# MetaCast: A Self-Driven Metaverse Announcer Architecture Based on Quality of Experience Evaluation Model

Zhonghao Lin
The Chinese University of Hong
Kong, Shenzhen
Shenzhen, Guangdong, China
zhonghaolin@link.cuhk.edu.cn

Haihan Duan
The Chinese University of Hong
Kong, Shenzhen
Shenzhen, Guangdong, China
haihanduan@link.cuhk.edu.cn

Jiaye Li
The Chinese University of Hong
Kong, Shenzhen
Shenzhen, Guangdong, China
jiayeli@link.cuhk.edu.cn

Xinyao Sun
Matrix Labs Inc
Vancouver, BC, Canada
University of Alberta
Edmonton, AB, Canada
asun@matrixlabs.org

Wei Cai*
The Chinese University of Hong
Kong, Shenzhen
Shenzhen, Guangdong, China
caiwei@cuhk.edu.cn

## ABSTRACT

Metaverse provides users with a novel experience through immersive multimedia technologies. Along with the rapid user growth, numerous events bursting in the metaverse necessitate an announcer to help catch and monitor ongoing events. However, systems on the market primarily serve for esports competitions and rely on human directors, making it challenging to provide 24-hour delivery in the metaverse persistent world. To fill the blank, we proposed a three-stage architecture for metaverse announcers, which is designed to identify events, position cameras, and blend between shots. Based on the architecture, we introduced a Metaverse Announcer User Experience (MAUE) model to identify the factors affecting the users' Quality of Experience (QoE) from a human-centered perspective. In addition, we implemented *MetaCast*, a practical self-driven metaverse announcer in a university campus metaverse prototype, to conduct user studies for MAUE model. The experimental results have effectively achieved satisfactory announcer settings that align with the preferences of most users, encompassing parameters such as video transition rate, repetition rate, importance threshold value, and image composition.

## CCS CONCEPTS

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

## KEYWORDS

Metaverse, Announcer, Quality of Experience, Human-centered Computing, Virtual Cinematography

*Wei Cai is the corresponding author (caiwei@cuhk.edu.cn).

## 1 INTRODUCTION

Metaverse is an emerging concept that is considered to be the next-generation Internet in which the users can completely immerse themselves in a 3D virtual space through interactive technologies. Currently, there are various directions of metaverse development, such as gaming-based (*Roblox*, *Minecraft*, *Fortnite*, blockchain-based [3] (*Decentraland* and *The Sandbox*), or extended reality-based [32] such as *Meta Horizon Worlds*. The common idea behind these projects is to provide a virtual world for users to interact with. As the metaverse attracts more users, the demand for an announcer system emerges [21]. Due to the sheer scale of the metaverse, users cannot cover it entirely with their own avatars. In this case, a metaverse announcer that broadcasts events in real-time can help users stay informed about ongoing events in the expansive virtual world.

However, most existing video game observers mainly serve in esports competitions, such as *Counter-Strike: Global Offensive (CS: GO)*, *League of Legends (LOL)*, and *PlayerUnknown's BattleGround (PUBG)*, etc. The esports live-streaming teams generally consist of a director and several observers. The observers monitor in-game data to keep track of ongoing events and capture highlight moments through a free camera. Subsequently, the directors select a video stream from all the observers' captured streams and push it to the audiences [27]. However, as a parallel universe to the real world, the metaverse should be accessible anytime to anybody in the world, and so does its observer. This poses a challenge as human-driven observers are unable to afford 24-hour delivery in the practical metaverse. Currently, some researchers have developed virtual cinematographic systems to address specific needs, such as plot-based storytelling [6], cinematic sequence [23], and complex crowd tracking [14]. Nonetheless, there is a lack of a general framework to fit the wide variety of situations occurring in the metaverse.

Zhonghao Lin, Haihan Duan, Jiaye Li, Xinyao Sun, & Wei Cai

Therefore, in this work, we proposed a three-stage metaverse announcer architecture consisting of the Event Manager, Prose Storyboard Language (PSL) [33] Interpreter, and Camera Controller. The Event Manager captures newsworthy events from the metaverse, such as crucial user actions or crowd gatherings. After determining the targets to be announced, the PSL Interpreter will choose several shot specifications for the event and translate them to corresponding camera positions. At last, the Camera Controller takes charge of the path planning, including transition time and curve.

Moreover, this paper identifies and analyzes various factors affecting the users' experience of a metaverse announcer, mainly containing two subjective factors: the accuracy of catching events and perceived video quality. For quantitative evaluation, we build a quality of experience (QoE) model, named Metaverse Announcer User Experience (MAUE) model, which comprises four objective factors: the transition time between shots, frequency of switching views, threshold value of important events, and image composition.

To validate the proposed metaverse announcer architecture, we implemented an announcer system named *MetaCast* in a university campus prototype. It can automatically capture events, film at an appropriate position, and drive the camera smoothly between shots. To determine suitable parameters for the MAUE model, we prepared video segments with different influencing factor settings and conducted user studies. Participants rated their satisfaction with the video segments through a 5-point Mean Opinion Score (MOS). Based on our experimental data, we obtained a series of announcer system settings that align with most users' perceptions.

Moreover, we also noticed that the results show significant variation among different users, e.g., someone prefers the bird's-eye view with a long duration, while someone prefers frequent switches in different events. To address this problem, we divided the experience into two stages. In *S1*, *MetaCast* adapts the settings from the previous experimental results. We invited 20 users to rate the performance, which shows that the selected parameter settings are palatable to most users. For *S2*, we designed a real-time feedback system for users to express their preferences during their long-term daily use. The users' feedback demonstrates that *MetaCast* could be subtly transformed to fit their preferences.

The major contributions of this paper are concluded as follows: 1) we proposed a three-stage architecture of metaverse announcer, which fills the need of metaverse applications; 2) we identified factors affecting metaverse announcer users' experience and present a quantitative QoE evaluation model; 3) we implemented an automatic announcer system for a university campus metaverse prototype named *MetaCast* based on the proposed three-stage architecture; 4) we conducted user studies to explore the factors affecting QoE and obtained a general setting for most users.

## 2 RELATED WORK

### 2.1 Intelligent Virtual Camera System

Many researchers have attempted to build a self-driven intelligent camera system in virtual worlds [2, 13, 24], and some of them integrated camera control into the gameplay module to provide a novel experience [25, 39]. For example, Halp et al. [17] presented a camera engine for computer games. They emphasized controlling the camera by constraints and managed the tradeoff between

constraint satisfaction and frame coherence. Hugh McCabe and James Kneafsey [30] proposed a virtual cinematography system for first-person shooter (FPS) games. They applied a finite state machine (FSM) to choose different shots in particular situations. The seminal work [5] introduced the Declarative Camera Control Language (DCCL) as a general tool for describing idiom-based solutions for cinematography problems. Film idioms are solutions for obtaining good cinematography and editing in a range of predefined situations [15, 19]. After that, Ronfard et al. [33] introduced prose storyboard language (PSL) to define the syntax and semantics of a high-level shot description language. Based on PSL, Galvane et al. [15] presented a novel importance-driven approach to generate cinematic replays for dialogue-based 3D video games rather than idiom-based ones. Due to the effectiveness of PSL, it will also be used in this paper to describe high-level shot specifications. On the other hand, Lino and Christie [29] dug into the area of camera control and proposed Toric Space, an efficient interval-based search technique for automated viewpoint computation of two or three targets. The technique has been reused in [14] to create dedicated steering behaviors for automatically computing shots of crowd simulations. The studies mentioned above have significant contributions in their target areas and have partly inspired the camera control techniques in this paper, but there needs to be a framework to fit different metaverse. Therefore, this paper proposes a complete and general framework for metaverse announcers to fill the gap.

### 2.2 Quality of Experience on Metaverse Projects

With the popularity of the concept of the metaverse, one of many focuses of researchers has been to evaluate users' QoE in metaverse-related areas [7, 8, 28]. For instance, as Porcu et al. [32] discussed in their initial analysis of the QoE in the metaverse, human and system are the main factors potentially impacting the QoE of metaverse applications. When referring to virtual reality technologies, discomfort symptoms such as cybersickness is considered to evaluate the QoE. In [34], the author presented the results of an explicit (questionnaire-based) and implicit (physiology-based) evaluation of the immersive wheelchair simulator. Han et al. [18] conducted within-subject repeated-measures experiments in an immersive environment based on head-mounted displays (HMD), which explored the correlation between viewing experience and comprehensive factors. On the other hand, the system influential factors concern the network, infrastructure, hardware, and multimedia content. Schatz et al. [36] presented a first approach toward subjective QoE assessment for omnidirectional video streaming, including the stalling impact of HMD-based omnidirectional video and the difference between HMD-based and traditional television-based video with respect to stalling. Wang and Dey [40] analyzed the factors affecting the QoE of cloud server-based mobile gaming, including game genres, video encoding factors, and the wireless network condition. Moreover, they developed a prototype for real-time measurement of Mobile Gaming User Experience (MGUE) using controlled subjective testing. Nevertheless, the existing works have not conducted research on assessing the QoE of announcers in metaverse projects. Hence, this paper analyzes the factors potentially influencing users' perceived experience of metaverse announcers and presents the results of questionnaire-based measurement.

# 3 METAVERSE ANNOUNCER ARCHITECTURE

In this section, we present a three-stage architecture of metaverse announcers, including the composition and the overall workflow. As illustrated in Figure 1, the announcer is composited of three stages: Event Manager, PSL [33] Interpreter, and Camera Controller. The Event Manager is responsible for capturing ongoing events raised by users' behaviors. It will elect one event to announce and transmit details about the core avatars of the event to the PSL Interpreter. The PSL Interpreter selects several shot specifications and calculates the corresponding camera positions and rotations. Once the Camera Controller receives the target, it will move smoothly between each shot by the preset blend method with a suitable transition time.
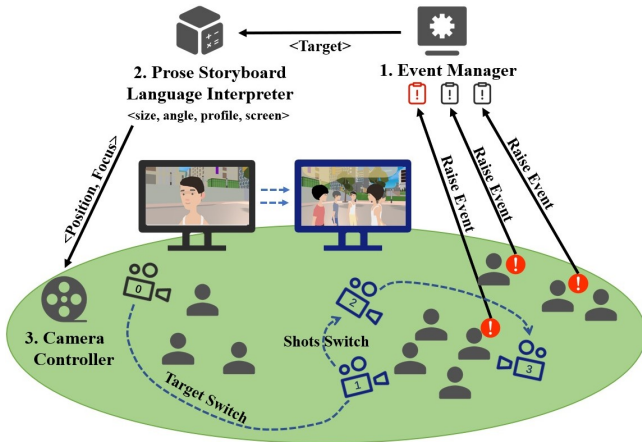


**Figure 1: Metaverse Announcer System Architecture**

## 3.1 Event Manager

The Event Manager is responsible for fetching important events both locally and globally in real time. Local events are those triggered by the significant actions of a few avatars. Since users' behavior can vary widely in different types of metaverse, prefabricated events triggered by specific users' actions such as McCabe et al.'s approach [30] cannot fit our expectations as a general framework for metaverse announcers. Hence, we formulate the user's behavior as a function with different weights for different actions,

$$Importance = w_1 * a_1 + w_2 * a_2 + ... + w_n * a_n (n > 0)$$

where $a_n$ represents different actions and $w_n$ are corresponding weights for each action. As a general formulation, it allows designers to specify weights that are appropriate for their metaverse, such as assigning a high weight to User-Generated Contents (UGC) [9] in gaming-based metaverses like *Minecraft* and *Roblox*. In blockchain-based metaverse like *The Sandbox*, the transaction of Non-Fungible Tokens (NFT) [4] should be regarded as an integral action while the token amount can partially determine the weight.

Global events, on the other hand, are triggered by large gatherings of avatars in the virtual world. To monitor these events, we propose calculating the regional population density and applying an adaptive weight coefficient. By doing so, we can efficiently balance the ratio of local and global events announcements.

## 3.2 Prose Storyboard Language Interpreter

PSL is a high-level shot description language used to define the syntax and semantics [33], which effectively inspires our architecture design. Table 1 shows the subset of PSL we focus on. In this subset, <Subject> represents the target avatar of the shot. <Angle> denotes the relative height between the camera and the subject's face. While <Size> refers to the relative distance, from Extreme Closeup (ECU) to Extreme Long Shot (ELS). <Profile> expresses the relative position of the camera with respect to the subject. For example, <Profile = 3/4 right> means the camera is placed at 45 degrees to the right of the direction the subject is facing. <Screen> stands for the target avatar's position on the user's screen.

**Table 1: Simplified PSL Grammar (Shot Specification)**

| |
|---|
| <Composition>::= <Angle, Size, Profile> on <Subject>[<Screen>] |
| <Subject> ::= <Object> |
| <Angle>::= Low | Eye | High |
| <Size>::= ECU | BCU | CU | MCU | MS | LS | ELS |
| <Profile>::= Front | 3/4 Right | Right | 3/4 Back Right | Back | 3/4 Back Left | Left | 3/4 Left |
| <Screen>::= Left | Center | Right |

Designers can specify the mapping table for different kinds of metaverse to customize shots. The high-level shot language can be transformed into camera positions and rotations. The input is the target avatar's position and the shot specification. It starts from the avatar's basic point, normally at its lower abdomen. Using <Angle>, we can determine the height of our target camera position. On this basis, we can get a directional vector defined by <Profile>. Move the camera at a distance determined by <Size> in this direction, and then we can get the camera position. The focus point begins from the midpoint of the target avatar's face, and the point should move a certain distance to the left or right according to <Screen> and <Size>. After that, control the camera to look at the focus point, then we can obtain the expected shot.

When constructing a shot, there is a well-developed system in photography for us to follow. For example, the "rule of thirds" is a longstanding principle in photography and has proven to be useful [26]. It guides us to divide the frame up into 3*3 grids and place the focus point along one of the third lines or on one of the four intersections of the third line. Designers can adjust the mapping table to ensure most image compositions obey this rule. In addition, the "look room principle" [20] suggests giving extra room in the direction the subject is facing. Hence some specifications can be filtered out, such as <Profile=right, Screen=right>. Without giving enough looking room, the main avatars may look like they are talking to a wall, which is weird to the audience. After these filters, PSL Interpreter sends well-designed shots to the Camera Controller.

## 3.3 Camera Controller

The Camera Controller manages the camera's movement between shots, including the blend time and blend method. Similar to the shot filter, there are several basic rules in cinematography that guide camera control. For instance, during a 2-person dialogue, a filmmaker should establish the scene with a wider shot (e.g., long shot) containing both characters [1, 16], which is also known as the

"rule of introduction". Moreover, the well-known "180-degree rule" states that the camera should remain on one side of the imaginary axis between two conversing actors to prevent viewer confusion and disorientation [22]. By adhering to these rules, we can eliminate certain shot combinations.

We assume three modes for metaverse announcers: bird's-eye view mode, first-person view, and third-person view [38]. In normal cases, the camera moves as a dolly in preset traces to patrol around the metaverse and observe the digital world from a macro perspective. When global events happen, such as gathering crowds, the camera will move toward the nearest point on the trace and look at the target area to announce the event. When a local event occurs with only one target, the camera will switch to the first-person view of the target avatar, while local events with multiple targets, such as a conversation between several users, will trigger the third-person view mode with the shots decided by the PSL Interpreter.

## 4 QUALITY OF EXPERIENCE MODEL

In this section, we start by identifying and analyzing influential factors affecting the Metaverse Announcer User Experience (MAUE). After that, we developed a quantitative model with four perspectives to measure the perceived quality of users, named MAUE model.

### 4.1 Influential Factors

Currently, existing work has discovered the QoE of network factors in mobile games, where poor network conditions will degrade the QoE [40]. Different from the existing work, our study aims to identify specific factors highly related to metaverse announcers rather than the influencing factors like network conditions (bandwidth, delay, etc.). We assume that metaverse announcers are ideal real-time systems with limited network delay. The most fundamental role of metaverse announcers is to precisely capture newsworthy events and broadcast them to the appropriate audience in time. Therefore, as depicted in Figure 2, the MAUE model primarily depends on two subjective factors: the accuracy of catching events and the quality of received video content, determined by the image composition of each shot and the smoothness of switching views.
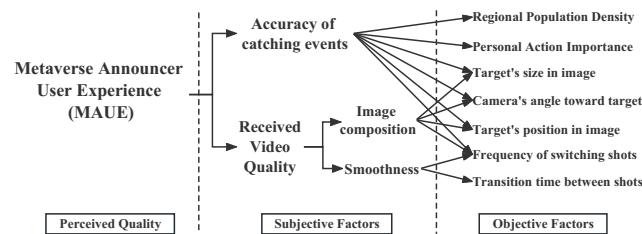


**Figure 2: Objective and Subjective factors affecting MAUE**

The subjective factors are determined by several objective factors. For instance, events are classified as global or local events. We use regional population density and personal action weighting functions to represent the importance, respectively. The image composition of shots can significantly influence users' perceived video content quality. We expect the camera to aesthetically capture the key elements of the event. Therefore, we quantify the image

composition into three factors: the target's size, the camera's angle towards the target, and the target's position in the image. These factors can be represented by four parameters <Size>, <Angle>, <Profile>, and <Screen> based on PSL [33]. The smoothness of the video is determined by the frequency of switching views and the transition time between shots. Intuitively, an excessively low frequency leads to a boring experience, while a high frequency may cause dizziness and confusion. Similarly, the transition time between shots influences smoothness. Switching shots without any transition frame probably confuses the user, and users may lose focus when the transition time is too close to the shot duration. In Section 6, we will explore the parameter selection through quantitative user studies.

### 4.2 Model Description

The Mean Opinion Score (MOS), a subjective score ranging from 1 (unacceptable) to 5 (excellent), has been widely used to measure the QoE [7, 22]. In this paper, we adopt MOS to measure MAUE and develop a quantitative model. As shown in Figure 2, MAUE depends on objective factors: the threshold importance value of raising events, image composition, transition time between shots, and frequency of switching views. Therefore, we formulate the MAUE model as follows:

$$MAUE = f(Transition, Repetition, Importance, Composition)$$

**Transition.** In virtual environments, natural camera motion is crucial for a positive user experience. The transition between different camera shots is a significant factor influencing MAUE. The transition consists of several parts: transition method, curve, and time. The transition method refers to the trace of the camera's movement, such as moving linearly from point 1 to point 2, along with a spherical trajectory or a collision-free visible aware path proposed by [31]. Violation of the "180-degree rule" will disorient viewers and weaken the user experience. Regarding the transition curve, various possibilities exist, such as linear, ease in hard out, hard in ease out, etc. In this paper, we primarily focus on exploring the influence of transition time on MAUE by conducting user studies to determine an appropriate transition time for metaverse announcers.

**Repetition.** Repetition's impact on MAUE is evident in popular esports competitions, where directors alternate views to maintain suitable video repetition rates. For FPS games, first-person views of valuable players, battle replays, and the bird's-eye views of firefight areas [27] are common. In Multiplayer Online Battle Arena (MOBA) games such as *League of Legends*, players control the actor from the isometric perspective [41], and the in-game observing system can even provide a 3D view of highlight moments to the audience to decrease repetition, which is not accessible to the players themselves. This study quantifies repetition as view-switching frequency, seeking a balanced value for metaverse announcer users.

**Importance.** As a critical aspect of metaverse announcers, accurately and effectively capturing events, which impact MAUE, is tied to selecting an appropriate importance threshold value for raising events. A fixed value cannot handle diverse situations within the digital world. For instance, during a grand celebration event, the number of online users increases rapidly. An Event Manager with an importance threshold value designed for daily use will be

triggered frequently since interactions between a large number of users generate a lot of events. Conversely, if a high threshold value is applied for daily use, no event deserves announcing to the users, which will lead to a dull experience. Therefore, the threshold value should be dynamically adjusted to handle diverse situations.

**Composition.** Image composition influences the MAUE in two aspects: aesthetic composition improves the received video quality, and logical composition highlights the relationships between the characters, allowing users to grasp the key points at once. We adopt some basic photography rules discussed in Section 3.2 to avoid poor composition. After filtering out, approximately 150 kinds of image composition remain. However, it is hard to distinguish which one is better since they all obey the basic rules. Therefore, we adopt the 5-point MOS to collect participants' rank of the specifications. After that, we can build a database with the rank and score of each specification.

## 5 SYSTEM IMPLEMENTATION

### 5.1 Campus Metaverse Prototype

CUHKSZ-Metaverse [10] is a blockchain-driven university campus metaverse prototype developed by *Unity* Engine, which has a low-poly style campus digital twin for our students to explore, interact, and conduct social activities. This metaverse establishes a connection with the real world, as the player's physical location influences the rate of acquiring digital tokens. The main gameplay is based on UGC creation [11, 12], with players able to create voxel-style objects and mint them as NFT using the built-in editor. Overall, CUHKSZ-Metaverse exhibits most characteristics of a metaverse.

### 5.2 Details of *MetaCast*

We implemented a practical metaverse announcer demo called *Meta-Cast* in CUHKSZ-Metaverse, following the architecture proposed in Figure 1. The Event Manager calculates the importance of each player's behavior every 10 seconds. Players' action in CUHKSZ-Metaverse mainly contains walking, running, jumping, talking, creating NFTs (pets and crafts) using the UGC editor, and NFT transactions. We evaluate the actions by the following factors: move distance, move speed, length of spoken words, number of voxels, and volume of NFT transaction. We attach corresponding weights to each factor respectively so that the importance of the player's action follows a normal distribution. In our blockchain-driven metaverse, the gameplay mainly focuses on UGC creation and NFT transactions [11, 12], so their weights are higher than others. The importance threshold value for raising events dynamically adjusts based on the number of online users to ensure that the frequency of raising events is stable. After receiving the targets raised from the Event Manager, the PSL Interpreter will choose three shot specifications randomly, pass them to the photography rule filter, and calculates the camera's position and rotation. The Camera Controller drives the camera to the target position smoothly in a collision-free trace. Moreover, we preset the shot duration as 5 seconds and the blend curve as ease in and out.

MetaCast offers three view modes: Bird's-Eye View, First-Person View, and Third-Person View modes, as illustrated in Figure 3. The Bird's-Eye View mode is designed for daily patrols and global events broadcast. As shown in Figure 3(c), the camera observes the digital

world from a bird's-eye view providing viewers with a macro picture, such as the distribution of users. Under normal circumstances, the camera moves as a dolly in a preset rounded trace, allowing viewers to observe the entire metaverse scene. During events, the camera switches to the first-person mode for a single target or the third-person mode for multiple targets, filming the event in sequence by randomly selecting three shot specifications.



(a) First-Person View    (b) Third-Person View    (c) Bird's-Eye View

**Figure 3: Three View Modes in MetaCast**

It is important to note that the content presented through Meta-Cast is viewed on a 2D screen, and thus the experiment does not consider issues such as dizziness associated with head-mounted display (HMD) devices. This decision is based on the necessity of screen-based displays even in reality-based metaverse applications. For example, people in the real world may need to observe events happening within the metaverse through a large 2D screen. Moreover, metaverse players using HMDs can opt to observe activities in other areas through a virtual 2D display screen.

## 6 EXPERIMENTS

### 6.1 Experimental Settings

To maintain a stable experimental environment, we used behavior-tree-driven Non-Player Characters (NPCs) in the metaverse prototype. NPCs follow these steps: walk to a random destination, chase a random nearby NPC, have a conversation, wait 0 to 10 seconds, and choose a new destination randomly. By fixing the random seed value, we ensure that events in the metaverse remain consistent and reproducible. Hence, user studies can be conducted using controlled subjective testing.

For the experimental participants, we recruited a total of 28 participants. Among them, twenty-five were young adults (two between 18-20 y.o. and twenty-three between 21-29 y.o.), followed by two participants between 30-39 y.o., and one between 40 and 49 y.o. The gender distribution is as follows: eight are female, and twenty are male, reflecting the current user distribution of CUHKSZ-Metaverse prototype, where most are college-aged students.

### 6.2 Evaluation on Transition

*6.2.1 Experimental Setup.* The video segment about a conversation between NPCs is used as the experimental material. As previous research [16] found, the average duration time of a shot in animation genre films is 5 seconds. Since avatars in CUHKSZ-Metaverse are animation-style models with limited facial expressions, we classify the content of our metaverse announcer as animation and choose 5 seconds as the duration time of a shot. In a video plot, there are three 5-second shots with different PSL specifications. We changed the transition time between each shot to 0/1/2/3/4/5s to conduct the user study. We chose 5 seconds as the boundary condition since the transition time should not exceed the duration of the shot itself.

Zhonghao Lin, Haihan Duan, Jiaye Li, Xinyao Sun, & Wei Cai

As shown in Figure 4, we chose three video plots filmed in different places of the CUHKSZ-Metaverse to avoid the influence of the image content.



(a) Video 1: a chat between two NPCs in the "Start-Up Zone"



(b) Video 2: a gathering of four NPCs in front of the "Library"



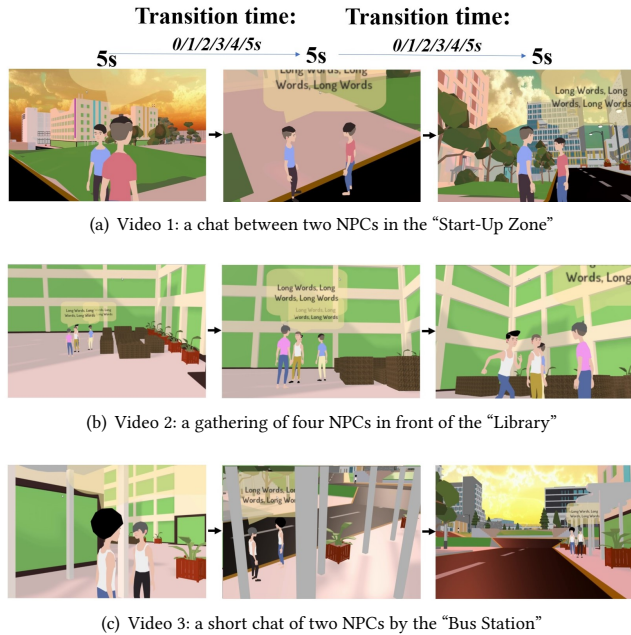(c) Video 3: a short chat of two NPCs by the "Bus Station"

**Figure 4: Experiment Setting: Shots of Video 1, 2, 3**

There is a total of 18 videos (3 plots each with 6 groups of parameters) in the transition experiment. For the convenience of reading, we named them as Transition-plot-parameter. For instance, Transition-3-0 represents the 0-second transition version of plot 3. In simple terms, Transition-3-0 is a 15-second video segment, where each of the three frames in Figure 4-Video3 stays for 5 seconds, and frame 1 will switch to frame 2 without any transition. For the user study, participants were first asked to watch Transition-1-n ($0 \leq n \leq 5$) sequentially. They annotated the QoE level of their last watched video regarding the question "How was your feeling about the transition speed between shots of the last watched video?" using a 5-point MOS, which denotes the QoE from low to high: poor (1), bad (2), fair (3), good (4), and excellent (5). After that, they continued to watch and rank Transition-2-n ($0 \leq n \leq 5$) with a two-minute rest between each plot to avoid the visual fatigue of the participants.

*6.2.2 Result and Analysis.* We calculated the average MOS for each video and each participant's rating, as shown in Figure 5.

Figure 5(a) displays the average score for different transition times, peaking at 2 seconds, while Figure 5(b) presents the distribution of user satisfaction, also showing the highest median score at 2 seconds. To ascertain that our results are efficacious, we performed Kruskal-Wallis Rank Test on the ordinal data, the result presented significant differences across the groups *(H=93.152, p<0.01)*. Furthermore, we identified that Group Transition-2 had a significantly higher user rating than Transition-1 *(Z=-3.733, p<0.01)* and Transition-3 *(Z=-2.274, p<0.01)* by Wilcoxon Signed Rank Tests. For the 0-second cut, some participants mentioned that it cannot express the relation between shots, such as the relative position, where they even need to think about which angle the camera is shot from and how the camera moves from the last shot to the current
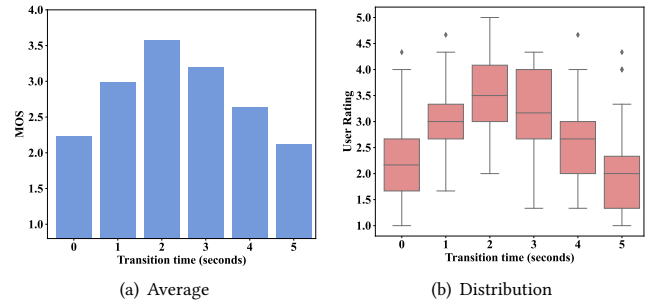


(a) Average  (b) Distribution

**Figure 5: Experiment Results on Transition**

shot. The 1-second version expressed the relative position of shots better, but the fast camera movement caused dizziness, leading to a decrease in users' quality of experience. Regarding the long transition time versions, they annoy the viewers due to the slow transition processes. The 5-second version scored lower than the 0-second version, indicating that most viewers preferred no blending over very slow blending between shots. In short, a 2:5 transition-to-shot-duration ratio offers the optimal QoE for metaverse announcer viewers.

## 6.3 Evaluation on Repetition & Importance

In this subsection, we examine the impact of video repetition on users' QoE, quantifying video repetition as the frequency of view-switching in the metaverse announcer. A low view-switching frequency may result in monotonous content, causing visual boredom and aesthetic fatigue. To some extent, the importance threshold value determines the switching frequency: a low threshold value may bring more events to be announced, while a high value could result in no events being broadcast. Due to the strong relationship between repetition and importance, we combine them and conduct the experiment to explore the optimal view-switching frequency.

*6.3.1 Experimental Setup.* Similar to the last experiment, we prepared three video segments with different plots by randomly setting the generation of NPCs, including number, spawn place, and suit. Each segment lasts for one minute, and we created six versions of each segment, each with a different frequency of view mode switching (1/2/3/4/5 times). We chose 5 times as the boundary condition since we set the duration of events to be 10 seconds. Therefore, in that case, the camera is constantly broadcasting different events. For low-frequency cases, the camera remains in the bird's-eye view for most of the time. The content of the segments can be described as switching between view modes to observe different events as shown in Figure 3. The video content is as follows:

**Video1**: Bird's-Eye view (Start-up Zone) -> Two-avatar Conversation (University Gate) -> One-avatar Running (College A) -> Two-avatar Conversation (University Gate) -> Four-avatar Gathering (Library A) -> Four-avatar talking (Library B)

**Video2**: Bird's-Eye view (Library A) -> Two-avatar Conversation (University Gate) -> Two-avatar Conversation (University Gate) -> One-avatar Walking (Start-up Zone) -> One-avatar Walking (College A) -> One-avatar Walking (Student Center)

**Video3**: Bird's-Eye view (Teaching Buildings) -> Two-avatar Conversation (Library A) -> One-avatar Running (College A) ->

Three-avatar Gathering (Teaching Building) -> One-avatar Walking (College A) -> One-avatar Running (Student Center)

We applied the same form in Section 6.2 to name the video segment. The content of Repetition-3-2 is that the camera keeps the bird's-eye view mode for ten seconds and then switches to announce a two-avatar conversation at Library A. After that, it returns to the bird's-eye view for ten seconds and switches to the first-person view of the user running in front of College A. For the rest of the video, the camera patrols in bird's-eye view mode. The participants rated the QoE level of their last watched video through the MOS scale regarding the question "How was your feeling about the frequency of switching views of the last watched video?". A two-minute rest was set between each plot to avoid visual fatigue.

*6.3.2  Result and Analysis.* The results from the 28 participants are presented in Figure 6.
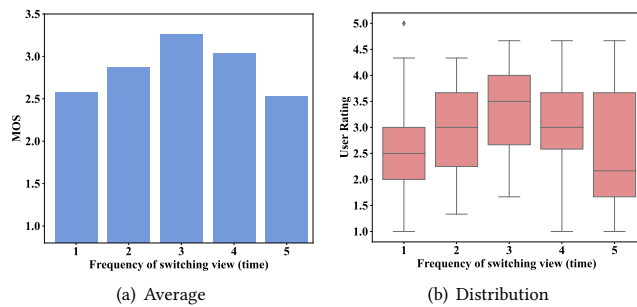


(a) Average            (b) Distribution

**Figure 6: Experiment Results on Repetition**

Both the average and median scores indicate that, during a one-minute announcer, switching the view three times obtains the highest score in the experiment. To ascertain that our results are efficacious, we performed Kruskal-Wallis Rank Test on the ordinal data, and the result indicated significant differences across the groups *(H=23.693, p<0.01).* Additionally, we have identified that Group Repetition-3 has a significantly higher user rating than Repetition-2 *(Z=-3.733, p<0.01),* Repetition-4 has a significantly higher user rating than Repetition-5 *(Z=-3.828, p<0.01)* by Wilcoxon Signed Rank Tests. However, we can observe the distribution of the MOS scores of the participants. For experiments with frequencies 3/4/5, the peak score is approximately 4.5, which implies that certain participants have a strong preference for these frequencies. In contrast, the minimum scores for frequencies 4 and 5 are as low as 1, indicating that some participants have a marked aversion to these frequencies. This reveals a considerable variation in participants' preferences for different frequencies. We conducted interviews with participants exhibiting diverse preferences. Some argued that during a brief period, there are few events that deserve the focus, leading them to favor the bird's-eye view of the world over focusing on unimportant events. While, a participant who assigned an average score of 4.67 to frequency 5 expressed that, compared to the monotonous bird's-eye view, he would rather explore what users are doing in various parts of the metaverse. Given the substantial preference gap among users, considering the median and lower bound, frequency 3 emerges as the optimal choice for most metaverse announcer users. In conclusion, the majority of users prefer metaverse announcers that allocate equal time to local and macro events, with a frequency that divides attention evenly between the two.

*6.3.3  Extension to Importance.* As mentioned above, the experiment video is organized as follows: a 60-second video, which catches events every ten seconds. Therefore, the result of frequency 3 can be interpreted in another way: most users prefer to use the metaverse announcers at the frequency that one out of every two fetch events can hit. Considering the difference in user behavior in each metaverse, we can generally assume the importance $I$ of users' behavior follows the normal distribution, $I \sim N(\mu, \sigma^2)$. In this case, the importance threshold value should be determined in real time according to the online user number. For instance, there are ten users currently in the metaverse. If we set the threshold value at 6.6%, the probability that at least one of the ten users' behavior deserves announcing is 50.5%. From the perspective of mathematical expectation, we can hit an important event every two fetches. The dynamic threshold percentage $i$ can be calculated by the number of online users $N$ and the frequency ratio of hitting event $f$ (our experiment result is $\frac{1}{2}$) by:

$$i = 1 - \sqrt[N]{f}, 0 < i < 1, 0 \leq f \leq 1$$

This algorithm is suitable for most kinds of metaverses, as long as the developer can implement the importance algorithm that conforms to the normal distribution.

## 6.4  Evaluation on Composition

*6.4.1  Experimental Setup.* Based on previous research [16], animated films tend to use MCU shots instead of CU due to the lack of detail in animated faces compared to human faces. Motivated by the study, we truncated ECU/BCU/CU on the <Size> parameter of PSL. And we had 3 <Angle>, 4 <Size>, 8 <Profile>, and 3 <Screen>, totally 288 combinations. As we mentioned in Section 3.2, we used photographic rules such as look room [20] to filter out unsatisfactory shot specifications, resulting in 153 shots. Twelve participants were asked to rank the image composition of random specifications in this experiment. Each of them ranked 153 images organized by different specifications in different scenarios. Each time they wanted to switch to the next picture, they were asked to rate "How was your feeling about the image composition of the last watched image?" with a 5-point MOS scale.

*6.4.2  Result and Analysis.* We collected 1836 ratings from 12 participants, as shown in Table 2. The Eye-LS-Right-Left specification received the highest rating, achieving a perfect score of 5. In simple terms, the camera takes a long shot at the subject's right side of the face while the subject is placed on the left side of the screen as illustrated in Figure 7(a).



(a) Eye-LS-Right-Left            (b) High-MS-Back-Left

**Figure 7: Image Composition Samples**

We can find that it is a commonly used specification [15, 16] in movie filming for a two-person conversation: One person facing right is placed on the left side of the screen while the other facing

Zhonghao Lin, Haihan Duan, Jiaye Li, Xinyao Sun, & Wei Cai

left is positioned on the right side of the screen. The QoE ratings are mapped into two groups of high and low according to Schatz et al.'s user accessibility threshold [35]. Ratings higher than 3.5 [7] are classified as the high group, while those below are assigned to the low group. After that, 42 specifications from rank 111 to rank 153 are classified as unsatisfactory image composition and filtered out. We identified several findings that may benefit future research: High-Level and MS (Medium Shot) is not a good combination as depicted in Figure 7(b). All the last three specifications have a combination of High-level and MS. Moreover, the highest MOS for any combination involving high-level and MS is 2.67, indicating that all of them are divided into unsatisfactory image composition.

**Table 2: Top 5 Specifications of Image Composition**

| Rank | Angle | Size | Profile | Screen | MOS |
|------|-------|------|---------|--------|-----|
| 1 | Eye | LS | Right | Left | 5 |
| 2 | Eye | MCU | 34Right | Left | 4.83 |
| 3 | Low | LS | Left | Right | 4.67 |
| 4 | Eye | MS | Left | Right | 4.67 |
| 5 | Eye | LS | 34Right | Left | 4.67 |
| ⋮ | | | ⋮ | | |
| 110 | High | ELS | 34Leftback | Right | 3.5 |
| ⋮ | | | ⋮ | | |
| 151 | High | MS | Back | Right | 1.83 |
| 152 | High | MS | Front | Right | 1.83 |
| 153 | High | MS | Back | Left | 1.67 |

### 6.5 Validation in *MetaCast*

In this section, we assess the performance of *MetaCast* using a two-stage validation approach inspired by Alexandre et al. [7]. The first stage, *S1*, is to build a general model for most users using the results of the above experiments. As we mentioned in Section 6.3.2, we find that users' preferences can vary widely. Therefore, we have *S2* to customize the announcer according to individual preferences through continuous feedback. This ensures a personalized metaverse announcer for each user. Moreover, *MetaCast* is designed in two forms: (1) for on-metaverse users, when events happen, the video from the camera will be displayed in the top-right corner as shown in Figure 8; (2) for offline viewers, *MetaCast* provides a window for observing the digital world by screen-based displays.

During the *S1* stage, the importance threshold value for raising events changes dynamically based on the number of online users to ensure that the frequency of raising events stays at 50%. The PSL Interpreter will randomly choose three of the 110 specifications classified as the good group following Table 2. Then, the Camera Controller can drive the camera to the target position smoothly, and each shot stays for 5 seconds, and the transition time is set to 2 seconds. We invited 20 participants who have attended previous experiments to evaluate their overall feeling about the *S1* stage of *MetaCast* using MOS. We received an average score of 4.1 from the participants, which indicates a significant improvement compared to the previous experiments. It demonstrates the effectiveness of the proposed MAUE model.

In the player's daily use (*S2*), the Ecological momentary assessment (EMA) [37] system is triggered at each event. The users are
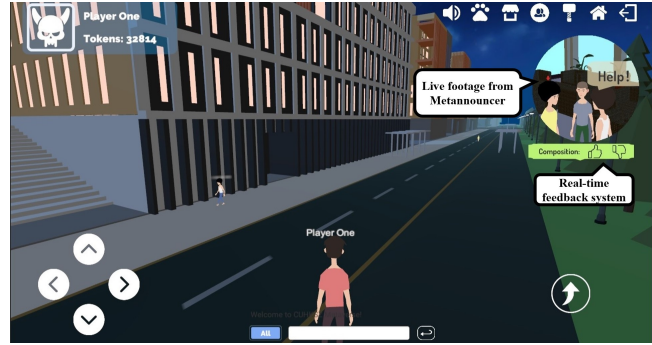


**Figure 8: Validation in CUHKSZ-Metaverse**

asked to provide feedback with thumbs-up or thumbs-down emoji on the image composition as shown in Figure 8. Similarly, the feedback for the transition time and the frequency of switching views is asked every ten minutes with a speed-up or slow-down emoji. The system remains unchanged if the user does not provide feedback. Otherwise, the value can be changed dynamically according to the user's feedback.

## 7 CONCLUSION AND FUTURE WORK

This paper presents a three-stage metaverse announcer architecture to automatically catch and broadcast events. We developed a quantitative MAUE model and conducted user studies to explore the relationship between factors and MAUE. Our findings reveal that a 2:5 transition-to-shot-duration ratio receives the highest user satisfaction. The majority of users prefer an equal balance between local and global events. The importance threshold for events should adjust dynamically based on the online user count. We also provided a table of ranked specifications for image composition. Consequently, our practical demo *MetaCast* received positive user feedback, confirming the effectiveness of our proposed architecture.

The present study acknowledges certain limitations. On one hand, a larger, more diverse participant pool would strengthen the findings. On the other hand, the experimental conditions could be more rigorous; however, we opted for shorter experiments due to concerns about participant fatigue during longer video sessions. Moreover, a single-blind study in which participants are kept unaware of the experimental sequence is better. In future work, we intend to examine long-term feedback for metaverse announcer customization and explore AI-based approaches, including machine learning for event capture and classification, to enhance system performance. Furthermore, while the present study effectively enables the display of virtual 2D screens within HMD-based immersive environments, we aim to expand our scope, focusing on improving the integration of metaverse announcers and HMDs, addressing the challenge of dizziness induced by first-person perspective transitions during announcements, and ultimately, developing a more immersive metaverse announcer experience.

## ACKNOWLEDGMENTS

# REFERENCES

[1] D. Arijon. 1991. *Grammar of the Film Language.* Silman-James Press. https://books.google.com.tw/books?id=6bQlAQAAIAAJ

[2] William H Bares, Joël P Grégoire, and James C Lester. 1998. Realtime constraint-based cinematography for complex interactive 3d worlds. In *AAAI/IAAI.* 1101–1106.

[3] Wei Cai, Zehua Wang, Jason B. Ernst, Zhen Hong, Chen Feng, and Victor C.M. Leung. 2018. Decentralized Applications: The Blockchain-Empowered Software System. *IEEE Access* 6 (17 September 2018), 53019–53033. https://doi.org/10.1109/ACCESS.2018.2870644

[4] Hongzhou Chen and Wei Cai. 2023. How Information Manipulation on Social Media Influences the NFT Investors' Behavior: A Case Study of Goblintown. Wtf. *IEEE Transactions on Computational Social Systems* (2023).

[5] David B Christianson, Sean E Anderson, Li-wei He, David H Salesin, Daniel S Weld, and Michael F Cohen. 1996. Declarative camera control for automatic cinematography. In *AAAI/IAAI, Vol. 1.* 148–155.

[6] Edirlei E. S. de Lima, Cesar T. Pozzer, Marcos C. d'Ornellas, Angelo E. M. Ciarlini, Bruno Feijó, and Antonio L. Furtado. 2009. Virtual Cinematography Director for Interactive Storytelling. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology (ACE '09).* 263–270. https://doi.org/10.1145/1690388.1690432

[7] Alexandre De Masi and Katarzyna Wac. 2022. Less annoying: quality of experience of commonly used mobile applications. In *Proceedings of the 13th ACM multimedia systems conference.* 86–95.

[8] Hongyang Du, Bohao Ma, Dusit Niyato, and Jiawen Kang. 2022. Rethinking quality of experience for metaverse services: A consumer-based economics perspective. *arXiv preprint arXiv:2208.01076* (2022).

[9] Haihan Duan, Yiwei Huang, Yifan Zhao, Zhen Huang, and Wei Cai. 2022. User-generated content and editors in video games: Survey and vision. In *2022 IEEE conference on games (CoG).* IEEE, 536–543.

[10] Haihan Duan, Jiaye Li, Sizheng Fan, Zhonghao Lin, Xiao Wu, and Wei Cai. 2021. Metaverse for Social Good: A University Campus Prototype. In *Proceedings of the 29th ACM International Conference on Multimedia (Virtual Event, China) (MM '21).* Association for Computing Machinery, New York, NY, USA, 153–161. https://doi.org/10.1145/3474085.3479238

[11] Haihan Duan, Zhonghao Lin, Xiao Wu, and Wei Cai. 2023. MetaCube: A Crypto-Based Unique User-Generated Content Editor for Web3 Metaverse. *IEEE Communications Magazine* (2023).

[12] Haihan Duan, Xiao Wu, and Wei Cai. 2022. Crypto-dropout: To create unique user-generated content using crypto information in metaverse. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP).* IEEE, 1–6.

[13] Ahmad Zainul Fanani, Supeno Mardi Susiki Nugroho, and Mochamad Hariadi. 2016. Virtual Camera Control Based on Cinematographic Rules of Camera Shot and Camera Angle. *International Review on Computers and Software (IRECOS) Vol 11* (2016).

[14] Quentin Galvane, Marc Christie, Rémi Ronfard, Chen-Kim Lim, and Marie-Paule Cani. 2013. Steering behaviors for autonomous cameras. In *Proceedings of motion on games.* 93–102.

[15] Quentin Galvane, Rémi Ronfard, Marc Christie, and Nicolas Szilas. 2014. Narrative-driven camera control for cinematic replay of computer games. In *Proceedings of the Seventh International Conference on motion in games.* 109–117.

[16] Alex Hall, Amy Pavel, Alyosha Efros, and Maneesh Agrawala. [n.d.]. *A-tool-for-computational-analysis-of-Narrative-Film.* https://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-102.pdf

[17] Nicolas Halper, Ralf Helbing, and Thomas Strothotte. 2001. A camera engine for computer games: Managing the trade-off between constraint satisfaction and frame coherence. In *Computer Graphics Forum*, Vol. 20. Wiley Online Library, 174–183.

[18] Feilin Han, Ying Zhong, and Minxi Zhou. 2022. Evaluating the Effect of Cinematography on the Viewing Experience in Immersive Environment. In *2022 IEEE International Conference on Multimedia and Expo (ICME).* IEEE, 1–6.

[19] Li-wei He, Michael F Cohen, and David H Salesin. 1996. The virtual cinematographer: A paradigm for automatic real-time camera control and directing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques.* 217–224.

[20] Timothy Heiderich. [n.d.]. *Cinematography Techniques: The Different Types of Shots in Film.*

[21] Hyun-joo Jeon, Ho-chang Youn, Sang-mi Ko, and Tae-heon Kim. 2022. Blockchain and AI Meet in the Metaverse. *Advances in the Convergence of Blockchain and Artificial Intelligence* 73 (2022).

[22] George V Kachkovski, Daniil Vasilyev, Michael Kuk, Alan Kingstone, and Chris NH Street. 2019. Exploring the effects of violating the 180-degree rule on film viewing preferences. *Communication Research* 46, 7 (2019), 948–964.

[23] Kaveh Kardan and Henri Casanova. 2008. Virtual cinematography of group scenes using hierarchical lines of actions. In *Proceedings of the 2008 ACM SIGGRAPH symposium on Video games.* 171–178.

[24] James Kneafsey. 2005. Camerabots: Cinematography for games with non-player characters as camera operators. (2005).

[25] Yana Knight, Héctor P Martínez, and Georgios N Yannakakis. 2013. Space maze: Experience-driven game camera control. (2013).

[26] Michael Koliska and Klive (Soo-Kwang) Oh. 2021. Guided by the Grid: Raising Attention with the Rule of Thirds. *Journalism Practice* 0, 0 (2021), 1–20. https://doi.org/10.1080/17512786.2021.1916402 arXiv:https://doi.org/10.1080/17512786.2021.1916402

[27] Jiaye Li, Minghao Li, Zikai Alex Wen, and Wei Cai. 2022. Understanding the Challenges of Team-Based Live Streaming for First-person Shooter Games. In *2022 IEEE Games, Entertainment, Media Conference (GEM).* 1–6. https://doi.org/10.1109/GEM56474.2022.10017960

[28] Zhonghao Lin, Haihan Duan, Zikai Alex Wen, and Wei Cai. 2022. What Features Influence Impact Feel? A Study of Impact Feedback in Action Games. In *2022 IEEE Games, Entertainment, Media Conference (GEM).* 1–6. https://doi.org/10.1109/GEM56474.2022.10017782

[29] Christophe Lino and Marc Christie. 2015. Intuitive and efficient camera control with the toric space. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 1–12.

[30] Hugh McCabe and James Kneafsey. 2006. A virtual cinematography system for first person shooter games. In *Proceedings of International Digital Games Conference*, Vol. 35.

[31] Thomas Oskam, Robert W Sumner, Nils Thuerey, and Markus Gross. 2009. Visibility transition planning for dynamic camera control. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation.* 55–65.

[32] Simone Porcu, Alessandro Floris, and Luigi Atzori. 2022. Quality of Experience in the Metaverse: An Initial Analysis on Quality Dimensions and Assessment. In *2022 14th International Conference on Quality of Multimedia Experience (QoMEX).* 1–4. https://doi.org/10.1109/QoMEX55416.2022.9900897

[33] Rémi Ronfard, Gandhi Vineet, and Laurent Boiron. 2013. The prose storyboard language. In *AAAI Workshop on Intelligent Cinematography and Editing*, Vol. 3.

[34] Débora Pereira Salgado, Ronan Flynn, Eduardo Lázaro Martins Naves, and Niall Murray. 2022. A questionnaire-based and physiology-inspired quality of experience evaluation of an immersive multisensory wheelchair simulator. In *Proceedings of the 13th ACM Multimedia Systems Conference.* 1–11.

[35] Raimund Schatz, Sebastian Egger, and Alexander Platzer. 2011. Poor, Good Enough or Even Better? Bridging the Gap between Acceptability and QoE of Mobile Broadband Data Services. In *2011 IEEE International Conference on Communications (ICC).* 1–6. https://doi.org/10.1109/icc.2011.5963220

[36] Raimund Schatz, Andreas Sackl, Christian Timmerer, and Bruno Gardlo. 2017. Towards subjective quality of experience assessment for omnidirectional video streaming. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX).* 1–6. https://doi.org/10.1109/QoMEX.2017.7965657

[37] Arthur A Stone and Saul Shiffman. 1994. Ecological momentary assessment (EMA) in behavorial medicine. *Annals of behavioral medicine* (1994).

[38] Laurie N. Taylor. 2002. VIDEO GAMES: PERSPECTIVE, POINT-OF-VIEW, AND IMMERSION. https://api.semanticscholar.org/CorpusID:2603188

[39] Lin Wang, Zhonghao Lin, and Wei Cai. 2021. Heraclitus's Forest: An Interactive Artwork for Oral History. In *Proceedings of the 29th ACM International Conference on Multimedia.* 1452–1453.

[40] Shaoxuan Wang and Sujit Dey. 2009. Modeling and Characterizing User Experience in a Cloud Server Based Mobile Gaming Approach. In *GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference.* 1–7. https://doi.org/10.1109/GLOCOM.2009.5425784

[41] Max Watson et al. 2015. A medley of meanings: Insights from an instance of gameplay in League of Legends. *Journal of Comparative Research in Anthropology and Sociology* 6, 01 (2015), 225–243.