A First Look at Immersive Telepresence on Apple Vision Pro

Ruizhi Cheng*
George Mason University
Fairfax, VA, USA
rcheng4@gmu.edu

Nan Wu*
George Mason University
Fairfax, VA, USA
nwu5@gmu.edu

Matteo Varvello
Nokia Bell Labs
Murray Hill, NJ, USA
matteo.varvello@nokia.com

Eugene Chai
Nokia Bell Labs
Murray Hill, NJ, USA
eugene.chai@nokia-bell-labs.com

Songqing Chen
George Mason University
Fairfax, VA, USA
sqchen@gmu.edu

Bo Han
George Mason University
Fairfax, VA, USA
bohan@gmu.edu

Abstract

Due to the widespread adoption of "work-from-home" policies, videoconferencing applications (*e.g.,* Zoom) have become indispensable for remote communication. However, they often lack immersiveness, leading to "Zoom fatigue" and degrading communication efficiency. The recent debut of Apple Vision Pro, a mobile headset that supports "spatial personas", offers an immersive telepresence experience. In this paper, we conduct a first-of-its-kind in-depth and empirical study to analyze the performance of immersive telepresence with FaceTime, Webex, Teams, and Zoom on Vision Pro. We find that only FaceTime provides a truly immersive experience with spatial personas, whereas others still operate 2D personas. Our measurements reveal that (1) FaceTime delivers semantic data to optimize bandwidth consumption, which is even lower than that of 2D personas for other applications, and (2) it employs visibility-aware optimizations to reduce rendering overhead. However, the scalability of FaceTime remains limited, with a simple server-allocation strategy that potentially leads to high network delay for users.

CCS Concepts

• **Networks** → **Network measurement**; • **Computing methodologies** → **Mixed / augmented reality**.

Keywords

Network Measurement, Immersive Telepresence, Apple Vision Pro

ACM Reference Format:
Ruizhi Cheng*, Nan Wu*, Matteo Varvello, Eugene Chai, Songqing Chen, and Bo Han. 2024. A First Look at Immersive Telepresence on Apple Vision Pro . In *Proceedings of the 2024 ACM Internet Measurement Conference (IMC '24), November 4–6, 2024, Madrid, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3646547.3689006

1 Introduction

Remote communication is indispensable in contemporary life, even in the post-pandemic era, as evidenced by ~90% of meetings involving remote participants in 2024 [59]. Existing remote communication systems predominantly rely on traditional 2D video-based conferencing. These platforms often lack the ability to convey social signals such as eye contact and body language, leading to inefficient communication [51] and so-called "Zoom fatigue" [66].

Figure 1: (a) Spatial persona on FaceTime *vs.* (b) 2D persona on Webex.

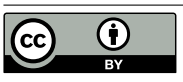
Immersive telepresence is a game changer in remote communication by offering engaging and interactive experiences and is widely recognized as the top use case in the forthcoming 6G [25, 60]. Despite its promise, the commercial availability of immersive telepresence systems has been limited. Although several tech giants have launched a few projects on immersive telepresence [39, 44, 53], with arguably the earliest one dating back to 2016 [53], they largely remain internal endeavors with almost no public access. Meanwhile, academic research in this area typically focuses on in-lab prototypes [28, 34, 37]. The recent debut of Apple Vision Pro [5], a mixed reality (MR) headset that supports "spatial persona", as shown in Figure 1 and introduced in §2, marks a significant milestone in immersive telepresence. Vision Pro allows users to pre-capture their personas, which are 3D human models capable of tracking their hand and head movements in real time.

In this paper, we conduct, to the best of our knowledge, the first measurement study to dissect the functioning and performance of immersive telepresence, focusing on four videoconferencing applications (VCAs) for Vision Pro: Apple FaceTime [12], Cisco Webex [24], Microsoft Teams [50], and Zoom [8]. We summarize our key findings as follows.

• All VCAs assign a server near the initiating user of a telepresence session, potentially leading to >100 ms network delays even when all users are located in the US.

• Only FaceTime offers a truly immersive telepresence experience with spatial personas. Moreover, its bandwidth consumption (<0.7 Mbps) is even lower than other platforms that deliver 2D personas (*e.g.,* >4 Mbps on Webex). The reason is that FaceTime benefits from emerging semantic communication [19], instead of directly streaming 3D content or 2D video.

*These authors contributed equally to this work.



IMC '24, November 4–6, 2024, Madrid, Spain

ACM ISBN 979-8-4007-0592-2/24/11
https://doi.org/10.1145/3646547.3689006

Figure 2: Cameras on Apple Vision Pro.

• The delivery of a spatial persona does not support rate adaption, mainly due to its employment of semantic communication. This is because semantic communication requires all semantic data to be fully delivered for accurate reconstruction, making it challenging to adapt to varying bandwidth conditions [19].

• Spatial personas on FaceTime leverage visibility-aware optimizations [32] to decrease rendering time by up to 59%. Yet, these optimizations are not exploited to reduce bandwidth consumption.

• The scalability of FaceTime remains limited. As the number of users grows, its CPU/GPU processing time increases correspondingly, and the bandwidth consumption rises almost linearly. The GPU processing time reaches ~9 ms per frame when there are five users, close to the 11.1 ms requirement for 90 frames per second (FPS) rendering on Vision Pro [9]. This explains why FaceTime currently supports a maximum of only five spatial personas [14].

Our findings contribute to a comprehensive understanding of the current design and development of immersive telepresence systems and their performance bottlenecks. The source code and data used in this paper are available at https://github.com/felixshing/IMC2024VisionPro. This work has been approved by the institutional review board (IRB) and does not raise any ethical issues.

2 Background

Spatial Persona *vs.* 2D Persona. In immersive telepresence, a persona is a dynamic digital representation of a participant that facilitates interactions with others. Apple Vision Pro's personas capture users' face, hand, and eye movements to make remote communication engaging. It offers two representations: spatial personas and 2D personas. Figure 1(a) shows the spatial persona on FaceTime. It can be viewed from different angles in real time, providing an immersive and interactive experience. As of the time of our measurement study (April 2024), we found that the spatial persona is available on only FaceTime. In contrast, the personas on other applications are still 2D, as shown in Figure 1(b) for Webex. It is generated for a static viewport, functioning as if recorded by a virtual camera in these applications that mimics the selfie camera. This means when a user moves, the display of remote participants' 2D personas does not change accordingly.

Mobile MR Headsets blend digital content with the real world, offering interactive experiences that bridge virtual and physical spaces. Optical see-through devices, such as Microsoft HoloLens 2 [2] and MagicLeap 2 [3], allow users to directly view their environment with digital overlays projected via transparent lenses. On the other hand, video see-through headsets, such as Meta Quest 3 [4] and Apple Vision Pro [5], capture the surrounding environment through their cameras and then display the imagery combining digital and real-world content on their screens. Figure 2 shows the cameras on Apple Vision Pro. The main cameras on the front provide a see-through view of the real world, and the tracking cameras sense the user's position and neighboring objects. The *TrueDepth* cameras can be used to pre-capture the spatial persona offline, and the downward cameras monitor the user's face. Additionally, the internal IR cameras track the user's eyes to offer better experiences, such as enabling eye contact in immersive telepresence.

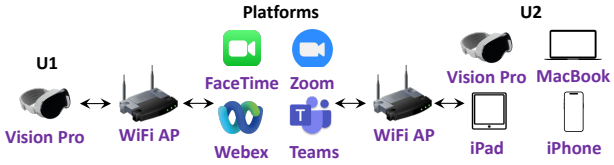


Figure 3: Measurement setup with two users, U1 and U2.

3 Experimental Setup

In this section, we describe the VCAs under investigation, the testbed setup, and the performance metrics of our measurement experiments conducted in April 2024.

3.1 Videoconferencing Applications

We investigate four popular VCAs: Apple FaceTime [12], Cisco Webex [24], Microsoft Teams [50], and Zoom [8]. We choose Apple FaceTime because it supports spatial personas [14], enabling an immersive experience for Vision Pro users. The other three applications have been extensively studied by the research community [18, 33, 45, 48, 58, 65] and are available on Vision Pro.

3.2 Testbed & Data Collection

Figure 3 shows our experimental setup. Unless otherwise mentioned, our experiments involve two users, U1 and U2. U1 is always equipped with Vision Pro, whereas U2 uses Vision Pro, MacBook, iPad, or iPhone. Most experiments are conducted with both users wearing Apple Vision Pro. U2 uses other devices when we test traditional 2D video calls on FaceTime for the protocol (§4.1) and throughput (§4.2) analysis. All devices are updated to the latest version of their operating system. U1 and U2 are connected to two different WiFi access points (APs), each with an average throughput of more than 300 Mbps. We use Wireshark [67] on each AP to capture and analyze network traffic. To assess the performance and resource utilization of Vision Pro, we use Xcode [15] to pair it with a dedicated MacBook where we run Apple's RealityKit tool [9].

Similar to a prior study [48], we collect telepresence statistics using the tools provided by Zoom [72], Webex [23], and Teams [49]. We measure network latency by running TCP `pings` [62] between our WiFi APs and Apple servers for FaceTime, since the servers block regular ICMP pings. We verify that no background process exists on the devices during our experiments. As our measured platforms are primarily designed for video conferencing, users in our experiments are instructed to engage in natural conversations and movements, simulating a typical meeting environment. We repeat each experiment at least five times, and each session lasts at least 120 seconds. In the following, we describe the performance metrics that we study.

• *Throughput:* We measure the throughput of these applications involving up to five participants, which is the maximum number of supported spatial personas on Vision Pro [14].

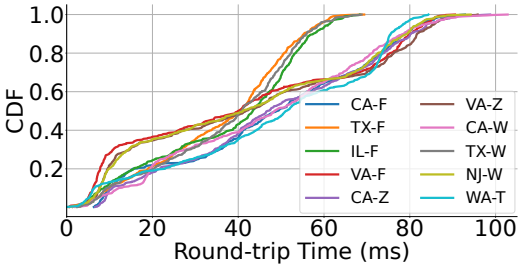


Figure 4: Round-trip time between FaceTime (F), Zoom (Z), Webex (W), and Teams (T) servers and test users. The server locations are indicated by their abbreviations: CA (California), TX (Texas), IL (Illinois), VA (Virginia), NJ (New Jersey), and WA (Washington State).

• *Display Latency:* We measure the difference in display latency between rendering real-world objects and the spatial personas of remote users. Recall that Vision Pro is a video see-through headset. It captures and renders real-world content, and then integrates it with the rendered spatial persona (§2). Thus, we can record the content displayed on the headset to measure this latency.
• *Frame Rate and Rendering Time for Each Frame*: The target FPS of Vision Pro is 90 [9]. We measure CPU/GPU processing time for each frame to identify bottlenecks if a frame misses its deadline.
• *Visual Quality*: On Vision Pro, the 3D model of a spatial persona is represented as mesh [54]. The visual quality of a mesh is influenced by the number of triangles, which are connected to form the geometry of the 3D model. For 2D personas, we resort to measuring the video resolution as done in previous work [18, 45, 48, 58, 65]. For both metrics, the higher they are, the higher the rendering overhead, and the better the visual quality.

4 Measurement Results

4.1 Server Infrastructure

Geolocation. We first investigate the server locations and network latency of the four VCAs. Since Vision Pro is available in only the US as of April 2024, we set up clients in nine different locations across the Western (three), Middle (three), and Eastern (three) US. For each experiment, these nine clients randomly join a VCA in different orders with different types of device. We use MaxMind [46] and ipinfo.io [35] to geolocate the servers we identify. Both tools return the same geolocation results for all tested servers.

We find that FaceTime, Zoom, Webex, and Teams operate four (Virginia, Illinois, California, and Texas), two (Virginia and California), three (New Jersey, California, and Texas), and one (Washington State) server(s) in the US, respectively. Zoom and FaceTime rely on peer-to-peer (P2P) communication, with data transmitted directly between users without involving a server, when there are only two users in a session, except for both users using Vision Pro on FaceTime. We ascertain that none of the servers employ *anycast* [47] by using the approach adopted by prior work [22]. All VCAs consistently assign a server that is closest to the initiating user of each telepresence session. For example, if a user in the Eastern US initiates a session, the server will always be in the Eastern US (if available), regardless of the locations of other participants.

Figure 4 presents the round-trip time (RTT) between the servers of the four VCAs and our test clients. We observe that even though all users and servers are located within the US, the RTT between them can still exceed 100 ms, as observed with the California server of Webex. For servers situated on the west coast (California and Washington State) and the east coast (Virginia and New Jersey), the RTT can be >80 ms when users are located on the opposite coast. Positioning servers in the middle of the US, such as Texas and Illinois, can potentially reduce the maximum RTT, with all RTTs falling below 70 ms. This is because the central location ensures a shorter distance to both coasts compared to the distance between the east and west coasts. However, this strategy may decrease the percentage of clients experiencing low RTTs, given that the majority of the US population resides on the east and west coasts [16], and the lowest RTTs occur when servers are located nearby. For example, only 20% of RTTs for the Texas server of FaceTime are below 20 ms, compared to 38% for its Virginia server.

Implications (1): The above results reveal that the straightforward solution of allocating a single server for all users can result in high network latency. This issue could become more pronounced when users are distributed across continents. For example, the one-way propagation delay between Europe and Asia may already exceed 100 ms [68], the threshold for maintaining a high quality of experience (QoE) in immersive telepresence [40]. A viable solution would be to deploy geo-distributed servers to ensure that each client connects to a nearby server, while inter-server connections are established by a high-speed private network to reduce RTT [22].

Protocols. When all users wear Vision Pro, FaceTime delivers the content via QUIC [63], different from prior studies [52] that reported its use of RTP [57]. However, the transmissions between Vision Pro users and non-Vision Pro users revert to RTP. We verify that its Payload Types (PTs) field, which indicates the audio and video codecs [52, 56], remains consistent with that in traditional 2D video calls on FaceTime. This may be because non-Vision Pro users are unable to render spatial personas. Thus, Vision Pro pre-renders the spatial persona and delivers it with 2D video. The other three applications continue to rely on RTP, even when all participants use Vision Pro, probably because their personas are 2D (§2).

4.2 Throughput Analysis

We next examine the throughput of Vision Pro for the four VCAs. By analyzing the uplink and downlink traffic of each VCA, we find that their servers are primarily used for data forwarding. Thus, the throughput of Vision Pro can be considered mainly as the data rate required by the spatial persona. For FaceTime, we compare the throughput of the spatial persona and that of the 2D persona, as their underlying protocols (§4.1) and types of delivered content (§4.3) are different.

Figure 5 shows our measurement results for two-user experiments, with the 95th, 75th, 25th, and 5th percentiles, median (red bar), and mean (blue dot). Surprisingly, the throughput of a spatial persona is the lowest with ~0.7 Mbps, while the throughput of a 2D persona for FaceTime is ~2 Mbps. Our further analysis (§4.3) indicates this is because FaceTime employs the semantic communication paradigm [19] to optimize bandwidth consumption for spatial personas. Among 2D personas for other applications, Webex consumes the highest bandwidth (>4 Mbps), while Zoom requires only ~1.5 Mbps. This is mainly because of their different resolutions

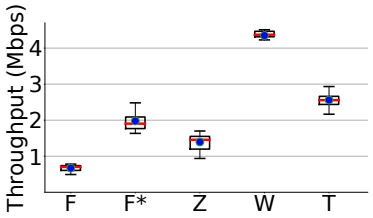


Figure 5: Throughput of FaceTime with spatial persona (F), FaceTime with 2D persona (F*), Zoom (Z), Webex (W), and Teams (T) with two participants. Blue dots represent mean values.

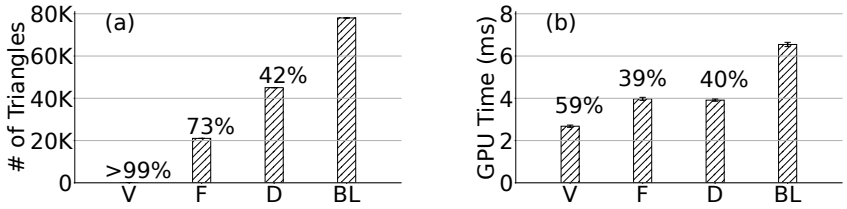


Figure 6: Number of triangles (a) and GPU processing time per frame (b) for the rendered spatial persona with various optimizations: viewport adaptation (V), foveated rendering (F), and distance-aware (D). The baseline (BL) is when the user stares at the spatial persona at a close distance of half a meter. The optimizations reduce the number of rendered triangles, leading to decreased GPU rendering time.

for 2D personas (1920×1080 on Webex *vs.* 640×360 on Zoom). Additionally, the different video compression approaches utilized by these applications may affect bandwidth consumption [52]. Note that while the 2D persona has a background, as shown in Figure 1(b), we observe that it is static and consistent across different applications, suggesting that it does not need to be delivered.

4.3 What is Being Delivered?

Immersive telepresence could use three approaches for 3D content delivery: 1) direct streaming [28, 37], 2) pre-rendering to 2D video before delivery [42], and 3) delivery of semantic information [19]. We next examine which method is used for spatial personas on FaceTime, the only VCA currently supporting this feature (§2).

Direct 3D Data Streaming. This approach involves sending the 3D model of the spatial persona to the receiver, who then renders it for display. Due to the data-hungry nature of 3D data, this approach may consume excessive bandwidth (*e.g.,* >1 Gbps [53]).

The RealityKit tool [9] shows that the 3D mesh of a spatial persona consists of 78,030 triangles, representing the complexity of the mesh [54]. To estimate the bandwidth requirements for streaming the spatial persona on Vision Pro, we select ten different meshes of human heads from Sketchfab [1], with the number of triangles varying from ~70K to ~90K. We compress these meshes using Draco [27], a 3D data compression tool widely used in telepresence systems [28, 37], and stream them at 90 FPS, the target frame rate of Vision Pro (§3.2). We find that the bandwidth consumption is 108.4±16.7 Mbps, even without texture (*i.e.,* the surface details of 3D mesh [54]), drastically higher than ~0.7 Mbps consumed by a spatial persona (Figure 5). It follows that the spatial persona is currently not delivered using the 3D mesh format.

Streaming of 2D Video. When the delivered content is 2D video, it could be directly captured by the sender or rendered from the spatial persona of the sender (*e.g.,* according to the predicted future viewport of the receiver [42]). We find that the content is not the video captured by the sender as a change in the sender's appearance (*e.g.,* a sticker on the face) is not communicated to the receiver.

Next, we investigate if the delivered content is the pre-rendered spatial persona. We cannot use throughput measurements to determine this, given that we do not have direct access to the resolution of the rendered spatial persona. A distinguishing characteristic of delivering pre-rendered immersive content is that the display latency between the local real-world objects and the spatial persona at the receiver (§3) should be influenced by the network delay. For example, if the network latency is high and the receiver changes the viewport, it will significantly delay the display of the pre-rendered spatial persona of the sender for the new viewport.

To measure the difference in display latency, we record the content displayed on U1's Vision Pro, which includes both local real-world objects and U2's spatial persona. We let U1 abruptly change the viewport to observe a different portion of U2's spatial persona from a new angle. For example, before changing the viewport, U1 views U2's spatial persona from the front, where only its front face is visible. Then, U1 quickly shifts the viewport to the left, observing one side of U2's spatial persona, including the entire ear. As U1 changes the viewport, we measure the time taken to render the newly emerged real-world objects and U2's spatial persona to determine the difference in display latency. We use Linux `tc` [6] to introduce extra network delays ranging from 0 to 1,000 ms between U1 and U2. Our experiments indicate that the measured difference in display latency remains consistent (<16 ms), suggesting that the delivered content is not 2D video captured/rendered by the sender.

Delivery of Semantic Information. Semantic communication is an emerging content delivery paradigm. For immersive telepresence, it involves sending only the meaningful semantic data of remote users to the receiver, who then reconstructs the 3D representation (*e.g.,* mesh) of remote users using the received data [19].

For the human body, keypoints represent a primary choice for conveying semantic information [19]. Given that the spatial persona primarily includes the head and hands (Figure 1), we explore the bandwidth requirement for delivering keypoints in these areas to verify whether semantic communication is the used method. Specifically, we utilize a ZED 2i RGB-D camera [7] to capture a video of 2,000 frames containing the head and hand regions of the user. We employ the widely used 68 facial keypoints from *dlib* [38] and 21 hand keypoints from *OpenPose* [17]. As the spatial persona primarily tracks the eye and mouth areas for facial expressions, as well as hand movements, we compress the 32 (mouth & eyes) + 2×21 (hands) = 74 extracted keypoints using LZMA [30] and stream them at 90 FPS. The average throughput is 0.64±0.02 Mbps, close to the bandwidth consumed by a spatial persona (0.67 Mbps on average). This suggests that FaceTime utilizes semantic communication to optimize bandwidth consumption for spatial personas.

Although semantic communication consumes less bandwidth than 2D/3D content streaming, it leads to challenges in supporting

rate adaptation. We conduct experiments by using Linux `tc` [6] to constrain the bandwidth. When the uplink bandwidth is 0.7 Mbps, the spatial persona becomes unavailable, with "poor connection" displayed on the screen. This may be because semantic data must be fully delivered for successful content reconstruction [19]. While currently, the bandwidth consumption for spatial personas on FaceTime is relatively low, rate adaption may still be necessary. The spatial persona does not yet provide a fully immersive experience, such as a photorealistic representation. At present, it captures only the head and hands and relies on an avatar-based model rather than lifelike representations of actual users (Figure 1). Achieving a high-quality, full-body immersive representation could demand significantly higher throughput than what we have observed [22].

Implications (2): We identify that the spatial persona on FaceTime utilizes semantic communication to reduce bandwidth consumption. However, semantic communication is not a silver bullet for immersive telepresence and has its own technical challenges. For example, since it requires all semantic data to be successfully delivered for reconstruction [19], semantic communication is not resilient to data loss and makes rate adaptation challenging. Additionally, as semantic information is inherently sparse, the reconstructed content may suffer from a loss of fidelity. Conversely, other streaming approaches, such as direct 3D data streaming and 2D video streaming, have their own limitations, as discussed earlier. A potential solution is to have servers intelligently select different streaming approaches for each client based on their available network and computational resources [42].

4.4 Visibility-aware Optimization

Visibility-aware optimizations can drastically reduce communication and computing overhead in immersive video streaming [32, 55, 69]. However, there is limited research on their adoption in commercial products. To fill this critical gap, we investigate the potential deployment of various visibility-aware optimizations for spatial personas on FaceTime. We analyze the number of triangles of rendered meshes for a spatial persona, indicative of its visual quality (§3.2), along with CPU/GPU processing time and bandwidth consumption. As a baseline, we consider U1 viewing U2's spatial persona from the closest distance (approximately half a meter), at which U2's entire persona just fits within U1's screen. In this scenario, no visibility-aware optimization should be applied.

We experiment with the spatial personas for the following possible optimizations: 1) viewport adaptation, 2) foveated rendering, 3) distance-aware optimization, and 4) occlusion-aware optimization. We find that the first three optimizations are employed to reduce the number of rendered triangles and thus decrease GPU processing time, as shown in Figure 6. Next, we will detail our conducted experiments and discuss the potential for further optimizations.

Viewport Adaptation processes only content in the user's viewport [32, 55]. We verify whether Vision Pro adopts it for spatial persona by having U1 turn the head to make U2's spatial persona out of U1's viewport. Our results show a decrease in the number of rendered triangles, from 78,030 to 36, and a 59% reduction in GPU rendering time per frame, from 6.55±0.11 ms to 2.68±0.05 ms.

Foveated Rendering benefits from the human visual system [64] to render with the highest visual quality for only foveal content around the center of the eye gaze and lowers the quality toward the periphery [29]. In our setup, U2's spatial persona appears at the left corner of U1's viewport when U1 gazes toward the right corner, placing U2's persona in U1's peripheral vision. This results in a 73% reduction in the number of rendered triangles (21,036), and a 39% decrease in GPU rendering time per frame (3.97±0.07 ms).

Distance-aware Optimization adjusts the rendered 3D content based on viewing distance [32]. We vary the viewing distance from half a meter to ten meters in increments of half a meter. Beyond three meters, a lower quality spatial persona is displayed, with the number of rendered triangles reduced by 42% to 45,036, and the GPU rendering time per frame reduced by 40% to 3.91±0.05 ms.

Occlusion-aware Optimization reduces the quality or omits the rendering of occluded content [32]. We experiment with five Vision Pro users, U1 through U5, and arrange U2 to U5 in a line, with U1 observing the rest from the front. If occlusion-aware optimization is implemented, the spatial personas of U3 to U5 should not be rendered on U1's Vision Pro, as they are occluded by U2. However, compared to the case where all users are visible, we do not observe a reduction in the number of rendered triangles and GPU processing time for U1, indicating that this optimization is not adopted.

Despite the adoption of several visibility-aware optimizations by Vision Pro for spatial persona, this does not translate into a reduction in bandwidth consumption and CPU processing time compared to scenarios without these optimizations. It suggests that optimizations are applied solely at the rendering stage but not during content delivery. The lack of bandwidth optimization might explain why the CPU processing time remains unchanged, since the CPU on Vision Pro is tasked with processing the received data, as indicated by the RealityKit tool [9].

Implications (3): Our measurements indicate that FaceTime employs several visibility-aware optimizations to reduce computational overhead for spatial personas. However, it has not yet implemented occlusion-aware optimizations, which could be beneficial when multiple users and/or objects are present within the same scene. Moreover, these visibility-aware optimizations do not benefit the data transmission stage. Nevertheless, implementing such optimizations to reduce bandwidth consumption is feasible. For example, if the content is known to fall outside of a receiver's viewport, it could be omitted from delivery to conserve bandwidth [32, 55, 69]. These optimizations could be further applied in immersive telepresence systems to reduce bandwidth consumption.

4.5 Scalability Analysis

We finally investigate the scalability of spatial personas on FaceTime by measuring the throughput and rendering overhead as the number of users increases. Specifically, we have at most five Vision Pro users joining a telepresence session, the maximum number currently supported by FaceTime [14]. The available bandwidth for each user is at least 100 Mbps. Figure 7 shows the number of rendered triangles, CPU/GPU processing time, and downlink throughput as a function of the number of concurrent users.

Although increasing the number of spatial personas almost linearly raises the average number of rendered triangles, the 5th percentile for five users remains almost the same as that for three users, as shown in Figure 7(a). This can be attributed to the visibility-aware

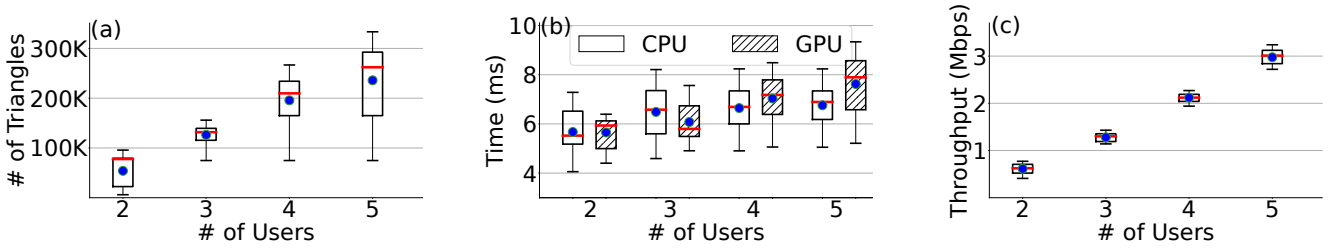


Figure 7: Number of rendered triangles (a), CPU/GPU processing time (b), and downlink throughput (c) of spatial personas for FaceTime with the number of users varying from 2 to 5.

optimizations adopted by FaceTime (§4.4). For instance, as the number of spatial personas increases, some of them may appear in the peripheral regions of the visual field, which will be displayed as a low-quality mesh with few triangles due to foveated rendering.

Despite the implementation of various visibility-aware optimizations, the GPU rendering time still increases by an average of 34.9% from two users (5.65±0.69 ms) to five users (7.62±1.29 ms), with the 95th percentile >9 ms, as shown in Figure 7(b), which is close to the rendering deadline (*i.e.,* ~11 ms for 90 FPS). This likely explains why FaceTime currently supports a maximum of five spatial personas. We also observe the CPU processing time increases by an average of 19.2% from two users (5.67±0.69 ms) to five users (6.76±1.29 ms). Figure 7(c) reveals that the downlink throughput of spatial personas almost linearly increases with the number of users. This is because the server just simply forwards the data (§4.1).

Implications (4): The scalability issues related to resource utilization and bandwidth consumption for spatial personas on FaceTime significantly impede its ability to support a large number of participants. A potential solution to address such scalability issues is to utilize remote rendering by offloading the GPU-intensive rendering process to the cloud [22]. By having the server handle rendering, even with many concurrent users, the server can render them into a 2D video frame. This ensures that the transmitted data remains independent of user numbers, mitigating scalability issues.

5 Discussion

Fully-automated Measurement Experiments. To the best of our knowledge, no existing tool can automatically play back predefined user inputs on Vision Pro. Thus, we resort to manual experiments in this study. A potential method for automating experiments is to attach Vision Pro to a robotic arm [71]. However, this may cause the spatial persona not to function, as it needs to track users' facial changes. We plan to build open-source tools for Vision Pro to facilitate automated and large-scale crowd-sourced measurement experiments in the wild.

Content Decryption. To know exactly the delivered content for a spatial persona, a promising solution is to decrypt the content. However, FaceTime utilizes QUIC [63] to deliver spatial persona (§4.1), which is encrypted by TLS 1.3 [70]. As spatial persona is end-to-end encrypted [10], simply utilizing the man-in-the-middle attack cannot get the TLS certificate, and thus it is challenging to decrypt the content. Instead of relying on content decryption, analyzing IP headers [58] and packet transmission patterns [48] may help better understand the delivered content for spatial persona.

Other Use Cases. This paper focuses on immersive telepresence, a major use case of remote collaboration. Vision Pro also facilitates other use cases such as collaborative whiteboards [11] and shared entertainment experiences (*e.g.,* playing games and watching movies) [13], which we plan to explore in the future.

6 Related Work

Network Measurements on VCAs. In recent years, there has been a growing research interest in measuring the network performance of VCAs [18, 33, 36, 45, 48, 52, 58, 65]. For example, Varvello *et al.* [65] build a large-scale testbed to facilitate the evaluation of videoconferencing performance in the wild. Sharma *et al.* [58] utilize IP/UDP headers for QoE estimation of VCAs. In this paper, we measure the performance of immersive telepresence with these VCAs on Apple Vision Pro.

Measurement of Immersive Applications. Existing studies on the performance of immersive applications have focused on immersive video streaming [71], Web-based extended reality (XR) [41], and social virtual reality (VR) platforms [20–22, 43]. For instance, MetaVRadar [43] correlates the network traffic of social VR applications with user activities. Liu *et al.* [41] investigate Web-based XR platforms accelerated by WebAssembly [31]. This paper measures spatial personas that improve telepresence experiences.

Telepresence Systems are increasingly gaining attention in the industry (*e.g.,* Holoportation [53] from Microsoft, Project Starline [39] from Google, and Codec Avatar [44] from Meta) and academia (*e.g.,* MetaStream [28], FarfetchFusion [40], and MeshReduce [37]). Moreover, the human-computer interaction community has developed in-lab prototypes for specific use cases, such as conducting remote surgeries [26] and teaching physical tasks [61]. In this paper, we measure commercial telepresence systems on Apple Vision Pro.

7 Conclusion

This paper presents a first-of-its-kind in-depth and empirical measurement study of immersive telepresence on Apple Vision Pro. Driven by the counter-intuitive results that the required bandwidth of the immersive spatial persona is even lower than its 2D counterpart, we conduct a comprehensive analysis of the delivered content. We find that spatial personas utilize semantic communication to optimize bandwidth consumption, which, however, leads to challenges for employing rate adaptation. Moreover, we dissect the visibility-aware optimizations and the scalability issue of spatial persona. We hope that our findings can shed light on the design practices of emerging immersive telepresence systems.

Acknowledgment

We thank the anonymous reviewers and our shepherd for their insightful feedback, as well as the participants of our experiments for their valuable contributions. This work was partially supported by the National Science Foundation under Grants CNS-2007153, CNS-2212296, and CNS-2235049.

References

[1] 2012. Sketchfab. https://sketchfab.com. [accessed on 09/11/2024].
[2] 2019. Microsoft HoloLens 2. https://www.microsoft.com/en-us/hololens. [accessed on 09/11/2024].
[3] 2022. Magic Leap 2. https://www.magicleap.com/magic-leap-2. [accessed on 09/11/2024].
[4] 2023. Meta Quest 3. https://www.meta.com/quest/quest-3/. [accessed on 09/11/2024].
[5] 2024. Apple Vision Pro. https://www.apple.com/apple-vision-pro/. [accessed on 09/11/2024].
[6] 2024. Linux tc Man Page. https://linux.die.net/man/8/tc.
[7] 2024. ZED 2i. https://www.stereolabs.com/zed-2i/l. [accessed on 09/11/2024].
[8] 2024. Zoom Meetings. https://zoom.us/.
[9] Apple. 2024. Analyzing the performance of your visionOS app. https://developer.apple.com/documentation/visionOS/analyzing-the-performance-of-your-visionOS-app.
[10] Apple. 2024. Apple Vision Pro Privacy Overview. https://www.apple.com/privacy/docs/Apple_Vision_Pro_Privacy_Overview.pdf. [accessed on 09/11/2024].
[11] Apple. 2024. Create and manage Freeform boards on Apple Vision Pro. https://support.apple.com/guide/apple-vision-pro/create-and-manage-freeform-boards-tan5281cfbb6/visionos.
[12] Apple. 2024. Make or receive a FaceTime call on Apple Vision Pro. https://support.apple.com/guide/apple-vision-pro/make-or-receive-a-facetime-call-tan440238696/visionos.
[13] Apple. 2024. Use SharePlay in FaceTime calls on Apple Vision Pro. https://support.apple.com/guide/apple-vision-pro/use-shareplay-in-facetime-calls-tan15b2c7bf9/visionos.
[14] Apple. 2024. Use spatial Persona (beta) on Apple Vision Pro. https://support.apple.com/guide/apple-vision-pro/use-spatial-persona-tana1ea03f18/visionos.
[15] Apple. 2024. Xcode. https://developer.apple.com/xcode/. [accessed on 09/11/2024].
[16] United States Census Bureau. 2020. United States Census. https://www.census.gov/2020results. [accessed on 09/11/2024].
[17] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of IEEE/CVF CVPR*.
[18] Hyunseok Chang, Matteo Varvello, Fang Hao, and Sarit Mukherjee. 2021. Can You See Me Now? A Measurement Study of Zoom, Webex, and Meet. In *Proceedings of ACM IMC*. https://dl.acm.org/doi/abs/10.1145/3487552.3487847
[19] Ruizhi Cheng, Kaiyan Liu, Nan Wu, and Bo Han. 2023. Enriching Telepresence with Semantic-driven Holographic Communication. In *Proceedings of ACM HotNets*.
[20] Ruizhi Cheng, Nan Wu, Songqing Chen, and Bo Han. 2022. Reality Check of Metaverse: A First Look at Commercial Social Virtual Reality Platforms. In *Proceedings of IEEE Workshop for Building the Foundations of the Metaverse (Metabuild), co-located with IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*.
[21] Ruizhi Cheng, Nan Wu, Songqing Chen, and Bo Han. 2022. Will Metaverse be NextG Internet? Vision, Hype, and Reality. *IEEE Network* 36, 5 (2022), 197–204.
[22] Ruizhi Cheng, Nan Wu, Matteo Varvello, Songqing Chen, and Bo Han. 2022. Are We Ready for Metaverse? A Measurement Study of Social Virtual Reality Platforms. In *Proceedings of ACM IMC*.
[23] Cisco. 2024. Check the Audio and Video Statistics of Your Cisco Webex Meeting. https://help.webex.com/en-us/article/nmghd9e/Check-the-Audio-and-Video-Statistics-of-Your-Cisco-Webex-Meeting.
[24] Cisco. 2024. Webex. https://www.webex.com/.
[25] Chamitha De Alwis, Anshuman Kalla, Quoc-Viet Pham, Pardeep Kumar, Kapal Dev, Won-Joo Hwang, and Madhusanka Liyanage. 2021. Survey on 6G Frontiers: Trends, Applications, Requirements, Technologies and Future Research. *IEEE Open Journal of the Communications Society* 2 (2021), 836–886.
[26] Danilo Gasques, Janet G Johnson, Tommy Sharkey, Yuanyuan Feng, Ru Wang, Zhuoqun Robin Xu, Enrique Zavala, Yifei Zhang, Wanze Xie, Xinming Zhang, et al. 2021. ARTEMIS: A Collaborative Mixed-Reality System for Immersive Surgical Telementoring. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*. https://doi.org/10.1145/3411764.3445576
[27] Google. 2024. Draco 3D Data Compression. https://google.github.io/draco/. [accessed on 09/11/2024].
[28] Yongjie Guan, Xueyu Hou, Nan Wu, Bo Han, and Tao Han. 2023. MetaStream: Live Volumetric Content Capture, Creation, Delivery, and Rendering in Real Time. In *Proceedings of ACM MobiCom*.
[29] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D graphics. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 1–10.
[30] Apoorv Gupta, Aman Bansal, and Vidhi Khanduja. 2017. Modern Lossless Compression Techniques: Review, Comparison and Analysis. In *Proceedings of IEEE International Conference on Electrical, Computer and Communication Technologies*.
[31] Andreas Haas, Andreas Rossberg, Derek L Schuff, Ben L Titzer, Michael Holman, Dan Gohman, Luke Wagner, Alon Zakai, and JF Bastien. 2017. Bringing the Web up to Speed with WebAssembly. In *Proceedings of ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*. https://doi.org/10.1145/3062341.3062363
[32] Bo Han, Yu Liu, and Feng Qian. 2020. ViVo: Visibility-Aware Mobile Volumetric Video Streaming. In *Proceedings of ACM MobiCom*.
[33] Jia He, Mostafa Ammar, and Ellen Zegura. 2023. A Measurement-Derived Functional Model for the Interaction Between Congestion Control and QoE in Video Conferencing. In *Proceedings of International Conference on Passive and Active Network Measurement (PAM)*.
[34] Keiichi Ihara, Mehrad Faridan, Ayumi Ichikawa, Ikkaku Kawaguchi, and Ryo Suzuki. 2023. HoloBots: Augmenting Holographic Telepresence with Mobile Robots for Tangible Remote Collaboration in Mixed Reality. In *Proceedings of ACM Symposium on User Interface Software and Technology (UIST)*. https://doi.org/10.1145/3586183.3606727
[35] ipinfo.io. 2024. https://ipinfo.io/. [accessed on 09/11/2024].
[36] Bart Jansen, Timothy Goodwin, Varun Gupta, Fernando Kuipers, and Gil Zussman. 2018. Performance Evaluation of WebRTC-based Video Conferencing. *ACM SIGMETRICS Performance Evaluation Review* 45, 3 (2018), 56–68.
[37] Tao Jin, Mallesham Dasa, Connor Smith, Kittipat Apicharttrisorn, Srinivasan Seshan, and Anthony Rowe. 2024. MeshReduce: Scalable and Bandwidth Efficient 3D Scene Capture. In *Proceedings of IEEE Conference Virtual Reality and 3D User Interfaces (VR)*.
[38] Davis E King. 2009. Dlib-ml: A Machine Learning Toolkit. *The Journal of Machine Learning Research* 10 (2009), 1755–1758.
[39] Jason Lawrence, Danb Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G Desloge, Tommy Fortes, Eric M Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, et al. 2021. Project Starline: a High-fidelity Telepresence System. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–16.
[40] Kyungjin Lee, Juheon Yi, and Youngki Lee. 2023. FarfetchFusion: Towards Fully Mobile Live 3D Telepresence Platform. In *Proceedings of ACM MobiCom*.
[41] Kaiyan Liu, Nan Wu, and Bo Han. 2023. Demystifying Web-based Mobile Extended Reality Accelerated by WebAssembly. In *Proceedings of ACM IMC*. https://doi.org/10.1145/3618257.3624833
[42] Yu Liu, Puqi Zhou, Zejun Zhang, Anlan Zhang, Bo Han, Zhenhua Li, and Feng Qian. 2024. MuV2: Scaling up Multi-user Mobile Volumetric Video Streaming via Content Hybridization and Sharing. In *Proceedings of ACM MobiCom*.
[43] Minzhao Lyu, Rahul Dev Tripathi, and Vijay Sivaraman. 2024. MetaVRadar: Measuring Metaverse Virtual Reality Network Activity. In *Proceedings of ACM SIGMETRICS*.
[44] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. 2021. Pixel Codec Avatars. In *Proceedings of IEEE/CVF CVPR*.
[45] Kyle MacMillan, Tarun Mangla, James Saxon, and Nick Feamster. 2021. Measuring the Performance and Network Utilization of Popular Video Conferencing Applications. In *Proceedings of ACM IMC*. https://dl.acm.org/doi/abs/10.1145/3487552.3487842
[46] MaxMind. 2024. https://www.maxmind.com/en/home. [accessed on 09/11/2024].
[47] Trevor Mendez, Walter Milliken, and Craig Partridge. 1993. Host Anycasting Service. RFC 1546. https://www.rfc-editor.org/info/rfc1546 [accessed on 09/11/2024].
[48] Oliver Michel, Satadal Sengupta, Hyojoon Kim, Ravi Netravali, and Jennifer Rexford. 2022. Enabling Passive Measurement of Zoom Performance in Production Networks. In *Proceedings of ACM IMC*.
[49] Microsoft. 2024. Monitor Call and Meeting Quality in Microsoft Teams. https://support.microsoft.com/en-us/office/monitor-call-and-meeting-quality-in-microsoft-teams-7bb1747c-d91a-4fbb-84f6-ad3f48e73511.
[50] Microsoft. 2024. Teams. https://www.microsoft.com/en-us/microsoft-teams/group-chat-software.
[51] Timothy Neate, Vasiliki Kladouchou, Stephanie Wilson, and Shehzmani Shams. 2022. "Just Not Together": The Experience of Videoconferencing for People with Aphasia during the Covid-19 Pandemic. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*.
[52] Antonio Nistico, Dena Markudova, Martino Trevisan, Michela Meo, and Giovanna Carofiglio. 2020. A Comparative Study of RTC Applications. In *Proceedings of IEEE International Symposium on Multimedia (ISM)*.
[53] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*

(UIST).

[54] Jingliang Peng, Chang-Su Kim, and C.-C. Jay Kuo. 2005. Technologies for 3D Mesh Compression: A Survey. *Journal of Visual Communication and Image Representation* 16, 6 (2005), 688–733.

[55] Feng Qian, Bo Han, Qingyang Xiao, and Vijay Gopalakrishnan. 2018. Flare: Practical Viewport-Adaptive 360-Degree Video Streaming for Mobile Devices. In *Proceedings of ACM MobiCom*.

[56] Henning Schulzrinne and Stephen Casner. 2003. RTP Profile for Audio and Video Conferences with Minimal Control. RFC 3551. https://rfc-editor.org/rfc/rfc3551.txt [accessed on 09/11/2024].

[57] Henning Schulzrinne, Stephen L. Casner, Ron Frederick, and Van Jacobson. 2003. RTP: A Transport Protocol for Real-Time Applications. RFC 3550. https://rfc-editor.org/rfc/rfc3550.txt [accessed on 09/11/2024].

[58] Taveesh Sharma, Tarun Mangla, Arpit Gupta, Junchen Jiang, and Nick Feamster. 2023. Estimating WebRTC Video QoE Metrics Without Using Application Headers. In *Proceedings of ACM IMC*. https://doi.org/10.1145/3618257.3624828

[59] Switchboard. 2024. 40+ meeting statistics you need to know in 2024. https://www.switchboard.app/learn/article/meeting-statistics-2024. [accessed on 09/11/2024].

[60] Faisal Tariq, Muhammad RA Khandaker, Kai-Kit Wong, Muhammad A Imran, Mehdi Bennis, and Merouane Debbah. 2020. A Speculative Study on 6G. *IEEE Wireless Communications* 27, 4 (2020), 118–125. https://doi.org/10.1109/MWC.001.1900488

[61] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. In *Proceedings of ACM Symposium on User Interface Software and Technology (UIST)*. https://doi.org/10.1145/3332165.3347872

[62] Michael C. Toren. 2024. tcptraceroute(1) - Linux man page. https://linux.die.net/man/1/tcptraceroute. [accessed on 09/11/2024].

[63] Jean-Marc Valin, Koen Vos, and Tim Terriberry. 2021. QUIC: A UDP-Based Multiplexed and Secure Transport. RFC 9000. https://datatracker.ietf.org/doc/html/rfc9000 [accessed on 09/11/2024].

[64] Christian J Van den Branden Lambrecht and Olivier Verscheure. 1996. Perceptual Quality Measure Using a Spatiotemporal Model of the Human Visual System. In *Proceedings of Digital Video Compression: Algorithms and Technologies*. https://doi.org/10.1117/12.235440

[65] Matteo Varvello, Hyunseok Chang, and Yasir Zaki. 2022. Performance Characterization of Videoconferencing in the Wild. In *Proceedings of ACM IMC*. https://doi.org/10.1145/3517745.3561442

[66] Wikipedia. 2022. Zoom fatigue. https://en.wikipedia.org/wiki/Zoom_fatigue. [accessed on 09/11/2024].

[67] Wireshark. 1998. https://www.wireshark.org/. [accessed on 09/11/2024].

[68] WonderNetwork. 2024. Global Ping Statistics. https://wondernetwork.com/pings.

[69] Nan Wu, Kaiyan Liu, Ruizhi Cheng, Bo Han, and Puqi Zhou. 2024. Theia: Gaze-driven and Perception-aware Volumetric Content Delivery for Mixed Reality Headsets. In *Proceedings of ACM MobiSys*.

[70] Xumiao Zhang, Shuowei Jin, Yi He, Ahmad Hassan, Z Morley Mao, Feng Qian, and Zhi-Li Zhang. 2024. QUIC is not Quick Enough over Fast Internet. In *Proceedings of ACM Web Conference (WWW)*.

[71] Chao Zhou, Zhenhua Li, and Yao Liu. 2017. A Measurement Study of Oculus 360 Degree Video Streaming. In *Proceedings of ACM on Multimedia Systems Conference*. https://doi.org/10.1145/3083187.3083190

[72] Zoom. 2024. Accessing Meeting and Phone Statistics. https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0070504.