Spatial Video Streaming on Apple Vision Pro XR Headset

Guodong Chen*
Northeastern University
Boston, MA, USA
chen.guod@northeastern.edu

Sizhe Wang*
Northeastern University
Boston, MA, USA
wang.sizh@northeastern.edu

Jacob Chakareski
New Jersey Institute Tech.
Newark, NJ, USA
jacobcha@njit.edu

Dimitrios Koutsonikolas
Northeastern University
Boston, MA, USA
d.koutsonikolas@northeastern.edu

Mallesham Dasari
Northeastern University
Boston, MA, USA
m.dasari@northeastern.edu

Abstract

Apple recently unveiled capturing spatial video experiences on an iPhone and their extended reality (XR) headset Vision Pro. Spatial videos can be viewed on immersive near-eye displays like the Apple Vision Pro for a more realistic experience with depth perception. However, streaming spatial videos encounters several challenges, such as bandwidth limitations, latency, and synchronization issues between multiple camera views. This position paper presents a comprehensive research agenda to address these issues and make spatial video streaming as ubiquitous as traditional online video for mobile systems and applications. We outline several research threads for exploration and discuss a series of novel ideas, including view-adaptive streaming strategies, multipath support, and QoE modeling, which we believe will become the fundamental components for future video experiences on mobile and wearable devices.

CCS Concepts

• **Human-centered computing → Ubiquitous and mobile computing**.

ACM Reference Format:
Guodong Chen, Sizhe Wang, Jacob Chakareski, Dimitrios Koutsonikolas, and Mallesham Dasari. 2025. Spatial Video Streaming on Apple Vision Pro XR Headset. In *The 26th International Workshop on Mobile Computing Systems and Applications (HOTMOBILE '25), February 26–27, 2025, La Quinta, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3708468.3711878

1 Introduction

In their early years, spatial videos, also known as stereoscopic or stereo 3D videos[1] and their more general form of multi-view video was considered the future of video communications because of their similarity to the human visual system, i.e., one camera view for each eye, providing depth perception [4]. However, spatial videos needed advanced display technologies and multi-camera capture pipelines that were not available on consumer-grade devices. They also faced significant challenges in processing even a single video stream because of the limited computing, storage, memory, and networking capacity. Thus, the content producers scaled back their investments in spatial video communications, turning instead to enhancing standard monocular videos. This status quo has changed today with advances in near-eye displays (i.e., XR headsets) and consumer devices (e.g., iPhones), which can capture, stream, and play spatial videos.

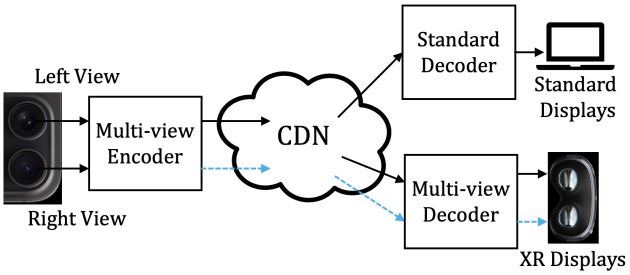


Figure 1: High-level architecture of a spatial video streaming system. Consumer-grade devices like the iPhone can now stream spatial videos and be viewed on XR headsets in 3D with a rich video experience.

Different from traditional videos, spatial videos contain two views and depth perception metadata, offering users a more immersive/realistic experience. While the additional view and metadata significantly increase the quality of experience for video watching and mobile communication, they also demand more advanced devices and spatial video streaming systems. The latter should transmit two high-quality views in parallel and inject the depth perception metadata correctly to ensure both views maintain visual clarity and present an accurate 3D effect at high resolution.

However, meeting the above requirements is not trivial. Our testing with the Apple TV streaming on Vision Pro revealed that spatial video bandwidth needs could reach as high as a few hundred Mbps (post-compression). Such bandwidths are not commonly available across the wide-area Internet for most of the world's population. Even in favorable network environments, maintaining such a high bitrate is challenging due to network variability problems [24].

*Both authors contributed equally to this research.
[1]Spatial video comprises two video streams– one for each eye, captured with two cameras separated by distance similar to that between the two human eyes. We use spatial, stereo, and stereoscopic video interchangeably throughout the paper.

HOTMOBILE '25, February 26–27, 2025, La Quinta, CA, USA

ACM ISBN 979-8-4007-1403-0/25/02
https://doi.org/10.1145/3708468.3711878

Over the past two decades, multi-view compression standards such as H.264 MVC [18] and MV-HEVC [29] have been developed. However, their implementation encoders are computationally expensive, limiting their practical application. As a result, most current existing spatial video streaming solutions still use standard codecs like H.264/5 [31] applied to each view independently. Moreover, rate-distortion optimized coding, streaming, multiple descriptions, and two-path delivery have been investigated [5, 6, 21]. More recently, Apple unveiled the first commercial real-time codec for on-demand and live stereo video streaming scenarios [2]. Additionally, multi-path networking protocols (e.g., MPTCP [14], MPQUIC [12] and FBDT [26]) that are suitable for streaming spatial videos and 360° videos have gained wider adoption. These technological advancements, along with the progress in XR headsets, make live spatial 3D video streaming possible in the near future.

This position paper outlines new research directions for spatial video streaming (SVS) that work on commodity devices such as phones and XR headsets. We propose several novel ideas that we envision will become building blocks of next-generation SVS systems. First, we discuss the need for a new quality of experience (QoE) models for spatial videos and demonstrate for the first time user preferences of monocular vs. stereoscopic video playback under different network conditions. Second, we explore different streaming strategies with bitrate adaptation algorithms for spatial videos, balancing the video quality and depth perception trade-offs under constrained networking conditions. Third, we propose a multi-path scheduling algorithm for spatial videos, where we schedule each view on a different network path (e.g., WiFi and LTE/5G). Figure 1 shows a high-level architecture.

We obtained preliminary results using an iPhone and Vision Pro XR headset [1]. To the best of our knowledge, we are the first to study an end-to-end spatial video streaming system on XR headsets. We report metrics like video quality of views and depth perception metrics such as disparity error under different network conditions. We also evaluate subjective user QoE by conducting a user study to understand the impact of network artifacts on stereoscopic vs. monocular videos.

2 Background and Related Work

Spatial videos provide depth perception by presenting two slightly different perspectives of the same scene to the viewer's left and right eyes. Spatial videos are captured with two time-synchronized cameras with a small horizontal distance between the cameras, emulating the binocular 3D human vision. Unlike other forms of 3D content, such as 360-degree videos [8] and volumetric videos [19], spatial 3D videos consist of two 2D video streams that together create a 3D effect. While 360-degree videos present their own viewport related challenges [15], they do not require depth perception information and are less sensitive to artifacts, making them easier to manage. Consumer-grade devices like the iPhone 15 Pro and Vision Pro can now capture spatial videos. These videos can be viewed using various techniques, such as anaglyph 3D glasses, autostereoscopic displays, and XR headsets (Vision Pro and Quest3).

Commercial content providers today (e.g., YouTube, Apple TV+, Netflix, etc.) mainly stream monocular video with MPEG-DASH [17], HLS and WebRTC using codecs such as H.264/AVC [31] or H.265 [28].

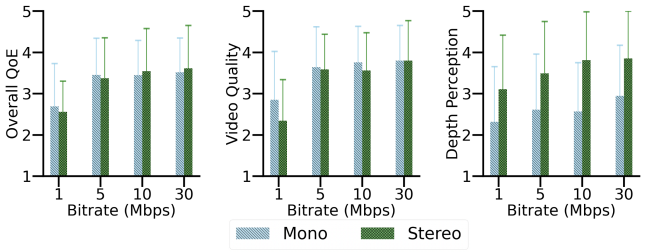


Figure 2: Impact of different spatial video bitrates on user experience. Users streamed the videos on Vision Pro Headset and rated their experience from 1 to 5. The key takeaway is that at lower bitrates, users favor standard monocular video over spatial 3D video, while at higher bitrates, the preference shifts to spatial video.

To encode spatial videos, several extensions are built on top of these standard codecs, such as H.264/MVC (Multiview Video Coding) [18] and MV-HEVC (Multi-View HEVC) [29].

Much of the early work in spatial video streaming focused on efficiently compressing the two views[2] of the video to exploit the redundancy across the two views of the spatial video [13]. While these coding approaches have shown tremendous success in achieving high compression efficiency, they were computationally very expensive. As a result, many spatial video streaming implementations used a DASH-based streaming strategy by compressing both views separately while sacrificing the bandwidth and/or quality [27]. Another line of research has focused on modeling quality and overall Quality of Experience (QoE) specifically for spatial videos [7]. However, most of these works focused on evaluating user experience on external walled displays (e.g., 3D TVs or autostereoscopic displays) and did not study the impact of network artifacts such as the impact of bitrate of individual views on depth perception, latency, and network variability.

3 Spatial Video Insights on Apple Vision Pro XR Headset

While there are studies on the QoE assessment of spatial videos viewed on traditional displays [30], to the best of our knowledge, there is no study on the viewing experience of mono and spatial videos under bitrate constraints on XR headsets. To address this gap, we conducted a subjective user study on user preferences for mono and spatial videos encoded with the same bitrate ladder. We developed a spatial video player app for Vision Pro, which includes features like stereoscopic playback and user rating interfaces designed with simple interactions.

We ask users for a five-point score ranging from 1 (Bad) to 5 (Excellent) based on three criteria: visual quality, depth perception, and overall QoE. A total of 15 non-expert subjects participated in the study. Each user wore the headset and watched 40 video clips (15 seconds each) consisting of mono and spatial videos at four different quality levels, presented in random order. The five selected scenes covered spatial and temporal complexity. Figure 2 shows results.

[2]we refer to each video stream of the spatial video as a view

Observations. We found multiple takeaways from this study that we believe will drive the design of future spatial video streaming strategies. First, users can perceive depth in both mono and spatial videos when viewed on the VisionPro headset, with spatial videos consistently achieving higher experience for depth perception. Second, visual quality is higher for mono videos over spatial videos across various bitrates. This preference can be attributed to the fact that compressing mono videos, which contain only one view, allocates more bitrate for quality improvement, thereby enhancing visual quality perception. Finally, we observe a notable shift in overall QoE from low to high bitrates. At lower bitrates (e.g., 1 Mbps and 5 Mbps), users favor mono videos over spatial videos. However, this preference shifts towards spatial videos at higher bitrates. Such a finding introduces new research problems for spatial video adaptive streaming in order to achieve the best QoE: whether to prioritize streaming stereoscopic videos at low quality or opt for monocular videos with relatively higher quality. We are currently conducting a larger-scale study to gain an in-depth understanding of network artifacts on spatial video streaming.

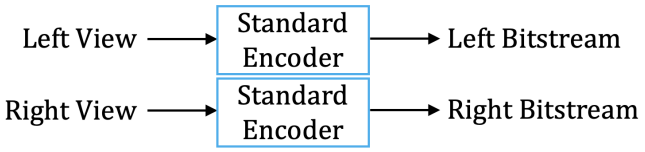
4 Research Agenda

At a high level, our research agenda is focused on intelligently decoupling the two views of spatial videos to enable live adaptive streaming. We introduce a series of key ideas addressing the challenges in streaming strategies, multi-path streaming, and QoE modeling for spatial video streaming on near-eye XR displays. Specifically, we examine two streaming strategies (§4.1): a traditional DASH-based approach that encodes and streams two viewpoints independently and a more integrated layered encoding method using multi-view codecs like MV-HEVC, which leverages inter-view redundancy for efficiency. We introduce a new problem of quality imbalance between two views and propose a super-resolution-based quality enhancement method to manage bitrate differences between the two views. Next, we present a dynamic multi-path scheduling strategy that optimizes the use of multiple network paths in real-time at the view and packet levels (§4.2). Furthermore, we present a comprehensive model to assess spatial video QoE, focusing on the overall depth perception, visual quality of both views and the impact of their differences (§4.3).

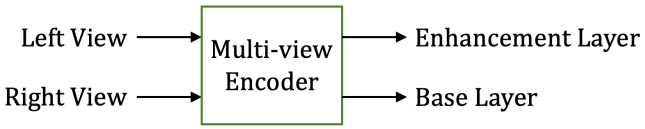
4.1 Streaming Strategies for Spatial Videos

There are two different strategies in the literature for streaming spatial videos. The first approach is based on a traditional DASH (dynamic adaptive streaming over HTTP) [17] streaming standard where both viewpoints are encoded separately, and the bitrate adaptation for each view is performed independently (see Figure 3(a)). The second approach is based on multi-view encoding algorithms [18, 29] where both viewpoints are encoded jointly into two layers (see Figure 3(b)).

4.1.1 DASH-based spatial video streaming. The DASH-based approach provides a significant advantage in spatial video streaming by enabling the use of standard codecs to encode and decode both views independently and in parallel. This allows us to use existing streaming infrastructure without any modifications, ensuring compatibility and flexibility.



(a) DASH-based spatial video streaming



(b) Layered spatial video streaming

Figure 3: Two strategies for streaming spatial videos: a) The DASH-based system encodes and adaptively streams the two views independently without exploiting the redundancy across the views; b) The layered streaming encodes the stereo views jointly and streams in an incremental layered fashion.

However, this approach introduces multiple fundamental limitations, primarily the increased bandwidth requirement. Since each viewpoint is encoded separately, the total bandwidth needed is effectively double that of standard videos, although both viewpoints overlap significantly. Additionally, the dynamic adaptive feature of DASH can result in each view being encoded and streamed at different bandwidths. When network congestion occurs, DASH may select a lower quality for one view to prevent buffering. This can result in a significant disparity between the quality of the two views, such as one view being encoded at a very low bitrate (e.g., 1 Mbps) and the other at a higher bitrate (e.g., 15 Mbps). Such imbalances can significantly decrease the overall visual QoE [20]. Moreover, the stereoscopic 3D effect may be disrupted due to the quality difference between the two views, beyond a certain threshold.

To address these issues, we propose a novel client-side super-resolution-based bitrate adaptation method that effectively manages bitrate/quality differences between the views. Our adaptive bitrate (ABR) algorithm deliberately allocates different bitrates to each view— primary with high quality and secondary with low quality. To balance the quality difference between the two views, i.e., improve the quality of the secondary view, we propose a super-resolution (SR)-based quality enhancement technique. We believe this method can significantly improve the quality of the lower-resolution view by leveraging the substantial similarities between the two views. The key insight here is the ability to recover the quality of the secondary view accurately *using the primary view as a prior*. In the past, SR has shown tremendous benefits in enhancing the quality in the case of monocular video streaming [3, 8, 25]. The process involves upsampling a low-resolution image using machine learning methods. In our approach, the high-resolution primary view and the lower-resolution secondary view together can recover a high-resolution secondary view effectively, thus maintaining a balanced video quality across the views and overall depth perception. We will explore different ways of using SR— an offline model trained per video and streamed online or a universal model that is trained and shipped to clients offline.

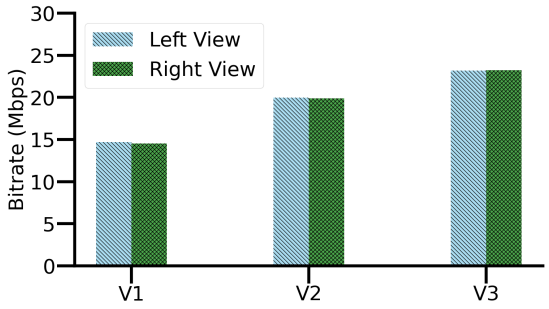


Figure 4: Apple's MV-HEVC encoding results in almost equivalent bitrates for both views across different videos.

There are several other open questions we plan to explore. Specifically, the impact of resolution differences between streams on SR performance is unclear. Understanding how these variations affect the SR process is a key focus of our study. We will also investigate the effect of stream arrival delays on SR quality, as synchronization issues may arise. Additionally, we will analyze the overhead of the SR algorithm in terms of delay, memory usage, and power consumption to ensure feasibility for resource-constrained devices.

4.1.2 Layered spatial video streaming. The above strategy, while effectively adapting spatial video bitrates, suffers from compression inefficiency because of separate encoding of the views. Unlike the DASH approach, layered spatial video streaming exploits redundancy between views to achieve lower bitrates for the secondary view while maintaining the same quality as the primary view. Our studies found that, in theory, using a state-of-the-art multi-view encoder— MV-HEVC [29] can achieve up to 40% bitrate savings on the secondary view compared to the primary view.

However, the use of multi-view codecs, like MV-HEVC, has significant challenges. These codecs are computationally demanding to achieve optimal bitrate savings, making them unsuitable for real-time applications. The complexity primarily arises from the complex inter-view frame prediction process that identifies similar blocks of pixels across views to compute residuals. This process is analogous to motion estimation in temporal frames, but it is conducted across views. Additionally, the lack of mature spatial video technology in the past meant that no hardware solutions were developed to support multi-view codecs, complicating their deployment. To circumvent this complexity, many existing solutions opt for a less intensive frame prediction approach, both temporally and across views, at the expense of compression efficiency. For instance, despite deploying MV-HEVC in Apple products like the iPhone and Vision Pro headset, we find almost no difference in bitrates between the left and right views. However, for MV-HEVC encoding standard, the second view should reference frames that are in the main view, which is reasonable to save 40% or more for the second layer. So, the almost identical bitrate for both views contradicts the fundamental advantage of multi-view coding methods, showing that current MV-HEVC is not mature. We have experimentally observed this outcome in our evaluation by capturing spatial videos on an iPhone and Vision Pro for 20 different scenes. We illustrate it in Figure 4 for three representative videos.

To address these challenges, we propose a content-aware adaptation strategy that trades the frame prediction process between views and temporal frames for optimal compression while maintaining coding speeds suitable for real-time applications. This strategy is based on a key insight: *the computational complexity of frame prediction primarily arises from the high residual information (i.e., difference) between frames.* Larger differences require more intensive computation, while smaller differences are less demanding.

4.2 Multi-path Spatial Video Streaming

Today's mobile devices widely support multiple network interfaces (e.g., WiFi, Cellular, Satellite, etc). Leveraging multi-path transport protocols (e.g., MPTCP [14] or MPQUIC [12]) allows applications to use network interfaces simultaneously and significantly improve user QoE. There has been extensive work on using multi-path for conventional video streaming [22, 23]. While many of these techniques can be directly applied to spatial video streaming, we identify several new opportunities for further improving spatial video QoE based on tighter application integration with MPQUIC. We note that advanced wireless links like FSO [16] provide very high speeds that are sufficient for streaming high-quality videos but are far from practical to realize them in realistic settings.

4.2.1 View vs. packet-level scheduling. A straightforward method for streaming spatial video involves sending the primary view over the faster path and the secondary view over the slower one. This approach is advantageous for two main reasons: Firstly, it ensures the primary view arrives and is decoded before the secondary view. By the time the secondary view reaches the application, the primary view is already prepared as a reference for decoding the secondary view. Secondly, this method avoids the classic head-of-line (HoL) blocking problem, where packets arriving on the slower path create a bottleneck for those on the faster path, as the application needs to process packets from both paths. Essentially, the stereoscopic video application requires the primary view first, then the secondary, which effectively balances out the flow.

However, the above approach is effective only if the network paths consistently maintain one as slow and the other as fast, which is rarely the case in practice. Network path latency and bandwidth can vary significantly. Furthermore, when using more than two paths, it becomes necessary to schedule packet-level rather than view-level, in which case we encounter the aforementioned HoL blocking problem.

To address the above issues, we propose a NAL (network abstraction layer) unit-level multi-path scheduling strategy that dynamically allocates application-level video packets (i.e., NAL units) across multiple paths based on real-time network conditions. NAL units are fundamental units of compressed video data standardized in H.264/AVC and MV-HEVC codecs. We schedule an entire NAL unit on a given path rather than scheduling transport-level packets as in conventional multi-path streaming strategies. Each frame in these codecs typically comprises several NAL units. The rationale for this strategy hinges on the independence of NAL units for each view. This independence allows for prioritizing the transmission of NAL units from the primary view across multiple paths before those from the secondary view. Also, since packets within a single NAL unit are interdependent for decoding, streaming an entire NAL

unit over a single path can effectively mitigate the HoL blocking issue that delays the decoding of NAL units. The benefit of this approach compared to existing multi-path streaming methods such as XLINK [32] and Chorus [22] is that we do not encounter the redundancy overhead due to packet re-injection in addressing the HoL blocking problem since we are prioritizing the primary view across all paths, to ensure its arrival before the secondary view.

We also explore a classical problem of mismatch between the transport layer's sending rate and the application layer's selected video bitrate based on the ABR algorithm because they are done separately [22]. The key questions we explore here include—How can ABR algorithms be assisted in making accurate bitrate decisions for spatial content under multi-path scenarios? How can transport performance consistently meet the QoE requirements of spatial video streaming in dynamic network conditions? We explore a joint scheduling strategy that predetermines transmission decisions for both streams before the actual data transmission begins. This schedule will be computed based on transport, which informs the ABR algorithms of the expected throughput for both paths proactively and allows adjustments based on transport-level bandwidth predictions. We will incorporate a two-way feedback control loop to facilitate the exchange of information between the client and server. This allows the client-side ABR algorithm to predict throughput based on the server's pre-determined scheduling decisions and enables the server packet scheduler to adjust its decisions dynamically during transmission.

4.3 Network Impact on Spatial Video QoE

The ultimate goal of this research agenda is to improve QoE in spatial video streaming. To maximize QoE, we must model the key factors of spatial videos affected by network artifacts. Unlike conventional monocular videos, depth perception is fundamental to spatial videos. Other metrics like video quality are extensively explored for monocular videos [9–11] but still require further investigation for spatial videos to understand the impact of the relative quality difference of each view. The networking community has under-explored these factors.

Impact of network artifacts on depth perception: Spatial videos are just two 2D videos captured from two viewpoints; the camera has no true depth. The users are given a sense of depth by projecting the views to the left and right eye. Since there is no explicit depth map here, we plan to use the *disparity* map as a proxy, which inversely relates to depth, i.e., high disparity indicates that objects are closer to the viewer, and vice versa. The disparity can be defined as the horizontal distance between each pixel in both views. We can obtain a disparity map by computing it for all pixels. Figure 5 shows the effect of different bitrates on disparity maps.

Our key observation here is that any change in the video quality of the spatial video can affect depth perception, leading to a corruption in disparity ($\mathcal{D}_d$). When viewing such spatial videos with corrupted disparity, we observe the dislocation of the objects in the scene, creating unnatural depth conflicts under poor network conditions. We compute the PSNR (peak signal-to-noise ratio) of disparty maps extracted from stereo videos at different bitrates using a reference disparity map extracted from the original stereo video, as shown in Figure 6. The figure shows that reduced bitrates

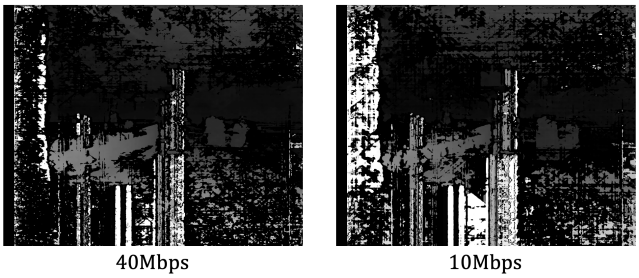


Figure 5: Disparity maps extracted from two spatial videos at different bitrates. Significant depth details are missing at low bitrate compared to high bitrate, which creates depth conflicts when viewed on XR headsets.

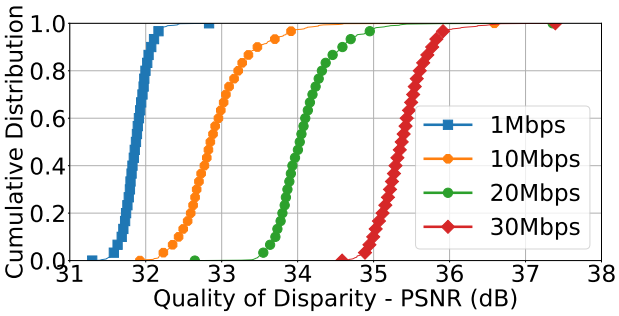


Figure 6: Effect of bitrate on disparity quality between two spatial views measured in PSNR (dB): Lower bitrates degrade disparity quality in stereoscopic videos, as seen by the decrease in PSNR values across rates from 30 Mbps to 1 Mbps.

result in poorer disparity quality, with a significant decrease in PSNR (by 3.5dB) from 30 Mbps to 1 Mbps.

Video quality imbalance between spatial views: The distortion in video quality for each view for different bitrates can be calculated separately (say $\mathcal{D}_l$ for left view and $\mathcal{D}_r$ for right view) using the standard video quality metrics that are used for monocular videos (e.g., PSNR, SSIM, or VMAF). Let $\mathcal{D}_v$ be the overall visual distortion observed by the users. A straightforward way to compute $\mathcal{D}_v$ is by averaging $\mathcal{D}_l$ and $\mathcal{D}_r$. This makes sense only if each view is encoded at symmetric quality, i.e., the same bitrate (or quality) for both views. However, for bitrate adaption, spatial views must encoded asymmetrically (i.e., different bitrate for each view). In this case, the overall visual distortion must be modeled nonlinearly such as using harmonic mean or weighted power mean.

We explore two approaches: 1) harmonic mean of distortions and 2) weighted power mean. The harmonic mean is particularly sensitive to smaller values, which makes it suitable for emphasizing the impact of the view with the lower quality in the overall distortion: $\frac{2}{\frac{1}{\mathcal{D}_l}+\frac{1}{\mathcal{D}_r}}$. This is heavily influenced by the worse quality of the two views. On the other hand, the weighted power mean allows for adjusting the emphasis on higher or lower distortions for each view in a more generalized way and is given as $\left(\frac{\alpha\cdot\mathcal{D}_l^p+(1-\alpha)\cdot\mathcal{D}_r^p}{2}\right)^{\frac{1}{p}}$,

where α is a weighting factor between left and right view, and p adjusts the focus on distortions, e.g., if $p > 1$, higher distortions are emphasized. If $p < 1$, lower distortions are emphasized. If $p = 1$, it simplifies to an arithmetic mean. It becomes the Euclidean norm if $p = 2$.

Modeling overall distortion: We combine the above disparity and visual distortion metrics to obtain the overall user-perceived distortion in quality using the following model: $\log(1 + \mathcal{D}_d) + \lambda_v \cdot \mathcal{D}_v$, where λ_v is a weight pertains to the visual quality of the spatial video, accounting for distortions in the left and right views. Small changes in disparity at different depths can have a disproportionately large effect on perceived depth. Disparity distortions can vary widely depending on content and encoding quality. We use a logarithmic function that helps compress a large range of disparity values into a smaller, more manageable scale.

Other factors: Multiple other factors must be considered during bitrate adaptation that we consider in our future work. For example, a virtual screen is used to view a spatial video on an XR headset. Unlike traditional displays, users can adjust the virtual screen at different distances from the eye. Similarly, users can change the virtual screen resolution dynamically. The ABR algorithm must adapt video resolution based on the distance and virtual screen resolution to utilize network bandwidth effectively. We will also consider other metrics like stalls and startup delays that are commonly explored in conventional streaming.

5 Conclusions

We present a comprehensive research agenda for spatial video streaming on emerging XR headsets. Our approach comprises view and packet adaptive bitrate algorithms, multipath support, and a QoE model, and tackles key challenges in delivering high-quality spatial 3D video content under constrained network settings. We are currently building a holistic live spatial video streaming system integrating all of our proposed methods with system-level optimization methods.

6 Acknowledgements

This work was supported by the National Science Foundation under grants CNS-2106150, CNS-2032033, CNS-2346528, and CNS-2340283. We thank the anonymous reviewers and the shepherd Robert LiKamWa for their thoughtful and constructive feedback.

References

[1] Apple Vision Pro. https://www.apple.com/apple-vision-pro/.
[2] Spatial Videos. https://developer.apple.com/videos/spatial-computing, Apr 2024.
[3] D. Baek, M. Dasari, S. R. Das, and J. Ryoo. dcSR: practical video quality enhancement using data-centric super resolution. *ACM CoNEXT*, 2021.
[4] J. Chakareski. Adaptive multi-view video streaming: Challenges and opportunities. *IEEE Communications Magazine*, 51(5):94–100, May 2013.
[5] J. Chakareski. Wireless streaming of interactive multi-view video via network compression and path diversity. *IEEE Trans. Communications*, 62(4):1350–1357, Apr. 2014.
[6] J. Chakareski, V. Velisavljević, and V. Stanković. User-action-driven view and rate scalable multiview video coding. *IEEE Trans. Image Processing*, 22(9):3473–3484, Sept. 2013. special issue on 3D Video Representation, Compression, and Rendering.
[7] Z. Chen, W. Zhou, and W. Li. Blind stereoscopic video quality assessment: From depth perception to overall experience. *IEEE Transactions on Image Processing*, 27(2):721–734, 2017.
[8] M. Dasari, A. Bhattacharya, S. Vargas, P. Sahu, A. Balasubramanian, and S. R. Das. Streaming 360-degree videos using super-resolution. In *Proc. IEEE INFOCOM*, 2020.
[9] M. Dasari, K. Kahatapitiya, S. R. Das, A. Balasubramanian, and D. Samaras. Swift: Adaptive video streaming with layered neural codecs. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pages 103–118, 2022.
[10] M. Dasari, S. Sanadhya, C. Vlachou, K.-H. Kim, and S. R. Das. Scalable ground-truth annotation for video qoe modeling in enterprise wifi. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pages 1–6. IEEE, 2018.
[11] M. Dasari, S. Vargas, A. Bhattacharya, A. Balasubramanian, S. R. Das, and M. Ferdman. Impact of device performance on mobile Internet QoE. In *Proceedings of the Internet Measurement Conference*, pages 1–7, 2018.
[12] Q. De Coninck and O. Bonaventure. Multipath quic: Design and evaluation. In *Proceedings of the 13th international conference on emerging networking experiments and technologies*, pages 160–166, 2017.
[13] S. A. El Mesloul Nasri, A. H. Sadka, N. Doghmane, and K. Khelil. Multiview Video Coding: A Comparative Study Between MVC and MV-HEVC. In *International Conference on Computer and Applications, Dubai, UAE*, pages 99–112. Springer, 2018.
[14] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure. Tcp extensions for multipath operation with multiple addresses. Technical report, 2013.
[15] H. Gupta, M. Curran, J. Longtin, T. Rockwell, K. Zheng, and M. Dasari. Cyclops: an FSO-based wireless link for VR headsets. In *Proceedings of the ACM SIGCOMM 2022 Conference*, pages 601–614, 2022.
[16] H. Gupta, M. Curran, J. Longtin, T. Rockwell, K. Zheng, and M. Dasari. Cyclops: An fso-based wireless link for vr headsets. In *SIGCOMM*, 2022.
[17] ISO/IEC JTC 1/SC 29. Information technology – Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats, 2012. ISO/IEC 23009-1:2012.
[18] ITU-T and ISO/IEC. Advanced Video Coding for Generic Audiovisual Services, 2014. ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC).
[19] T. Jin, M. Dasari, C. Smith, P. Apicharttrisorn, A. Rowe, and S. Seshan. MeshReduce: Scalable and Bandwidth Efficient Scene Capture for 3D Telepresence. In *Proceedings IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2024.
[20] Y. Liu, S. Ci, H. Tang, Y. Ye, and J. Liu. Qoe-oriented 3d video transcoding for mobile streaming. *ACM TOMM*, 8(3s):1–20, 2012.
[21] Z. Liu, G. Cheung, J. Chakareski, and Y. Ji. Multiple description coding and recovery of free viewpoint video for multi-path wireless network streaming. *IEEE J. Selected Topics in Signal Processing*, 9(1):151–164, Feb. 2015. special issue on Visual Signal Processing for Wireless Networks.
[22] G. Lv, Q. Wu, Y. Liu, Z. Li, Q. Tan, F. Yang, W. Chen, Y. Ma, H. Guo, Y. Chen, et al. Chorus: Coordinating mobile multipath scheduling and adaptive video streaming. In *MobiCom*, pages 246–262, 2024.
[23] T. W. d. P. Paiva, S. Ferlin, A. Brunstrom, O. Alay, and B. Y. L. Kimura. A first look at adaptive video streaming over multipath QUIC with shared bottleneck detection. In *ACM Multimedia*, pages 161–172, 2023.
[24] U. Paul, V. Gunasekaran, J. Liu, T. N. Narechania, A. Gupta, and E. Belding. Decoding the divide: Analyzing disparities in broadband plans offered by major US ISPs. In *ACM SIGCOMM*, 2023.
[25] A. Sarkar, J. Murray, M. Dasari, M. Zink, and K. Nahrstedt. L3BOU: Low latency, low bandwidth, optimized super-resolution backhaul for 360-degree video streaming. In *2021 IEEE International Symposium on Multimedia (ISM)*, pages 138–147. IEEE, 2021.
[26] S. Srinivasan, S. Shippey, E. Aryafar, and J. Chakareski. FBDT: Forward and backward data transmission across multiple RATs for high quality mobile virtual reality 360-degree video streaming. In *Proc. Multimedia Systems Conf.*, pages 130–141, Vancouver, BC, Canada, June 2023. ACM.
[27] T. Su, A. Sobhani, A. Yassine, S. Shirmohammadi, and A. Javadtalab. A DASH-based HEVC multi-view video streaming system. *Journal of Real-Time Image Processing*, 12:329–342, 2016.
[28] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Trans. on circuits and systems for video technology*, 22(12):1649–1668, 2012.
[29] G. Tech, J. Wegner, T. Wiegand, and G. Sullivan. Overview of the Multiview and 3D Extensions of High Efficiency Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(1):35–49, 2016.
[30] H. Urey, K. V. Chellappan, E. Erden, and P. Surman. State of the art in stereoscopic and autostereoscopic displays. *Proceedings of the IEEE*, 99(4):540–555, 2011.
[31] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7):560–576, 2003.
[32] Z. Zheng, Y. Ma, Y. Liu, F. Yang, Z. Li, Y. Zhang, J. Zhang, W. Shi, W. Chen, D. Li, et al. Xlink: Qoe-driven multi-path quic transport in large-scale video services. In *SIGCOMM*, pages 418–432, 2021.