TELECOM INFRA PROJECT®

QoE Framework Application to Telepresence and Volumetric Video

Table of Contents

Authors and Contributors

Javed Rahman

Technology Development Strategist, T-Mobile USA.

javed.2.rahman@t-mobile.com

Ricardo Serrano Gutierrez

Core and Service Platforms Engineer, Telefónica

ricardo.serranogutierrez@telefonica.com

Françoise Blouin

QoE Engineering, Meta Platforms Inc.

francoisb@meta.com

Xinli Hou

Connectivity Technologies and Ecosystems Manager, Meta Platforms Inc.

xinlihou@meta.com

1. Executive Summary

Real Time Immersive Telepresence use case is a perfect fit for the evolving world of digital connectivity and realism, known as the metaverse. Aided with media types such as volumetric video data and network performance levels that allow scalability, its digital spaces are capably filled with content that can be freely and fully interacted with. This enables immersive telepresence to build a niche for itself by letting people meet to socialize, work, play and explore. Thus, through real time generation and completeness in attributes which match real world shape and form, this use case lies at the intersection between video conference type remote collaboration and conventional face-to-face meetings to conveniently spur metaverse growth.

The development of volumetric video based immersive media technology is helping to fill in the voids within the afore mentioned digital spaces with daily life scenes and objects we are familiar with or purposefully need. This capability to increasingly pair comfort or convenience with function has led the way towards envisioning digital setups that can be built to uniquely bring people together. Contrast this to 2D video conferencing, where overcoming various forms of disconnect such as that with the lack of detail or for not being able to achieve proper perspective when trying to capture the essence of a narrative under presentation, remains quite challenging. This can lead to the inability to achieve adequate emotional fulfillment. On the other end, face to face meetings also come with their own set of feasibility related challenges – unavoidable demands on schedule or other impact due to remote locations. Volumetric video, through any of the representations used for its realization be it point clouds, polygon meshes or voxel arrays, and the related attributes, allows freedom to navigate freely in a captured scene with all six degrees of viewing options (6 DoF). This ability to bring another person's physical space into one's own and comfortably interact visually to appreciate a full 3D perspective, allows for the customization of various telepresence scenarios.

While volumetric video content creation processes have over time converged in approach and techniques, being able to successfully distribute with scale such video representations allow end users to adequately experience the multisensory 3D media-

based environment. Additionally, devices or setups which enable proper consumption of the delivered content also have critical contributions to the user experience equation. The display capabilities of various device types can influence the image quality available, the extent to which images can be visualized, how immersive the scenes can be organically, and the number of objects representable, including human subjects.

Transmitting volumetric video content however requires large amounts of data. Live streaming it to materialize a use case should therefore be supported by a network capable of carrying moderately large bandwidth content as required for multiple streaming objects within a scene. Hence there needs to be fast and efficient compression algorithms. Such a network should also exhibit a low latency performance to enable rapid interaction.

To ensure the right levels of quality of experience while consuming volumetric video-based applications and services, a wireless network -based transmission system for content distribution is critical and in-turn bears the burden of meeting quality of service expectations. However, from an end-to-end (E2E) perspective, performance metrics of volumetric video-based services also need to factor in impacts from the content creation end where cloud-based processing environments involve rendering and encoding. Similarly, the end-user consumption side performance depends on device display capabilities.

All considered together, the Quality of Experience (QoE) associated with volumetric video based use cases needs to take into consideration specifics related to its exact implementation while working within bounds established by a framework of parameter classes to adequately reflect user satisfaction and perception of this immersive experience.

2. Introduction

Live telepresence for real-time remote collaboration utilizing volumetric technology capabilities is introduced in this document as one of the already proposed use cases for realization through a Metaverse Ready Network. For remote collaboration to be appealing and successful, the *sense of presence* between two individuals is a vital necessity to capture for retaining aspects of experience when together face to face. To support such expectations, clear hearing, comfortable viewing, and body movement to fully demonstrate visual attributes of objects of common interest augmented by one's own relative movement in 3D space, must be properly realized.

Volumetric video is a media format that represents three-dimensional content for playback and real time experience via traditional flat screens (including smart phones), 3D displays, and emerging extended-reality (XR) platforms, as used in augmented and virtual reality (AR/VR). By fusing captures of critical vantage points from a multiple camera setup, it brings the following attributes to a desired telepresence session for adoption as a successful use case -

- Depth Perception
- Flexible Viewing Angles
- Flexible Movement
- Spatial Audio
- Low Latency for Real Time Application
- Low Bandwidth for Scale

Handheld flat screen devices allow touch screen usage to rotate the view of the object to realize several of the above listed visual attributes to present different point of views. With the other more advanced display devices, objects can be rendered and overlaid, placed in a virtual room, and projected in shape into the physical space in front. All allow users to then move around and within the scene with 6DoF.

This document identifies factors and related metrics to make possible the evaluation and qualitative assessment respectively of volumetric video content delivered for the purposes of enabling the use case under consideration. They can be applied against the performances measured of an end to end system designed to support real time

streaming for an end user experience that is both fully immersive and interactive. For such an experience to occur *remotely* between two individuals with capabilities to freely explore within the immediate 3D physical space of an *untethered* environment, identified metrics will define the relevant requirements of the delivered volumetric video's qualitative attributes. A wireless network which then capably supports the deployment of such a distribution system to realize the desired end user QoE, will be considered *metaverse ready.*

To that end, specific QoE metrics reflecting the broader intent of the QoE Framework Sub-group of Telecom Infra Project's (TIP) Metaverse Ready Networks (MRN) Project Group under development are identified and prioritized within system implementation so that user experience satisfaction is achieved. This will subsequently identify and define application and network-delivery QoS metrics required to realize the desired QoE. This document therefore only briefly identifies (Table 7) fundamental QoS metrics that are essential to deliver the proper use case application and system performance.

3. Use Case Description:

Live 3D Telepresence – Enabling collaboration by supporting the visual sharing of information between two remotely located persons, through an immersive and interactive experience.

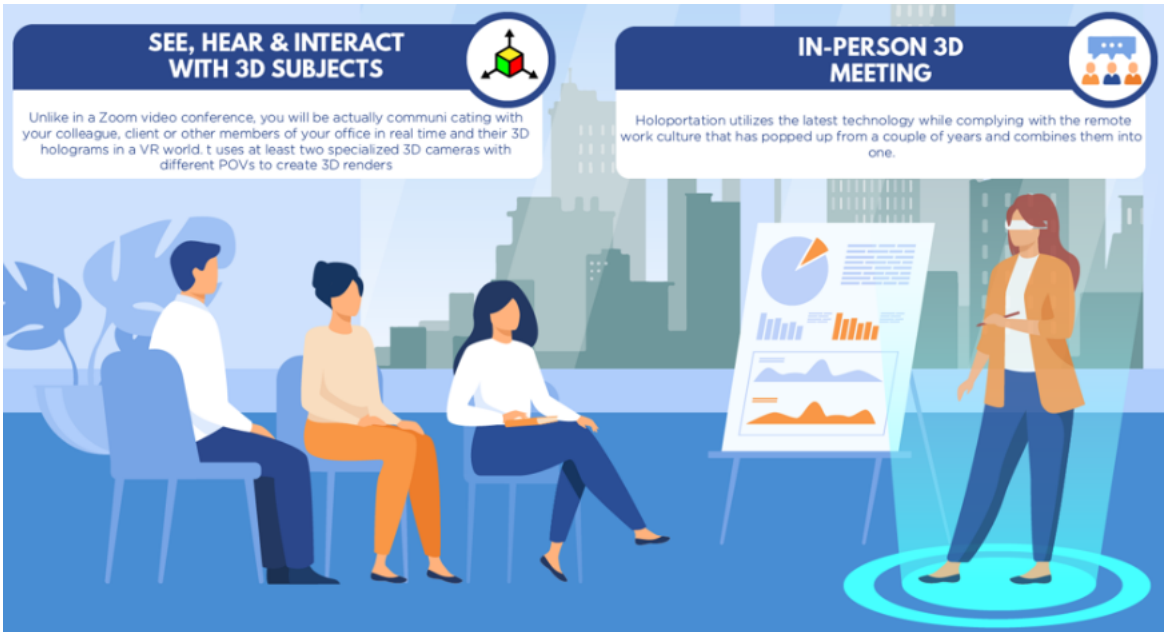


Fig. 2: Conceptual representation of "Visitor" end experience of Volumetric Video based Telepresence use case (Source: Forbes.com)

Telepresence is a mode of communication between groups of people located in two or more geographically separate rooms, such as offices or lounges, by connecting those spaces virtually. Traditional 2D systems today are limited in capability as the image of the remote location viewed by participants from another location is the same to all. This is fundamentally different from in person face-to-face meetings where each person sees the surroundings from his or her unique point of view and is not limited to the fixed and only viewpoint of a remote camera.

Real time interaction that is immersive, is the key lure of this proposed use case. It is an extension to objective specific applications that already exist, such as remote assistance. Based on camera setup configuration and locations, such a telepresence experience can

also be both symmetric and bi-directional. To truly distinguish it from remote assistance, the setup should be conducted in a symmetric fashion - being fully volumetric on both sides and allowing visual interaction that is immersive at the same level to one another.

While experiencing real-time telepresence a system user (PARTICIPANT) will also be capable of seeing the surroundings within the scene origination point. Visual communication and capabilities, along with scene exploration through enhanced visualization (SELF LOCOMOTION [1]) is a necessity. This includes capability of natural movement in the 3D space, situational awareness, and capture of gaze direction. Natural human representation (size and shape-wise) at the host end and its presentation to another user (PARTICIPANT) through immersive solutions however require new and specific technologies, tailored implementation methodologies, and a scalable delivery system.

Volumetric video media type is the key enabling technology for re-creating the telepresence use case scene capture from the host end and making possible its consumption at the user end. Critical encoding techniques are consequently required that perform in real time with compression ratios making possible the delivery of captured content to the intended user at scale.

3.1 Basic Components of Use Case Scenario Setup

The following content highlights requirements and modes of implementation specific to the use case envisioned in this document to enable the proper match-up of QoE metrics subsequently presented.

3.1.1 Essentials

- Person to Person – 1 individual at each end (Remote location)
- User (Primary initiator) and Visitor (Secondary participant)
- Presence – Virtual (Hologram, preferably photo-realistic vs human-realistic [Lacking texture details])
- Bi-directional
- Live / Real time (specific to this use case)
- Rendering -
- Image Generation

- Rapid animation for optimized body movement and gap enhancement
- Additional Rendering – scene extension (background or secondary object, for example)

3.1.2 Layout and Immersion Capture

3.1.2.1 Required Immersive Components

- Visual Communication [*I see you*] – YES (Transmission of the visual representation of a user/person to enable visual non- verbal communications such as stare, look, (hand) gesture based pointing to objects in the other scene or one's own surrounding etc.). Seeing the other person (visitor) is key to transmit non-verbal communication cues, such as pointing to objects within the scene rendered of the visitor's space.
- Remote Presence perception [*I see what you see*] – YES (Visit – become familiar with the other's surrounding. Transmission in real time the visual representation of the surroundings of the user/person. The other person can see the physical environment)
- Shared Immersion [*I am with you*]– NO (Meet in a common environment)
- Embodied Interaction [*I control objects*] – NO (Action of a user is represented within the system and allows user to interact with it – for example, affect an object in the scene)

3.1.2.2 Layout

User end layout (Key attributes)

- Studio Camera capture [360 degree space through *collection of* cameras arranged in a circle and at different levels – not a single omni-directional camera setup]
- Local Video Capture - 360 degrees
- For fully immersive 6 DoF user experience)
- Feedline to processor [On-premises Aggregator]
- Radio Access – 5G uplink/downlink
- Number of Cameras – See table 6A. Minimum value required to enable the capture of whole surface volumes of objects.

User End "Visitor" view – In what form the visitor is shown to the user?

- Real time video stream from above described collection of cameras – YES.
- CGI avatar image that is animated by the application engine – NO.
- User End World view – What surroundings of the visitor's space the user sees.
- Fully immersive video capture of actual physical surrounding – YES.
- Limited CGI pre rendering (and static) of actual physical surrounding – NO.
- Full VR – NO.

3.2 Volumetric Video Technology – The Enabling Technology

Volumetric video is an emerging type of multimedia content. Unlike conventional video formats and 360 degree videos that are 2D, every frame in a volumetric video consists of a 3D scene created by a point cloud or a polygon mesh. The 3D characteristics together provide the most critical feature required to enjoy any volumetric video based experience – that is for the user to be capable of engaging in motion that has six degrees of freedom (6DoF). While watching a volumetric video, users can move with six degrees of freedom (6DoF) along three rotational dimensions by changing viewing direction in yaw, pitch, and roll and three translational dimensions by altering viewpoint position in X, Y, and Z. For example [2], in a chemistry class, students can investigate a molecular structure from various directions by moving around its volumetric display, when the instructor is explaining the properties of a chemical element.

Steps involved with creating, delivering and appropriately consuming volumetric video include,

- 3D acquisition & re-construction through constant recording to accurately and in real-time estimate the scene objects' actual 3D shapes.
- Efficient encoding technique suited for limited delivery bandwidth enabling transmission of the 3D model to the viewing location.
- 3D display device through which content is experienced to its fullest range or closest to.

Sense of Presence

A critical feature while experiencing telepresence using volumetric video is to capture the appropriate level of presence. Presence can be defined as "the subjective experience of being in one place or environment even when one is physically situated in another." [3] Volumetric video media enable its user to have the sense of actually being in the

remote environment. This is called **Spatial Presence** [1] and is made possible through the ability to re-situate oneself within the re-constructed scene location. In reality, it is the attainment of perceptual immersion.

The other component providing a sense of presence has to do with the perception of being together with another (user with host) being. **Co-presence** exists when a person senses that there is another person in the environment. This promotes interaction, but being able to incorporate directional sound makes the sense of co-presence within the real-virtual space even more convincing.

How does volumetric video provide the sense of presence – by being immersive.

- It has to be 3D.

Volumetric video media type generation and consumption is integral to experiencing an immersive 3D telepresence use case. Through 6DoF movement capability, a user of this type of a telepresence system can not only freely look around in any direction of the scene by changing the yaw, pitch, and roll of the viewing direction, but also mimic real world walking motion within the application environment through change in translational position within the 3D space.

- 6DoF ability – Vital for generating immersive environment
- What makes it interactive - Gaze following, hand gestures etc. are two means of interaction specifically tied to the use case under discussion. Gaze depiction is made possible through a virtual ray cast line and hand gestures are represented using the 3D mesh of the user's hand.

The ability to generate geometric models and dynamically shape them with incoming real time captures allows the system to constantly render new scenes and make available novel viewpoints for each participant. These can also be updated to appropriately reflect visual changes from user movement in their own 3D space.

How can it be made real time - Camera based depth capture of scene objects instantly undergo 3D reconstruction with real time updating through continuous recording and processing.

3.3 Volumetric Video Creation and Delivery System for 3D/Immersive Experience

The volumetric video capture process enables a person being recorded to be physically recreated onto an actual 3D space. A studio setup typically entails the volumetric capture equipment set, high bandwidth interconnection for data transfer, live viewing and control, and a host of dedicated resources to support a full production workflow, especially for non-real time production (for playback). Otherwise, the workflow includes capture, 3D re-construction, and a successive series of processing intensive pipeline jobs including compression and high resolution volumetric rendering, which prepares the immersive media for delivery and consumption.

At the center of the studio setup is a synchronized multi-camera video recording system. A typical configuration involves connecting several cameras to a 10GbE network switch and using a single PC as a control station. Serial Digital Interface (SDI) monitoring equipment can be used for low-latency monitoring. Client software is provided for the control station, or an SDK can be used to integrate commands into custom software applications. The client software configures the cameras as needed, initiates recordings, and displays live video for real-time monitoring. Once recordings are completed, the cameras begin transmitting their files to a network-accessible shared folder, where they can be accessed by third party processing software.

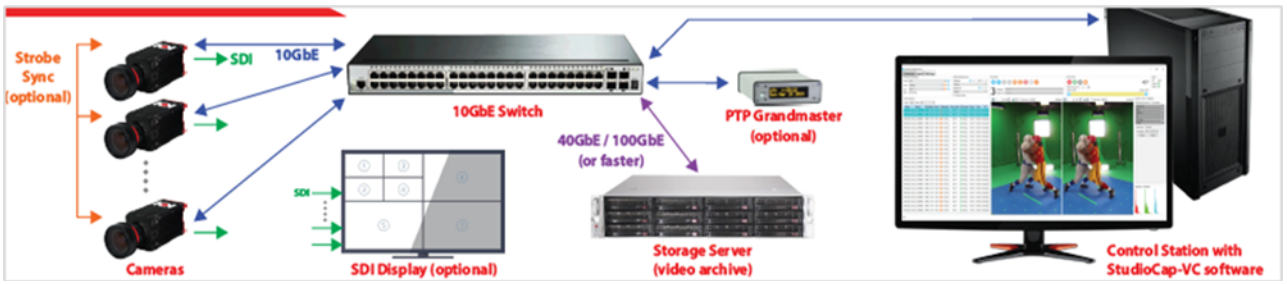


Figure 3: Volumetric Video Studio Capture Components & Interconnection

High resolution cameras with CMOS image sensors capable of capturing frame images with resolutions greater than 20 MP, are required, especially for photo-realistic asset development. It also allows for the generation of high 3D point cloud density and a subsequent mesh representation for high quality animation and scene rendering. Camera frame capture rate (shooting speed) is usually at 30 FPS minimum.

Cameras are grouped in pairs to allow capture of stereoscopic perspective (stereo pairing). This enables depth perception. From each pair depth maps are computed, and all are fused together into a 3D model. If the cameras are arranged in 3 rings of different heights, occlusion is effectively avoided from the source capture itself.

Figure 4: Volumetric Video Capture Environment (Studio Layout)

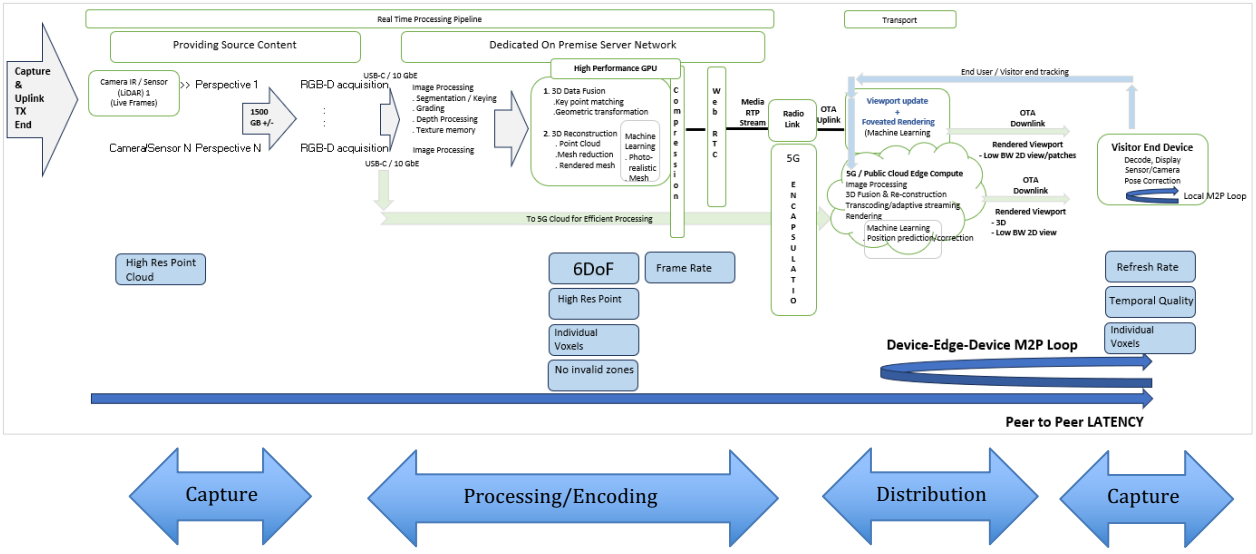


Figure 5: One Representation of and End-To End Volumetric Video Streaming Systems

High-resolution cameras with 4K resolution or greater can deliver holograms of their captured performers with finely detailed and life-like textures. Representing minute surface details such as hair, skin and fabric aids in the improvement of the overall realism factor of the immersive experience being produced. For real-time volumetric video systems supporting the telepresence use case, higher resolution cameras are required for creating more realism in the interactive exchanges.
A multi-planar volumetric video capture, distribution and re-construction system also has 4 key rate definitions.

a. Capture frame rate – To successfully create a 3D data model of volumetric video user as he or she moves.
b. 3D volumetric frame rendering rate [5] – The volumetric image frame buffer update rate.
c. 3D volumetric frame dataset transfer rate – How many volumetric image frames per second are transferred to the volumetric display.
d. 3D volumetric image display refresh rate – The projection system's re-constructed volumetric image optical refresh rate or simply the screen refresh rate of 2D display screens.

3.4 Display System Requirements

A volumetric video display system shall be capable of the retaining 3 of the following capabilities of the real physical human it is representing –

1. Display user movement with 6 Degrees of Freedom (6DoF)

The above is required to enable this use case such that principal user can teach visitor by exploiting the full dimensional movement of object possible in real world. By displaying full motion of objects or lack of relative to proper functioning, user can educate visitor through display of damaged surfaces or lack of function. Visitor can visually grasp the problem as if he or she were seeing it right in front. The full impact of displacement as is possible, can be presented.

2. 3D shape presentation

One of the key advantages of volumetric video is its ability to capture the full range of motion and expressions of a person or object, which can be viewed from any angle in 3D

space. 3D representation lets a viewer understand details of an object, verify accuracy of its shape, understand how multiple parts interact to assemble a product etc. Depending upon position of viewer, he or she will see different parts of scene objects – viewpoint upgrade.

3. Immersion

Immersion allows the ability to capture presence, as if visitor is right behind the display area, for example. Spatial audio reinforces this concept. Depending upon the display technology, either the scene can be moved around to reveal a previously occluded object, or walking around exposes previously hidden objects or partially obscured objects.

In addition, the following attribute is required to make the use case more appealing from the interaction point of view –

- Near Photo-Realistic Hologram

To the user and visitor alike, the other appears in extremely lifelike form, with detailed facial expressions, realistic body movements, and even lifelike hair and clothing. The end product is a lifelike reproduction of the real person.

4. State of Art

Stereoscopic display aided devices such as the Magic Leap AR glasses or Meta's Quest VR headsets have made possible to experience volumetric video content ubiquitously and in a cost effective manner. However, it is through a few very capable and well established capture and production systems that production studios and setups across the globe are making it possible to provide volumetric video services to meet the increasing demand.

Studio Facility	Volumetric Video Capture System
4Dviews	HOLOSYS
Avatar Dimension	MS Mixed Reality Capture
Canon VVS-K	Free Viewpoint Video System
Dimension Studio	MS Mixed Reality Capture
Evercoast	Mavericks
Room	Volucam
Volucap GmbH	VoluCap Max

Such capture and production systems mainly distinguish themselves by varying camera brand selections, number of cameras, streaming formats, processing techniques etc. Here, encoding happens mainly in non-real time, allowing the attainment of higher compression ratios.

FormaVision, Google's Starline, and MATSUKO are 3 other volumetric video content creators whose product can be streamed live in real time. Obviously then, content resolution could be slightly lower. However, this provides with the option to interact with the delivered content, thus supporting 2-way collaboration.

Other companies such as Hungarian HoloVizio and Australia based Voxon have developed innovative 3D display devices using light field technology and a light rendering engine that projects light points onto a swept surface, respectively. Such volumetric technologies bring digital content "to life" without the need to wear glasses and headgear. This in turn allows collaboration through a genuinely immersive environment with viewing angle limited only by the human eye's field of view.

5. QoE/QoS Metrics

This section defines the concept of QoE associated with the volumetric video technology based use case discussed herein. QoE will be characterized and described from a user's perspective with measurement sources indicated, and how display system capabilities can further influence them.

5.1 Volumetric Video Use Case QoE Fundamentals & Measurement Opportunities

Volumetric video consumption provides a multi-dimensional experience. Use cases based off the collection of media types which define volumetric video therefore provide experiences that are predominantly evaluated through users' subjective ***perception*** of the overall quality (created by a combination of those experiences).

Each experience type lets end user to capture a sense of "feeling" (subjective implication) based on all or some of the following captures - a) immersion/presence achieved, b) level of satisfaction, and c) user's attitude *towards the interaction or experience.*

While consuming volumetric video based content, the user's **human responses** encompass both primitive and non-primitive "feelings," which can be considered to have 3 components –

I. Immersion – Level of presence felt. [Possibility to freely explore the immersive content].
II. Level of Satisfaction - With the performance of the 3D hologram/system. Most likely functionality wise, from the user's perspective. [Use of HMD for example].
III. Attitude towards the interaction - Interaction is a functional action, and hence the trigger for attitude development. It shapes the decision (or take off things) that inspires future use.

The above 3 in some combination or another impact the 3 dimensions of experiences that make up the composite experience with any volumetric video based use case.

5.1.1 The 3 Types of Experiences

- **Emotional experience**
  - Highly subjective. Response driven by both immersion and attitude towards the interaction, thereby generating a sentimental arousal.

- **Cognitive experience**
  - Response driven primarily through interaction. It is a mental process that helps to gain knowledge and understanding through thinking, reasoning, remembering etc. For a volumetric video experience, a cognitive experience is highlighted by the ease and depth with which information is exchanged.

- **Perceptual experience**
  - Perception of overall experience. A perceptual experience provides the basic components for the formation of a concept. Perceptual quality can be determined by the physical characteristics of the media and their interaction with the user's physiology leading to a quality judgement. Therefore, corresponding responses are driven primarily by immersion or as with a volumetric video experience – the sense of presence.

5.1.2 What To Measure

Measure a RESPONSE that is based on –

a) Subjective components – a) Symptoms of motion sickness – nausea + disorientation, b) different dimensions of presence – perceptual realism + level of engagement + spatial and social presence, c) Eye strain or headaches, d) distraction These components ARE NOT measured, but user feedback is a rating instead.

   Overall Experience (including usage interest) rating – Users are enthusiastic about overall system and agree that this type of system would be a useful extension to methods of collaboration and communications and would like to use it for the implemented use case.

   Sense of Spatiality rating – For the video hologram mainly. Could also apply to the overall VR content as well. A proper sense of space alike the physical world in terms of position, size, shape, and direction.

Reference Object Representation rating – Quality of the reference objects in a scene. This is an optional evaluation. Such objects help with creating a perspective (orientation, especially) shared between users at both ends.

Comprehensibility of Body Language and Gaze Communication – While communications through in-session spoken language depends on audio quality and sound level, challenges may appear more with understanding correctly and in-time what is indicated through gestures or gaze based communications. It is likely that gaze communications are no better than head movement.

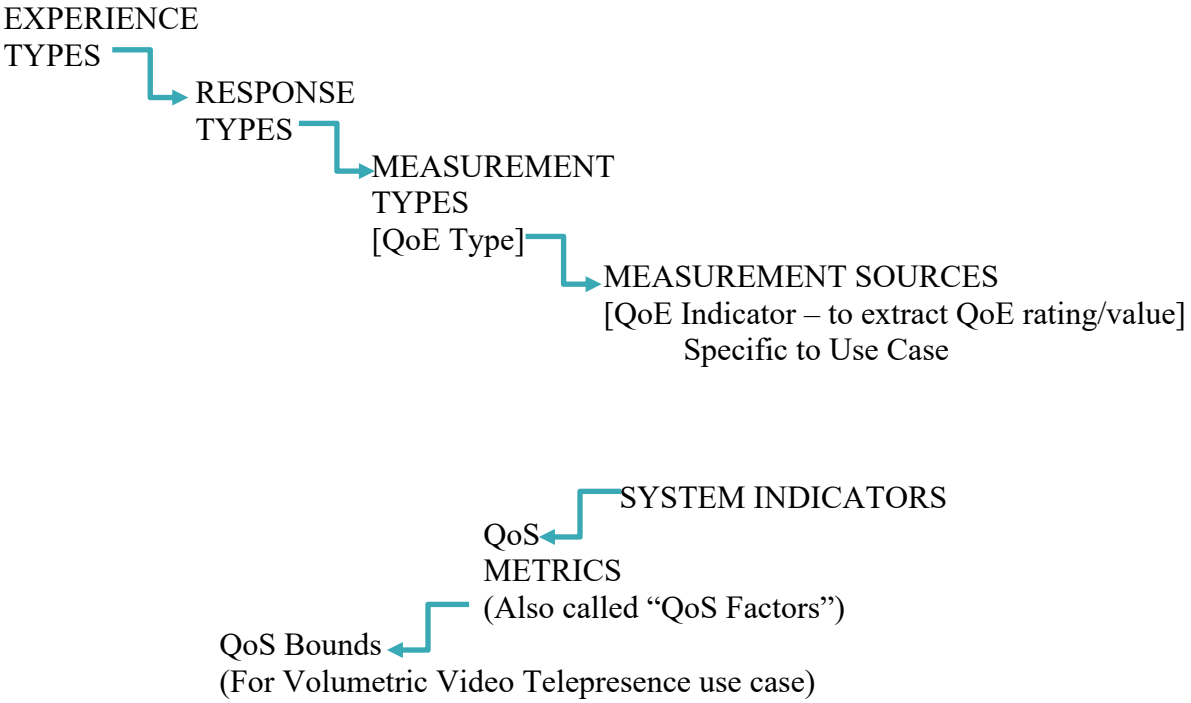
Illusion of Physical Co-Presence – Captured through high visual quality of a 3D avatar or hologram. It appears that the remote participant is locally present without being able to differentiate against a real human. A complete and accurate reconstruction and its presentation may be less critical if primary objective is more communication centric. In such a case recognizing subtle cues or body language is more critical than high visual quality.

b) **Performance** of the volumetric asset in achieving its intended goal. While this can be specifically defined for a 3D avatar as measurement of its accuracy and speed of response to user input or its ability to guide a user through a task, for a real time telepresence use case it primarily entails speed of verbal response or hand gesture to another user's action. As such, network latency performance is a dependency. During collaboration, if content sharing between the two users is heavily data centric due to the richness in detail (resolution), network bandwidth availability becomes critical.

c) **Usability** of the volumetric representation – extent of body movement (accuracy in identifying using pointing gesture), level of photo-realism, audible and human tone, detail from all angles etc. High usability allows high INTERACTION.

d) **User Behavior** – Characterized by level of anxiety/trust, confidence, interest and engagement or attention level, emotional arousal, comfort etc.
   i. Eye movement
   ii. Heart rate
   iii. Facial expressions
   iv. Eye strain /headaches

5.1.3 QoE Definition Structure Taxonomy



Past use cases were limited in definition by a single type of media that delivered the content. Metaverse use cases, particularly volumetric video based influence users in a multi-fold of ways that require broadening of experience ranges identified together with their sources or stimuli. QoE assessments need to consider technical factors contributed to by the supported network delivering the experience, along with user perception factors (dependent on levels of immersion primarily) and other remaining factors that also influence users but are more application centric.

Table 1: Defining Use Case Quality of Experiences

EXPERIENCE TYPES	RESPONSE TYPES	MEASUREMENT TYPES [QoE type]	MEASUREMENT SOURCES [QoE Indicators]
Emotional	*Immersion Attitude*	i. Subjective **ii.** User Behavior	• Overall Experience rating • Phantom Arrays [Causing Eyestrain / Causality Effect] • Stroboscopic Effects [Impaired Vision / Causality Effect] • Sickness & Headaches

			• Observable Factors - Eye movement, Facial expressions, Heart Rate.
Cognitive	*Attitude Satisfaction*	i. Usability ii. Performance	• AI/ML animation to enhance movement [Body movement – pointing gesture]. • Voice quality and tone. • 3D viewpoints update. • Phantom Arrays [Errors in visual understanding/ Causality Effect] • Response time due to Temporal/Spatial artifacts [Causality Effect]
Perceptual	*Immersion Satisfaction*	i. Usability ii. User Behavior iii. Subjective	• Flicker [Lighting quality at source environment impacts usability / Causality Effect] • Physical Co-Presence rating • Observable Factors - Eye movement, Facial expressions, Heart Rate. • Overall Experience rating • Stroboscopic effects [Interference Levels upon local light sources - impacts clarity of understanding from target viewer's perspective / Causality Effect] • Flicker [Distraction Levels from visual unsteadiness / Causality Effect]

The above QoE experience types and corresponding measurement sources appear to also fall in line with the TIP-MRN QoE and QoS Framework Development sub-group's identification of necessary QoE categories required to make any Metaverse based use case flourish.

Based on Figure 3 (QoE metrics breakdown in three fundamental categories) of the TIP-MRN QoE Framework sub-group's technical paper *An Approach to Quality of Experience Engineering*, we confirm the following relationship –

Table 2: QoE Type to QoE Category mapping

QoE Type	QoE Category
Emotional	Human
Cognitive	Human, System, Context
Perceptual	Human, System

5.1.4 Additional Critical "Measurement Sources"

1. QoE Indicator - CONTINIOUS MOTION PARALLAX
    a. "Experience Types" influenced:
        i. Emotional: YES
        ii. Cognitive: NO
        iii. Perceptual: YES
    b. DESCRIPTION: Smooth update of viewpoint object including objects that are closer. Displacement of user recorded for all objects in scene should be accurate relative to each other, and not updated with discrete movement, as user moves. It is possible to look into or behind objects, and in doing so, hidden details become visible while other parts of the displayed object disappear, as it happens with real 3D experience.
    c. QoE ailment: 3D image does not jump between the views. No "bumpy" experience or loss of focus.
    d. QoE Requirement: Should apply to both horizontal and vertical perspectives.

2. QoE Indicator - Contradiction in Eye Convergence w/ Focusing.
    a. "Experience Types" influenced:
        i. Emotional: YES
        ii. Cognitive: NO
        iii. Perceptual: NO
    b. DESCRIPTION:
    c. QoE ailment: Disorientation, seasickness, discomfort
    d. QoE Requirement: Use technique or technology to modify the eye convergence to reduce eye accommodation and enhance the comfort in viewing 3D video.

3. QoE Indicator – Capability to address individual voxels.
    a. "Experience Types" influenced:

i. Emotional: NO
ii. Cognitive: NO
iii. Perceptual: YES

b. DESCRIPTION: Position dependent effects. The point of a given view does not move if the viewer is moving user is moving and is exactly there where it seems to be. Voxels with non-isotropic radiation profiles.
c. QoE ailment: Inaccuracy
d. QoE Requirement: Accuracy in creating user position dependent effects.

4. QoE Indicator – Ensuring Absence of Invalid Zones in Field of View
   a. "Experience Types" influenced:
      i. Emotional: NO
      ii. Cognitive: YES
      iii. Perceptual: PARTLY?
   b. DESCRIPTION: The volumetric content can be seen over entire field of view.
   c. QoE ailment: Incomplete collaboration.
   d. QoE Requirement: Generate all light beams to make the 3D view of the displayed object visible in the whole field of view.

5. QoE Indicator – Wide viewing angle (optional - viewing distance [tbd])
   a. "Experience Types" influenced:
      i. Emotional: YES
      ii. Cognitive: YES
      iii. Perceptual: YES
   b. DESCRIPTION: Required so that realistic 3D models can be presented with natural behavior capture.
   c. QoE ailment: TBD
   d. QoE Requirement: 360 degrees horizontal / 180 degrees + for vertical

6. QoE Indicator – Generation of points anywhere in FoV
   a. "Experience Types" influenced:
      i. Emotional: YES
      ii. Cognitive: YES
      iii. Perceptual: YES
   b. DESCRIPTION: Required so that realistic 3D models can be presented with natural behavior capture.
   c. QoE ailment: TBD
   d. QoE Requirement: 360 degrees horizontal / 180 degrees + for vertical

5.1.5 Causality Effects - Other Measurement Sources and their Influence on QoE

Undesirable motion artifacts that are perceptible adversely impact QoE. Their impact can be minimized through higher refresh rates at the device end. Based on use case, they can be made tolerable to the human viewing experience even at frame rates almost 50% of what is needed to entirely eliminate them.

Temporal Light Artifacts [4] – Should be absent, or if present, be so at levels that minimally influence the QoE in an adverse sense. Such artifacts are undesired changes in visual perception, induced by a light stimulus sourced out of the system setup, whose luminance or spectral distribution fluctuates with time to the user.

- Temporal Aliasing (Flicker) – Perception of visual unsteadiness induced by light stimulus whose luminance or spectral distribution fluctuates with time.
    - Can create visual unsteadiness that is distracting.
    - Can reduce the immersive level if scene includes light sources such as lamps, interior lighting, and stationary assets within the scene that are under study as part of the use case, with fixed levels of brightness.
    - May be limited in effect for a real time tele-presence use case setup.

- Temporal Aliasing (Stroboscopic Effect) - Change in motion perception induced by a light stimulus whose luminance or spectral distribution fluctuates with time.
    - Can cause user eye strain and headache leading to impaired vision (temporary) and therefore performance.
    - Effects can interfere with local light sources.
    - Both other effects will influence immersion and interaction capability.

- Phantom Array – Perception of a spatially extended series of light spots when the user's eye moves across a light source that fluctuates with time.
    - Adversely effects cognitive performance and can create feelings of sickness.

5.1.6 QoS Metrics

To provide the desired user QoE, volumetric video transmission involves large amounts of data to adequately represent 3D objects and scenes. Therefore, system implementation of the capture, delivery and consumption processes requires proper optimization without sacrificing the perceptual quality and other experience components. With that point in mind, the following table maps QoE influencing system level indicators to relevant QoS metrics that would need to be met to provide the overall quality of experience.

Table 3: QoS/QoE Relationship

System Indicators	Factors (QoS Metrics)
Immersive Level Indicator	Degree of Freedom (Application metric)
Streaming Quality Indicator	Bit Rate (NW Metric)
Real-World Feel Indicator	Texture Quality (Content metric) Point Density (Content metric)
Application Performance Indicator	Loading Delay (Application metric)
Latency Quality Indicator	Motion to Photon Latency (NW metric) Glass to Glass Latency (NW metric) Motion to Motion Latency – RTT (NW metric)
Media Quality Indicator	Frame Rate (Application metric) FOV (Application metric)

5.1.7 QoE - Volumetric Video Display System

Ideally desirable – A display system enabling the rendering of most if not all optical cues needed for human visual capabilities to perfectly see projected 3D images of real world objects.

However, such a setup to support 3D viewing may be cost prohibitive and burdensome upon network QoS implementation to meet the QoE needed for its users. Because volumetric video media type based applications do not necessarily require a volumetric display system, any available “3D display system” can be used, particularly the ubiquitously available stereoscopic technique based VR headsets or AR glass-wear. However, not all device types support features with capabilities to deliver the desired levels of QoE.

5.1.7.1 Supporting Display Technologies

5.1.7.1.1 Restricted 3D rendering (Stereoscopic)

- Virtual headsets (near eye displays) – Oculus 2, Magic Leap 2
- 3D enabled mobile application / 3D movie theatre.

Modern head-mounted stereoscopic displays (AR/VR headsets) stabilize the eye position relative to the display screens, allowing them to render a 3D scene with a significantly smaller number of required rays. As stereoscopic displays can quite seamlessly integrate existing imaging and rendering techniques, there kind for 3D display devices have been commercially developed the most. The most basic design of a stereoscopic display works by showing a separate planar image to each eye to create a stereo vision, while maintaining a decent spatial resolution. This is, for instance, the case with many commercial head-mounted displays (HMDs), such as Oculus Quest 2 and HTC Vive Flow, but such a design-lacks proper focusing cues.

Emerging head-mounted display products use large numbers of full parallax video views to provide the user a strong immersion impression through a small, head-mounted optical display in front of each eye (binocular HMD). However, relative to smart phones, they are inferior display resolution wise, and in general restricted to a less than desirable field of view.

QoE challenges associated with comfort (Impacting **emotional** and **perceptual** experience types)

- Inconvenient
- Hygiene concern
- Dizziness (Lacking focusing cues) – Due to “Vergence-Accommodation Conflict” symptoms.
- Interruption of conversation

QoE challenges associated with comfort (Impacting emotional and **perceptual** experience types)

- Limited viewing angle

Table 4: Essential QoE Support Capabilities (Stereoscopic Devices)

QoE Indicator	QoE Influenced	Supportability
Continuous Motion Parallax	No discrete changes in image with user position movement, especially to maintain horizontal perspective	Required
User eye convergence and focusing match [No contradiction]	Disorientation, sickness/dizziness	Required [6]
Voxels can be individually addressed	Image shape or relative position does not change with user motion. 3D rendering accuracy.	N/R
3D points can be generated anywhere in the FoV	• Additional objects behind primary image, including hidden images. • Additional objects in FoV	Both required
Wide viewing angle (Degrees)	Functional support for Collaboration themed use case. Minimal Requirement is required.	110 (Horizontal) * 95 (Vertical) 110 (Diagonal)
Eye or Head tracking	For free movement. MTP becomes a factor.	Required
No invalid zones in FoV	Incomplete image if enough laser sources are not present or supporting material is ineffective (non-uniform excitation)	Required
Direction selective light emittance by each pixel	Clarity of image – capture of realism in natural view. All light beams as necessary are present	Required
High Refresh Rate	Required to avoid disorientation Higher required to enable collaboration with high interactivity. (Validate against 144 Hz and unchanged QoE with lower rates).	Required – 120 Hz min. (144 Hz recommended)
Display Resolution (x, y pixels)	Clarity and sharpness of image.	4096 x 2160

* For extended peripheral vision allowing the capture of the immersive world

5.1.7.1.2 Auto Stereoscopic

- Headset/glass-wear free
- Limited viewing angle

Includes – a) Volumetric display, b) Light Field display, and c) Holographic display

5.1.7.1.2.1 Volumetric Display

Volumetric displays offer an inherently different mode of interaction, providing the opportunity for a group of people to gather around the image, which is a 3D display, and

interact in a natural manner without 3D glasses or headset. In its true sense, the 3D object that is re-created, does not exist on a screen, but instead, exists physically in a volume or 3D space. The 3D image is given shape complete with surface attributes through forms of emission, scattering, or illumination within a well-defined physical space outlined by x, y, z coordinates.

The following 3D characteristics must be ensured by volumetric display –

- Object is viewable from all directions.
- Has perspective to distinguish parts of the object closer to the viewer from those behind.
- Has precise focal depth to ensure focus (focus cues)
- Exhibits motion parallax.

Limitations with volumetric displays –

- Difficult to show occlusions as voxels appear semi-transparent
- While object is indeed viewable from all directions, as light is uniformly emitted or reflected in all directions, surface appearance is not viewing angle dependent.
- The depth of the object or scene is limited to the physical display volume. Such devices have a larger viewing zone, but with limited depth.

Display Type 1: “Static Volume” volumetric display (Addressed media projection – gas, liquid, solid)

- Headset/glass-wear free
- High speed laser transmitting photons (vs. rasterized scanning)
- Viewing angle – Wide

Key requirements and design components (Implementation)

- Basic Concept: A static volume volumetric display system based off an addressed media projection is implemented by transmitting multiple beams of laser with differing wavelengths across the spread of voxels on a host material. The host material may be a choice of either gas or solid predominantly. When done at an appropriately high rate and with a large array of lasers to specifically (with address) define a correspondingly large number of voxels, a 3D volumetric image is created with a wide field of view, with no flicker, and with a high resolution.
- Based upon w infra-red laser sources for each voxel targeted. The two lasers shoot

photons with differing wavelengths that become incident upon an optically active material. The atom, ion or molecule that represents each voxel is consequently excited.

- The photon from the first laser defines the targeted voxel for the successive laser(s). It is therefore called the "Addressing Source" laser and has a wavelength λ_1. This photon excites the atom to elevate its energy from the ground state level to the first excited state level.
- The photon from the second laser with wavelength λ_2 further excites the same voxel to provide it with the appropriate display exhibit as is represented by the corresponding point of the original 3D device it is re-creating. It is therefore referred to as the "Imaging Source." The excitation energy of the targeted voxel increases further.
- For a properly selected optically active material, the total energy absorbed is subsequently radiated out (minus the internal losses) as the ion/atom returns to its ground energy state. This results in the generation of a visible photon, manifested as a spot of light within the volume of the display material.
- A full image is created by "addressing" multiple volume pixels (the voxels).

Critical enablers required to ensure QoE:

- No Flicker - As the light emitted from each voxel disappears quickly after the addressing source is terminated, the entire image must be refreshed at a sufficient rate to the deceive the human eye into seeing a cohesive image.
    - No flicker. Ensure fast refresh rate (30Hz – 100 Hz).
- High Resolution – A large array size for the laser sources enables the system to address many voxels, thereby enabling the generation of a high resolution image.
- Large Viewing Range – The host material selected for the implementation of the display shall be capable of generating multi-color imagery and not be limiting the viewing range of the image displayed.

Table 5A: Essential QoE Support Capabilities (Static Volume Devices)

QoE Indicator	QoE Influenced	Supportability
Continuous Motion Parallax	No discrete changes in image with user position movement, especially to maintain horizontal perspective	Required – both vertically and horizontally
User eye convergence and focusing match [No contradiction]	Disorientation, sickness/dizziness	Required
Voxels can be individually addressed	Image shape or relative position does not change with user motion. 3D rendering accuracy.	Required
3D points can be generated anywhere in the FoV	• Additional objects behind primary image, including hidden images. • Additional objects in FoV	• May be technology restricted. • May be possible with limited scope
Wide viewing angle	Functional support for Collaboration themed use case	Required
No Eye or Head tracking dependency	Latency impact averted, and allows free movement	Required
No invalid zones in FoV	Incomplete image if enough laser sources are not present or supporting material is ineffective (non-uniform excitation)	Required
Direction selective light emittance by each voxel	Clarity of image – capture of realism in natural view. All light beams as necessary are present	Required
High Refresh Rate	Required to avoid disorientation Higher required to enable collaboration with high interactivity.	60 Hz (Minimal for human eye FPS processing capability)
Resolution (x, y, z) in voxels	Clarity and sharpness of image.	50 million [7] (60 Hz refresh rate)

Display Type 2: "Swept Volume" volumetric display (Layered implementation – Multi-layered)

- Headset/glass-wear free
- High speed digital light engine
- High speed optimized volume rendering engine
- Viewing angle – unlimited. True volumetric.

Implementation

- Basic concept: Project millions of voxels (light points) onto physical space per second.
- 3D data from captured physical world object is layered (sub-sets of data). Layer count should be very high >> 100. Each layer is projected one at a time onto a high speed reciprocating screen. Due to "persistence of vision" the human eye blends the images together, to perceive what then is a 3D image, viewable from any angle.

Critical enablers required to ensure QoE:

- High speed software code that converts 3D geometry into a 3D array of voxels. 3D geometry created externally that is to be rendered is sliced into digital horizontal cross sections and then projected in sequence. Photons belonging to each sequence of emission are diffused upon impact on selected screen to create the physical cross section (snapshot) of the 3D object. The very fast rate of transmission creates the impression of the 3D object from the perspective of human vision through the phenomenon known as "Persistence of Vision." The human eye blends hundreds of layers together to capture a true three-dimensional representation.
- Very high speed projection rate of voxels onto the diffuser scale – 4000 FPS.
- No chromatic aberration which causes blur in re-created 3D image. This is particular at edges and will impact emotional and perceptual experience types.
- No optical distortion. This will impact perceptual and cognitive experience types.
- Resolution required: 200 - 500 million voxels [8],[2]

Table 5B: Essential QoE Support Capabilities (Swept Volume Devices)

QoE Indicator	QoE Influenced	Supportability
Continuous Motion Parallax	No discrete changes in image with user position movement, especially to maintain horizontal perspective.	Required – both vertically and horizontally
User eye convergence and focusing match [No contradiction]	Disorientation, sickness/dizziness	Required
Voxels can be individually addressed	Image shape or relative position does not change with user motion. 3D rendering accuracy.	Required
3D points can be generated anywhere in the FoV	• Additional objects behind primary image, including hidden images. • Additional objects in FoV.	• May be possible with limited scope. • May be technology restricted.
Wide viewing angle	Functional support for Collaboration themed use case	Required (Better supported than static volume display)
No Eye or Head tracking dependency	Latency impact averted, and allows free movement	Required
No invalid zones in FoV	Incomplete image if enough laser sources are not present or supporting	Required

	material is ineffective (non-uniform excitation)	
Direction selective light emittance by each voxel	Clarity of image – capture of realism in natural view. All light beams as necessary are present	N/R
High Refresh Rate	Required, to enable collaboration with high interactivity.	120 Hz FPS, minimal. Better supported than Static Volume
Resolution (x, y, z) in voxels	Clarity and sharpness of image.	200 – 500 million (Initial) 4K min. (x, y) on 100 layers (Advanced)

5.1.7.1.2.2 Light Field Devices

Future TIP-MRN work activity item.

Light Field Display devices (also known as *super multi-view displays)* use a very high number of video views to create a replica of the original field of light while typically providing full motion parallax, even if not necessarily in the same amount for the horizontal and vertical directions. No eyewear is needed and, the "vergence-accommodation conflict" is less of an issue.

5.1.7.1.2.3 Holographic Devices

Future TIP-MRN work activity item.

5.2 Proposed QoE and QoS Target Metrics Using Stereoscopic Device

The following tables present QoE requirements relevant to various stages of this use case implementation. Requirements are captured for system components and various QoE indicators to positively influence user QoE. The proposed values focus mainly upon initial values to provide satisfactory levels of experience. Advanced values where proposed are based upon incremental and industry direction and reflect highly desirable quality of experience levels. Tests and data capture will provide additional means to identify or validate advanced values.

Table 6A: QoE Influencing Requirements applicable to "Capture" stage of Volumetric Video based Live and Real Time Telepresence use case

Requirement	Proposed Initial Value	Proposed Advanced Value
Number of Cameras	9 min. *	*Future Work Item*
Camera Depth Output	16 bits [9]	*Future Work Item*
Uniform Depth Range	0.3 – 2 m [9]	*Future Work Item*
Depth Step	1 mm	*Future Work Item*
RGB Sensor/Frame Resolution	1080 p 1920 x 1080	4K 4096 x 2160
Pixel Format	4:4:4	*Future Work Item*
Bits/Color Component (RGB)	3 Bytes total [10]	*Future Work Item*
Frame Rate (in Static Light)	30 FPS min.	60 FPS min.
Frame Rate (in flicker/strobing)	60 FPS min.	200 FPS min.
Tracking Frequency (Sensor)	1000 Hz [11]	*Future Work Item*

* Minimum camera count per H-plane. Human eye FoV[14] is 210° and using 42° FoV cameras with aim to minimize crosstalk due to depth-sensor overlap, an additional 4 will be needed, totaling nine. This will enable the capture of whole surface volumes of objects.

Table 6B: QoE Influencing Requirements applicable to "Production" stage of Volumetric Video based Live and Real Time Telepresence use case

Requirement	Proposed Initial Value	Proposed Advanced Value
Rendering (Frame Output) Rate	60 Hz [11]	90 Hz [11]
PCL Points per Frame (By Fidelity)	100k – Low [12] 988 k – High [12] [13]	*Future Work Item*
3D Mesh Size	Dependent upon above	*Future Work Item*
3D Point Representation	12 Bytes	*Future Work Item*
Depth Resolution	1280 x 720 min	*Future Work Item*
Direction Selective Light Emittance	Capability Requirement	*Future Work Item*

Table 7: Essential QoS Requirements for Evaluation

Requirement	Proposed Initial Value
Latency - Scene in Viewport Update Delay (Fig. 5 for reference) • 1- way peer to peer latency / end-to-end • User to visitor • Pose to Render to Photon • Comparable to Glass (Camera) to Glass (HMD) Latency • Viewport Independent Delivery (Fixed Viewport)	=< 100 ms * (Over Metaverse Ready 5G Network)
Latency User Head Movement to Display Update Delay (Fig. 5 for reference) • Motion to Photon Delay • Viewport Dependent • Sensor to optimized pose prediction /correction to rendering to Display • Optimized Rendering latency < 10 ms (With pose prediction, correction)	< 20 ms**
Downlink Bandwidth (Consumption End Content Delivery)	50 Mbps ***
Uplink Bandwidth (Capture End)	*Future Work Item* (Architecture Dependent Recommendations)
Start-up-Delay (Application Metric)	< 2s
Stall Duration / Re-buffering (Application Metric) [15]	< 10ms

* Utilizing appropriate 5G QoS class.

** This can be the time to display a pre-rendered or based upon pre-rendered volumetric content (3D content may be already available or partially available). Update viewport to match visitor's 6DoF shift.

*** Point Cloud Size = 100,000 & FPS = 30. 4K texture resolution with H.265 compression

Table 8: QoE Influencing Requirements applicable to "Consumption" stage of Volumetric Video based Live and Real Time Telepresence use case

QoE Indicator	QoE Influenced	Experience Type Response Type QoE Type	Proposed Value
Display Resolution	Visual Quality (Objective)	Cognitive + Perceptual Satisfaction Usability	Initial - 4K 4096 x 2160 Advanced - 8K min. or comparable for PPD = 60 w/ H-FoV = 210 deg.
FoV	Visual Quality (Objective)	*As above.*	Initial – 110 deg. Advanced – 210 deg. (Horizontal) [14]
Spatial Pixel Density	Visual Quality (Objective)	*As above.*	3200 PPI min. [16] 6000 PPI (Advanced) [17]
Angular Pixel Density (PPD)	Visual Quality (Objective)	*As above.*	30 (Initial) 60 (Advanced) [17]
Brightness	See Through Capability (Objective)	Cognitive Attitude Usability	200 - 500 nits [20] Variable as required Wide FoV
Display Refresh Rate	Normal Visual Fatigue (Subjective)	Emotional + Cognitive Attitude / Satisfaction Subjective / Usability	120 Hz min. [20] 144 Hz Preferred [18] 360 Hz min. (Advanced) [18]
Display Refresh Rate	With Judder or non-smooth motion. To Tackle Blur. With Multiple	Emotional + Perceptual Attitude / Satisfaction Subjective / User-Behavior	144 Hz [19]

	Imaging. (Subjective)		
Viewport Drift	Visual Quality (Subjective)	Emotional + Perceptual Immersion / Satisfaction Usability / User-Behavior	< 0.1m [15]
Viewport Smoothness	Visual Quality (Subjective)	*As above*	< 0.01m [15]
Perturbation (L2 Norm w/ countermeasure)	Visual Quality (Subjective)	*As above*	< 0.05 [21]

5.3 QoS Measurements in "One-to-Many" Holograms Use Case

This section presents an implemented Volumetric Video use case scenario based on one-to-many holograms, and related QoS measurements obtained from performed tests. The aim is to better understand necessary QoS values which meet stated QoE expectations. The front-end of the application will capture images of the user's head and torso using the device camera. The application's back-end will operate on a Telco Edge and Cloud environment and produce 3D images from the front-end's captured images. These 3D images will then be used to create holograms that can be sent to other users' devices, enhancing the interactive call experience by displaying the hologram of the user captured on the sender device.

In terms of hologram characteristics, the images are taken with the front camera of a mobile phone. As a result, the hologram is a lifelike representation of head and torso which was captured by the sender's device. The hologram's resolution is "adequate" and the required throughput does not increase significantly with the resolution.
Two approaches have been taken regarding the fluidity of the received hologram. The fluidity depends on the Frames per Second (FPS) performed. Therefore, the tests in Table 9 were conducted at 15 and 30 FPS, which provided good fluidity and resolution for both test cases. The number of FPS required is directly proportional to the number of holograms computed per second, which in turn affects the required network

throughput. The throughput increases linearly with each hologram. As Table 9 shows, for 30 FPS, 15Mbps is sufficient, a throughput that is more than feasible within 5G architectures.

Latency also affects the QoE and this does not depend on the number of holograms constructed but mainly in the computing capacity of the Edge and the network performance.

Table 9: Performed QoS measurements for One-to-many holograms use case.

Source: MATSUKO

		Capture Device (iPhone) (Color / Depth)	Connectivity (Upload- Closest Node)		Inter-MNO networking	Backend/Reconstructor (Color / Depth)	Connectivity (Download - Access)		Viewer Device (# of Vertices)
			FPS: 15	FPS: 30			FPS: 15	FPS: 30	
Hologram Resolution Bandwidth ranges are controlled via WebRTC bitrate controller (minimum requirement)	Low	(128 x 96) / (128 x 96)	2-3 Mb/s	4-6 Mb/s		(128 x 160) / (128 x 160)	3-4 Mb/s	6-8 Mb/s	No. of vertices: 15k-20k
	Mid	(256 x 192) / (256 x 192)	3-4 Mb/s	6-8 Mb/s		(256 x 320) / (256 x 320)	4-6 Mb/s	8-12 Mb/s	No. of vertices: 50k-80k
	High	(1024 x 768) / (256 x 192)	4-6 Mb/s	8-12 Mb/s		(1024 x 1280) / (256 x 320)	5-7 Mb/s	10-14 Mb/s	No. of vertices: 50k-80k
Latency (Expected) (high resolution)	4G (Cloud)	(256 x 192) / (256 x 192)	20-40 ms	40-60 ms	25-45 ms	30 ms	20-40 ms	40-60 ms	roundtrip: 190-470 ms
	5G (Cloud)	(1024 x 768) / (256 x 192)	5-25 ms	5-25 ms	20-40 ms	30 ms	10-30 ms	15-35 ms	roundtrip: 130-320 ms
	5G (EDGE)	(1024 x 768) / (256 x 192)	5-25 ms	5-25 ms	10-30 ms	30 ms	5-25 ms	5-25 ms	roundtrip: 110-270 ms
	Wi-Fi (EDGE)	(1024 x 768) / (256 x 192)	5-25 ms	5-25 ms	10-30 ms	30 ms	5-25 ms	5-25 ms	roundtrip: 110-270 ms

As previously mentioned, a higher FPS delves into smoother hologram movement, while increasing resolution allows for better image quality, including the ability to see finer details such as skin pores. A frame rate of 15 FPS and high resolution already can provide a sufficient minimally desirable quality of experience in this regard. Further enhancements should focus on improving the contours and edges of the images, such as the neck, ears, and hair, which can be achieved through AI development. This will

result in an even more realistic hologram.

Based on tests conducted and the above measurement results obtained with TIP- MRN partner Telefonica, MATSUKO will launch a multiparty holographic meeting in early 2024 which supports multiple holograms and interactions among them. This is made possible through advanced capabilities of the 5G network. Such measurements are also being considered as a basis for adopting the application within the 6G infrastructure as part of work under the 6G-XR project where innovative 6G application performances are being validated.

7.Network Enablers

The most notable enablers are as follows:

1. **Quality on Demand (QoD) API - as defined in CAMARA**

   (GitHub - camaraproject/QualityOnDemand: Repository to describe, develop, document and test the QualityOnDemand API family).

   This API suite provides clients with the option to tailor network capacities according to their needs. To meet the demands of Volumetric Video applications, it is crucial to have adaptive network capabilities that can respond to user requirements. Therefore, it is imperative to offer developers of these sorts of services a standardized interface for accessing such capabilities.

2. **Slicing Capabilities:** One of the novel features of 5G networks is Slicing, which allows the establishment of exclusive end-to-end connectivity streams for individual traffic and tasks. Volumetric Video is an applicable example for this capability as it calls for dedicated and separate networking resources to uphold isolated connectivity for the particular service from the rest of the internet traffic flowing through the provider's network. This way, different connectivity profiles can be defined for Volumetric Video services.

3. **Edge Cloud:** Computation and processing of Volumetric Video application data necessitates powerful computational nodes. Additionally, latency is a critical factor that should be considered for enhancing QoE. Consequently, the Edge Cloud provides a dual benefit by providing high-performance computing nodes that are geographically distinct from the point of data capture yet offer low latency.

These enablers are crucial in fulfilling the requirements of volumetric video applications, providing standardized interfaces, separated networking resources, and enhanced

computational nodes to improve quality of experiences through improved quality of service.

References

[1] Perez, P., Gonzalez-Sosa, E., Gutierrez J., Garcia N. (2022). Emerging Immersive Communication Systems: Overview, Taxonomy, and Good Practices for QoE Assessment, In *Frontiers in Signal Processing '22,* July 7, 2022.

[2] https://voxon.co/making-molecular-chemistry-easy/

[3] Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. Presence: Teleoperators and Virtual Environments, 7 (3), 225–240. https:// https://doi.org/10.1162/105474698565686 Abstract

[4] https://www.designlights.org/wp-content/uploads/2021/02/dlc-shm-2016_measuring-flicker.pdf

[5] https://www.photonics.com/Articles/Real-Time_Volumetric_3D_Imaging_Technology/a58372

[6] Boo H., Lee Y.S., Yang, H., Matthews, B., Lee, T.G., Wong, C.W. (2022) Metasurface wavefront control for high-performance user-natural augmented reality waveguide glasses. In *Nature, '22,* April 06, 2022.

[7] Refai, Hakki H., Static Volumetric Three-Dimensional Display, *Journal Of Display Technology,* 5 (10), October 2009.

[8] Geng, J. (2013). A Volumetric 3D display based on a DLP projection engine. In *Displays, '13,* Issue 1, January, 39-48.

[9] https://dev.intelrealsense.com/docs/depth-image-compression-by-colorization-for-intel-realsense-depth-cameras#2-depth-image-colorization

[10] Qian, F., Han B., Pair J., Gopalkrishnan, V., (2019). Toward Practical Volumetric Video Streaming On Commodity Smartphones. In *HotMobile '19*, February 27-28, 2029.

[11] https://medium.com/@DAQRI/motion-to-photon-latency-in-mobile-ar-and-vr-99f82c480926

[12] 3GPP TR26.998 V17.0.0 (2022-03): "Support of 5G glass-type Augmented Reality / Mixed Reality (AR/MR) devices."

[13] Lee, K., Yi J., Lee Y., Choi S., Kim Y.M., (2020). GROOT: a real-time streaming system of high-fidelity volumetric videos. In *MobiCom '20, Proceedings of the 26th Annual International Conference on Mobile Computing and Networking,* April 2020, Article No. 57, pg. 1-14. https://doi.org/10.1145/3372224.3419214

[14] Cuervo, E., Chintalapudi, K., Manikanta, K. (2018). Creating the Perfect Illusion: What will it take to Create Life-Like Virtual Reality Headsets? In *HotMobile '18,* February 12-13, 2018.

[15] Liu, Y., Han B., Narayan, A., Zhang, Z, (2022). Vues: Practical Mobile Volumetric Video Streaming Through Multiview Transcoding. In *MobiCom '22,* October 24-28, 2022.

[16] https://inquisitiveuniverse.com/2021/01/10/pixel-density-explained/

[17] https://sid.onlinelibrary.wiley.com/doi/10.1002/msid.1378 : Advanced VR and AR Displays: Improving the User Experience (Juhwa Ha, Sangho Kim, Daeho Song, Sangho Park, Jaebeom Choi)

[18] Dixit A., Sarangi S. R., Minimizing the Motion-to-Photon-delay (MPD) in Virtual Reality Systems (2023), In *ArXiv '23,* vol. 2301.10408, https://api.semanticscholar.org/CorpusID:256231063

[19] A. Mackin, K. C. Noland and D. R. Bull, "The visibility of motion artifacts and their effect on motion quality," *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, 2016, pp. 2435-2439, DOI: 10.1109/ICIP.2016.7532796.

[20] https://vr-compare.com/ar

[21] Tang Z., Feng X., Xie Y., Phan H., Guo T., Yuan B., Wei S., VVSec: Securing Volumetric

Video Streaming via Benign Use of Adversarial Perturbation (2020), Proceedings of the 28th ACM International Conference on Multimedia, Oct. 2020, pages 3614-3623. https://doi.org/10.1145/3394171.3413639

List of Abbreviations

Abbreviation	Definition	Description
6DoF	Six Degrees of Freedom	Movement in 3 dimensional space – x, y, z axis together with pitch, yaw, and roll.
DLP	Digital Light Processing	Technology to experience projected images.
FoV	Field-Of-View	Angular extent of the observable world that is seen at any given moment.
MRN	Metaverse Ready Networks	TIP project group developing solutions and architectures for enhancing network efficiencies to deliver immersive user experience.
MTP	Motion To Photon	Latency - Length of time between the user performing a motion (e.g., turning the head to the left) and the display showing the appropriate content for that particular motion (e.g., the content on the head mounted display (HMD) subsequently moving to the right).
PCL	Point Cloud	Set of data points in a 3D coordinate system.
QoD	Quality of Demand	Network feature enabling telecom operator clients and application developers to have more precise control over connectivity quality of certain services so that user experience is enhanced, regardless of operator to which customer subscribes to.
QoE	Quality of Experience	A measure of user experience with a service or application.
QoS	Quality of Service	Quality of service is the description or measurement of the overall performance of a service, such as a telephony or computer network, or a cloud computing service, particularly the performance seen by the users of the network.
TIP	Telecom Infra Project	Global community of companies and organizations working together to accelerate the development and deployment of open, disaggregated, and standards-based technology. solutions.