low social density space, participants had ample room to move and could stand at their desired distances within room to move and adjust as needed. Figure 2 shows the proxemic positions for the high and low social density spaces, demonstrating how individuals used personal space in these settings. In line with previous work, we saw that participants avoided collisions in the intimate zone and maintained standing distances in the social and public zones where possible. When constrained in the high social density space, participants maintain social distances, with some collisions in the personal zone. When given space to move, more relaxed proxemics were maintained, with participants spreading across the social and public proxemic zones.

We extended proxemic analysis from previous studies of digital proxemics [4, 12, 15, 28] by adding analysis of how hands aid in defining personal space and how social density might change how people maintain and communicate personal space.

Figure 3 shows a relationship between social density and how people use virtual space and their hands. The position of others relative to the field of view is shown in greyscale, with the proxemic zones shown as concentric rings. In a zoomed view of the intimate zone, the hands are shown in purple (left hand) and orange (right hand) while speaking and not speaking. The facilitator data is excluded from this analysis.

Figure 3 (a) and (d) show results in line with previous research on digital proxemics. In the low social density room, participants tended to stand within the social proxemic zone, with some participants in the personal space, and limited collisions in the intimate space. Participants also tended to keep others within their field of view, and rarely stand with someone directly behind them.

Within the high social density room, there were tight restrictions on personal space and people behaved differently. There are virtually no collisions in the intimate zone, and participants were arranged around the border of the personal zone. Participants has a strong tendency to keep others within their field of view, but the close standing distances forced by the space would make it impossible to see other participants in full. Hotspots are visible directly in front of each participant and to each side, showing the typical configuration in the high social density room to form a circle against the walls.

The hand movements are also affected by social density. Even though in physical reality participants were alone in their room and did not risk hitting anyone or invading their physical space with their hands, participants in the high social density room gestured closer to their own body, as shown in Figure 3 (e) and (f).

Within the low social density room, there was a greater likelihood to ignore intimate collisions, but there was also different use of the hands, in particular while speaking. Gesturing extends further out from the body, sometimes extending beyond the intimate zone. While people often use their hands in interesting ways while speaking [16], looking at position alone does not give a clear picture of how this translates to virtual environments. Hands are also used to claim personal space and add expression while speaking.

4.1.2 Conversation Task: Turning Taking, Silence, and Crosstalk. Participants completed a tightly and loosely structured conversation task, which we designed to elicit stable and unstable conversations for our analysis. For the two groups where the facilitator data is present, we calculated conversational metrics as speaker turn duration, silence duration, and cross-talk duration[1], as shown in Figure 5.

We see differences between the tightly structured story task and the loosely structured truths task. In particular, the challenges for turn-taking can be seen in the increased duration of cross talk during these tasks. The story task was especially challenging in the high social density room, where the conversation involved especially short utterances, long durations of silence, and longer periods of cross talk.

Another metric commonly used in interaction analysis is speaker participation and conversation balance [19]. We visualise this as speaker rank, which describes how dominant speakers were during these conversational tasks. The speaker rank calculates the proportion of time each speaker spent speaking and ranks these in order from most to least for each group, as shown in Figure 4. The facilitator's speech is excluded from this metric.

Figure 4 shows a pattern across all groups, where one speaker often held the floor for roughly half of the time, followed by the next dominant speaker up to a third of the time, with decreasing participation until the least active speaker who made few if any utterances. It was surprising that even in the Truths task, where turning-taking was partially moderated by the facilitator, we see the same pattern of dominance and fall-off in speaker participation. Previous work has looked at influencing speaker participation through visual interventions [19], and similarly found these patterns of behaviour difficult to influence and control.

4.2 Questionnaire Results

4.2.1 Quality of Interaction. Questionnaires were administered after each task, allowing to compare the effect of the task on the QoI, as well as increasing the statistical power when comparing the two room configurations. A total of 96 scores were collected per item. Prior to running any comparison, the normality of the scores was tested using the Kolmogorov-Smirnov test. The null hypothesis was rejected ($p < 0.001$); thus, non-parametric tests were used on the scores. Regarding the impact of the task on the QoI, a Wilcoxon test for paired samples revealed no significant statistical difference between the tasks for any of the items on the

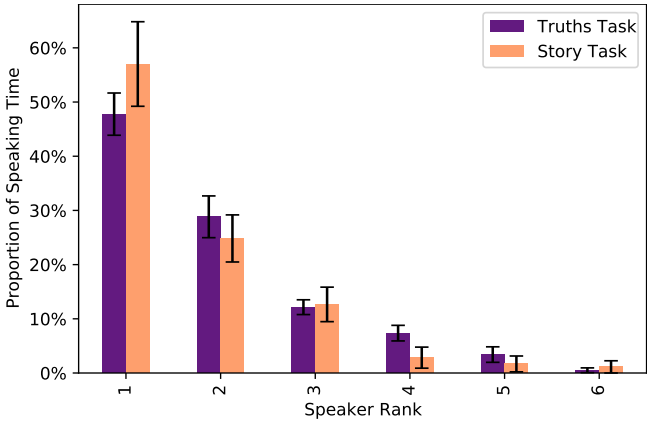


Figure 4: The speaker ranks takes the proportion of time spent speaking for participants in each group and ranks them from most to least time spent speaking. Speaker dominance is very consistent between groups.

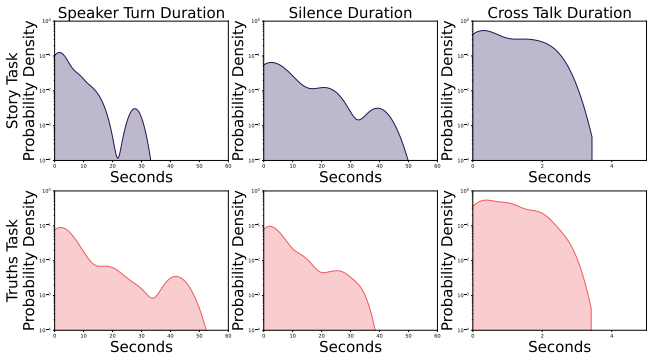


Figure 5: Conversational metrics for the four groups which included the facilitator. Speaker turn duration describes the time a speaker can hold the floor with trailing silence excluded. Silence duration describes the duration of silence when no speaker is active. Cross-talk duration describes the duration when two or more people are speaking simultaneously. Y-axis shows probability density using log scale.

questionnaire ($p > 0.05$ for all items). We thus focused our analysis on comparing the effect of the different room configurations on the items. Figure 6 depicts the violin plot of the score distribution for every item in the QoI questionnaire. Please note that the scales were inverted, when necessary, to ease readability; in all cases, higher scores signify positive responses. The plot shows the distribution of the scores, along with the median value, separately for the low social density configuration (left, darker colour) and the high social density configuration (right, lighter colour). In general, we can observe high values for almost all the items in the questionnaire, with median values between 5 and 7; the notable exception is for the item "Involvement", for which results are fairly spread, with a median value of 4. This might be due to the phrasing of the question ("How much did your experience in the virtual environment seem consistent with your real-world ones?"), which

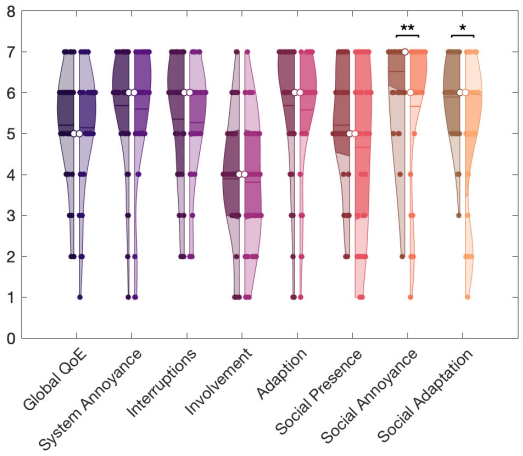


Figure 6: Violin plot showing the results of the QoI questionnaires (higher values are better). Left and right sides refer to the low density and high density room configurations, respectively. Statistical significant difference, when present, is indicated with asterisks on top (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$).

led to confusion among the participants. A Mann-Whitney test for unpaired samples performed on the items reveals no statistical difference for 6 out of the 8 items in the questionnaire (Global QoE: $Z = -0.061, p = 0.952, r = 0.006$; System Annoyance: $Z = 0.023, p = 0.982, r = 0.002$; Interruptions: $Z = -0.525, p = 0.599, r = 0.054$; Involvement: $Z = 0.336, p = 0.737, r = 0.034$; Adaption: $Z = 0.890, p = 0.374, r = 0.091$; Social Presence: $Z = -1.131, p = 0.258, r = 0.115$), whereas the difference was found to be significant for Social Annoyance ($Z = -2.621, p = 0.009, r = 0.268$) and Social Adaptation ($Z = -2.320, p = 0.020, r = 0.237$), with small to medium effect sizes in both cases. Results seem to indicate that the high social density room configuration led to disruption in terms of listening to other participants (Social annoyance: "I was able to understand my partners' talking") and cooperating with them (Social Adaptation: "My partners and I worked together well"), with respect to the low social density room configuration, which offered more space to move and a more comfortable setting.

4.2.2 Presence, Immersion, Social Presence. Questionnaires for Immersion, Presence, and Social Presence were administered after completing both tasks, leading to 48 scores per item. In all cases, we grouped the items according to the original reference, and we compared the distributions associated with the high and low density room configurations.

Figure 7a displays the distribution of the scores associated with the factors "Immersion" and "Person-Virtual Environment (VE) Interaction". The values were averaged to ease comparison between different factors, and ordered so that higher scores indicate a positive outcome. We can see that, for both room configurations, scores are favorable for both factors, indicating high immersion levels. A normality check conducted using the Kolmogorov-Smirnov test rejected the null hypothesis; thus, we employed non-parametric statistical tests. A Mann-Whitney U-test revealed no significant differences between the room configurations for Immersion ($Z = 0.8418, p = 0.3999, r = 0.1215$) nor for Person-VE Interaction ($Z = -0.2100, p = 0.8336, r = 0.0303$).

Figure 7b depicts the distribution of the scores gathered using the Social Presence questionnaire, aggregated in the factors "Quality of Interaction", "Presence/Immersion", and "Social Meaning". Please note that for this questionnaire, the Likert scale ranged from 1-5. As with the other questionnaires, items were averaged per factor and scaled so that higher values would indicate positive outcomes. We can see that high values were reported for factors "Presence/Immersion" and "Social Meaning", corroborating the findings of the previous questionnaires. Values associated with "Quality of Interaction" sit slightly lower. The low scores stem mainly from the answers given to the first two questions, which mapped more to an emotional connection: "I was able to feel my partners' emotions during the experience", and "I was sure that my partners often felt my emotion". It is worth noting that the Social Presence questionnaire was originally formulated to be used with pairs that knew each other, and validated with a task (namely, photo sharing) which allows for an emotional connection between people. Conversely, our tasks were not designed to foster an emotional bonding between participants, which explains why low scores were given to the corresponding questions. Test using Kolmogorov-Smirnov rejected the normality hypothesis, and the Mann-Whitney U-test applied to the data failed to reject the null hypothesis for all factors (Quality of Interaction: $Z = 0.1497; p = 0.8810; r = 0.0218$; Presence/Immersion: $Z = -0.4051; p = 0.6854; r = 0.0591$; Social Meaning: $Z = -1.4946; p = 0.1350; r = 0.2180$).

Finally, figure 7c shows the violin plot of the scores associated with the items in the iGroup Presence questionnaire, grouped in the factors "Presence", "Social Presence", "Involvement", and "Experienced Realism". As for the previous questionnaire items, scores within each factor were averaged to facilitate comparison, and they were flipped if necessary to always map higher scores to a more positive outcome. We can see high values for all the factors in the questionnaire, except for the "Experienced Realism", which received lower scores on average (median value $M = 3$ and $M = 3.25$ for the small and big room configuration, respectively). The low scores associated with realism can be explained by considering that our environment is not a faithful replica of the real world, instead being designed as rather barebone and simplified. Thus, it is reasonable that users gave low scores to questions such as "The virtual world seemed more realistic than the real world", and "How real did the virtual world seem to you". As for the previous case, we ran a normality test using Kolmogorov-Smirnov, which rejected the hypothesis of normality. The Mann-Whitney U-test we applied to the data failed to unveil any significant difference between the room configurations, for all factors (Presence: $Z = -0.4239; p = 0.6716; r = 0.0618;$

Social Presence: $Z = -1.6233; p = 0.1045; r = 0.2368$; Involvement: $Z = -0.3312; p = 0.7405; r = 0.0483$; Experienced Realism: $Z = -0.9726; p = 0.3307; r = 0.1419$).

4.2.3 Simulator Sickness. The SSQ was administered before and after the VR experiment, as suggested by the literature [5]. For each participant, three factors, namely Nausea N, Oculomotor disturbance O, and Disorientation D, were computed, according to the original reference, and the total score was computed by summing the three factors and applying a weight of 3.74 [17]. The original paper indicates the following ranges for interpreting the scores: negligible (< 5), minimal (5 – 10), significant (10 – 15), and concerning (15 – 20). In general, we witness a marked increase in symptoms after the VR test. The reported results for Nausea are negligible before the test ($N = 4.56$), but they turn significant after the test ($N = 10.78$); for Oculomotor disturbance, they are minimal before the test ($O = 8.73$), and they become significant after ($O = 13.18$); similarly, for Disorientation, values are significant before the test ($D = 10.29$), but they become concerning after ($D = 19.06$). The total score before the test is equal to 88.21, whereas, after the test, it increases to 160.94. The difference between symptoms before and after the experiment is significant, according to a Wilcoxon signed rank test, only for Nausea ($Z = -2.2515, p = 0.0244, r = 0.2373$), but not for Oculomotor disturbance ($Z = -0.9613, p = 0.3364, r = 0.1013$) or Disorientation ($Z = -1.6820, p = 0.0926, r = 0.1773$), and neither for the total score ($Z = -1.8705, p = 0.0614, r = 0.1972$).

We also test for statistical differences between the room configurations, to see whether they had an effect on the sickness symptoms, using a Mann-Whitney U-test. Results indicate no significant differences in Nausea ($Z = 0.5891; p = 0.5558; r = 0.0621$), Oculomotor disturbance ($Z = 1.8261, p = 0.0678, r = 0.1925$), or Disorientation ($Z = -0.9189, p = 0.3581, r = 0.0969$), nor for the total score ($Z = 0.4423, p = 0.6583, r = 0.0466$).

4.3 Qualitative Results

The qualitative results were analysed from the focus group transcripts using a three stage coding process. The transcript data includes 155 minutes of discussion with the 48 participants in groups of 6. Participant identifiers are shown as the group number and participant ID.

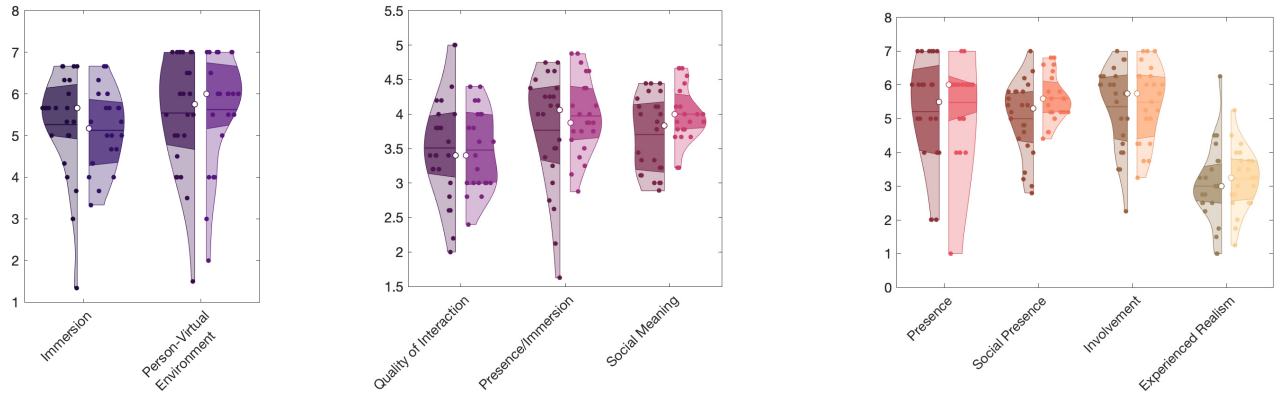
4.3.1 Social Density: Visceral Reactions to Personal Space. Participants had visceral reactions to personal space and how this impacted their experiences in the high or low social density spaces. By design, the high social density space gave each participant limited personal space. The space was described as "too small," "crowded," and even "claustrophobic." Participants in the high social density space were concerned about being too close to others, struggled to find space for themselves, and described the sensation of constantly wanting to step back. The sense of feeling constrained in the high social density space is also visible in the hand tracking data shown in Figure 3, where participants gestured closer into their own bodies even though their physical space was not constrained.

The option to use space more freely in the low social density space gave participants more ways of expressing themselves in the space. For example, Participant 3D stated that "I kind of move closer when accusing someone just to see how they react" when describing the Truths conversation task. Participant 5C and 7B commented on how moving towards others while they were speaking helped them to better hear the speaker. The ability to use space freely in the low social density space afforded these expressive and playful actions, which were not possible in the uncomfortable close distances of the high social density space.

Participants described how challenging it could be to keep others within their field of view in the high social density space, where the proportion of the body that is visible is a well established issue in face-to-face interactions [6, 12]. In the high social density space, participants would be standing within the *personal* proxemic zone, preventing visibility of complete avatars or the whole group simultaneously. Participant 6A noted particular difficulties in the high social density space, stating that "it was too small. I tried to look at people... but there was always someone in the way." In contrast, in the low social density space, participants benefited from arranging themselves specifically so they could see others. When given freedom to stand at one's desired distance relative to others, participants spread comfortably across the personal and social zones, as shown in Figure 3. This enabled better visibility of other participants, and resulted in a better quality of interaction.

4.3.2 Conversation: Who's Line is it Anyway? Conversation in VR presented specific challenges for turn taking and demonstrating attention. The importance of non-verbal cues in social VR has been explored thoroughly in related research [1, 2, 20], with different capabilities for non-verbal communication. In our evaluation, non-verbal communication was limited to head, hand, and body movements and participants were specifically concerned about the challenges of turn-taking during the conversation tasks. Missing social cues influenced the ebb and flow of conversations. Participant 2C highlighted, "you don't want to interrupt somebody because you cannot see who's speaking." 7B revealed, "I kept on thinking who's gonna talk next." The challenge of uncertain timing and potential interruptions was further exemplified by participant 2B, "Felt like there was a bit too much silence. I was like, should I say something or not?" The anxiety of taking turns and the impulse to stay engaged were encapsulated by participant 7B, "Kept on thinking who's gonna talk next." The struggle of identifying gaps in the conversation to make a contribution was highlighted by participant 2D, "Don't want to interrupt somebody because you cannot see who wants to start speaking." The issue of timing and breaks in communication was evident as participant 8A mentioned, "Unnatural breaks... people are waiting to start speaking but nobody wants to start."

A key challenge to conversational turn-taking was displaying and interpreting attention while listening and speaking. 6E shared, "I was not sure if people are listening, I think it's a lack of eye contact, you don't know if other people are looking at you or not." Another participant (2B) underlined, "In the real world, you can see who's coming to talk." The simple signal of head gaze was effective for some participants. Participant 1D stated that "when somebody is speaking, you the people are really looking and giving attention." Participant 2B stated that "I did feel like people were listening to me because all the heads were coming to me." Although head gaze contains less signal than eye gaze when demonstrating attention



(a) Immersion questionnaire **(b) Social presence questionnaire** **(c) iGroup presence questionnaire**

Figure 7: Violin plot showing the results of the immersion, social presence, and iGroup presence questionnaires, aggregated per item (higher values are better). The left and right sides refer to the low density and high density room configurations, respectively.

[25], this was useful for participants in understanding if others were paying attention.

5 Conclusion

This paper explores social density as a factor in social interactions in a social immersive environment. Our between subjects study explored social factors like proximity, conversation, and quality of interaction in a high and low social density space, demonstrating how people interact differently across these kinds of spaces. This mixed methods approach combines quantitative and qualitive methods to provides new insights into how interaction unfolds in social immersive environments.

Acknowledgments

This work was supported through the European Union Horizon Europe research and innovation programme, under grant agreement No 101070109 (TRANSMIXR) and the ERC Consolidator Grant FUSION, proposal 101126024, (funded by the UK Horizon guarantee scheme EPSRC project EP/Z000432/1).

All of the code, data, and tools for analysis used in this paper are openly available on GitHub: https://github.com/julierthanjulie/social-density-xr

References

[1] A. Abdullah, J. Kolkmeier, V. Lo, and M. Neff. Videoconference and embodied vr: Communication patterns across task and medium. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29, 2021.

[2] N. Aburumman, M. Gillies, J. A. Ward, and A. F. d. C. Hamilton. Nonverbal communication in virtual reality: Nodding as a social signal in virtual interactions. *International Journal of Human-Computer Studies*, 164:102819, 2022.

[3] A. Baum and S. Koman. Differential response to anticipated crowding: Psychological effects of social and spatial density. *Journal of Personality and Social Psychology*, 34(3):526, 1976.

[4] S. Benford, D. Snowdon, A. Colebourne, J. O'Brien, and T. Rodden. Informing the design of collaborative virtual environments. In *Proceedings of the 1997 ACM International Conference on Supporting Group Work*, GROUP '97, pp. 71–80. Association for Computing Machinery, New York, NY, USA, Nov. 1997. doi: 10.1145/266838.266866

[5] P. Bimberg, T. Weissker, and A. Kulik. On the usage of the simulator sickness questionnaire for virtual reality research. In *2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*, pp. 464–467. IEEE, 2020.

[6] J. Gehl. *Life between buildings: using public space*. The Danish Architectural Press, 2001.

[7] M. Gelfand. *Rule makers, rule breakers: Tight and loose cultures and the secret signals that direct our lives*. Scribner, 2019.

[8] M. J. Gelfand, J. L. Raver, L. Nishii, L. M. Leslie, J. Lun, B. C. Lim, L. Duan, A. Almaliach, S. Ang, J. Arnadottir, Z. Aycan, K. Boehnke, P. Boski, R. Cabecinhas, D. Chan, J. Chhokar, A. D'Amato, M. Subirats, I. C. Fischlmayr, R. Fischer, M. Fülöp, J. Georgas, E. S. Kashima, Y. Kashima, K. Kim, A. Lempereur, P. Marquez, R. Othman, B. Overlaet, P. Panagiotopoulou, K. Peltzer, L. R. Perez-Florizno, L. Ponomarenko, A. Realo, V. Schei, M. Schmitt, P. B. Smith, N. Soomro, E. Szabo, N. Taveesin, M. Toyama, E. V. de Vliert, N. Vohra, C. Ward, and S. Yamaguchi. Differences between tight and loose cultures: A 33-nation study. *Science*, 332(6033):1100–1104, 2011. doi: 10.1126/science.1197754

[9] E. Goffman. *Behaviour in Public Places: Notes on the Social Organization of Gatherings*. Free Press of Glencoe, NY, USA, 1963.

[10] S. W. A. Group. *From lizarding to lingering: how we really behave in public spaces*. The Guardian, 2020.

[11] M. J. Guitton. Swimming with mermaids: Communication and social density in the Second Life merfolk community. *Computers in Human Behavior*, 48:226–235, 2015. doi: 10.1016/j.chb.2015.02.004

[12] E. T. Hall. The Hidden Dimension: man's use of space in public and private The Bodley Head. *London, Sydney, Toronto*, 121, 1969.

[13] B. Hillier and J. Hanson. *The social logic of space*. 1988. doi: 10.4324/9780429450174-9

[14] S. Hudson, S. Matson-Barkat, N. Pallamin, and G. Jegou. With or without you? interaction and immersion in a virtual reality experience. *Journal of Business Research*, 100:459–468, 2019.

[15] T. Iachini, Y. Coello, F. Frassinetti, and G. Ruggiero. Body space in social interactions: a comparison of reaching and comfort distance in immersive virtual reality. *PloS one*, 9(11):e111511, 2014.

[16] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, vol. 7. CUP Archive, 1990.

[17] 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*, 3(3):203–220, 1993.

[18] J. Li, Y. Kong, T. Röggla, F. De Simone, S. Ananthanarayan, H. De Ridder, A. El Ali, and P. Cesar. Measuring and understanding photo sharing experiences in social virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. Association for Computing Machinery, New York, NY, USA, 2019.

[19] J. V. Li, M. Kreminski, S. M. Fernandes, A. Osborne, J. McVeigh-Schultz, and K. Isbister. Conversation balance: A shared vr visualization to support turn-taking in meetings. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491101.3519879

[20] D. Maloney, G. Freeman, and D. Y. Wohn. " talking without a voice" understanding non-verbal communication in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25, 2020.
[21] R. Moore, E. Hankinson Gathman, and N. Ducheneaut. From 3D space to third place: The social life of small virtual spaces. *Human Organization*, 68(2):230–240, 2009. ISBN: 0018-7259 Publisher: Society for Applied Anthropology.
[22] R. Oldenburg. *The great good place: Cafes, coffee shops, bookstores, bars, hair salons, and other hangouts at the heart of a community*. Da Capo Press, 1999.
[23] T. W. Schubert. The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness. *Z. für Medienpsychologie*, 15(2):69–71, 2003.
[24] D. Thery and B. F. Katz. Auditory perception stability evaluation comparing binaural and loudspeaker Ambisonic presentations of dynamic virtual concert auralizations. *The Journal of the Acoustical Society of America*, 149(1):246–258, 2021. ISBN: 0001-4966 Publisher: Acoustical Society of America.
[25] O. Špakov, H. Istance, K.-J. Räihä, T. Viitanen, and H. Siirtola. Eye gaze and head gaze in collaborative games. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3317959.3321489
[26] W. H. Whyte. The social life of small urban spaces. *Publisher: Conservation Foundation Washington, DC*, 1:466–468, 1980.
[27] J. Williamson, J. Li, V. Vinayagamoorthy, D. A. Shamma, and P. Cesar. Proxemics and Social Interactions in an Instrumented Virtual Reality Workshop. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pp. 1–13. Association for Computing Machinery, New York, NY, USA, May 2021. doi: 10.1145/3411764.3445729
[28] J. R. Williamson, J. O'Hagan, J. A. Guerra-Gomez, J. H. Williamson, P. Cesar, and D. A. Shamma. Digital proxemics: Designing social and collaborative interaction in virtual environments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. Association for Computing Machinery, New York, NY, USA, 2022.

A Data Collection

A cloud server with a single HTTP POST endpoint was set up for collating all the user data gathered during an experiment session. The server is implemented using the Go programming language to achieve adequate performance and keep system load to a minimum. Running as a background process, the server listens for incoming POST requests on TCP port 6000. Further, all requests are handed to the sever through a Nginx reverse proxy, which also takes care of CORS policy validation to allow the Mozilla Hubs clients to communicate directly with the data collection server via AJAX.

Upon reception of a request on the right endpoint, a streaming JSON decoder is instantiated, ready to receive the payload body of the POST request. Once the entire payload is received and validated, the program checks for the presence of all required fields. If all required fields are present, the decoded payload is converted to a comma-separated format and appended to a compressed CSV file using a streaming GZip compressor. The server also adds a UNIX timestamp to each record, which can be used correct possible time drift and/or inaccuracies in the timestamps received from the clients. To prevent file corruption through concurrent access, the write operation is guarded by a mutex. After a successful write, the request handler returns a message with HTTP status 200 to the client.

If the submitted data did not pass validation, the server returns an error with HTTP status 400 to the client. If the data could not be written to the file, an error with HTTP status 500 is returned.

Through the use of GZip compression, the data collected during a typical session, which amounts to about 2 GB, can be compressed to about 500 MB, thus keeping storage space use to a minimum. Further, through the use of a streaming compressor, the file handle can be held onto without having to close and reopen the file for every request.

B Metrics Gathered from Browser Environment

`timestamp`	Device's UNIX timestamp in milliseconds
`fps`	Current frame rate
`uuid`	UUID
`user_agent`	Device user agent
`isBrowser`	Device type
`isLandscape`	Device orientation
`isWebXRAvailable`	VR availability
`avatarID`	Avatar ID
`isHeadsetConnected`	Headset connection status
`isRecording`	Recording status
`pathname`	Current URL
`urlQuery`	Query section of the URL

C Metrics Gathered from DOM Tree

`isLoaded`	Has user finished loading
`isEntered`	Has user joined room
`isFlying`	Is user flying
`isVisible`	Is user visible
`isSpeaking`	Is user speaking
`isMuted`	Is user muted
`volume`	Current user volume
`rigPosX, rigPosY, rigPosZ`	Avatar position (X, Y, Z)
`rigDirectionX, rigDirectionY, rigDirectionZ`	Avatar direction (X, Y, Z)
`rigQuatX, rigQuatY, rigQuatZ, rigQuatW`	Avatar quaternion rotation (X, Y, Z, W)
`povPosX, povPosY, povPos`	POV position (X, Y, Z)
`povDirectionX, povDirectionY, povDirectionZ`	POW direction (X, Y, Z)
`povQuatX, povQuatY, povQuatZ, povQuatW`	POV quaternion rotation (X, Y, Z, W)

Curating with Technology: How to Bring Old Fashion Back to Life in Museum Exhibitions

Karolina Wylężek
Centrum Wiskunde & Informatica
Amsterdam, the Netherlands
karolina.wylezek@cwi.nl

Irene Viola
Centrum Wiskunde & Informatica
Amsterdam, the Netherlands
irene.viola@cwi.nl

Pablo Cesar
Centrum Wiskunde & Informatica
Amsterdam, the Netherlands
Delft University of Technology
Delft, the Netherlands
p.s.cesar@cwi.nl

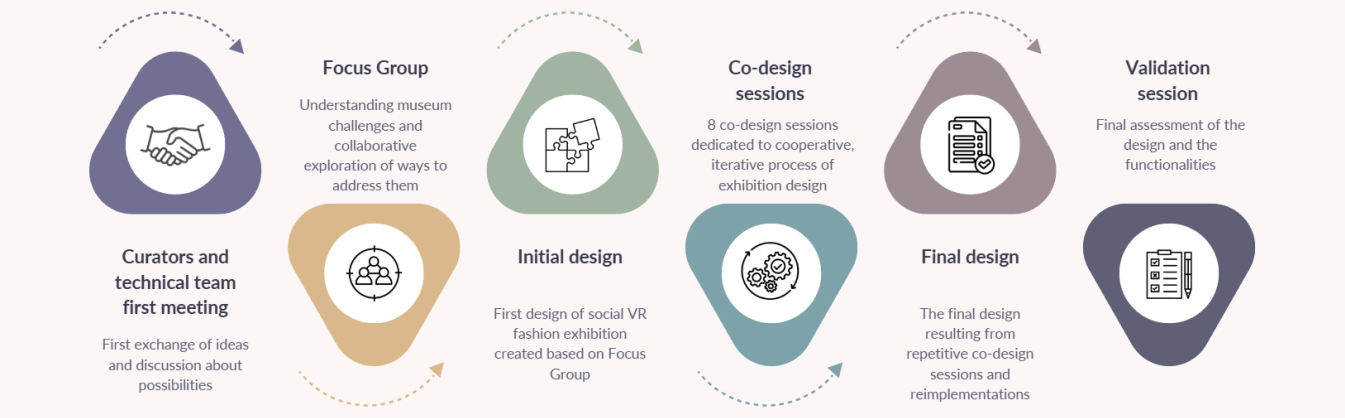


Figure 1: Timeline of the social VR fashion exhibition design engaging curators and technical experts.

Abstract

Social museums, constantly challenged by changing visitors' needs, are beginning to adopt technology in order to enrich guests' experiences. However, designing an exhibition that incorporates digital tools is not easy - it requires a new approach and expertise in both cultural heritage and technology. At the same time, there is a lack of clear guidance on how to effectively design digitally enhanced exhibitions. In this work we follow a human-centric approach, which engages both museum curators and technical experts throughout all stages of the exhibition design. The process, presented in Figure 1, starts with a focus group with curators (N = 4) aiming at understanding the current museum challenges and exploring ways to address them. Based on the workshop results, an initial design is prepared, which is later reiterated during 8 co-design sessions (N = 15). The final design is validated during the validation session (N = 6), resulting in a set of requirements important for social VR fashion exhibition design. The study provides insights for curators into how exhibitions of the future could look like and guidelines on how to design such an exhibition, engaging the technology team throughout the whole process.

IXR '25, Dublin, Ireland

ACM ISBN 979-8-4007-2051-2/2025/10
https://doi.org/10.1145/3746269.3760419

CCS Concepts

• **Human-centered computing** → **User centered design**; **Virtual reality**; *Walkthrough evaluations.*

Keywords

social virtual reality; VR design; exhibition design; user-centered process; social museum

ACM Reference Format:
Karolina Wylężek, Irene Viola, and Pablo Cesar. 2025. Curating with Technology: How to Bring Old Fashion Back to Life in Museum Exhibitions. In *Proceedings of the 3rd International Workshop on Interactive eXtended Reality (IXR '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746269.3760419

1 Introduction

In the 20th century, museums underwent a metamorphosis. What changed compared to previous centuries was the primary focus: it shifted from solely conserving and maintaining the exhibits to prioritizing the visitors' needs. During the 1st International Workshop on Ecomuseums and New Museology, held in Quebec City on October 13, 1984, the basis of the concept of a social museum was established. Since then, the museum has become a place of cultural democracy, social dynamism, openness and interactivity [23]. This new approach brought many new challenges, one of them becoming more attractive to visitors. In order to achieve that, the museums need to: become more interactive, fulfill the need of co-production, e.g. by consumer's active participation in the experience, fulfill the need of engagement, e.g. by immersion, and fulfill

the need of personalisation - tailoring the experience to meet user's needs through customisation, interaction and technology [25].

The switch to a social museum was not the end of the institution's transformation. Nowadays, museums have started to use digital technologies to further attract and satisfy visitors, emerging into so-called digital social museums. Digital technologies help create interactive exhibitions and "improve the relationship between the museum and the user" [23], addressing many of the issues social museums have. This study focuses on social Virtual Reality as a technology that can address the above-described challenges by introducing interactivity, thereby ensuring active participation of visitors in the exhibition, enabling access to fragile items, and assuring sociality in the experience. The chosen exhibition topic, historical fashion, can particularly benefit from the interactivity provided by VR. Visitors tend to seek more physical contact with fashion artifacts than with other objects, as clothing is among the items we keep closest to our bodies. This creates a more intimate bond and a stronger desire for interaction [28].

To take advantage of the opportunities that social VR brings, a carefully selected approach towards designing the exhibition is needed. Based on the interviews conducted with curators, museums still follow a path of exhibition design, in which they engage a technical team after the whole idea is already developed, hiring them more as "contractors" rather than co-creators. This process may lead to the loss of many potential solutions that could emerge if the technical team was involved in the process from the beginning.

This article presents a new course of designing a digital social exhibition based on cooperation between the technical team and curators throughout the whole process. Moreover, it explores elements that should be included in a social VR fashion exhibition to address the challenges museums face and create a successful museum experience.

Hence, this work answers the following research question:
RQ1: How to design a social Virtual Reality fashion exhibition engaging museum curators and technical experts?

- *RQ1.1: What elements are important in the design of a social VR exhibition focused on fashion?*
- *RQ1.2: How to implement the social aspect together with all benefits coming with sociality into VR fashion exhibition?*

By following a new, human-centered design process that established constant cooperation between experts from both cultural and technical sides, we managed to find out which design elements are crucial for social VR fashion exhibition and what are the best practices while implementing them. This study resulted in the demo exhibition that was presented at IEEE VR 2025 [37].

2 Related Work

To create the exhibition, curators need to take many elements into consideration, from the general layout of the exposition to the smallest details of the objects' presentation. The exhibit consists of two key elements: its physicality and the accompanying information. It is important to bring the visitor close to the exhibit, taking into account both of these attributes - reducing the physical and informational distance between the visitor and the object [14].

2.1 Experience

The physical distance can be minimized by allowing interactions between the object and the visitor. During the visit to a museum, the visitors hope to have the possibility to touch and experience things by themselves. It is not only the interaction with the exhibit that is important. The whole environment should be responsive and able to engage the visitor, increasing the immersion in the experience. The audience should be entertained but also sometimes physically engaged and intellectually challenged, allowing for creativity while exploring the exhibition [14]. There are many ways of implementing interactions and many benefits coming from them. An example of the interaction implementation might be via games adjusted to the topic of the exhibition. Those games can induce in visitors the feeling of achievement while providing entertainment at the same time [9]. An interesting and important case is creating interactions not with the historical exhibits themselves but with objects related to them. Those interactions often make people more interested in the main artifact, at the same time introducing another very important element in exhibition design - context [9].

It is important to mention the significance of the exhibition layout - the way people move through the space has considerable influence on the order visitors watch the exhibits, and hence their interpretation of the exhibition [33] [22]. The research suggests though that the visitors themselves like to have the freedom to decide how they will move around the museum space [14]. One example of how to accommodate the public freedom with curating the story is creating a main path, which will always be followed by the visitors, but will also include sub-paths with small experiences within the main branch, giving more freedom of choice [33]. The space layout can also induce social behaviors in people. For example, some spaces produce an effect in which people re-encounter each other, which makes the museum visit more socially exciting [22].

Sociality has a significant influence on the exhibition experience. The companion of a museum visit has a huge impact on how the individual perceives and interprets the exhibits. Sometimes the interaction with their companions can even determine if they notice something at all [22]. Social presence is proven to increase immersion [15], improve learning [26], boost well-being [30], increase engagement, and help to interact with the exhibits [25]. Some researchers argue that the social component in the museum visit is even more important than any other factor [10].

2.2 Context

The context in the case of a social museum exhibition has many angles. Firstly, the context in the information given around the exhibit - it might be the political or social situation in which the piece was created, or the reference to contemporary times. The visitors are very sensitive to the informational context while creating their opinions and perspectives [14]. Not only what the information is, but also how it is given matters. The narration should be fitted to the museum theme (e.g. patriotic or scientific) so it creates a complete whole and does not disturb the perception of the exhibition [14]. The way the museum space looks like has a huge impact on how much the audience enjoys the exhibition as well and should match the exhibition topic [14]. It also influences the emotions people develop towards the exhibit and the memory retention [32] [7].

Of similar importance are the "additional" objects that are in close neighborhood to the exhibits. They can give an intended meaning to the exhibition, that would not be visible if those objects were not included [22]. Archive materials placed around the exhibit are effective in giving information about it: what was its purpose, how was it used, what materials it is made from and so on. Giving people the context of the times they live in helps them understand the artifacts they look at and their purposes. [24].

2.3 Learning

In the case of passing information, context is only one of many aspects that should be taken into account while designing the social museum exhibition. Firstly, text is not the most effective way of communicating the message. People tend to skip reading panels in the museum, and even if they do, they are often not really attentive to what they read [32]. Adding other visual means (like pictures or posters) and auditory information is not enough. The solution is to introduce a multisensory experience - trying, touching, feeling, hearing and seeing. The information should also be given in a way that encourages visitors to think and use their imagination, which is both: pleasurable and educational [9].

2.4 Fashion artifacts

In the case of fashion exhibitions, it is very important to display the exhibit in an appropriate way. Firstly, it should be possible to see it from all sides - clothes look different in the front and in the back, and to fully appreciate them, visitors need to see the object from all angles [24]. Also, seeing garments from up-close enhances the visitor experience [36] [35]. It not only results in the excitement of seeing valuable objects from a really small distance, but also allows to see details, which is really important, as some garments might look completely different from afar [24]. One element that often restricts visitors from seeing exhibits in detail is the glass around the objects. Getting rid of the glass can strongly benefit the exhibition experience by allowing a better view of the exhibit, avoiding ugly light reflections and creating an increased sense of intimacy between the visitor and the garment [24].

2.5 Social Virtual Reality

Many of the challenges described above can be addressed using social VR. In VR the user can freely manipulate the object, which makes the experience of interacting with clothes more "real". The decision of which part of the clothing to show and how to present it would not need to be taken as the visitor could rotate the piece, disassemble it into parts, put it together, or zoom it in and out.

Another benefit of VR is the possibility to constantly display all of the exhibits. Historic clothes are prone to destruction and need to be regularly taken out of the exhibition for conservation [28]. The VR version of the exhibits would be available all year round. Moreover, it would be possible to show the visitors pieces that are never available to them in the real world because, for example, they are too delicate to be displayed in the museum [26]. VR can also provide interactivity in delivering information.

As social VR experiences are proven to positively influence the relationship between people [12], we can suspect that a social VR museum might have a big impact on building and improving connections between individuals. Moreover, as most people go to museums in pairs or groups [10], creating an experience designed for more than one person would be more fit for the context of their visit.

Figure 2: Focus Group with museum curators.

3 Focus Group

Designing social museum exhibitions, fashion exhibitions and VR experiences have well-established, scientifically proven processes and requirements. However, to our knowledge, no research has yet been conducted on how to create a successful experience combining all of these elements. To define how to do that, a focus group exploring the design process of the Social Virtual Reality Fashion Exhibition was organized. The picture from the focus group is shown in Figure 2.

3.1 Methodology

Four museum curators, out of whom three with a focus on fashion heritage, were invited to take part in the focus group. Based on the literature review, six goals were set to achieve during the workshop:

(1) Learning about challenges that fashion museums encounter nowadays.
(2) Gathering ideas about how we could answer those challenges using social Virtual Reality.
(3) Setting the requirements for the exhibits that should be chosen for the social Virtual Reality exhibition.
(4) Discussion on how to present exhibits and information.
(5) Brainstorming about interactions (user-object and user-user).
(6) Open discussion about all other design elements.

The workshop lasted 2,5 hours and consisted of a short social VR technology introduction and interactive exercises. The exercises were prepared based on the knowledge gathered during the literature review stage. The meeting resulted in ideas written on upfront prepared worksheets and a recording of the discussions. The data was transcribed and analyzed based on Constructivist Grounded Theory [8]. The data analysis mapping is presented in Figure 3.

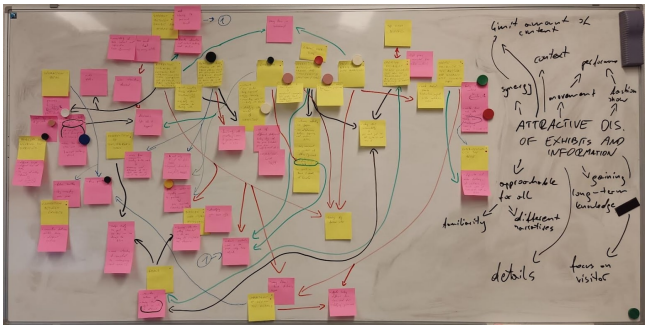


Figure 3: Focus Group data analysis.

3.2 Results

As a result of the analysis, a few concepts important for the research emerged: context, learning, experience, and vulnerability. Two of those - learning and experience - manifested as goals that the museum would like to achieve through the social VR exhibition. Vulnerability was chosen to be the main topic of the exhibition and the requirement in the selection process of the garments.

3.2.1 Experience. During the focus group, many aspects important for the User Experience emerged. Firstly, the synergy between exhibit, information and interactions turned out to be an important topic. Curators were convinced that a good balance and well-designed connection between those three elements can lead to a very pleasant, interesting and easily approachable experience. Also connection between the exhibits turns out to be of high importance. It helps to create a consistent story across the exhibition, which in turn has a positive effect on the whole experience [13] [21].

Another idea that could enhance the experience was creating a performance with the museum objects playing as props. Curators mentioned the possibility of making a fashion show or presenting to the public the "work behind the scenes" - the restoration process that normally happens behind closed doors. One of the experts mentioned an example of the exhibition she saw at the Victoria and Albert Museum in London - "Hollywood Costume"[1]. She mentioned it was an amazing experience as she could see the costumes integrated into the iconic movie scenes, and admire how they are used as props. Both of these examples, the stories and the creation of performance using the exhibits, lead to a very important element that strongly influences the experience - context [1].

Finally, according to curators, the possibility of seeing the exhibits better and interacting with them would have a beneficial influence on the experience. During the focus group, the problem experts often identified was that many museums struggle to be visitor-friendly: they are big, with many exhibits, and hence overwhelming. The way of presenting the objects is often very static and the information is communicated by long blocks of text. The interactivity might provide solutions to those problems: it might break down the routine and make the presentation of the exhibits and information more entertaining. The research already confirmed that interactivity has a huge impact on user experience and users are strongly attracted to the interactive features [31]. However, too many interactive elements can also result in losing consistency and introducing chaos into the exhibition [2]. As already mentioned, it is important to maintain the synergy and good proportions between interactivity, information and the exhibits themselves.

Another issue is crowds in museums. They can cause physical discomfort and irritability, noise and the violation of distance norms that often result in frustration [13]. When the museum is filled with people, it makes it hard to see the exhibits, which can ruin the experience. It is important to provide a good visibility of the exhibit, as it strongly impacts the amount of attention people pay to the object [18]. Moreover, seeing small details that are usually not visible from behind the glass and experimenting, maybe even discovering something about the exhibits, might improve the experience drastically. Those last findings have a significant influence on yet another important for museums area - learning.

3.2.2 Learning. As mentioned above, experience has a significant influence on visitors' learning. Watching a performance is proven to make the audience build interest in the underlying topic and makes gaining knowledge more effective compared to traditional learning methods [16] [27]. According to the literature, performing in the act and taking an active part in creating it is even more effective in sparking curiosity and developing a long-term understanding of the subjects explored in the performance [4] [17]. Using social VR could make taking an active part in historical clothing-related performances accessible for everybody, which, in the physical world, is usually not possible.

Having better access to the exhibits and seeing the details also improves learning. When there are no crowds in the museum, it is easier to investigate the object without feeling stressed out about taking too much time in front of the exhibit [13]. Lack of physical distractions, like noise and limited physical space, also helps to focus on the exhibit. When adding to that the possibility of seeing details that are difficult to spot or even not visible to the naked eye, the interest and possibilities of learning grow drastically.

3.2.3 Context. During the focus group curators emphasized the relevance of context - connecting the exhibits to events, times, locations, stories and contemporary times. Those links should enable visitors to establish emotional connections with the objects. Together with improved experience, the benefit of gaining knowledge appears. Relating the exhibits to contemporary times might improve understanding of past events, help people connect with them and treat them more seriously, which in turn can educate people on how to make (socially) responsible choices nowadays [3].

Context might be introduced to the exhibition in many ways. Let's start with performances that are a perfect opportunity to build context around the exhibit. While creating the performance the authors have the freedom of shaping the environment around the object due to their needs. They have various tools: scenography, props, lights, music, story, and more, that can make the world around the exhibit more authentic and present. During the focus group, one of the curators described a performance-based museum exhibition as follows: "It was like stepping into another world and seeing the garments and how the garments were worn, and how the actors performed into them". She also mentioned an exhibition that

[1] http://www.vam.ac.uk/content/exhibitions/exhibition-hollywood-costume/about-the-exhibition/

took part in Antwerp, in which the movement specialist and dancer worked with old garments to create an accurate fashion show. They studied the garments, pictures and times of the objects, and then replicated the way of walking of people from this period who were actually wearing this type of clothes in their everyday lives. She said it was very impressive to see how different the movement of the body was and how the material movement corresponded to this old way of walking. This is a clear, really interesting example of how the context can be used to show the information about the object that otherwise is difficult to uncover and how much positive influence it has on the experience. It is also important to note the benefit of knowledge retention here - in the end, she still remembers those exhibitions and claims they were one of the best she had seen.

Creating a story is not only achieved by performance. The curators can rely on connections between exhibits that might build the context of events, people and times. This method promotes a deeper visitor engagement with the exhibits, encouraging them to perceive these exhibits as integral components of a broader, interconnected narrative or theme, rather than as standalone items, which in turn positively influences experience and learning [5]. This is also of special relevance for social VR exhibitions, as the technology would allow the use of exhibits that are not placed at the same location but connect together and create a consistent story.

3.2.4 Fashion artifacts and vulnerability. The last, very important conclusion from the focus group workshop is the main characteristic of the garments to be chosen for the exhibition - their vulnerability. Museums own pieces that are so fragile that cannot be put on display or even touched by the curators. Placing these objects in a social VR fashion exhibition not only contributes to the preservation of the fashion and the possibility of showing people the lost garments, but also has high educational value. The garments' models can offer valuable perspectives on the initial structure and appearance of old textiles. This can enrich the understanding and admiration of these artifacts, both among experts and the broader audience, as well as provide a research tool for historians [6] [11].

Vulnerability is also a story itself. Presenting people with artifacts that cannot be touched anymore can convey a sense of transience, showing that objects are also not eternal. There are multiple studies showing that strong emotions are triggered when people perceive vulnerability in other individuals [11]. One could assume that showing the story about the fragility of objects would also cause at least a slight emotional reaction, which is desired to create a better user experience.

3.3 Design requirements

The focus group provided many requirements and important for the experience design elements:

(1) **The exhibits should be connected to each other through a well-curated story.**
(2) **The synergy between exhibits, information about them and interactions should be maintained.**
(3) **Context and interactions should be incorporated for a good user experience.**
Context, based on the results from the experiment, can be introduced using performance, showing elements familiar to the visitors or creating a story. For the interactions, curators listed two types of interactions with the objects: direct (directly impacting the presented piece, like rotation or virtual try-outs) and indirect (more focused on elements around the garments, like games or knowledge quizzes). The idea of virtual try-outs of garments was heavily discussed, however was finally dismissed based on the problem with historical accuracy: people in the 19th century had different height, shape of the body and way of walking. Showing the participants dress from these times displayed around their bodies and saying that is how they would look like in the 19th century would be a lie.
(4) **The way the exhibits are displayed and the information is given should be approachable for everybody.**
To achieve that, the concept of familiarity could be used by, for example, introducing various narratives that fit a broader audience. The curators pointed out the importance of showing the details of the exhibits, introducing context while presenting the object and giving information, but at the same time limiting the amount of content to not overwhelm the visitor. The exhibition should be focused on the visitors and allow them to gain long-term knowledge.
(5) **The scans of real garments should be used.**
Instead of using models created based on the garments, we should use the scans of actual objects. This way the visitors would be given an opportunity to see actual pieces, with all the elements exactly the way they are in the physical objects and curators would gain a chance to further study the artifacts.
(6) **While selecting the artifacts, the priority should be given to fragile objects.**
The objects selected for the exhibition should be (or soon become) too fragile to be physically displayed in the museum, so that by placing them in a VR exhibition we could still show them to the public.

All of those elements have been taken into account while designing the environment. However, during many of the co-design sessions with stakeholders taking part in the project, some of the ideas changed, were adjusted or completely dropped, while other concepts emerged. Nevertheless, many of the elements determined by the focus group were directly implemented into the final design of the experience.

4 Social VR Fashion Exhibition Design

4.1 Methodology

The focus group with curators introduced the background knowledge on what is important in the design of the experience. However, to ensure the best possible outcome, we decided to take a user-centered design approach. During the seven months of the design and development process, eight co-design sessions took place. Their goal was to consult the experts and users at each stage of the design and development process, adjusting the project based on the outcomes of these discussions. A total of fifteen people participated in the meetings in various configurations. Among the participants, there were six curators, four of whom with expertise in fashion exhibitions, an experience designer specialized in fashion, five social XR experts and three users. After the final design was ready,

the validation session was organized. The diagram showing the timeline of the design process is shown in Figure 1.

4.2 Design decisions

The starting point of our process was collecting the exhibits. In the experience, the scans of garments are used, which are not so easily available at the moment. We were limited to choose mainly from the objects that were already scanned. Only a few garments were scanned purposefully for this project to make the story more coherent. Thus, the research to find the themes for the exhibition was based on the garments that were available. After collecting the exhibits and conducting the research the curators working within the project and the designer specialized in fashion created the story connecting all of the collected exhibits into one whole. The scanned garments play the role of star exhibits. They lead the narrative, and other artefacts further expand on the exhibition story. The curators and the designer also shared with us information about the exhibits that could be placed in the social VR experience. An advantage that social VR exhibition has is the freedom to shape the space. Here it is possible to arrange the exhibition rooms in any desired way. There are no physical or financial restrictions; the only limitation is the time that is needed to develop the space. It was decided that the exhibition will consist of three rooms, each of them providing a different context. In each of those rooms a sub-story of the exhibition is told by the artefacts, together creating a bigger and more complex narrative. After creating the space the garments were placed in the rooms. After that, to introduce sociality into the experience, it was integrated with VR2Gather [34] which provides real-time user representation using pointclouds and is designed to run with Unity.

4.2.1 Experience design. The design that was concluded from all of the co-design meetings consisted of a training area and three exhibition rooms. Each of the exhibition rooms introduces a different context in a form of various design styles: a modern museum, a neutral gallery room and a historical room.

For the majority of people, visiting a museum is a social activity. It is important that the virtual museum also allows to visit the exhibition with a companion and fulfills the need for social interaction. Hence, the experience is designed for two or more participants.

After deciding on the general concept, context introduction and sociality of the exhibition, the experience timeline was created. The participants start the visit with a training where they learn how to use controllers to move around and interact with the environment. After getting familiar with the controls, the participants start their journey through three exhibition rooms. The order of the rooms was designed to introduce as little confusion as possible - hence the neutral room is in the middle between the modern and historical.

After the training area, the participants move on and explore the first exhibition room. Here, at the beginning of their path, they encounter a virtual mirror, in which they can see what they look like - for this experience, their virtual bodies are created using point cloud volumetric video, which allows their virtual selves to look exactly like their real bodies. The realistic representation of their bodies positively influences the users' immersion and sense of presence in the environment [34]. After the participants are finished looking in the mirror, they move on to the exhibits' display. There, two garments are shown. The users can closely look at and interact with them. The visitors are also provided with the basic information about the exhibits. When the participants decide that they have already explored the room and want to move to the next space, each of them needs to click the teleportation button assigned to them. This action moves them to the second exhibition room, which is designed as a neutral space. There, similarly to the first room, the visitors explore the space, interact with the exhibit and read information about it. When they are ready to go, one of them clicks the teleportation button and both visitors are moved to the last room - the historical space. Here they once again explore the space, inspect and interact with the garment and get information about it. After finishing those activities, one of the participants clicks the teleportation button and they are moved back to the training area. This is the end of the fashion exhibition experience.

4.2.2 Design of the rooms and interactions with exhibits. The design of the first room, the modern room, was inspired by a real museum, namely the Netherlands Institute for Sound and Vision (NISV) building, and it is based on the design from another project in which our team cooperated with NISV [29].
Two exhibits that are placed in this room are displayed on a wooden background. The space itself is huge and pretty dark, so a well-defined display was needed to make the exhibits well-visible. For the same reason stronger lights are pointing to the two garments. In between them one can find the information display with two interaction panels attached on both sides. The information display contains text with basic facts about the garment and a picture of it. At the bottom part of the display, one can find a button with an arrow, which allows to switch to the next information page (describing the second garment). The interaction panels consist of two rotation buttons, zoom-in, zoom-out and reset, which allow to manipulate the garments. Apart from the exhibition part, the room consists of two other important zones: training and mirror areas.

The second room was designed to have a neutral atmosphere. The colors are maintained in the grayscale, the space does not include any decoration except for the posters related to the exhibition topic. The space visible behind the window was left unchanged from the Unity's default blurred horizon and blue sky, not to disturb the neutrality of the room. The exhibit is located in the corner of the room and is placed on a pedestal. In front of it one can find an interaction panel. The interactions are exactly the same as in the previous room. The information is displayed on a black screen placed on a similar panel as the interaction buttons.

The last room was designed to recall a 19th-century house, as most of the garments chosen to be displayed in this exhibition come from the 19th century. The floor is wooden, the walls are covered with Victorian-era floral wallpaper and the ceiling is decorated with elegant tiles. Around the room one can find many objects from the epoch: a big, wooden wardrobe, a decorative chandelier or a gold-covered clock. One of the more interesting parts is the cabinet containing arsenic rocks. The reason for the presence of this additional exhibit is the history of the dress presented in the room - to achieve the deep, green color of the garment, arsenic was used. Behind the window the panorama shows Amsterdam townhouses. The exhibit is located in the corner of the room on a wooden pedestal. The interaction panel is placed on the wall

next to the exhibit. Similarly to the one in the previous room, it has zoom-in, zoom-out, rotate and reset options. The information about the garment is displayed on a semi-transparent black screen hanging in the air on the right side of the space. The designs of all of the rooms are presented in Figure 4 and the interaction panels in Figure 5.

4.2.3 Virtual museum exhibits. The garments chosen for the social VR fashion exhibition all share one important characteristic: they are very fragile, usually even too vulnerable to be displayed in the physical exhibition. The garments have been scanned using photogrammetry to ensure that they will be perfect representations of the real garments, preserving their exact shape, structure and appearance. All of the garments come from the 19th century, except for one, which was created at the beginning of the 20th century.

4.2.4 Interactions between visitors. There are three types of interaction between visitors in the experience: guided interaction, natural interaction and environment-triggered interaction. The first type, guided interaction, is represented in the experience by the teleportation mechanism in the first room. To move to the next space two of the visitors need to cooperate, as each of them has to click one of the teleportation buttons. Natural interaction is every communication, verbal and nonverbal, that emerges between the users on its own, simply because of the fact that they are in the same virtual space where they can see and hear each other. The last type of interaction - environment-triggered interaction - describes all of the actions the users take together that are directly influenced by the elements of the virtual world. An example of such interaction is trying to adjust the size of the dress by one of the participants while the other one plays the role of a model.

5 Validation

5.1 Methodology

After the design was ready and the experience was implemented, the evaluation session was organized. In this experiment, similarly to the focus group session, there were some participants whose focus lies on fashion (4 people) and some having expertise in different areas (2 people). The session aimed at discussing the design of the environment, the presentation of the objects and information about them and deciding on how to improve the whole experience. The session was divided into two parts. Firstly, the participants tested the experience in pairs. The think-aloud protocol [20] was adapted for this part of the experiment. To address the specific, important issues, some additional questions were prepared. In order not to disturb the experience, the participants were first given time to explore the room and then, before they moved to the next space, the prepared questions (per room) were asked. The questions are shown in Appendix A. For this part of the validation the sound and video recordings of the computer screen showing what the participants were looking at were taken. The reason for that was the need to know what element of the experience the participant was talking about at a particular moment.

The second part of the validation session was based on discussion. Each of the participants was asked what they think about the experience, what they find well-made and what should be improved. After each given opinion, there was a small discussion about the aspect pointed out by the speaker. The discussions were audio-recorded.

After the session the data was transcribed and analyzed using Constructivist Grounded Theory [8].

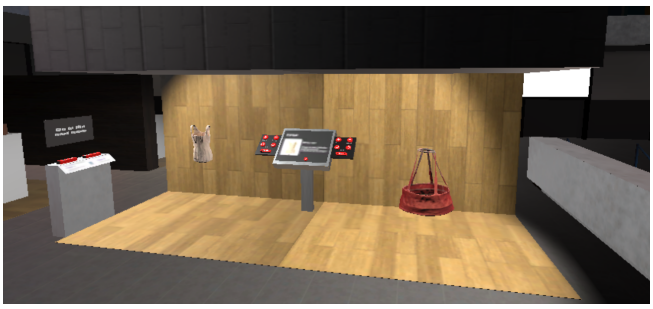
5.2 Results

Overall, the experts were impressed by the design of the experience. They liked the diversity of the rooms, found interactions enjoyable and the whole experience interesting.

The data shows that a key goal for museums to achieve through social VR fashion exhibition is to share knowledge and deepen understanding of the objects and their history. It can be achieved by allowing to present and interact with objects that cannot be presented in the physical world anymore and introducing the possibility of passing information in a more interesting way. However, as the audience is expected to be very diverse, we need to be careful to ensure the visitors pay attention and do not miss the story that the curators want to tell them. Some of the users might be mainly interested in the novelty of the technology and hence focused on the way social VR works rather than the story presented in the exhibition. Other ones can have problems with using controllers which can be disturbing and cause them to miss (part of) the narrative. During the expert validation session a few solutions to this problem have been proposed.

Firstly, synergy has the potential to help attract people interested mainly in the novelty of technology also to other elements of the experience, like, for example, the story. This time the experts mentioned mainly the connection and a good balance between technology, exhibits, narrative, space and room design (aesthetics). All of these elements should be logically tied together creating one whole, in which particular elements do not fight with each other but rather complement each other. The important element - the story - also needs to play by this rule. It is better for it to be simple, but well executed and fitting together with the whole exhibition. While talking about the synergy the participants gave examples of visually connecting the exhibit with the surrounding or adding decorations connected to the main exhibit.

Another solution to the problem of variety in the target audience is keeping the environment, interactions and the story simple. However, there are also some challenges associated with this idea. Firstly, as the curators repeated many times: "Mirroring the real world is a bad idea. You would never have an experience that is better than the real world.". According to the participants we should create something that does not resemble reality - the best idea is going beyond the schemes instead of recreating the actual museum. The problem that appears here is that creating something simple and straightforward, but at the same time creative and out of our world is difficult. Those descriptions even exclude each other up to some level - for some participants being in a weird environment might be overwhelming. It is necessary to find a balance between something simple and user-friendly and something interesting that ensures a great user experience. As it turned out during the experiment, to address this dilemma another approach is needed. The context can be used to gradually introduce the visitor to the virtual world. The visitor should start their virtual museum visit in a room designed to mirror the physical space they are currently in.



(a) Modern room design.

(b) Historical room design.

(c) Historical room design - additional objects.

(d) Neutral room design.

Figure 4: The designs of 3 rooms created for the social VR museum exhibition.

(a) Interaction panel - historical room.

(b) Interaction panel - modern room.

(c) Interaction panel - neutral room.

Figure 5: Interaction panels.

This, according to the experts taking part in the validation session, should allow for a smooth transition from the real to the virtual world. The participants will be able to learn how to interact with the environment in a safe space, without feeling of being overwhelmed. The second room should have context fitted to the topic of the exhibition. This space should already be different from the room in which visitors are physically present, but still pretty normal - something that could exist in the real world. However, some "unnatural" elements are slowly being introduced there, like floating elements or a playful way of giving information. Finally, the third room can feel unreal. During the session the participants gave a few ideas of what this space could look like: "oniric", "crazy", darkness with light just on the exhibits or space without boundaries surrounded by a never-ending horizon. According to the experts, it should make it easier for the visitors to get the full message. As the story and information will be given gradually together with visitors' growing control skills and understanding of the virtual space, they should not be so distracted by the technology or handling the controllers, and hence be able to catch more of the prepared narrative.

6 Conclusion

The presented work results in a workflow for creating digital exhibitions, which engages the museum curators and technical experts throughout the whole process. This ensures that the potential of the technology in the project is fully utilized. The study also provides a list of requirements that need to be met to ensure the exhibition fulfills important for the museum goals: satisfying user experience and successful knowledge transfer to the visitors. The next step is to build the final exhibition based on validation results, present it at the museum for public feedback, and conduct a user study on how contextual elements affect experience, learning, and social interaction. Naturally, this specific exhibition reflects the vision of the curators we worked with; however, the process and methodology are generalizable to other cultural institutions. User engagement will be measured using a questionnaire designed for cultural events, adapted to the context of the exhibition [19].

Acknowledgments

This work was supported in part by a grant from 5Dculture.

References

[1] Jennifer D Adams. 2007. The historical context of science and education at the American Museum of Natural History. *Cultural Studies of Science Education* 2 (2007), 393–440.

[2] Sue Allen and Joshua Gutwill. 2004. Designing with multiple interactives: Five common pitfalls. *Curator: The Museum Journal* 47, 2 (2004), 199–212.

[3] Keith C Barton and Linda S Levstik. 2004. *Teaching history for the common good.* Routledge.

[4] Kenan Bas and Esen Durmus. 2019. Pre-Test the Effect of Teaching Social Studies Course through Performing Arts on the Students' Academic Achievement and Permanence of Their Knowledge. *International Journal of Education and Literacy Studies* 7, 2 (2019), 107–121.

[5] Leslie Bedford. 2016. *The art of museum exhibitions: How story and imagination create aesthetic experiences.* Routledge.

[6] Sarah A Bendall. 2019. The case of the "French Vardinggale": A methodological approach to reconstructing and understanding ephemeral garments. *Fashion Theory* 23, 3 (2019), 363–399.

[7] David Brieber, Marcos Nadal, and Helmut Leder. 2015. In the white cube: Museum context enhances the valuation and memory of art. *Acta psychologica* 154 (2015), 36–42.

[8] Kathy Charmaz. 2017. Constructivist grounded theory. *The journal of positive psychology* 12, 3 (2017), 299–300.

[9] Liu Danyun and Cheong Yeong Jiun. 2016. Historical cultural art heritage come alive: Interactive design in Taiwan palace museum as a case study. In *2016 22nd International Conference on Virtual System & Multimedia (VSMM).* IEEE, 1–8.

[10] Stéphane Debenedetti. 2003. Investigating the role of companions in the art museum experience. *International Journal of Arts Management* (2003), 52–63.

[11] Anton JM Dijker. 2010. Perceived vulnerability as a common basis of moral emotions. *British Journal of Social Psychology* 49, 2 (2010), 415–423.

[12] Guo Freeman and Dane Acena. 2021. Hugging from a distance: Building interpersonal relationships in social virtual reality. In *ACM international conference on interactive media experiences.* 84–95.

[13] Christina Goulding. 2000. The museum environment and the visitor experience. *European Journal of marketing* 34, 3/4 (2000), 261–278.

[14] Daria E Jaremen and Andrzej Rapacz. 2018. Cultural events as a method for creating a new future for museums. *Turyzm* 28, 1 (2018), 25–33.

[15] Myunghwa Kang and Ulrike Gretzel. 2012. Effects of podcast tours on tourist experiences in a national park. *Tourism Management* 33, 2 (2012), 440–455.

[16] Brian Kisida, Laura Goodwin, and Daniel H Bowen. 2020. Teaching history through theater: The effects of arts integration on students' knowledge and attitudes. *AERA Open* 6, 1 (2020), 2332858420902712.

[17] Laure Kloetzer, Simon Henein, Ramiro Tau, Susanne Martin, and Joëlle Valterio. 2020. Teaching through performing arts in higher education: Examples in engineering and psychology. (2020).

[18] Jakub Krukar and Ruth Conroy Dalton. 2020. How the visitors' cognitive engagement is driven (but not dictated) by the visibility and co-visibility of art exhibits. *Frontiers in psychology* 11 (2020), 350.

[19] Sueyoon Lee, Irene Viola, Silvia Rossi, Zhirui Guo, Ignacio Reimat, Kinga Ławicka, Alina Striner, and Pablo Cesar. 2024. Designing and evaluating a vr lobby for a socially enriching remote opera watching experience. *IEEE Transactions on Visualization and Computer Graphics* 30, 5 (2024), 2055–2065.

[20] Clayton Lewis. 1982. Using the'thinking-aloud'method in cognitive interface design. *Research Report RC9265, IBM TJ Watson Research Center* (1982).

[21] Anna Lundqvist, Veronica Liljander, Johanna Gummerus, and Allard Van Riel. 2013. The impact of storytelling on the consumer brand experience: The case of a firm-originated story. *Journal of brand management* 20 (2013), 283–297.

[22] Sharon Macdonald. 2007. Interconnecting: museum visiting and exhibition design. *CoDesign* 3, S1 (2007), 149–162.

[23] José Manuel Mas and Abel Monfort. 2021. From the social museum to the digital social museum. *ADResearch: Revista Internacional de Investigación En Comunicación* 24 (2021), 8–25.

[24] Lesley Ellis Miller, Victoria, and Royaume-Uni). Albert museum (Londres. 2017. *Balenciaga: Shaping Fashion.* V & A Publishing.

[25] Joanna Minkiewicz, Jody Evans, and Kerrie Bridson. 2014. How do consumers co-create their experiences? An exploration in the heritage sector. *Journal of marketing management* 30, 1-2 (2014), 30–59.

[26] Louis Nisiotis, Lyuba Alboul, and Martin Beer. 2020. A prototype that fuses virtual reality, robots, and social networks to create a new cyber–physical–social eco-society system for cultural heritage. *Sustainability* 12, 2 (2020), 645.

[27] David T Overton and Maria Chatzichristodoulou. 2010. The teaching of science through the performing arts. *Procedia-Social and Behavioral Sciences* 2, 2 (2010), 3871–3875.

[28] Alexandra Palmer. 2008. Untouchable: Creating desire and knowledge in museum costume and textile exhibitions. *Fashion Theory* 12, 1 (2008), 31–63.

[29] Ignacio Reimat, Yanni Mei, Evangelos Alexiou, Jack Jansen, Jie Li, Shishir Subramanyam, Irene Viola, Johan Oomen, and Pablo Cesar. 2022. Mediascape xr: A cultural heritage experience in social vr. In *Proceedings of the 30th ACM International Conference on Multimedia.* 6955–6957.

[30] Jessie Sun, Kelci Harris, and Simine Vazire. 2020. Is well-being associated with the quantity and quality of social interactions? *Journal of Personality and Social Psychology* 119, 6 (2020), 1478.

[31] Alistair Sutcliffe and Jennefer Hart. 2017. Analyzing the role of interactivity in user experience. *International Journal of Human–Computer Interaction* 33, 3 (2017), 229–240.

[32] Clara Swaboda. 2019. The art of display: How exhibition context influences the appreciation of artworks in museums. *Unpublished master's thesis* (2019).

[33] Kali Tzortzi. 2007. Museum building design and exhibition layout. In *Proceedings of the 6th International Space Syntax Symposium, Istanbul, Turkey*, Vol. 1215. 072.

[34] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, and Pablo Cesar. 2023. Vr2gather: A collaborative, social virtual reality system for adaptive, multiparty real-time communication. *IEEE MultiMedia* 30, 2 (2023), 48–59.

[35] Gordana Vrencoska. 2015. Museum Fashion Exhibitions: The fashion designer as an artist and new paradigms of communication with the audience. *New space in art and science* 515 (2015), 528.

[36] Louise Wallenberg. 2020. Art, life, and the fashion museum: for a more solidarian exhibition practice. *Fashion and Textiles* 7, 1 (2020), 17.

[37] Karolina Wylężek, Irene Viola, Pablo Cesar, Jack Jansen, Suzanne Mulder, Dylan Eno, Wytze Koppelman, Marta Franceschini, Marco Rendina, and Ninke Bloemberg. 2025. Fashion Beneath the Skin-a Fashion Exhibition Experience in Social Virtual Reality. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW).* IEEE, 1650–1651.

A Appendix one

Room 1:

(1) Environment:
  (a) What do you think about the design of the room?
  (b) Is the lighting good? What do you think about colours?
(2) Object:
  (a) Are the exhibits of a good size?
  (b) Are they placed well? (position, direction)
  (c) What do you think about the interactions?
  (d) Are the buttons easy to use?
(3) Information:
  (a) What do you think about the information panel?
  (b) What do you think specifically about its: position in the room, size, readability and appearance?
(4) What do you think about the other objects that are placed in the room?
  (a) Do they fit in?
  (b) Are they positioned well?
  (c) What about their size and appearance?
(5) Navigation:
  (a) Is the moving to the next room mechanism and task division clear?

Room 2:

(1) Environment:
  (a) What do you think about the design of the room?
  (b) Is the lighting good? What do you think about colors?
(2) Object:
  (a) Is the exhibit of a good size?
  (b) Is it placed well? (position, direction)
  (c) What do you think about the interaction panel? Its size and positioning?
(3) Information:
  (a) What do you think about the information panel?
  (b) What do you think specifically about its: position in the room, size, readability and appearance?

(4) What do you think about the other objects that are placed in the room?
   (a) Do they fit in?
   (b) Are they positioned well?
   (c) What about their size and appearance?
(5) Navigation:
   (a) Is it clear how should you go on with your visit? Is the teleportation button well positioned and signalled?

Room 3:

(1) Environment:
   (a) What do you think about the design of the room?
   (b) What do you think about lighting and colours?
(2) Object:
   (a) Is the exhibit of a good size?
   (b) Is it placed well? (position, direction)
   (c) What do you think about the interaction panel? Its size and positioning?
(3) Information:
   (a) What do you think about the information panel?
   (b) What do you think specifically about its: position in the room, size, readability and appearance?
(4) What do you think about the other objects that are placed in the room?
   (a) Do they fit in?
   (b) Are they positioned well?
   (c) What about their size and appearance?
   (d) Is it clear what is shown in the showcase?