Real-Time Generative Augmented Reality Systems: Design Principles for Geometry-Grounded, Context-Aware Generation

Julien Caldrex, Elise Morvane

2025-12-20

Abstract

Generative models have recently demonstrated impressive capabilities in image, video, and 3D content creation. However, deploying such models in real-time augmented reality (AR) systems remains a major challenge due to strict latency constraints, the need for spatial consistency with the physical world, and the requirement for continuous context awareness. This paper presents a systems-level analysis of real-time generative AR, arguing that explicit 3D geometry is a necessary foundation for stable, interactive, and context-aware generation on wearable and mobile platforms. We analyze architectural trade-offs across representations, memory, control, and latency, and distill design principles from recent geometry-grounded generative systems.

1 Introduction

Diffusion-based and foundation models have rapidly advanced the state of generative media, enabling high-quality image and video synthesis from text or multimodal prompts [2, 4, 11]. These advances raise the prospect of generative augmented reality systems that can dynamically create, modify, and contextualize virtual content in the physical world.

However, translating offline generative capabilities into real-time AR remains nontrivial. Classical AR research has long emphasized the importance of low latency, spatial alignment, and perceptual stability for user comfort and immersion [1, 3]. Generative models, in contrast, are typically designed for offline inference with minimal real-time constraints. Bridging this gap requires rethinking generative pipelines from a systems perspective.

Recent work suggests that explicit 3D representations provide a crucial bridge between generative modeling and real-time AR requirements [6, 15]. In this paper, we synthesize insights from generative modeling, 3D vision, and AR systems to articulate design principles for real-time generative AR.

2 System Requirements for Real-Time Generative AR

Real-time generative AR systems differ fundamentally from offline generation along several axes.

2.1 Latency and Responsiveness

Motion-to-photon latency above approximately 20 milliseconds can cause discomfort, disorientation, and loss of presence in head-mounted AR systems [5, 12]. Generative AR pipelines must therefore be designed around strict latency budgets, often requiring decomposition of computation into lightweight on-device components and heavier asynchronous processes.

Large diffusion models alone are insufficient under these constraints. Representation efficiency and incremental updates become central system concerns [7].

2.2 Spatial Alignment and Consistency

Unlike video generation, AR content must remain spatially anchored to the physical environment. Errors in alignment or geometry lead to perceptual drift and break immersion [1]. Explicit modeling of scene geometry is therefore not optional but foundational.

2.3 Persistent Context and Memory

AR experiences frequently span multiple interactions and sessions. Systems must remember virtual entities, spatial anchors, and user preferences over time. This requires persistent latent state beyond frame-level conditioning [6].

3 Geometry-Grounded Representations

Explicit geometry decouples spatial structure from appearance, enabling stable alignment and efficient rendering.

3.1 Representation Trade-offs

Neural Radiance Fields provide high-fidelity geometry but are computationally expensive to update and render [10]. In contrast, 3D Gaussian Splatting offers a compact and explicit representation with real-time rasterization [7].

Table 1 illustrates why Gaussian-based representations are particularly well-suited for interactive AR scenarios.

Table 1: Comparison of Generative Representations for Real-Time AR

Representation	Latency	Consistency	Editability
2D Latent Diffusion	High	Low	Medium
NeRF-based	High	High	Low
3D Gaussian Splatting	Low	High	High

Table 2: Design Dimensions in Generative AR Systems

Dimension	Offline Generation	Real-Time AR
Latency Budget	Seconds	Milliseconds
Spatial Anchoring	Optional	Required
Persistent Memory	Limited	Essential
User Interaction	Post-hoc	Continuous

3.2 Incremental Updates

Geometry-grounded representations support localized updates without reprocessing the entire scene. This property is critical for interactive editing and context-aware augmentation.

4 Context-Aware Generation

Context awareness in AR extends beyond prompt conditioning. Systems must integrate environmental semantics, user intent, and historical interaction.

World-model-based approaches maintain a latent representation of the environment that evolves over time, enabling coherent generation across interactions [6]. In applied AR systems, this enables content that adapts to location, task, and narrative context [15].

5 Controllability and Physical Plausibility

Real-time AR requires predictable and interpretable control interfaces. Physics-aware constraints provide a natural mechanism for enforcing motion realism and interaction plausibility [13, 16].

Reinforcement learning–based alignment methods further stabilize behavior under repeated user interaction, particularly in multi-entity scenarios [9]. These methods complement geometry by operating at the policy and objective level.

6 Identity and Persistence

Persistent identity is essential for social and narrative AR applications. Memory-based identity modeling enables consistent virtual characters across time, viewpoint, and interaction context [14]. Geometry further stabilizes identity by anchoring shape and spatial extent.

Table 2 highlights how real-time AR imposes fundamentally different system requirements.

7 Applications

Geometry-grounded generative AR enables a wide range of applications:

- **Contextual Visualization**: Adaptive overlays in museums, education, and navigation [15].
- **Immersive Storytelling**: Persistent narrative elements embedded in physical space [6, 8].
- **Interactive Avatars**: Identity-consistent agents responding to user behavior [14].

8 Discussion and Open Challenges

Despite progress, several challenges remain. Robust geometry estimation under dynamic lighting, occlusion, and motion remains difficult [17]. Balancing generative flexibility with real-time constraints requires careful co-design across hardware, software, and model architecture [5].

Future work should explore hierarchical pipelines where lightweight geometry reasoning runs continuously on-device, while heavier generative components operate asynchronously.

9 Conclusion

We presented an expanded systems-level analysis of real-time generative AR, emphasizing the central role of explicit geometry, persistent memory, and controllable generation. Geometry-grounded approaches provide a principled foundation for stable, responsive, and context-aware AR systems capable of real-world deployment.

References

[1] Ronald Azuma. A survey of augmented reality. *Presence*, 1997.

[2] Omer Bar-Tal et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[3] Mark Billinghurst et al. A survey of augmented reality. *Foundations and Trends in Human-Computer Interaction*, 2015.

[4] Andreas Blattmann et al. Stable video diffusion. *arXiv preprint arXiv:2311.15127*, 2023.

[5] Jason Jerald. *The VR Book: Human-Centered Design for Virtual Reality*. ACM, 2015.

[6] Yixiao Kang, Yukun Song, and Sining Huang. Dream world model (dreamwm): A world-model-guided 3d-to-video framework for immersive narrative generation in vr.

[7] Bernhard Kerbl et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023.

[8] Blair MacIntyre and Mark Billinghurst. A decade of vr storytelling research. *IEEE TVCG*, 2024.

[9] Xinyu Meng et al. Identity-grpo: Optimizing multi-human identity preservation via reinforcement learning. *arXiv preprint arXiv:2506.18244*, 2025.

[10] Ben Mildenhall et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[11] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[12] Mel Slater. Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 2016.

[13] Richard Smith. Physics-based interaction for virtual environments. *IEEE Computer Graphics and Applications*, 2018.

[14] Yukun Song, Sining Huang, and Yixiao Kang. Temporal-id: Robust identity preservation in long-form video generation via adaptive memory banks.

[15] Yukun Song, Yixiao Kang, and Sining Huang. Context-aware real-time 3d generation and visualization in augmented reality smart glasses: A museum application.

[16] Yukun Song, Yixiao Kang, and Sining Huang. Vace-physicsrl: Unified controllable video generation through physical laws and reinforcement learning alignment.

[17] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.