When AI Becomes the Viewer: Rethinking Immersive Networking for Human–AI Interaction

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

Amitabh Varshney
University of Maryland
College Park, MD, USA
varshney@umd.edu

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

Abstract

The rapid advancement of foundation models is transforming how humans communicate and collaborate with generative AI (GenAI). Yet, despite remarkable progress in text, audio, and video interactions, today's GenAI platforms remain confined to 2D interfaces, lacking the spatial understanding and embodied interaction needed for natural collaboration in the physical world. We envision the next generation of human–AI interaction as *real-time, immersive 3D communication*, where humans and AI perceive, reason about, and interact through shared spatial environments. Realizing this vision requires more than extending existing immersive systems. It fundamentally changes the objectives of immersive networking, as the primary consumer of media content changes from *human viewers* to *foundation models*. In this position paper, we identify the key challenges introduced by this paradigm shift and propose a research agenda along two complementary directions: (1) rethinking immersive content delivery by optimizing uplink streaming for *AI perception and reasoning rather than human visual fidelity*; and (2) rethinking generative streaming by exploiting the *progressive and predictable characteristics of AI-generated 3D content* to enable low-latency downlink delivery. By establishing a networking and systems foundation for real-time human–AI 3D interaction, we aim to catalyze the convergence of immersive computing and embodied AI.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing** → **Mixed / augmented reality**; • **Information systems** → **Multimedia streaming**.



This work is licensed under a Creative Commons Attribution 4.0 International License.

HotNets '26, Salt Lake City, UT, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-3029-0/2026/11
https://doi.org/10.1145/3847290.3859945

Keywords

Real-time Human–AI Interaction, Extended Reality (XR), Generative AI, Foundation Models, 3D Gaussian Splatting

ACM Reference Format:
Ruizhi Cheng, Amitabh Varshney, and Bo Han. 2026. When AI Becomes the Viewer: Rethinking Immersive Networking for Human–AI Interaction. In *The 25th ACM Workshop on Hot Topics in Networks (HotNets '26), November 16–17, 2026, Salt Lake City, UT, USA*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3847290.3859945

1 Introduction

Recent advances in foundation models [4, 66] have transformed how humans communicate and collaborate with generative AI (GenAI). Existing platforms, including ChatGPT, Gemini, and Claude, support increasingly natural interactions through text, speech, images, and even live video [11, 19]. Despite this remarkable progress, these interactions remain fundamentally 2D, confined to screens and voice or video interfaces. Consequently, although today's human–AI communication is conversationally rich, it remains limited in perceiving spatial context, understanding human activities in the physical world, and responding via immersive, embodied interactions, thereby falling short of the natural collaboration enabled by shared 3D environments [8, 10, 14, 18].

Emerging work has begun to recognize that human–AI interaction calls for rethinking multimedia systems and networking [11, 12, 19, 20, 56]. In particular, AI video chat [56, 57] argues that networking should evolve from optimizing content for *human viewing* to *AI understanding*, echoing earlier video analytics systems that adapt video streams to scarce upstream bandwidth while preserving inference accuracy [22, 35, 64]. However, these systems operate on 2D video for either fixed analytics tasks or foundation models that respond with text or speech, rather than on a shared 3D scene. We argue that the next evolution is *real-time human–AI 3D interaction*, in which AI perceives rich spatial environments and responds with immersive, context-aware content.

Figure 1 illustrates this vision. Instead of exchanging only text, audio, or video, users and their surrounding environments are continuously captured and represented in 3D. Foundation models jointly reason over spoken instructions,



Figure 1: A paradigm of human–AI 3D interaction. In this healthcare training example, the foundation model understands the user's environment and generates immersive 3D guidance to assist needle insertion.

spatial context, human pose, gestures, and surrounding objects, generating immersive responses such as spatial guidance, embodied virtual assistants, or interactive 3D visualizations that seamlessly integrate into the user's physical environment. Such interactions enable significantly richer situational awareness, more natural communication, and tighter human–AI collaboration than today's 2D interfaces, benefiting use cases such as healthcare, sports training, industrial assembly, and interior design (§3).

Realizing this vision requires substantially more than extending existing immersive communication systems [13, 16, 17]. Compared with AI video chat [56, 57], human–AI 3D interaction introduces fundamentally richer scene representations, continuous spatial reasoning, and AI-generated immersive responses, giving rise to a new class of networking challenges. On the uplink, the network must efficiently deliver semantically meaningful 3D scene representations that preserve geometry, spatial relationships, and human motion, enabling foundation models to reason about dynamic environments. On the downlink, the network no longer transports completed media, but progressively streams AI-generated 3D content that is *spatially grounded, continuously refined, and tightly coupled with the inference process*. These shifts call for new abstractions for 3D content representation, semantic streaming, and edge-assisted generation.

In this position paper, we argue that immersive computing [15] must converge with GenAI to establish a networking and systems foundation for real-time human–AI 3D interaction. We identify two complementary research directions that naturally arise from this paradigm shift.

GenAI-aware Adaptive Dual-path Uplink Streaming (§4.1). Building on the observation that uplink content is primarily consumed by foundation models rather than human viewers [11, 19, 56], we argue that immersive streaming should prioritize *foundation-model utility* over visual fidelity. In contrast to AI video chat [56, 57], immersive interaction introduces substantially richer scene representations, significantly higher bandwidth demands [16, 36, 59, 60], and requires semantic adaptation over shared 3D scenes rather than individual camera frames. Accordingly, we propose a *focus-aware adaptive bitrate (ABR)* framework that infers semantically important 3D regions from user prompts and interaction context, instead of relying on human gaze [61], viewport prediction [25], or visual saliency [55]. Bandwidth is allocated to maximize the utility of the transmitted 3D content for foundation-model perception and reasoning. To further improve responsiveness, we propose a *dual-path* architecture that separates prompt-critical scene updates from deferred global scene synchronization. Unlike conventional layered video streaming [21, 51] and multi-path transport [26, 43], the two paths are organized according to their *semantic roles* in AI reasoning, enabling prompt-relevant content to reach the foundation model with minimal latency while maintaining a consistent global scene representation.

Hierarchical Edge-assisted Generative Downlink Streaming (§4.2). Traditional live and on-demand video streaming assumes that content has been fully encoded before transmission. In contrast, immersive downlink content is generated online by foundation models, making it inefficient to wait for complete generation before delivering any feedback. We therefore propose *progressively streaming AI-generated content* from coarse to fine, jointly coupling generation and transport to provide immediate visual feedback while finer details are synthesized incrementally. Unlike conventional progressive streaming [1, 7, 27, 44], which refines pre-encoded content through fixed hierarchies, our generation process dynamically adapts to the user prompt and interaction context. Moreover, because users may interrupt or redirect interactions at any time, generation can terminate early, avoiding wasted computation. To further reduce user-perceived latency, we propose *predictive edge-assisted content pre-generation and caching*. Rather than caching previously generated content [53], the edge predicts likely responses from prompt embeddings and scene semantics, streaming lightweight previews while the backend model completes the full-resolution output.

2 Background

Real-time Human–AI Communication. The evolution of foundation models is transforming human–AI interaction from text exchanges to real-time multimodal communication. Early systems primarily supported text, followed by audio interaction in which users issue spoken prompts and receive synthesized speech, resembling a phone call [11, 20]. In such interaction, the spoken exchange fosters the impression of conversing with another person rather than a machine. More recently, platforms such as ChatGPT and Gemini have introduced video interaction, allowing users to continuously stream their surroundings to GenAI for visual understanding

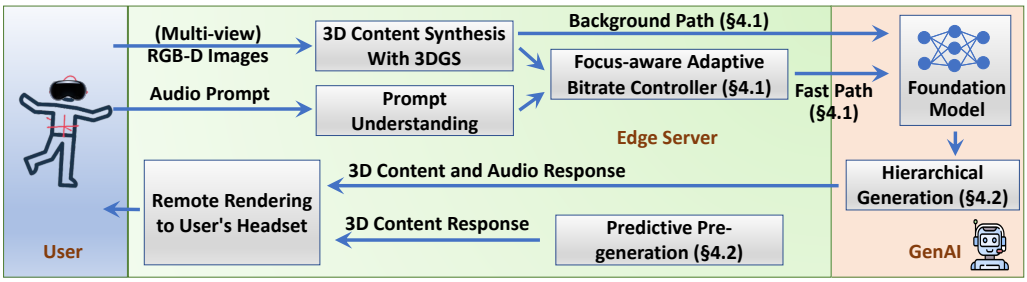


Figure 2: Workflow of human–AI 3D interaction.

and assistance [12, 19, 56]. This evolution mirrors the historical progression of human–to–human communication, from text messaging and voice calls to video conferencing and immersive telepresence [9, 13, 16, 17, 31]. Thus, we envision that the next logical step is human–AI *3D* interaction.

3D Foundation Models. Recent advances in 3D-aware foundation models have begun enabling language-guided understanding of spatial environments. For example, PointCLIP [65] extends 2D vision-language pre-training to point clouds through multi-view projections, while PointLLM [63] aligns 3D geometry with LLMs to support open-ended reasoning over point clouds. More recently, SplatTalk [50] and 4D LangSplat [34] leverage Gaussian splatting to jointly model geometry, appearance, and language in 3D and dynamic 4D scenes. These advances represent important steps toward embodied AI, but existing models remain primarily designed for scene understanding rather than real-time interactive communication. In particular, they are limited by static or scene-specific representations, preventing them from supporting continuous, low-latency human–AI interaction in dynamic, previously unseen environments.

3 Vision: Human–GenAI 3D Interaction

Figure 2 illustrates our vision of real-time human–AI 3D interaction. A user wearing an extended reality (XR) headset captures first-person observations and issues spoken prompts, capabilities already available on commercial devices such as the Apple Vision Pro [17]. Optional RGB-D cameras provide additional spatial context. An edge server aggregates these multimodal inputs, reconstructs a live 3D scene, and forwards it, along with the user's prompts, to a backend foundation model [12, 20]. The model jointly reasons over the spatial scene, human actions, and user intent to generate immersive 3D responses, such as holographic guidance or embodied virtual assistants, together with synchronized audio feedback. The generated content is streamed back through the edge server and rendered on the XR headset, completing a closed-loop interaction among the user, the physical environment, and the foundation model.

Healthcare training illustrates the advantages of this paradigm. As shown in Figure 1, the foundation model estimates needle orientation, tissue geometry, and hand pose, then overlays immersive guidance to visualize the desired insertion trajectory while providing spoken instructions. Unlike 2D video interaction, it continuously reasons about 3D geometry and delivers fine-grained corrections, such as adjusting insertion angles, maintaining tool alignment, and avoiding unsafe motions. In rehabilitation, the model can provide posture corrections and range-of-motion targets with progressive difficulty and safety limits. Similarly, in sports training, the model provides real-time feedback on body posture, joint alignment, and swing trajectory through immersive overlays. In industrial assembly and maintenance, the model can verify part poses and anchor step-by-step 3D instructions on the workpiece [48]. In interior design, the model can generate 3D furniture that fits the scanned room for the user to preview in place. These examples show the potential of human–AI 3D interaction for applications requiring embodied intelligence, spatial understanding, and natural collaboration.

We adopt 3D Gaussian splatting (3DGS) [47, 60] as the underlying scene representation because it provides an effective balance between rendering quality and efficiency. 3DGS explicitly models a scene as a collection of 3D Gaussian points, each parameterized by its position, orientation, scale, opacity, and color [32]. However, our proposed networking architecture is representation-agnostic and can be extended to other explicit 3D representations, including point clouds [25, 61] and meshes [6, 9, 16, 31].

4 Proposed Research Agenda

4.1 GenAI-aware Adaptive Dual-path Uplink Streaming

Research Problems. Traditional dynamic-scene representations are optimized for *human perceptual quality* (*e.g.,* photorealism and visual smoothness) [9, 13, 16, 36, 59, 60]. When the consumer is a *foundation model*, however, the objective

shifts fundamentally: the model does not need photorealistic content, but rather *semantically informative and computationally efficient tokens* that it can directly interpret for reasoning and action. To this end, the uplink transmission pipeline should be redesigned to optimize for the utility of foundation models, rather than for visual fidelity.

Technical Challenges and Proposed Research:

(1) *Adaptive Streaming Guided by Foundation-model Focus.* Streaming live 3D content is notoriously bandwidth-intensive (*e.g.,* >1 Gbps [38, 46]). It is therefore necessary to design adaptive bitrate (ABR) mechanisms [2, 40, 49] that can adjust transmission rates according to available network capacity without degrading the quality of interaction between the user and the foundation model.

In immersive content delivery, a common strategy for reducing bandwidth is to identify the consumer's "focus" and transmit only that region at high quality. For example, when the consumer is a human, the system can exploit properties of human vision, such as the user's viewport [25] or foveal region [61], to transmit the content only within the user's visual focus. However, in the uplink of 3D human–AI interaction, the consumer is not a human but a *foundation model.* Unlike humans, a foundation model does not perceive content visually; instead, it reasons over 3D data. This raises an important question: *what defines the "focus" of a foundation model?* Our insight is that the model's focus is inherently driven by the user's prompt. For example, in a medical-training scenario, if the user asks "Is the needle position correct?", the user's attention, and consequently the model's focus, should naturally center on the needle and its surrounding tissue. Hence, the system can allocate more bandwidth to transmit these regions at high quality while reducing the Gaussian point density in less relevant areas. Focus-driven streaming has been demonstrated to be a promising way to reduce bandwidth consumption, for example, by more than 50% for volumetric content consumed by human viewers [61] and for 2D video consumed by deep neural networks [22].

To achieve this, we propose a *focus-aware ABR framework* that anticipates the foundation model's focus on the edge and uses it to guide uplink scheduling and bandwidth adaptation, as shown in Figure 3. When receiving an audio prompt from the user, the edge server first performs lightweight speech recognition (*e.g.,* OpenAI Whisper) and converts the transcribed text into embedding vectors using a small text-embedding model consistent with the foundation model's language space (*e.g.,* the text encoder used in CLIP [42]). This produces a prompt embedding $\mathbf{q}_t$ that captures the user's current intent at time t. To estimate which regions of the 3D scene are most relevant to the user's intent, we need to represent each 3D Gaussian point with an embedding vector

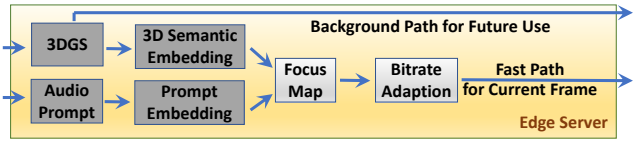


Figure 3: Paradigm of focus-aware adaptive bitrate framework and dual-path transmission in the uplink.

that facilitates computing its similarity to the prompt embedding. To enable this, our key idea is to attach a compact *semantic vector* $\{\mathbf{z}_{i,t}\}$ to each Gaussian point of the 3D scene, representing how that local region would be encoded in the *foundation model's feature space* (*e.g.,* "hand", "needle", or "tissue surface" in a medical-training scenario). Prior methods, such as LangSplat [41] and FMGS [67], embed these semantics into static 3DGS scenes through offline per-scene optimization, whereas live interaction requires extracting them within a tight per-frame latency budget (*e.g.,* 33 ms at 30 FPS [16]). In order to generate the semantics in real time, the edge applies a small pretrained vision model (*e.g.,* TinyCLIP [58]) to extract per-pixel features from the multi-view RGB-D streams. Features corresponding to the same Gaussian point are then fused across views and averaged over a short time window, producing a stable and view-independent semantic vector for each Gaussian point.

The edge then compares the prompt embedding $\mathbf{q}_t$ with the 3D semantic embedding vectors $\{\mathbf{z}_{i,t}\}$ to estimate which regions of the scene are most relevant to the user's intent. This relevance is computed through a similarity function between $\mathbf{q}_t$ and $\mathbf{z}_{i,t}$ (*e.g.,* cosine similarity), producing an attention score $A_t(i)$ for each Gaussian point: $A_t(i) = \alpha\, A_{t-1}(i) + (1-\alpha)\, \cos(\mathbf{q}_t, \mathbf{z}_{i,t})$, where α controls temporal smoothing across frames. This *dynamic focus map* allows the edge to preserve high attention on persistent regions of interest (*e.g.,* a surgical tool) while quickly adapting to the user's prompts. The resulting focus map is combined with real-time network measurements to perform ABR under the current bandwidth constraint. Specifically, regions with higher attention are represented using denser Gaussian points, while areas with lower attention are coarsened or temporarily omitted from transmission. By performing this focus estimation entirely at the edge, the framework ensures that limited uplink bandwidth is spent on 3D content that is most relevant for the foundation model's subsequent reasoning, achieving efficient and context-aware streaming.

(2) *Dual-path Uplink Transmission for Real-time and Background Data.* With the help of the proposed focus-aware ABR framework, the edge can prioritize transmitting the most relevant 3D regions when the available bandwidth is insufficient to deliver the entire scene at high quality. However,

selectively transmitting content introduces a new problem: regions that are not delivered in the current frame may later become relevant as the user's focus shifts. For example, in a medical-training scenario, a user who initially asks about "needle insertion angle" may later request "tissue contact area", information that was previously deprioritized. To ensure complete scene availability over time, the system must eventually transmit all 3D content, without interfering with real-time updates. Thus, we propose a *dual-path uplink architecture* consisting of a *fast path* for latency-critical updates and a *background path* for deferred scene synchronization.

The fast path transmits live 3D content that reflects the user's current focus, as determined by the ABR controller, while the background path delivers remaining 3D content that was temporarily deprioritized to ensure eventual scene completeness. The two paths, however, have fundamentally different transmission requirements and optimization goals. The fast path handles live, continuously generated content where latency is the dominant factor. Its design aligns with the real-time communication (RTC) paradigm [20, 37, 52, 54], which uses *frame-based transmission*: each frame is sent immediately once produced, minimizing interaction delay. In contrast, the background path carries a larger volume of already available data, where throughput efficiency becomes the priority. This is similar to on-demand streaming, which uses *chunk-based transmission* to deliver multiple data segments and maximize bandwidth utilization [30].

Widely deployed transport protocols are typically optimized for either RTC or reliable content delivery, but rarely both. A recent measurement study of commercial human–AI audio calling systems, such as ChatGPT and Gemini, revealed that today's systems naturally exhibit two distinct communication patterns [11]. Human-spoken queries are generated continuously and therefore favor latency-oriented, frame-based streaming with short packet interarrival times to preserve conversational responsiveness. In contrast, GenAI responses are tokenized and can be aggregated into segments before transmission. Consequently, GenAI response delivery more closely resembles on-demand streaming, where chunk-based transmission prioritizes reliability and bandwidth utilization over latency. These observations suggest that human–AI interaction comprises two traffic classes with distinct transport requirements [20]. Latency-sensitive traffic favors immediate frame-based delivery, whereas throughput-oriented traffic benefits from reliable chunk-based transmission. Efficiently supporting both within a single interactive session remains an open systems challenge.

A straightforward solution is to employ separate transport protocols for the two traffic classes, such as an RTC protocol for the fast path and a reliable transport for the background path. However, maintaining multiple transport connections increases implementation complexity, duplicates congestion-control and flow-management logic, and may introduce resource contention between heterogeneous transports. These drawbacks motivate a unified transport architecture that accommodates both traffic classes within a single connection while allowing each to retain its preferred semantics.

A promising direction is a *cross-layer transport architecture* that unifies frame-based and chunk-based delivery over a common transport substrate while differentiating their behaviors at the application layer. We advocate building this architecture atop QUIC [3], which provides a flexible transport foundation for multiplexing heterogeneous traffic within a single connection. For throughput-oriented background traffic, HTTP/3 naturally supports reliable and efficient content delivery. For latency-sensitive foreground traffic, however, QUIC's default reliable and in-order delivery semantics introduce unnecessary delays, much like TCP. To overcome this limitation, we propose leveraging the emerging Media-over-QUIC (MoQ) paradigm [12, 20, 23, 24], which exploits QUIC's datagram extension [39] to support lightweight, frame-based transmission without imposing strict reliability or ordering constraints. Together, HTTP/3 and MoQ provide complementary application-layer abstractions that enable efficient support for both communication paradigms over a unified QUIC transport.

As shown in Figure 3, under this cross-layer design, both transmission paths coexist within a single transport framework, avoiding competition at the transport layer. This paradigm also extends naturally to the downlink transmission that delivers GenAI responses. During the initial phase of response delivery, the downlink can adopt frame-based transmission over QUIC/MoQ to minimize user-perceived latency. Subsequently, as the remaining content becomes available, it can switch to chunk-based delivery over QUIC/H3 to maximize bandwidth utilization.

4.2 Hierarchical Edge-assisted Generative Downlink Streaming

Research Problems. Minimizing the end-to-end conversational latency is crucial in human–AI interaction to provide users with a natural and coherent conversational experience. This latency is defined as the time between the user finishing a prompt and the first GenAI response being perceived [11, 19, 20]. In RTC, maintaining sub-second end-to-end latency is essential for natural conversation [11, 19, 56]. This requirement means that the generation and delivery of 3D content must keep pace with real-time speech responses, ensuring synchronized visual and auditory feedback. While model-level optimizations (*e.g.,* distillation [45]) can shorten response generation time, achieving sub-second conversational latency also demands rethinking the delivery pipeline.

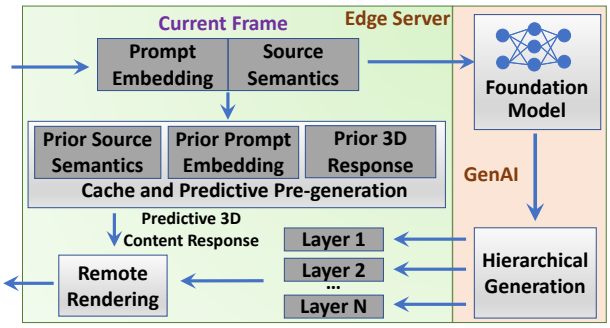


Figure 4: Paradigm of the hierarchical generation and streaming framework, as well as the predictive pre-generation and caching mechanism in the downlink.

Technical Challenges and Proposed Research:

(1) *Hierarchical Generative Streaming.* One unique feature of human–AI interactions, compared to human–to–human communication, is that they often exhibit abrupt, user-driven topic shifts [20]. Users may interrupt an ongoing response when they perceive that the reply deviates from their intent or contains factual or semantic errors (*e.g.,* "LLM hallucinations" [29]). In such cases, generating the entire 3D response before transmission is inefficient. It increases inference latency and, consequently, prolongs user-perceived latency.

To address this, we propose a *hierarchical generation and streaming* framework, as shown in Figure 4. The key idea is to design a *multi-layer 3D representation* that allows the model to generate and transmit the scene progressively from coarse to fine detail. Specifically, *Layer 1 (coarse preview):* the model first generates a sparse set of Gaussian points with a large spatial extent to capture the overall structure of the object. This coarse layer is sufficient for users to perceive the basic shape and intent of the response, and can be streamed and rendered immediately as the initial visual feedback. *Layer 2+ (progressive densification):* if the user does not interrupt the conversation, the model continues to generate additional Gaussian points that incrementally refine details. These higher-density layers are streamed sequentially (layer by layer) as they are generated, enhancing visual fidelity while preserving continuity with previously rendered layers.

To achieve this design, we highlight two key research problems and propose potential solutions. *(a) How to generate multiple layers of 3D representations.* Existing 3DGS generative models typically produce all Gaussian points simultaneously. To support hierarchical generation, the model should learn to first output a small number of large Gaussian points representing the basic shape, and then add smaller points to refine details. One promising direction is to introduce a *level-of-detail indicator* during training that allows the model to generate at different density levels. *(b) How to determine the granularity for each layer.* Even if the model can generate multiple layers, we still need to decide how many Gaussian points to include in each layer. A preview that is too coarse may confuse users, whereas one that is overly detailed undermines the goal of providing rapid feedback. A promising direction is to design an *adaptive controller* that dynamically determines the Gaussian point density. For instance, the controller can monitor the remaining latency budget and prioritize generating the most perceptually important parts first. This prioritization can be guided by the *importance map* trained via reinforcement learning, where the system learns to allocate point density to different regions based on user feedback.

(2) *Predictive Pre-generation and Caching at Edge.* To further reduce perceived delay, we propose leveraging the edge server for *predictive pre-generation and caching*, allowing likely responses to be prepared in advance of user requests (Figure 4). The key insight is that human–AI conversations often exhibit temporal and semantic continuity. For example, during an instructional or training session, users tend to ask follow-up questions about previous objects or gestures. By tracking recent interaction context, the edge server can predict potential future requests/responses. When the actual prompt arrives, the edge can instantly stream these pre-generated assets as an initial response, hiding model inference latency while the backend completes full generation.

To enable accurate and rapid response generation, the edge server must cache both the uplink and downlink contexts from prior interactions. Given that the downlink 3D content generated by the foundation model may be relatively small-scale, such as guidance overlays, and the edge server needs to pre-generate the content similar to what the foundation model will generate, it may be reasonable for the edge server to cache the raw generated content. We denote the set of Gaussian points of the generated 3D response as $\mathcal{G}_k = \{\mathcal{N}(\boldsymbol{\mu}_{i,k}, \Sigma_{i,k}, \mathbf{c}_{i,k})\}_{i=1}^{M_k}$. For the uplink content, however, caching the raw 3D scenes becomes infeasible because it contains a large 3D scene coming from the source. Instead, a practical direction is to cache the 3D semantic embedding $\mathbf{z}_k$, along with the user's prompt embedding $\mathbf{q}_k$ (§4.1).

For a new frame t, the edge first determines whether to trigger pre-generation. It compares $\mathbf{q}_t$ and $\mathbf{z}_t$ with all cached $\{(\mathbf{q}_k, \mathbf{z}_k)\}_{k=1}^{N}$. The similarity between frame t and a cached interaction k is computed as $\text{score}_k = \lambda \cos(\mathbf{q}_t, \mathbf{q}_k) + (1 - \lambda) \cos(\mathbf{z}_t, \mathbf{z}_k)$, where $\lambda \in [0, 1]$ balances linguistic and source-scene semantic relevance. We compute a similarity score for each cached item and pick $k^\star \triangleq \arg\max_{k \in \{1,\ldots,N\}} \text{score}_k$. If $\text{score}_{k^\star} \geq \tau$, the edge server triggers pre-generation using $(\mathbf{q}_{k^\star}, \mathbf{z}_{k^\star}, \mathcal{G}_{k^\star})$; otherwise, it waits for the backend response. The similarity threshold τ can be tuned empirically through user studies or A/B testing, balancing response

speed and user experience (*e.g.,* measured via mean opinion score, MOS).

Once pre-generation is triggered, the edge must quickly synthesize a plausible 3D response that aligns with the backend foundation model's eventual output. To achieve low latency, the edge should avoid invoking any heavy 3D generative model as used in the backend. Instead, it directly modifies the cached coarse 3D scene $\mathcal{G}_{k^\star}$ corresponding to the most similar past interaction. This can be achieved by adjusting the Gaussian parameters to reflect the new prompt intent. Formally, the edge performs a conditional transformation: $\widehat{\mathcal{G}}_t = f_\theta(\mathbf{q}_t, \mathbf{q}_{k^\star}, \mathbf{z}_{k^\star}, \mathcal{G}_{k^\star})$, where f_θ is a lightweight *transformation network* (*e.g.,* MLP) that predicts offsets $(\Delta\boldsymbol{\mu}_i, \Delta\Sigma_i, \Delta\mathbf{c}_i)$ for each Gaussian point to produce the updated first-layer preview $\widehat{\mathcal{G}}_t$.

The global model can first be pretrained on public multimodal datasets to learn general semantic-to-geometry mappings. To the best of our knowledge, no dataset currently exists for human–AI 3D interaction. We plan to collect and release such a dataset to support this emerging research direction. During deployment, the pretrained global model can be continuously fine-tuned at the edge server using cached triplets $(\mathbf{q}_k, \mathbf{z}_k, \mathcal{G}_k)$ from the user's past interactions, enabling personalized optimization.

5 Limitations and Discussion

We do not present an end-to-end prototype or experimental evaluation because the software ecosystem for human–AI 3D interaction is still in its infancy. The primary bottleneck is the lack of general-purpose foundation models capable of understanding and reasoning over *dynamic* 3D scenes. Existing models, such as PointCLIP [65], PointLLM [63], and SplatTalk [50], are designed primarily for *static* 3D scene understanding. More recent efforts, such as 4D LangSplat [34], extend foundation models to dynamic scenes through 4D Gaussian splatting, but remain largely *scene-specific*, requiring environment-dependent optimization and exhibiting limited generalization to unseen environments.

Nevertheless, we view these limitations as opportunities rather than barriers. Just as networking research on video streaming preceded today's large-scale video platforms, we believe networking and systems research for human–AI 3D interaction should begin before mature 3D foundation models become available. Many of the architectural questions raised in this paper, including semantic-aware immersive streaming [13, 16], AI-oriented transport abstractions [12, 20], hierarchical generative streaming, and edge-assisted inference, are largely orthogonal to the underlying foundation models and will remain central as future models evolve.

Another promising research direction is to balance the trade-off between computation and networking by deciding where each computation should run: on the device, on the edge server, or in the cloud. Because 3D content synthesis [33, 60] and foundation-model reasoning [56, 57] are computationally intensive, running them on the device remains largely impractical. Beyond the edge server shown in Figure 2, a more powerful cloud server is likely needed for the most demanding tasks, such as foundation-model reasoning and 3D generation. For example, generating a single 3D asset can take up to 15 s even on an NVIDIA A100 GPU [28]. However, involving the cloud adds wide-area transmission, which may become a bottleneck [62]. In contrast, device–edge connections generally have lower and more stable latency than device–cloud connections [5, 62]. Thus, when the edge server can handle a task, offloading it to the edge rather than the cloud often offers a better trade-off. For instance, 3D content synthesis from multi-view RGB-D images (Figure 2) can run on the edge server in real time [13, 16, 60]. Overall, balancing computation and networking requires treating the device, edge, and cloud as a whole and placing each computation according to its compute demand and the network conditions of each use case.

6 Conclusion

In this position paper, we envision a paradigm shift in human–AI interaction, from today's 2D text, audio, and video interfaces to real-time, immersive 3D collaboration. We argue that this evolution fundamentally changes the objectives of immersive networking: uplink content is no longer delivered primarily for human perception but for foundation-model understanding, while downlink content evolves from pre-rendered media to AI-generated immersive experiences. To address these emerging challenges, we outlined a research agenda with two complementary directions: GenAI-aware adaptive uplink streaming that prioritizes semantic utility over visual fidelity, and hierarchical edge-assisted generative downlink streaming that exploits the progressive nature of AI-generated content to reduce user-perceived latency. We hope this work stimulates discussion within the networking and systems community and inspires a new generation of architectures that seamlessly integrate immersive computing and foundation models, ultimately enabling the next era of embodied AI.

Acknowledgment

We thank the anonymous reviewers and our shepherd, Aruna Balasubramanian, for their insightful feedback, which has greatly strengthened this work. This work was partially supported by the National Science Foundation under Grants CNS-2212296, CNS-2235049, CNS-2235050, CNS-2402992, and CNS-2437203.

References

[1] Pierre Alliez and Mathieu Desbrun. 2001. Progressive Compression for Lossless Transmission of Triangle Meshes. *IEEE Transactions on Visualization and Computer Graphics* 7, 3 (2001), 195–202.

[2] Abdelhak Bentaleb, Bayan Taani, Ali C Begen, Christian Timmerer, and Roger Zimmermann. 2018. A Survey on Bitrate Adaptation Schemes for Streaming Media Over HTTP. *IEEE Communications Surveys & Tutorials* 21, 1 (2018), 562–585.

[3] Pengqiang Bi, Yifei Zou, Mengbai Xiao, Dongxiao Yu, Yijun Li, Zhixiong Liu, and Qun Xie. 2024. LiteQUIC: Improving QoE of Video Streams by Reducing CPU Overhead of QUIC. In *ACM MM*.

[4] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, et al. 2021. On the Opportunities and Risks of Foundation Models. *CoRR* abs/2108.07258 (2021). arXiv:2108.07258 https://arxiv.org/abs/2108.07258

[5] Batyr Charyyev, Engin Arslan, and Mehmet Hadi Gunes. 2020. Latency Comparison of Cloud Datacenters and Edge Servers. In *Proceedings of IEEE Global Communications Conference*.

[6] Guodong Chen, Filip Hácha, Libor Váša, and Mallesham Dasari. 2025. TVMC: Time-Varying Mesh Compression Using Volume-Tracked Reference Meshes. In *Proceedings of the ACM MMSys*.

[7] Yihang Chen, Mengyao Li, Qianyi Wu, Weiyao Lin, Mehrtash Harandi, and Jianfei Cai. 2026. PCGS: Progressive Compression of 3D Gaussian Splatting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

[8] Ruizhi Cheng, Jie Li, Songqing Chen, and Bo Han. 2025. Dissecting User Experience of Social Virtual Reality: A Tale of Five Platforms. *Proceedings of the ACM on Human-Computer Interaction* 9, 7, Article CSCW224 (2025).

[9] Ruizhi Cheng, Kaiyan Liu, Nan Wu, and Bo Han. 2023. Enriching Telepresence with Semantic-driven Holographic Communication. In *Proceedings of ACM HotNets*.

[10] Ruizhi Cheng, Erdem Murat, Lap-Fai Yu, Songqing Chen, and Bo Han. 2024. Understanding Online Education in Metaverse: Systems and User Experience Perspectives. In *Proceedings of IEEE VR*. 598–608.

[11] Ruizhi Cheng, Surendra Pathak, Guowu Xie, Matteo Varvello, Songqing Chen, and Bo Han. 2025. Hello, GenAI? Dissecting Human to Generative AI Calling. In *Proceedings of ACM IMC*.

[12] Ruizhi Cheng, Mohamed Mohsin Shaik, Guowu Xie, and Bo Han. 2026. Huni: A Networked Testbed for Real-time Human and Generative AI Interaction. In *Proceedings of ACM ImmerCom*.

[13] Ruizhi Cheng, Harshvardhan Takawale, Nan Wu, Nirupam Roy, Sennur Ulukus, Matteo Varvello, Eugene Chai, and Bo Han. 2026. LITE: Loss-resilient Immersive Telepresence with Multimodal Semantics. In *Proceedings of ACM SIGCOMM*. 628–641.

[14] 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 VRW (MetaBuild Workshop)*. 141–148.

[15] 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.

[16] Ruizhi Cheng, Nan Wu, Vu Le, Eugene Chai, Matteo Varvello, and Bo Han. 2024. MagicStream: Bandwidth-conserving Immersive Telepresence via Semantic Communication. In *Proceedings of ACM SenSys*.

[17] 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 ACM IMC*.

[18] 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*.

[19] Ruizhi Cheng, Guowu Xie, and Bo Han. 2026. Can You See Me, GenAI? Characterizing Video-Based Conversational Interaction with Generative AI. In *Proceedings of ACM IMC*. 574–588.

[20] Ruizhi Cheng, Guowu Xie, and Bo Han. 2026. Real-Time Human and Generative AI Interaction: Network Challenges and Opportunities. *IEEE Network* (2026). Early access.

[21] Mallesham Dasari, Kumara Kahatapitiya, Samir R. Das, Aruna Balasubramanian, and Dimitris Samaras. 2022. Swift: Adaptive Video Streaming with Layered Neural Codecs. In *Proceedings of USENIX NSDI*.

[22] Kuntai Du, Ahsan Pervaiz, Xin Yuan, Aakanksha Chowdhery, Qizheng Zhang, Henry Hoffmann, and Junchen Jiang. 2020. Server-Driven Video Streaming for Deep Learning Inference. In *Proceedings of ACM SIGCOMM*. 557–570.

[23] Zafer Gurel, Tugce Erkilic Civelek, Deniz Ugur, Yigit K Erinc, and Ali C Begen. 2024. Media-over-QUIC Transport vs. Low-Latency DASH: a Deathmatch Testbed. In *Proceedings of ACM MMSys*.

[24] Zafer Gurel, Tugce Erkilic Civelek, Atakan Bodur, Senem Bilgin, Deniz Yeniceri, and Ali C Begen. 2023. Media over QUIC: Initial Testing, Findings and Results. In *Proceedings of ACM Multimedia Systems*.

[25] Bo Han, Yu Liu, and Feng Qian. 2020. ViVo: Visibility-aware Mobile Volumetric Video Streaming. In *Proceedings of ACM MobiCom*.

[26] Bo Han, Feng Qian, Lusheng Ji, and Vijay Gopalakrishnan. 2016. MP-DASH: Adaptive Video Streaming Over Preference-Aware Multipath. In *Proceedings of ACM CoNEXT*.

[27] Hugues Hoppe. 1996. Progressive Meshes. In *Proceedings of ACM SIGGRAPH*.

[28] Hanzhe Hu, Tianwei Yin, Fujun Luan, Yiwei Hu, Hao Tan, Zexiang Xu, Sai Bi, Shubham Tulsiani, and Kai Zhang. 2025. Turbo3D: Ultra-fast Text-to-3D Generation. In *Proceedings of IEEE/CVF CVPR*. 23668–23678.

[29] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, et al. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[30] Te-Yuan Huang, Ramesh Johari, Nick McKeown, Matthew Trunnell, and Mark Watson. 2014. A Buffer-Based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service. In *Proceedings of ACM SIGCOMM*.

[31] Tao Jin, Mallesham Dasari, Connor Smith, Kittipat Apicharttrisorn, Srinivasan Seshan, and Anthony Rowe. 2024. MeshReduce: Scalable and Bandwidth Efficient 3D Scene Capture. In *Proceedings of IEEE VR*.

[32] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (ToG)* 42, 4 (2023), 1–14.

[33] Leshu Li, Jiayin Qin, Jie Peng, Zishen Wan, Huaizhi Qu, Ye Han, Pingqing Zheng, Hongsen Zhang, Yu Cao, Tianlong Chen, and Yang Katie Zhao. 2025. RTGS: Real-Time 3D Gaussian Splatting SLAM via Multi-Level Redundancy Reduction. In *Proceedings of IEEE/ACM MICRO*. 1838–1851.

[34] Wanhua Li, Renping Zhou, Jiawei Zhou, Yingwei Song, Johannes Herter, Minghan Qin, Gao Huang, and Hanspeter Pfister. 2025. 4D LangSplat: 4D Language Gaussian Splatting via Multimodal Large Language Models. In *Proceedings of CVPR*.

[35] Yuanqi Li, Arthi Padmanabhan, Pengzhan Zhao, Yufei Wang, Guoqing Harry Xu, and Ravi Netravali. 2020. Reducto: On-Camera Filtering for Resource-Efficient Real-Time Video Analytics. In *Proceedings of ACM SIGCOMM*.

[36] Kaiyan Liu, Ruizhi Cheng, Nan Wu, and Bo Han. 2023. Toward Next-generation Volumetric Video Streaming with Neural-based Content Representations. In *Proceedings of ACM ImmerCom*.

[37] Zili Meng, Tingfeng Wang, Yixin Shen, Bo Wang, Mingwei Xu, Rui Han, et al. 2023. Enabling High Quality Real-Time Communications with Adaptive Frame-Rate. In *Proceedings of USENIX NSDI*.

[38] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of ACM Symposium on User Interface Software and Technology (UIST)*. https://doi.org/10.1145/2984511.2984517

[39] Tommy Pauly, Eric Kinnear, and David Schinazi. 2022. An Unreliable Datagram Extension to QUIC. https://datatracker.ietf.org/doc/html/rfc9221. [accessed on 23-Oct-2026].

[40] Stefano Petrangeli, Viswanathan Swaminathan, Mohammad Hosseini, and Filip De Turck. 2017. An HTTP/2-Based Adaptive Streaming Framework for 360° Virtual Reality Videos. In *Proceedings of ACM International Conference on Multimedia*.

[41] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. 2024. LangSplat: 3D Language Gaussian Splatting. In *Proceedings of CVPR*.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of ICML*.

[43] Swetank Kumar Saha, Shivang Aggarwal, Rohan Pathak, Dimitrios Koutsonikolas, and Joerg Widmer. 2019. MuSher: An Agile Multipath-TCP Scheduler for Dual-Band 802.11ad/ac Wireless LANs. In *Proceedings of ACM MobiCom*.

[44] Yuang Shi, Géraldine Morin, Simone Gasparini, and Wei Tsang Ooi. 2025. LapisGS: Layered Progressive 3D Gaussian Splatting for Adaptive Streaming. In *Proceedings of the 2025 International Conference on 3D Vision (3DV)*.

[45] Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, et al. 2024. LLM Pruning and Distillation in Practice: The Minitron Approach. https://arxiv.org/abs/2408.11796.

[46] Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 2024. 3DGStream: On-the-Fly Training of 3D Gaussians for Efficient Streaming of Photo-Realistic Free-Viewpoint Videos. In *Proceedings of CVPR*.

[47] Yuan-Chun Sun, Yuang Shi, Cheng-Tse Lee, Mufeng Zhu, Wei Tsang Ooi, Yao Liu, Chun-Ying Huang, and Cheng-Hsin Hsu. 2025. LTS: A DASH Streaming System for Dynamic Multi-Layer 3D Gaussian Splatting Scenes. In *Proceedings of ACM MMSys*.

[48] Arthur Tang, Charles Owen, Frank Biocca, and Weimin Mou. 2003. Comparative Effectiveness of Augmented Reality in Object Assembly. In *Proceedings of ACM CHI*. 73–80.

[49] Farzad Tashtarian, Abdelhak Bentaleb, Hadi Amirpour, Sergey Gorinsky, Junchen Jiang, Hermann Hellwagner, and Christian Timmerer. 2024. ARTEMIS: Adaptive Bitrate Ladder Optimization for Live Video Streaming. In *Proceedings of USENIX NSDI*.

[50] Anh Thai, Songyou Peng, Kyle Genova, Leonidas Guibas, and Thomas Funkhouser. 2025. SplatTalk: 3D VQA with Gaussian Splatting. In *Proceedings of ICCV*.

[51] Yi-Zhen Tsai, Xuechen Zhang, Zheng Li, and Jiasi Chen. 2025. L3GS: Layered 3D Gaussian Splats for Efficient 3D Scene Delivery. In *Proceedings of ACM MobiCom*.

[52] Bo Wang, Xingxing Yang, Zili Meng, Yaning Guo, Chen Sun, Justine Sherry, Hongqiang Harry Liu, Mingwei Xu, and Jianping Wu. 2024. Zhuge: Toward Consistent Low Latency With Minimal Control Loop Delay. *IEEE/ACM Transactions on Networking* (2024).

[53] Fei Wang, Feng Wang, Jiangchuan Liu, Ryan Shea, and Lifeng Sun. 2020. Intelligent Video Caching at Network Edge: A Multi-Agent Deep Reinforcement Learning Approach. In *Proceedings of IEEE INFOCOM*.

[54] Shuoqian Wang, Mengbai Xiao, and Yao Liu. 2024. RoIRTC: Toward Region-of-Interest Reinforced Real-Time Video Communication. In *Proceedings of IEEE ICME*.

[55] Shibo Wang, Shusen Yang, Hailiang Li, Xidong Zhang, Chao Zhou, Chenren Xu, Feng Qian, Nabing Wang, and Zongben Xu. 2022. SalientVR: Saliency-Driven Mobile 360-Degree Video Streaming with Gaze Information. In *Proceedings of ACM MobiCom*.

[56] Jiangkai Wu, Zhiyuan Ren, Liming Liu, and Xinggong Zhang. 2025. Chat with AI: The Surprising Turn of Real-time Video Communication from Human to AI. In *Proceedings of ACM HotNets*.

[57] Jiangkai Wu, Zhiyuan Ren, Junquan Zhong, Liming Liu, and Xinggong Zhang. 2026. Artic: AI-oriented Real-time Communication for MLLM Video Assistant. In *Proceedings of ACM SIGCOMM*. 1649–1662.

[58] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. 2023. TinyCLIP: CLIP Distillation via Affinity Mimicking and Weight Inheritance. In *Proceedings of CVPR*.

[59] Nan Wu, Bo Chen, Ruizhi Cheng, Klara Nahrstedt, and Bo Han. 2025. NeVo: Advancing Volumetric Video Streaming with Neural Content Representation. In *Proceedings of ACM MobiCom*. 267–282.

[60] Nan Wu, Weikai Lin, Ruizhi Cheng, Bo Chen, Yuhao Zhu, Klara Nahrstedt, and Bo Han. 2025. Advancing Immersive Content Delivery with Dynamic 3D Gaussian Splatting. In *Proceedings of ACM HotMobile*.

[61] 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*.

[62] Mengwei Xu, Zhe Fu, Xiao Ma, Li Zhang, Yanan Li, Feng Qian, Shangguang Wang, Ke Li, Jingyu Yang, and Xuanzhe Liu. 2021. From Cloud to Edge: A First Look at Public Edge Platforms. In *Proceedings of ACM IMC*.

[63] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024. PointLLM: Empowering Large Language Models to Understand Point Clouds. In *Proceedings of ECCV*.

[64] Ben Zhang, Xin Jin, Sylvia Ratnasamy, John Wawrzynek, and Edward A. Lee. 2018. AWStream: Adaptive Wide-Area Streaming Analytics. In *Proceedings of ACM SIGCOMM*. 236–252.

[65] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. 2022. PointCLIP: Point Cloud Understanding by CLIP. In *Proceedings of CVPR*.

[66] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, et al. 2024. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *International Journal of Machine Learning and Cybernetics* (2024), 1–65.

[67] Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. 2025. FMGS: Foundation Model Embedded 3D Gaussian Splatting for Holistic 3D Scene Understanding. *International Journal of Computer Vision* 133, 2 (2025), 611–627.