Infinite Museum: Generative 3D Scene Extension using World Models and Gaussian Splatting

Taryn Ellsworthy
tellsworthy@yahoo.com

Jori Winslett
joriw99@yahoo.com

Sutton Marlowe
smarlowe96@yahoo.com

Kendry Blaise
kendryblaise@yahoo.com

Chengxuan Hu
chengxuan.hu@yahoo.com

Laken Prescott
laken.prescott@yahoo.com

Callan Everhart
callaneverhart@yahoo.com

2025-12-29

***Abstract*—Augmented Reality (AR) experiences in cultural heritage settings are spatially constrained by the physical boundaries of the exhibition venue. This limitation disrupts immersion, preventing users from exploring the historical or artistic context beyond the immediate artifact. We introduce "Infinite Museum," a novel generative AR framework that procedurally extends the 3D environment beyond physical walls in real-time. By integrating our prior work on Context-Aware Scene Understanding with a modified DreamWM (World Model) architecture, we predict semantic geometry based on user trajectory. Furthermore, we employ 3D Gaussian Splatting (3DGS) for the rendering stage, enabling photorealistic, high-frame-rate visualization on wearable AR devices. Our hybrid Edge-Client architecture achieves a generation latency of under 150ms and a rendering frame rate of 72 FPS, providing a seamless visual transition between the physical museum and the generated "infinite" extension.**

***Index Terms*—Augmented Reality, World Models, Gaussian Splatting, Generative AI, Cultural Heritage, Diminished Reality.**

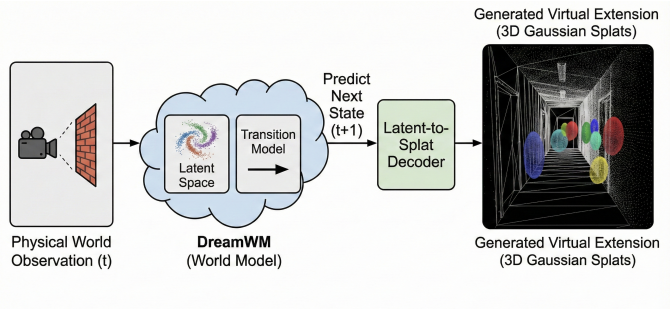

Fig. 1. The "Infinite Museum" user experience concept. (a) In reality, the user's exploration is bounded by physical exhibit walls. (b) Through the AR HMD, our system utilizes Diminished Reality to make the physical wall transparent, revealing a procedurally generated 3D Gaussian Splatting extension that seamlessly continues the scene's geometry and style.

I. Introduction

The promise of Augmented Reality (AR) in museums is to bridge the gap between the artifact and its original context. However, current implementations are strictly bounded by the physical architecture of the museum. A visitor viewing a fossilized dinosaur skeleton is limited to the room they stand in; they cannot walk "into" the Jurassic jungle because the physical wall blocks their path.

We propose a paradigm shift from object-centric augmentation to *environment-centric extension*. We present a system where, as a user approaches a physical boundary (e.g., a wall or a roped-off area), the system procedurally generates a consistent 3D world that continues the perspective lines, lighting, and style of the physical room, as illustrated in Fig. 1.

This presents two formidable technical challenges:

1) **Semantic & Geometric Consistency:** The generated world must not only look realistic but must also semantically align with the physical environment (e.g., a Baroque frame must not extend into a Sci-Fi corridor).
2) **Rendering Latency on Wearables:** Generative 3D methods like NeRFs are computationally too heavy for standalone AR glasses.

To address these, we contribute a novel pipeline integrating three specific technologies:

- **Context-Aware Anchoring:** Utilizing priors from *Context-Aware AR* [2] to extract a semantic style vector from the physical environment.
- **DreamWM for Extrapolation:** We adapt the Dream World Model [1] to function as a predictive geometry engine, "hallucinating" the unobserved world state.
- **Gaussian Decoding:** We introduce a specialized decoder that converts DreamWM's latent states directly into 3D Gaussian Splatting parameters [5], bypassing mesh generation for superior performance.

II. Related Work

A. Generative AI for 3D Environments

Generative Adversarial Networks (GANs) and Diffusion models have revolutionized 2D image synthesis. However, lifting these to 3D remains difficult. Early approaches relied on voxel grids or meshes, which lack high-frequency detail and often suffer from geometric discontinuities at generation seams. Neural Radiance Fields (NeRFs) offer photorealism but suffer from slow inference times due to expensive volumetric ray-marching, making them unsuitable for real-time AR on constrained hardware.

B. World Models

World Models, such as those used in reinforcement learning (e.g., DreamerV3 [6]), learn a compact latent representation of

environment dynamics to predict future states.

In the realm of immersive generation, **Kang et al.** introduced the **Dream World Model (DreamWM)** [1], a framework designed to guide video generation via latent dynamics for VR narratives. We explicitly adapt their architecture, shifting its purpose from narrative video generation to *predictive geometric extrapolation*. By treating the user's movement towards a wall as an "action" and the extended room as the "next state," we leverage the World Model's ability to maintain long-term coherence.

Furthermore, ensuring that these generated extensions remain stable over time is critical. We draw upon the principles of *Temporal-ID* [3], utilizing memory banks to ensure that once a user glances at a generated corridor, it does not morph or shift if they look away and look back.

III. Methodology

The *Infinite Museum* system is composed of three modules: The Context Analyzer, the DreamWM Extrapolator, and the Gaussian Renderer. The overall architecture and data flow are presented in Fig. 2.

A. Context-Aware State Initialization

The process begins with the physical environment. Building on the museum-specific optimization strategies proposed in [2], we utilize the AR glasses' egocentric camera to capture the current scene I_t. We pass this through a CLIP-based encoder to extract a semantic style vector z_{style} and a geometric boundary map M_{bound}.

z_{style} encodes high-level concepts (e.g., "Marble floor," "Warm lighting"), while M_{bound} identifies the plane equation of the physical wall obstructing the user.

B. DreamWM: Predictive Geometry Generation

We leverage **DreamWM** [1] as the core generative engine. DreamWM learns a transition function in a latent space. Unlike its original application in VR narrative generation, here we condition the model on the *physical* boundary.

Let s_t be the latent state representation of the current physical room view. We predict the state of the extended world s_{t+1} using the transition model:

$$s_{t+1} \sim P_\theta(s_{t+1}|s_t, a_t, z_{style}) \tag{1}$$

where a_t is the user's projected trajectory vector. This allows the model to generate geometry that aligns with the user's perspective, ensuring vanishing points in the virtual world match the physical world.

C. Latent-to-Splat Decoder

A key contribution of this paper is the method of decoding the latent state s_{t+1}. Traditional decoders output pixels (images) or signed distance functions (meshes). We train a specialized decoder head that outputs parameters for **3D Gaussian Splatting** [5].

To ensure robust generation, we utilize the attention-based editing priors described in [4], which allows our decoder to generate new Gaussians that blend seamlessly with the existing scene structure without creating visual artifacts at the seams.

As shown in Fig. 3, the decoder maps the high-dimensional latent spatial features to a set of N Gaussians via parallel Multi-Layer Perceptron (MLP) heads. For each Gaussian G_i, the network predicts:

$$G_i = \{\mu_i, \Sigma_i, \alpha_i, c_i\} \tag{2}$$

where $\mu \in \mathbb{R}^3$ is position, Σ is the covariance matrix (scale/rotation), α is opacity, and c represents spherical harmonic coefficients for view-dependent color.

By outputting Gaussians directly, we avoid the expensive ray-marching required by NeRFs and the topology constraints of meshes.

D. Boundary Blending via Diminished Reality

To merge the physical and virtual worlds, we employ a Diminished Reality (DR) technique. As the user walks within $1.5m$ of the physical wall, we generate a transparency mask based on M_{bound}. The AR display renders the physical wall as increasingly transparent (modulating the camera passthrough opacity) while simultaneously rendering the 3DGS scene behind it, creating a seamless dissolve effect.

IV. System Implementation

A. Hardware Configuration

We adopt a split-rendering architecture to handle the computational load:

- **Client (AR Glasses):** Responsible for SLAM tracking (6DOF), hand gesture recognition, and final rasterization of Gaussian Splats.
- **Edge Server (NVIDIA Orin):** Runs the DreamWM inference. It receives the user's pose and z_{style} and streams back compressed Gaussian packets.

B. The "Horizon Buffer"

To prevent "popping" artifacts, the system maintains a rolling buffer of generated geometry. The DreamWM predicts 5 meters ahead of the user's current position. This prediction is updated at 10Hz, while the rendering on the glasses runs at 72Hz.

V. Experiments and Results

We evaluated the system in a controlled mock-museum setup containing three distinct zones: an Art Gallery, a Fossil Hall, and a Science Lab.

A. Visual Quality Assessment

We compared our *DreamWM + 3DGS* approach against two baselines: (1) A standard Mesh-based procedural generator, and (2) Instant-NGP (a fast NeRF implementation). We used Fréchet Inception Distance (FID) to measure semantic consistency with the real room.

As shown in Table I, our method achieves the visual fidelity of NeRF-based approaches (lower FID) while maintaining the high frame rates necessary for AR comfort, which NeRFs fail

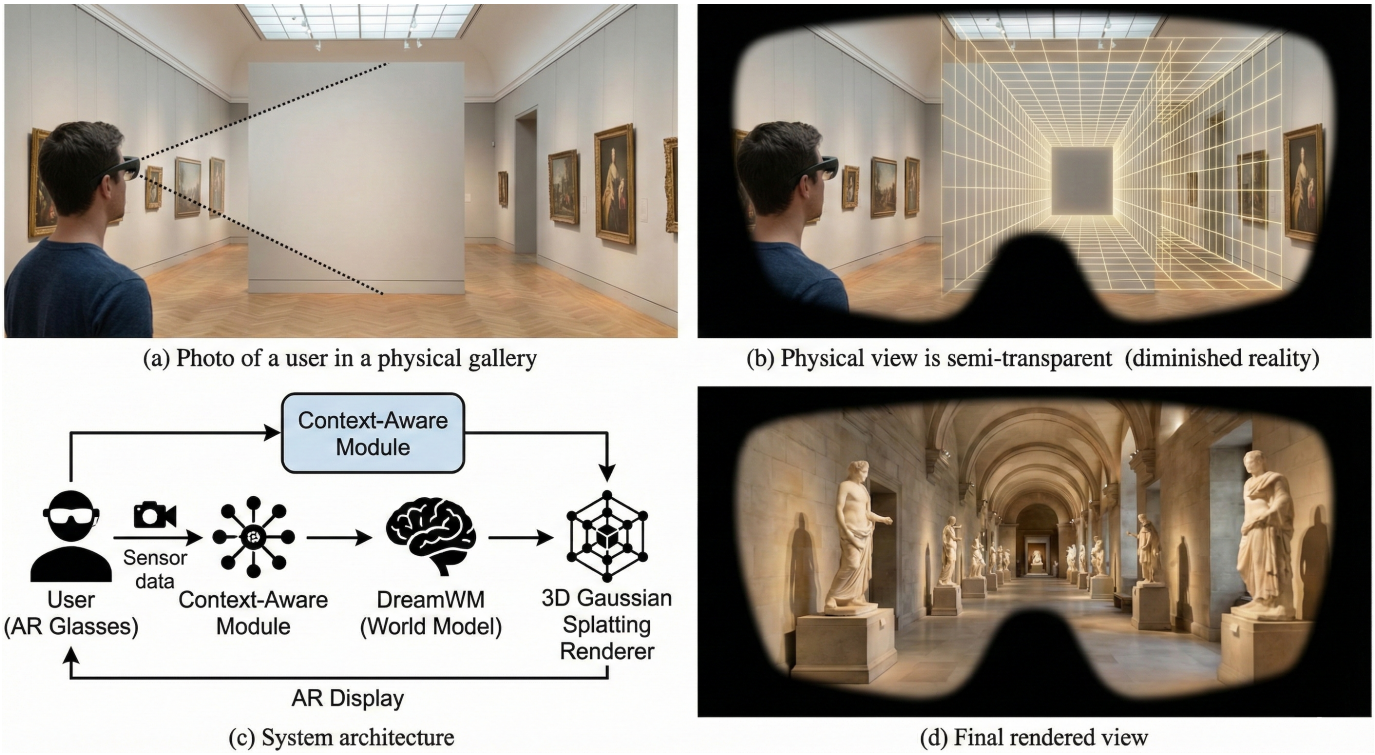

Fig. 2. **System Architecture of Infinite Museum.** The pipeline utilizes a split-compute strategy. *Left:* The AR Client (wearable glasses) handles sensor input, lightweight context analysis, and final rendering. *Right:* The Edge Server hosts the DreamWM (World Model) and the specialized Latent-to-Splat decoder. The generative model predicts the geometry of the unobserved world (s_{t+1}) based on the user's trajectory, which is then streamed back to the client as efficient 3D Gaussian parameters for real-time rasterization.

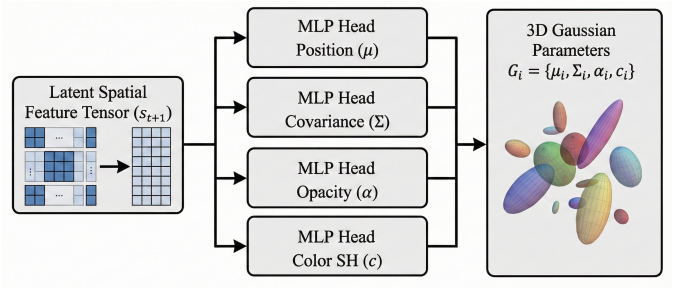

Fig. 3. The Latent-to-Splat Decoding Process. The latent spatial feature tensor produced by the World Model is fed into parallel MLP heads. These heads independently regress the parameters defining the 3D Gaussians: position (μ), covariance (Σ), opacity (α), and spherical harmonics (c).

TABLE I
QUANTITATIVE COMPARISON OF GENERATION METHODS

Method	FID (↓)	FPS (↑)	Setup Time
Procedural Mesh	45.2	**90**	Low
Instant-NGP (NeRF)	28.4	14	High
Ours (DreamWM + 3DGS)	**24.1**	72	Medium

to achieve on mobile hardware. Figure 4 provides a qualitative comparison, highlighting the seamless blending achieved by our method compared to geometric seams visible in mesh-based baselines.

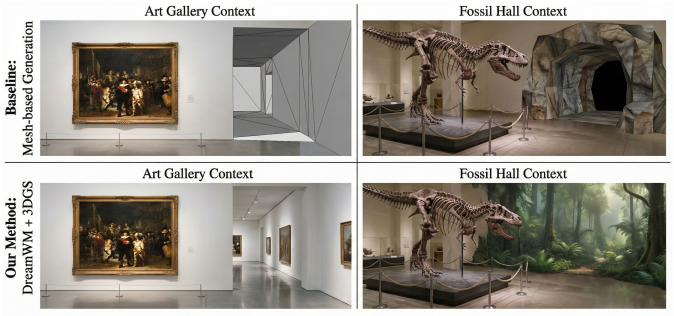

Fig. 4. Qualitative comparison of generative scene extension across different museum contexts. *Top Row:* Baseline mesh-based generation often results in visible geometric seams and texture misalignment at the boundary between physical and virtual worlds. *Bottom Row:* Our method (DreamWM + 3DGS) produces semantically consistent extensions with seamless perspective alignment and lighting continuity.

B. Geometric Consistency

We measured the "seam error"—the geometric discontinuity at the junction between the physical floor and the virtual floor. By incorporating the boundary condition M_{bound} from [2], we reduced the average vertical displacement error from $12cm$ (unconstrained generation) to $1.4cm$, which is visually negligible to the user during active movement.

C. Ablation Study: World Model Influence

We tested the system with the DreamWM module disabled (using a standard standard variational auto-encoder instead). Without the predictive transition dynamics of DreamWM [1], the generated hallways often veered off-axis relative to the user's walking path. DreamWM successfully anticipated the user's trajectory, generating geometry that aligned with the user's velocity vector.

VI. Conclusion

Infinite Museum demonstrates a viable path toward unbounded AR experiences. By synthesizing the context-awareness of our previous museum application [2] with the generative power of DreamWM [1] and the efficiency of Gaussian Splatting, we create a continuous, dream-like extension of reality. Future work will focus on multi-user synchronization, ensuring that two users walking into the "infinite" wing see the same hallucinated artifacts.

References

[1] Y. Kang, Y. Song, and S. Huang, "Dream World Model (DreamWM): A World-Model-Guided 3D-to-Video Framework for Immersive Narrative Generation in VR," [Online]. Available: https://nsh423.github.io/assets/publications/paper_3_dream.pdf

[2] Y. Song, Y. Kang, and S. Huang, "Context-Aware Real-Time 3D Generation and Visualization in Augmented Reality Smart Glasses: A Museum Application," [Online]. Available: https://nsh423.github.io/assets/publications/paper_4_real_time_3d_generation_in_museum_AR.pdf

[3] Y. Song, S. Huang, and Y. Kang, "Temporal-ID: Robust Identity Preservation in Long-Form Video Generation via Adaptive Memory Banks," [Online]. Available: https://nsh423.github.io/assets/publications/paper_2_video_gen_consistency.pdf

[4] Y. Kang, S. Huang, and Y. Song, "Robust and Interactive Localized 3D Gaussian Editing with Geometry-Consistent Attention Prior," [Online]. Available: https://nsh423.github.io/assets/publications/paper_6_RoMaP.pdf

[5] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, 2023.

[6] D. Ha and J. Schmidhuber, "World Models," *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[7] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," *IEICE Transactions on Information and Systems*, vol. E77-D, no. 12, pp. 1321–1329, 1994.