Generative AI for Immersive Video: Recent Advances and Future Opportunities

Kaiyuan Hu[1] , **Yili Jin**[1] , **Hao Zhou**[1] , **Linfeng Du**[1] , **Jiangchuan Liu**[2] and **Xue Liu**[1,3]
[1]McGill University
[2]Simon Fraser University
[3]Mohamed bin Zayed University of Artificial Intelligence
{kaiyuan.hu, yili.jin, hao.zhou4, linfeng.du}@mail.mcgill.ca, jcliu@sfu.ca, xue.liu@mcgill.ca

Abstract

Immersive video serves as a key component of eXtended Reality (XR) that aims to create and interact with simulated virtual or hybrid environments. Such a technology allows users to experience immersive sensations that transcend time and space, and meanwhile continuously providing training data for emerging technologies like Embodied AI. Thanks to the advancements in capturing, computing, and display, recent years have witnessed many excellent works for XR and related hardware or software systems. However, challenges like high creation cost, lack of immersion, and limited scalability hinder the practical application of immersive video services. Whilst recently emerged Generative Artificial Intelligence (GenAI) provides us with new insights in tackling existing challenges. In this paper, we conduct a comprehensive survey into the recent advances and future opportunities on how GenAI can benefit immersive video services. By introducing a systematic taxonomy, we meticulously classify the pertinent techniques and applications into three well-defined categories aligned with the pipeline of immersive video service: content creation, network delivery, and client-side display. This categorization enables a structured exploration of the diverse roles on how GenAI can benefit immersive video service, providing a framework for a more comprehensive understanding and evaluation of these technologies. To the best of our knowledge, this work is the first systematic survey of GenAI in XR settings, laying a foundation for future research in this interdisciplinary domain.

1 Introduction

Immersive video is a core component of eXtended Reality (XR), which encompasses technologies like panoramic video (360° video) and volumetric video. These technologies aim to create an immersive environment that allows users to create, explore, or interact in a way that transcends traditional video experience. For panoramic videos, by offering the sphere-shape panoramic frame, users are allowed to freely switch their viewport in three degrees (yaw, pitch, and roll). Volumetric videos further promote the user experience to up to six degrees of freedom (6-DoF), allowing users to switch their viewport while changing their position in the virtual space they are immersed in. With applications spanning across entertainment, education, remote collaboration, and healthcare, immersive video reveals the potential to transform the way we interact with the virtual world.

Despite the rapid advancement in capturing, transmission, and display technologies, several key challenges continue to hinder the widespread adoption and effectiveness of immersive videos. Unlike traditional video capture technologies, which are nearing perfection, immersive video creation faces significant hurdles, primarily due to the high costs, associated with the specialized hardware and software requirements, limiting the accessibility of immersive content creation. Additionally, the inherently large data size of immersive videos, particularly volumetric video, poses significant challenges for network transmission. Current network limitations often create bottlenecks, impeding the seamless delivery of online immersive video services. On the client side, rendering immersive content demands substantial computational resources, leading to high latency and reduced display quality, especially for users with less powerful hardware. These challenges collectively constrain the overall user experience and limit the scalability of immersive video applications.

Generative Artificial Intelligence (GenAI) has emerged as a transformative technology that leverages generative models to create, modify, and optimize content, including texts, images, videos, and other media. With its rapid progress, it has shown significant promise in various fields, ranging from content generation to real-time optimization and decision-making. Such features also bring new insights in tackling the existing challenges spanning across diverse fields. In the context of immersive video, GenAI plays a pivotal role in overcoming several challenges. For content creation, it can drastically reduce production costs by automating the generation of high-quality video content, such as 3D models, textures, and entire virtual environments. This not only lowers the need for expensive hardware and manual workflows but also democratizes immersive video creation, making it more accessible to creators in various sectors. In terms of network delivery, GenAI enhances the compression and encoding of immersive video data, optimizing storage and enabling faster transmission. By intelligently predicting and reconstructing video frames, GenAI reduces data size, alleviates network congestion, and improves scalability. Lastly, for client-side display, GenAI optimizes rendering techniques by applying super-resolution and neural rendering techniques. This improves visual quality and reduces the computational load on devices, providing smoother experiences across a wide range of devices, from mobile phones to high-end VR headsets.

In this paper, we conduct a comprehensive investigation into current advances and future opportunities of GenAI on how it can contribute to immersive video services. To facilitate a deeper understanding, we introduce a taxonomy that aligns with the immersive video service pipeline, offering a structured framework for exploring GenAI's potential contributions across content creation, network delivery, and client-side display. To the best of our knowledge, this is the first systematic survey of GenAI in immersive video contexts, establishing a foundation for future research in this interdisciplinary field.

2 Related Work & Motivation

Though many works have explored the service pipelines of immersive video, particularly in areas such as content creation, network delivery, and client-side display, few have addressed the intersection of GenAI with immersive video technologies. Currently, there are dozens of survey papers [Yaqoob *et al.*, 2020; Jin *et al.*, 2023a; Viola and Cesar, 2023; Jin *et al.*, 2024b] focusing on immersive video, covering topics like video encoding, real-time rendering, and VR/AR content delivery. However, none have comprehensively explored how GenAI can be leveraged to address the unique challenges within the immersive video pipeline. The integration of GenAI with immersive video systems presents new opportunities to tackle long-standing issues such as high production costs, large data sizes, and rendering inefficiencies. Given the rapid advancements in both fields, there is a clear gap in the literature when it comes to understanding how GenAI can reshape the future of immersive video services. This paper aims to fill this gap by providing a comprehensive survey of the current state of GenAI applications in immersive video, highlighting the promising intersections between these technologies and their potential to revolutionize immersive media experiences.

3 Taxonomy

The immersive video service pipeline can be divided into three key stages: *Content Creation*, *Network Delivery*, and *Client-side Display*. Each stage plays a crucial role in delivering high-quality immersive video experiences, and GenAI can significantly enhance these stages. This section provides a detailed taxonomy of how GenAI can contribute to each stage of the immersive video pipeline.

3.1 Content Creation

Content creation for immersive video services encompasses two interconnected processes: *Enhanced Content Capturing*, which refines real-world data acquisition, and *Generative Content Creation*, which synthesizes immersive content from textual or sparse inputs. Together, these processes ensure high-quality, adaptable, and scalable content tailored for immersive services like panoramic and volumetric video.

Enhanced Content Capture

This section focuses on advancing real-world data acquisition through AI-driven methods. Multi-modal sensor fusion, such as combining RGB, LiDAR, and depth data, helps reduce alignment errors and enhances the accuracy of dynamic scene reconstructions. Data enhancement techniques, including diffusion models, play a crucial role in denoising and completing incomplete 3D objects, preserving geometric details, and improving overall data quality. Additionally, methods like GAN-based super-resolution enhance immersive content created with low-quality data, significantly boosting visual fidelity and realism.

Generative Content Creation

GenAI techniques empower creators to generate immersive content from minimal inputs. Methods such as Neural Radiance Fields (NeRF) [Mildenhall *et al.*, 2021] enable the creation of photorealistic 3D scenes from 2D images, while diffusion-based approaches enhance 3D object and scene generation for volumetric video applications. Recent advancements in text-to-3D generation allow for the creation of detailed 3D models from textual descriptions, pushing the boundaries of creativity in virtual environments and interactive storytelling. These innovations streamline content generation, making it faster and more accessible for creators.

3.2 Network Delivery

Network delivery in immersive video services focuses on efficiently transmitting large-scale, high-quality content, such as volumetric video, over networks. This stage addresses two critical challenges: *GenAI-assisted Transmission Optimization*, which enhances compression and reduces data sizes, and *Semantic-based Transmission*, which prioritizes critical content to optimize bandwidth usage.

GenAI-assisted Transmission Optimization

GenAI improves the transmission of immersive video content by optimizing compression algorithms and reducing data sizes while maintaining content quality. Traditional video compression methods often struggle with the complexity of immersive videos, especially for volumetric content. GenAI-driven approaches, such as neural-based compression, address these limitations by using neural representations to compress data efficiently.

Semantic-based Transmission

Semantic-based transmission optimizes bandwidth usage by identifying and prioritizing critical regions of immersive video content. Certain segments, such as the viewer's focal area, are more important than peripheral regions, particularly in formats like panoramic video. Traditional transmission methods treat all parts of the video equally, which can lead to inefficient bandwidth use. While semantic transmission techniques extract semantic information to ensure higher fidelity transmission for key features while compressing less important regions more aggressively.

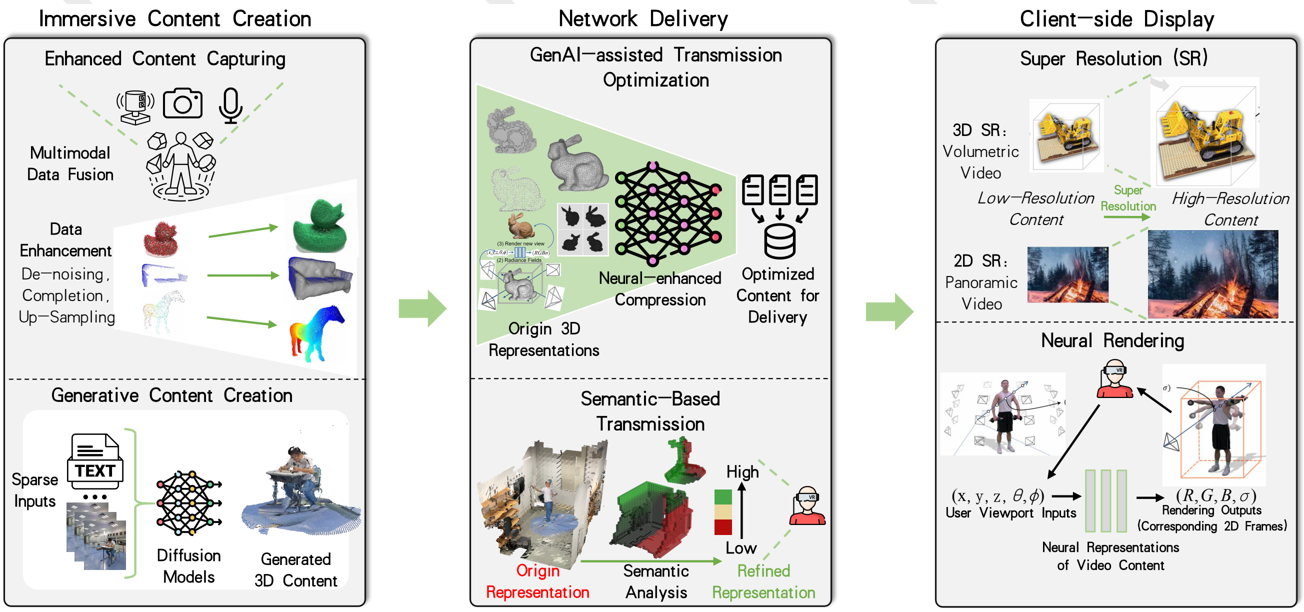

Figure 1: Overview of the Immersive Video Service Pipeline with GenAI Enhancements: This figure illustrates the three main stages of the immersive video service pipeline: **Content Creation** focuses on enhancing data acquisition and enabling generative content creation. **Transmission** optimizes data size and prioritizes important content using GenAI-assisted compression and semantic-based transmission. **Display** improves visual quality and rendering efficiency using super-resolution and neural rendering techniques.

3.3 Client-side Display

Client-side display for immersive video services focuses on rendering and presenting high-quality, responsive content to the user. This process involves two key aspects: *Super-Resolution (SR) for Display*, which enhances the quality of both 3D and 2D immersive video, and *Neural Rendering*, which optimizes the efficiency and realism of rendering processes for real-time immersive experiences.

Super-Resolution (SR) for Display

Super-resolution (SR) enhances the resolution and visual fidelity of immersive video content, in both volumetric and panoramic formats. 3D super-resolution techniques enhance 3D models like point clouds and meshes, refining spatial resolution and providing finer geometry details. In the case of panoramic videos, 2D super-resolution techniques are applied post-rendering to improve the clarity and detail of low-resolution frames. Advanced methods like temporal-consistent diffusion models ensure stability across video sequences, preserving consistency and enhancing the overall viewing experience.

Neural Rendering

Neural rendering leverages deep learning models to simulate and render 3D scenes more efficiently than traditional methods. By utilizing techniques like NeRF and Video Octree (VOctree) structures, neural rendering enables interactive, real-time rendering of volumetric video content. This approach enhances the rendering process by reducing memory overhead and computational load, allowing for photorealistic and editable 3D content. Additionally, hybrid frameworks that combine neural rendering with rasterization provide seamless integration for VR headsets, offering users an immersive experience with optimized real-time performance.

4 Recent Advances

Following the immersive video service pipeline outlined in Section 3, this section reviews recent advances in GenAI across the three key stages: content creation, network delivery, and client-side display. We first explore how GenAI is transforming content creation through efficient data acquisition and synthetic data generation. Innovations like NeRF and text-to-3D generation enable the creation of immersive content from minimal inputs. In network delivery, advancements in neural representations improve data transmission efficiency while semantic-aware compression optimizes bandwidth usage. Finally, we examine breakthroughs in client-side display, including neural-enhanced super-resolution and real-time neural rendering, which elevate visual fidelity and streamline the rendering process. By investigating the advances along the service pipeline, we highlight how GenAI addresses the unique challenges of immersive video services while uncovering the opportunities for future innovation.

4.1 Content Creation

Immersive video content creation serves as the foundation of the entire service pipeline. Traditionally, this process involves capturing video content from the real world using complex device setups, such as calibrated camera arrays. However, with the advent of GenAI technologies, the paradigm of immersive content creation has undergone a significant shift. Instead of relying solely on real-world capture, GenAI enables the direct generation of immersive content from simple inputs or prompts, revolutionizing the way we produce and interact with immersive media.

Enhanced Content Capturing

Capturing immersive video data from the real world has long been the cornerstone of traditional immersive video creation. However, this process often faces significant challenges, including high acquisition costs, sensor limitations, and data scarcity. Recent advancements in GenAI have revolutionized the way immersive video data is acquired, offering transformative improvements in quality, efficiency, and scalability. By leveraging cutting-edge AI techniques, such as diffusion models, transformers, and generative adversarial networks (GANs), GenAI enables more robust and cost-effective solutions for real-world data capturing.

In this part, we explore how GenAI enhances the acquisition of immersive video data through two key approaches: **sensor fusion** and **data enhancement**. These advancements not only address the limitations of traditional methods but also pave the way for new possibilities in immersive video creation.

Sensor Fusion across multiple modalities has always been one of the key challenges in immersive video capturing, particularly for volumetric video, which integrates diverse modalities such as RGB, depth, and spatial audio. Unlike traditional 2D media, volumetric video demands precise alignment and synchronization of heterogeneous data streams to reconstruct continuous high-fidelity 3D scenes. Traditional fusion methods, such as geometric calibration or handcrafted feature matching, often struggle with misalignment, noise amplification, and computational inefficiency, especially for dynamic scenes or unstructured environments.

Recent advances in GenAI have introduced new solutions to the aforementioned challenges. For instance, in LiDAR4D [Zheng *et al.*, 2024], a dynamic neural field framework is proposed for space-time LiDAR synthesis, combining LiDAR point clouds with RGB video frames to reconstruct 4D (3D+time) scenes. This framework leverages a spatiotemporal neural field to model dynamic objects and environments, addressing the challenges of temporal consistency and multimodal fusion in point cloud-based immersive video creation. Additionally, recent advancements in image fusion, such as the work by [Zhang *et al.*, 2025] on natural language-guided infrared and visible image fusion, leverage CLIP to guide the fusion process using natural language expressions. By integrating CLIP with a low-redundancy feature fusion network, such a method has shown great promise in improving the quality of fused images by reducing redundant features and enhancing focus on critical information.

Data Enhancement is critical to refining raw data captured from sensors. Leveraging GenAI technologies like GANs and diffusion models, the quality and fidelity of the raw data could be improved.

One of the most common challenges in real-world capturing is noise in depth data, often caused by environmental interferences such as surface reflection or absorption. While diffusion models have shown significant potential in denoising 3D point clouds, one of the most common data formats, recent advancements have further enhanced this capability. A notable development is P2P-Bridge [Vogel *et al.*, 2024], which introduces a novel framework that adapts Diffusion Schrödinger bridges to point clouds. Unlike traditional methods that predict point-wise displacements based on point features or learned noise distributions, P2P-Bridge learns an optimal transport plan between paired point clouds.

Building on this, another significant advancement is the Point Cloud Upsampling Diffusion Model (PUDM) [Qu *et al.*, 2024], which targets point cloud upsampling while simultaneously addressing denoising. PUDM treats sparse point clouds as conditions and iteratively learns the transformation relationship between sparse and dense point clouds. The model uses a denoising diffusion probabilistic approach, enhancing the quality of point clouds by removing noise and filling in missing details. Moreover, PUDM employs a dual mapping paradigm to improve feature discernment, learning complex geometric details in the point cloud without the need for additional upsampling modules. PUDM not only enhances point cloud quality but also enables high-quality arbitrary-scale upsampling during inference, marking a significant leap forward in point cloud refinement for immersive applications.

Another real-world capturing challenge is the presence of gaps or missing data due to occlusions and sensor limitations, which significantly degrade the reconstruction quality. One recent approach SDS-Complete [Kasten *et al.*, 2024] notably employs pretrained text-to-image diffusion models to guide the completion of missing parts of point clouds. SDS-Complete leverages semantic guidance provided by textural descriptions of objects, enabling it to generate missing surfaces while aligning with the known point cloud and the semantics of the object. By leveraging test-time optimization, SDS-Complete ensures that the generated points align with both the original point cloud and the global object characteristics, maintaining both overall geometrical accuracy and realism of the reconstructed 3D scene.

Generative Content Creation

GenAI has opened up innovative ways to create immersive video content, especially from sparse or minimal input data. By utilizing advanced generative models, GenAI can generate high-quality 3D content and dynamic scenes that offer an immersive experience, often from simple inputs such as images, sketches, or even texts. Below, we explore two key approaches that have seen significant advancements: Text-to-3D Generation and Image-to-3D Generation.

Text-to-3D generation converts textual descriptions into 3D content. Advances in models like Text2Mesh [Michel *et al.*, 2022] allows users to generate 3D scenes and objects from simple text prompts, enabling creators to build immersive environments without extensive 3D design skills. However, traditional solutions often suffer from slow convergence, missing details, or inaccurate 3D geometry. While one notable progress lies in DreamTime [Huang *et al.*, 2024], which solves these problems by improving the optimization process through diffusion-guided sampling. By utilizing diffusion-guided sampling, DreamTime better aligns the 3D optimization process, resulting in faster generation of more realistic, detailed 3D scenes.

Building upon the previous advancements in text-to-3D generation, another work GVGEN [He *et al.*, 2024] innovates the 3D content generation by employing a structured volumetric representation GaussianVolume, which organizes 3D Gaussian points into a fixed-volume structure. This organization allows for the capture of intricate texture details within a volume composed of a fixed number of Gaussians. To optimize this representation, GVGEN introduces the 'Candidate Pool Strategy', a unique pruning and densifying method that enhances detail fidelity through selective optimization. Furthermore, GVGEN utilizes a coarse-to-fine generation pipeline, which first constructs a basic geometric structure and then predicts complete Gaussian attributes, enabling the model to generate instances with detailed 3D geometry. This approach simplifies the generation process and empowers the model to produce more accurate and realistic 3D structures.

Image-to-3D Generation involves converting 2D images into 3D models, enabling the creation of immersive environments from single or multiple images. Recent advancements have significantly improved the quality and efficiency of this process. A notable progress is PC^2 [Melas-Kyriazi *et al.*, 2023], which reconstructs 3D shapes from a single RGB image using a conditional denoising diffusion process. This approach begins with a set of 3D points sampled from a Gaussian distribution and iteratively refines them to match the object's shape. The key innovation is projection conditioning, where local image features are projected onto the partially denoised point cloud at each diffusion step. This technique enables the generation of high-resolution, sparse geometries that align well with the input image, and it can also predict point colors after shape reconstruction. Due to the probabilistic nature of the diffusion process, the method can generate multiple plausible 3D shapes from a single image. This approach has shown significant improvements over previous methods, particularly in handling complex real-world data.

4.2 Network Delivery

Efficient data transmission is essential for smooth immersive video experiences, especially when dealing with emerging large-scale, high-quality content such as volumetric video. The fundamental challenge lies in the sheer size of the data generated by immersive content, which can include 3D models, neural representation models, or high-resolution panoramic video streams. As immersive content evolves and becomes more complex, the need for optimized transmission strategies becomes increasingly critical to ensure a seamless user experience.

GenAI-assisted Transmission Optimization

GenAI's rapid progress provides new solutions to optimize the data transmission pipeline for immersive video content. One of the primary ideas is through neural-enhanced compression techniques [Ma *et al.*, 2020; Jin *et al.*, 2025b]. Traditional video compression algorithms, such as H.265 [Pastuszak and Abramowski, 2015] or VP9 [Bienik *et al.*, 2016], often struggle to maintain high quality while reducing file size, especially with the complex, multi-dimensional data of immersive videos. By integrating GenAI into the transmission process, the compression performance can be significantly improved. A recent achievement by [Shi *et al.*, 2024] proposes an end-to-end pipeline for compressing volumetric video using neural-based representations. This approach encodes the differences between consecutive NeRFs, thus effectively capturing dynamic aspects of the scene and reducing data sizes while preserving key content details. Additionally, HiNeRV [Kwan *et al.*, 2023], proposes an implicit neural representation (INR) approach that integrates lightweight layers and hierarchical positional encodings. Unlike existing methods, HiNeRV combines depth-wise convolutional, MLP, and interpolation layers into a unified architecture, enabling simultaneous encoding of videos at both frame and patch levels for enhanced flexibility and efficiency. A tailored codec and training pipeline incorporating pruning and quantization further optimize performance retention during lossy compression.

These advancements showcase GenAI's transformative potential for immersive video delivery. By refining compression algorithms to prioritize dynamic scene elements and reduce redundancy, GenAI enables efficient transmission of high-quality content over existing networks. This eliminates the need for infrastructure overhauls while accelerating streaming speeds and enhancing user experiences, particularly for bandwidth-intensive formats like volumetric video. Such neural-driven methods ensure scalable, high-fidelity delivery of immersive media.

Semantic-based Transmission

Semantic-based transmission focuses on optimizing the transmission pipeline by identifying and prioritizing important regions of immersive video content, reducing redundancy and thus avoiding unnecessary bandwidth consumption. In immersive video services, certain segments, such as the viewer's focus area, are more critical than others as proved by [Hu *et al.*, 2023a; Hu *et al.*, 2023b]. Traditional transmission methods like [Qian *et al.*, 2019; Jin *et al.*, 2023b; Liu *et al.*, 2023] treat all parts of the video equally, which can lead to inefficient bandwidth use, especially in content like panoramic or volumetric videos where peripheral areas are less important.

A recent approach by [Xie *et al.*, 2024] leverages semantic communication (SemCom) techniques to enhance the efficiency of transmitting 3D point cloud data. This method extracts both local and global semantic information from the point clouds, enabling the transmission to focus on the most important features. The local semantic encoder extracts detailed information from specific regions of the point cloud, while the global semantic encoder captures the overall structure, ensuring both local detail and global context are preserved. By applying these techniques, the system ensures that critical data is transmitted with higher fidelity, while less significant regions can be more aggressively compressed, reducing overall data size without compromising quality.

For more challenging live streaming scenarios, LiveVV is proposed [Hu *et al.*, 2025b] to address the challenges of live volumetric video streaming by using scene segmentation and adaptive transmission. The system identifies dynamic content within the scene, prioritizing it for higher fidelity transmission, while less critical static elements are compressed more aggressively to reduce bandwidth usage. Additionally, Volumetric Video Adaptive Bitrate Streaming (VABR) is employed to dynamically adjust the streaming quality based on real-time network conditions, ensuring a smooth user experience.

4.3 Client-side Display

The display stage of immersive video content is pivotal to ensuring that the final output is both visually compelling and responsive. GenAI's ability to improve quality and optimize rendering has been transformative, addressing the computational challenges of rendering and delivering immersive video, especially in real-time. This section will explore the key ways in which GenAI is enhancing the display pipeline of immersive video services.

Super-Resolution (SR) for Display

GenAI models have made significant strides in applying super-resolution (SR) techniques to immersive video, improving the resolution and realism, especially to volumetric content, which typically requires higher computational resources for rendering at real-time speeds. In the context of volumetric video, 3D super-resolution is used to enhance point clouds and 3D meshes. Volumetric video often relies on lower-resolution 3D representations due to the high bandwidth consumption and computational cost of rendering. Applying 3D SR techniques improves these representations by increasing the spatial resolution of point clouds or mesh details, providing finer geometry details.

For instance, the study GaussianSR [Yu *et al.*, 2024] is proposed to leverage 2D diffusion priors learned from large-scale image data to enhance 3D super-resolution. By distilling 2D knowledge into 3D representations using Score Distillation Sampling (SDS), GaussianSR improves the resolution of 3D Gaussian primitives, leading to higher-quality synthesized views. This approach demonstrates the potential of integrating 2D diffusion priors into 3D SR for volumetric video, offering a pathway to enhance visual fidelity in immersive video services.

Panoramic video is another immersive video paradigm that comes with less degree-of-freedom. 2D SR techniques are often applied post-rendering to enhance the display quality of frames with low resolution. Leveraging the advances of GenAI methods, immersive video services can significantly enhance the quality of panoramic content, ensuring a clearer and more detailed viewing experience for users.

For instance, [Hu *et al.*, 2025a] introduce a method that represents each pixel as a continuous Gaussian field, allowing for the refinement and upsampling of encoded features through 2D Gaussian Splatting (2D-GS). This approach enhances the representation ability by establishing long-range dependencies and dynamically assigning Gaussian kernels to pixels, resulting in high-fidelity super-resolution with fewer parameters than existing methods.

Additionally, [Zhou *et al.*, 2024b] introduce a text-guided latent diffusion framework for video upscaling. This framework ensures temporal coherence through two key mechanisms: locally, it integrates temporal layers into U-Net and VAE-Decoder, maintaining consistency within short sequences; globally, it introduces a flow-guided recurrent latent propagation module to enhance overall video stability by propagating and fusing latent representations across entire sequences. This approach allows for flexible control over the balance between restoration and generation, enabling a trade-off between fidelity and quality.

By employing SR techniques in both 3D and 2D domains, immersive video services can deliver high-resolution content with lower costs, improving the overall display quality of volumetric and panoramic video while minimizing the demands on system resources.

Neural Rendering

Neural rendering is an emerging technique that uses deep learning models to simulate the process of rendering 3D scenes more efficiently. Unlike traditional rendering methods, which rely on physically accurate simulations of light and materials, neural rendering uses neural networks to generate realistic images based on learned patterns and approximations.

A novel work, NeuVV [Zhang *et al.*, 2022b], proposes a new approach to volumetric video rendering using NeRF, enabling immersive, interactive 3D video experiences. By incorporating factorization techniques like hyper-spherical harmonics (HH) decomposition and learnable basis representations, NeuVV enhances rendering efficiency and reduces memory overhead, crucial for real-time performance. A key innovation is its use of a Video Octree structure, which allows dynamic manipulation of video content—such as repositioning 3D performances and adjusting textures—at interactive speeds. This is paired with a hybrid neural-rasterization framework that integrates seamlessly with VR headsets, facilitating high-quality, real-time volumetric video rendering. NeuVV's ability to deliver photorealistic, editable content makes it highly relevant for the future of immersive media applications like virtual and augmented reality.

Another notable work, YuZu [Zhang *et al.*, 2022a], introduces a system for volumetric video streaming that optimizes both visual quality and bandwidth efficiency through adaptive 3D super-resolution. Unlike traditional volumetric streaming systems, which require transmitting dense 3D point clouds or meshes at high bandwidths, YuZu employs a novel split-rendering framework that offloads SR tasks to edge servers. By streaming low-resolution 3D content and applying lightweight 3D SR models at the client side, YuZu reduces bandwidth usage by 4.1× while maintaining perceptual quality comparable to native high-resolution streams.

A key innovation is the first QoE model for volumetric streaming, which quantifies user-perceived quality based on spatial resolution, temporal consistency, and rendering latency. This model enables YuZu to dynamically adapt SR parameters (e.g., spatial and temporal upsampling rates) in response to network fluctuations, achieving 23% higher QoE than static approaches. The system also introduces line-rate SR processing, leveraging GPU-CPU co-design to achieve real-time performance (60 FPS) on commodity hardware. Evaluations on volumetric datasets like 8i Voxelized Light Fields [Krivokuća *et al.*, 2018] demonstrate YuZu's ability to reduce client-side rendering latency by 35% while preserving visual fidelity, making it a scalable solution for immersive applications such as VR and telepresence [Jin *et al.*, 2025a; Duan *et al.*, 2025; Jin *et al.*, 2024a].

5 Challenges & Opportunities

This section will identify the challenges of GenAI-based immersive services, including scalability, latency demand, bandwidth bottleneck, and privacy concerns while exploring the potential opportunities to tackle these challenges.

5.1 Scalability

Scalability is a critical factor in the widespread application of immersive services. The capability to scale these applications efficiently across different networks, user devices, and environments is of great importance to ensure a seamless experience and accommodate the growing demand. In particular, scalability challenges include handling changing user numbers and loads, increasing data complexity and service demands, and maintaining performance across diverse hardware and network conditions. Scalability allows users to expand their immersive service demands without facing bottlenecks in performance or accessibility [Zink *et al.*, 2019], which is a very common request for gaming, enterprise collaboration, healthcare, or education. Meanwhile, it also has higher requirements for related GenAI techniques, indicating that the algorithms must have great scalability. For instance, the AI model should be able to handle changing amounts of inputs and generate desired content with acceptable latency.

5.2 Latency Demand

Immersive videos have stringent requirements for ultra-low latency connections, which rely on real-time data transmission and immediate feedback to create realistic and seamless environments. High latency can result in noticeable lag, motion sickness, and a diminished sense of presence [Chang *et al.*, 2020], making the reduction of latency a crucial challenge for developers and network providers. Meanwhile, despite the great potential of GenAI techniques in immersive services, they may lead to extra latency for the overall process. For example, existing studies by [Zhou *et al.*, 2024a] have explored the broad application of large language models (LLMs) in image and video processing, but LLMs may significantly increase the latency due to the huge number of parameters. The progress of 5G and beyond networks is a significant step toward achieving ultra-low latency in immersive applications [Hazarika and Rahmati, 2023], enhancing the real-time capabilities of immersive systems. Future advancements in network technology, such as 6G, mmWave, and Terahertz communications, are expected to further reduce latency and unlock new possibilities for immersive applications.

5.3 Bandwidth Bottleneck

Immersive video services rely on high-bandwidth connections to deliver high-quality and reliable user experiences. Specifically, immersive applications usually require real-time transmissions of high-resolution 3D graphics, spatial audio, and interactive elements in the immersive environment. Unlike traditional video streaming, immersive content must dynamically adjust to user movements and interactions, requiring significantly more bandwidth. Additionally, many immersive applications will offload processing requests to cloud servers due to computational resource constraints on local devices [Wu *et al.*, 2024]. Such task offloading will also increase network traffic and require extra bandwidth. Without sufficient bandwidth, users may encounter lag, buffering, and degraded video quality, degrading the quality of experience in immersive environments. Despite the satisfactory performance of GenAI-based techniques, they also require more computational resources, which indicates more requests for computational task offloading, e.g., offloading content generation tasks to cloud servers or edge nodes [Gül *et al.*, 2020]. To this end, multiple bandwidth optimization techniques have been proposed, including adaptive streaming, compression algorithms, network prioritization, edge caching, etc.

5.4 Privacy Concern

Privacy is a critical concern in the application of immersive video services [Jin *et al.*, 2024c]. For instance, the devices that provide immersive services can collect user behavioral patterns and personalized avatars can be used to steal personal financial information illegally. In addition, based on the data collected during immersive services, one can make inferences about the users regarding their location, occupations, interests, and behavior [Jin *et al.*, 2022]. The privacy concern may increase when GenAI technologies are involved, in which more detailed user data is collected for content creation and display, including body signals, facial features, and emotional responses. In particular, AI models usually require vast amounts of data for training and customization, which often includes sensitive personal information. This data can expose users to privacy risks, including identity theft and data breaches. Furthermore, companies may collect user inputs and interactions to refine their models without explicit consent, raising concerns about data ownership and transparency. GenAI also enables the creation of synthetic media in immersive video services, such as deepfakes [Mirsky and Lee, 2021] and AI-generated text [Sadasivan *et al.*, 2023]. These can be used to impersonate individuals, manipulate opinions, or spread false information. This problem may be alleviated by robust privacy measures such as strong encryption, access controls, and continuous security audits.

References

[Bienik *et al.*, 2016] Juraj Bienik, Miroslav Uhrina, Michal Kuba, and Martin Vaculik. Performance of h. 264, h. 265, vp8 and vp9 compression standards for high resolutions. In *2016 19th International Conference on Network-Based Information Systems (NBiS)*, pages 246–252. IEEE, 2016.

[Chang *et al.*, 2020] Eunhee Chang, Hyun Taek Kim, and Byounghyun Yoo. Virtual reality sickness: a review of causes and measurements. *International Journal of Human–Computer Interaction*, 36(17):1658–1682, 2020.

[Duan *et al.*, 2025] Xize Duan, Yili Jin, Lei Zhang, and Fangxin Wang. Semconf: A system for multiparty semantic video conferencing. In *Proceedings of the 35th Workshop on Network and Operating System Support for Digital Audio and Video*, NOSSDAV '25, page 71–77, 2025.

[Gül *et al.*, 2020] Serhan Gül, Dimitri Podborski, Jangwoo Son, Gurdeep Singh Bhullar, Thomas Buchholz, Thomas Schierl, and Cornelius Hellge. Cloud rendering-based volumetric video streaming system for mixed reality services. In *Proceedings of the 11th ACM multimedia systems conference*, pages 357–360, 2020.

[Hazarika and Rahmati, 2023] Ananya Hazarika and Mehdi Rahmati. Towards an evolved immersive experience: Exploring 5g-and beyond-enabled ultra-low-latency communications for augmented and virtual reality. *Sensors*, 23(7):3682, 2023.

[He *et al.*, 2024] Xianglong He, Junyi Chen, Sida Peng, Di Huang, Yangguang Li, Xiaoshui Huang, Chun Yuan, Wanli Ouyang, and Tong He. Gvgen: Text-to-3d generation with volumetric representation. In *European Conference on Computer Vision*, pages 463–479. Springer, 2024.

[Hu *et al.*, 2023a] Kaiyuan Hu, Yili Jin, Haowen Yang, Junhua Liu, and Fangxin Wang. Fsvvd: A dataset of full scene volumetric video. In *Proceedings of the 14th ACM Multimedia Systems Conference*, MMSys '23, 2023.

[Hu *et al.*, 2023b] Kaiyuan Hu, Haowen Yang, Yili Jin, Junhua Liu, Yongting Chen, Miao Zhang, and Fangxin Wang. Understanding user behavior in volumetric video watching: Dataset, analysis and prediction. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

[Hu *et al.*, 2025a] Jintong Hu, Bin Xia, Bin Chen, Wenming Yang, and Lei Zhang. Gaussiansr: High fidelity 2d gaussian splatting for arbitrary-scale image super-resolution. In *The 39th Annual AAAI Conference on Artificial Intelligence*, pages 3554–3562, 2025.

[Hu *et al.*, 2025b] Kaiyuan Hu, Yongting Chen, Kaiying Han, Boyan Li, Haowen Yang, Yili Jin, Junhua Liu, and Fangxin Wang. Livevv: Human-centered live volumetric video streaming system. *IEEE Internet of Things Journal*, 2025.

[Huang *et al.*, 2024] Yukun Huang, Jianan Wang, Yukai Shi, Boshi Tang, Xianbiao Qi, and Lei Zhang. Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. In *The Twelfth International Conference on Learning Representations*, 2024.

[Jin *et al.*, 2022] Yili Jin, Junhua Liu, Fangxin Wang, and Shuguang Cui. Where are you looking? a large-scale dataset of head and gaze behavior for 360-degree videos and a pilot study. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 2022.

[Jin *et al.*, 2023a] Yili Jin, Kaiyuan Hu, Junhua Liu, Fangxin Wang, and Xue Liu. From capture to display: A survey on volumetric video. *arXiv preprint arXiv:2309.05658*, 2023.

[Jin *et al.*, 2023b] Yili Jin, Junhua Liu, Fangxin Wang, and Shuguang Cui. Ebublio: Edge-assisted multiuser 360° video streaming. *IEEE Internet of Things Journal*, 10(17):15408–15419, 2023.

[Jin *et al.*, 2024a] Yili Jin, Xize Duan, Fangxin Wang, and Xue Liu. Headsetoff: Enabling photorealistic video conferencing on economical VR headsets. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7928–7936, 2024.

[Jin *et al.*, 2024b] Yili Jin, Junhua Liu, Kaiyuan Hu, and Fangxin Wang. A networking perspective of volumetric video service: Architecture, opportunities, and case study. *IEEE Network*, 38(6):138–145, 2024.

[Jin *et al.*, 2024c] Yili Jin, Wenyi Zhang, Zihan Xu, Fangxin Wang, and Xue Liu. Privacy-preserving gaze-assisted immersive video streaming. *IEEE Transactions on Mobile Computing*, 23(12):15098–15113, 2024.

[Jin *et al.*, 2025a] Yili Jin, Xize Duan, Kaiyuan Hu, Fangxin Wang, and Xue Liu. 3d video conferencing via onhand devices. *IEEE Trans. Circuits Syst. Video Technol.*, 35(1):900–910, 2025.

[Jin *et al.*, 2025b] Yili Jin, Jiahao Li, Bin Li, and Yan Lu. Neural image compression with regional decoding. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(3), 2025.

[Kasten *et al.*, 2024] Yoni Kasten, Ohad Rahamim, and Gal Chechik. Point cloud completion with pretrained text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[Krivokuća *et al.*, 2018] Maja Krivokuća, Philip A. Chou, and Patrick Savill. 8i voxelized surface light field (8ivslf) dataset. Input Document m42914, ISO/IEC JTC1/SC29 WG11 (MPEG), Ljubljana, July 2018.

[Kwan *et al.*, 2023] Ho Man Kwan, Ge Gao, Fan Zhang, Andrew Gower, and David Bull. Hinerv: Video compression with hierarchical encoding-based neural representation. In *Advances in Neural Information Processing Systems*, volume 36, pages 72692–72704, 2023.

[Liu *et al.*, 2023] Junhua Liu, Boxiang Zhu, Fangxin Wang, Yili Jin, Wenyi Zhang, Zihan Xu, and Shuguang Cui. Cav3: Cache-assisted viewport adaptive volumetric video streaming. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 173–183, 2023.

[Ma *et al.*, 2020] Siwei Ma, Xinfeng Zhang, Chuanmin Jia, Zhenghui Zhao, Shiqi Wang, and Shanshe Wang. Image and video compression with neural networks: A review. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1683–1698, 2020.

[Melas-Kyriazi *et al.*, 2023] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12923–12932, 2023.

[Michel *et al.*, 2022] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022.

[Mildenhall *et al.*, 2021] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[Mirsky and Lee, 2021] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1):1–41, 2021.

[Pastuszak and Abramowski, 2015] Grzegorz Pastuszak and Andrzej Abramowski. Algorithm and architecture design of the h. 265/hevc intra encoder. *IEEE Transactions on circuits and systems for video technology*, 26(1):210–222, 2015.

[Qian *et al.*, 2019] Feng Qian, Bo Han, Jarrell Pair, and Vijay Gopalakrishnan. Toward practical volumetric video streaming on commodity smartphones. In *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*, pages 135–140, 2019.

[Qu *et al.*, 2024] Wentao Qu, Yuantian Shao, Lingwu Meng, Xiaoshui Huang, and Liang Xiao. A conditional denoising diffusion probabilistic model for point cloud upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Sadasivan *et al.*, 2023] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

[Shi *et al.*, 2024] Yuang Shi, Ruoyu Zhao, Simone Gasparini, Géraldine Morin, and Wei Tsang Ooi. Volumetric video compression through neural-based representation. In *Proceedings of the 16th International Workshop on Immersive Mixed and Virtual Environment Systems*, pages 85–91, 2024.

[Viola and Cesar, 2023] Irene Viola and Pablo Cesar. Volumetric video streaming: Current approaches and implementations. *Immersive Video Technologies*, pages 425–443, 2023.

[Vogel *et al.*, 2024] Mathias Vogel, Keisuke Tateno, Marc Pollefeys, Federico Tombari, Marie-Julie Rakotosaona, and Francis Engelmann. P2p-bridge: Diffusion bridges for 3d point cloud denoising. In *European Conference on Computer Vision*, pages 184–201. Springer, 2024.

[Wu *et al.*, 2024] Yixuan Wu, Kaiyuan Hu, Qian Shao, Jintai Chen, Danny Z Chen, and Jian Wu. Teleor: Real-time telemedicine system for full-scene operating room. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 628–638. Springer, 2024.

[Xie *et al.*, 2024] Shangzhuo Xie, Qianqian Yang, Yuyi Sun, Tianxiao Han, Zhaohui Yang, and Zhiguo Shi. Semantic communication for efficient point cloud transmission. In *2024 IEEE Global Communications Conference, GLOBECOM*, pages 2948–2953, 2024.

[Yaqoob *et al.*, 2020] Abid Yaqoob, Ting Bi, and Gabriel-Miro Muntean. A survey on adaptive 360 video streaming: Solutions, challenges and opportunities. *IEEE Communications Surveys & Tutorials*, 22(4):2801–2838, 2020.

[Yu *et al.*, 2024] Xiqian Yu, Hanxin Zhu, Tianyu He, and Zhibo Chen. Gaussiansr: 3d gaussian super-resolution with 2d diffusion priors. *arXiv preprint arXiv:2406.10111*, 2024.

[Zhang *et al.*, 2022a] Anlan Zhang, Chendong Wang, Bo Han, and Feng Qian. YuZu: Neural-enhanced volumetric video streaming. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pages 137–154, 2022.

[Zhang *et al.*, 2022b] Jiakai Zhang, Liao Wang, Xinhang Liu, Fuqiang Zhao, Minzhang Li, Haizhao Dai, Boyuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Neuvv: Neural volumetric videos with immersive rendering and editing. *arXiv preprint arXiv:2202.06088*, 2022.

[Zhang *et al.*, 2025] Jundong Zhang, Kangjian He, Dan Xu, and Hongzhen Shi. Clip-based natural language-guided low-redundancy fusion of infrared and visible images. *IEEE Transactions on Consumer Electronics*, 2025.

[Zheng *et al.*, 2024] Zehan Zheng, Fan Lu, Weiyi Xue, Guang Chen, and Changjun Jiang. Lidar4d: Dynamic neural fields for novel space-time view lidar synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5145–5154, 2024.

[Zhou *et al.*, 2024a] Hao Zhou, Chengming Hu, Ye Yuan, Yufei Cui, Yili Jin, Can Chen, Haolun Wu, Dun Yuan, Li Jiang, Di Wu, Xue Liu, Charlie Zhang, Xianbin Wang, and Jiangchuan Liu. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. *IEEE Communications Surveys & Tutorials*, 2024.

[Zhou *et al.*, 2024b] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2535–2545, 2024.

[Zink *et al.*, 2019] Michael Zink, Ramesh Sitaraman, and Klara Nahrstedt. Scalable 360 video stream delivery: Challenges, solutions, and opportunities. *Proceedings of the IEEE*, 107(4):639–650, 2019.