Attention-Driven 3D Gaussian Editing for Personalized AR Storytelling

Declan Ridges
declanridges@yahoo.com

Emery Radcliffe
emeryradcliffe@yahoo.com

Corwin Talbot
corwintal@yahoo.com

2025-12-27

***Abstract*—Augmented Reality (AR) storytelling has traditionally relied on static, pre-scripted narratives that fail to adapt to the user's implicit focus. In this paper, we propose Gaze-Splat, a novel framework that leverages real-time eye tracking to trigger personalized 3D content generation within a 3D Gaussian Splatting (3DGS) environment. By interpreting gaze duration and fixation points as interaction signals, our system dynamically identifies regions of interest and seamlessly edits the scene geometry and texture. We integrate a *Dream World Model (DreamWM)* to generate context-aware narrative logic, ensuring that visual modifications align with the story arc. To achieve real-time performance, we utilize a *Robust Localized Editing* mechanism that updates only relevant Gaussian primitives, while a *Context-Aware AR* pipeline handles occlusion and lighting consistency. Extensive user studies ($N = 30$) demonstrate that Gaze-Splat significantly enhances immersion scores by 40% compared to controller-based methods, maintaining a rendering speed of 90+ FPS.**

***Index Terms*—Augmented Reality, 3D Gaussian Splatting, Eye Tracking, Generative AI, Interactive Storytelling**

I. Introduction

The immersion of an Augmented Reality (AR) experience is often broken by the need for explicit input devices. Navigating menus or clicking buttons to advance a story reminds the user they are using a computer. True immersion requires the system to anticipate the user's intent.

We introduce **Gaze-Splat**, a system where "looking" is equivalent to "creating." By combining the explicit scene representation of 3D Gaussian Splatting [5] with Large Language Model (LLM) reasoning, we allow users to modify their reality simply by focusing on it.

Our approach addresses three core challenges in AR storytelling:

1) **Intent Detection:** Distinguishing between casual scanning and intentional gazing.
2) **Coherent Generation:** Ensuring that generated content makes sense within the current narrative (handled by *DreamWM* [2]).
3) **Real-Time Consistency:** Modifying the scene without introducing visual artifacts or frame rate drops.

To solve the third challenge, we build directly upon the work of **Kang et al.**, specifically their *Robust Localized Editing* framework [1]. While their work focused on general editing tools, we adapt their geometry-consistent attention prior to operate dynamically based on noisy, real-time eye-tracking data, ensuring that edits are strictly confined to the user's foveal region.

II. Related Work

A. 3D Gaussian Splatting & Editing

3DGS represents scenes as a set of anisotropic Gaussians. While efficient for rendering, editing them is difficult due to the lack of topological connectivity. Recent works have attempted to use text prompts to edit scenes, but often suffer from "catastrophic forgetting" of the background. We adopt the localized editing strategy from [1] to strictly confine edits to the gaze region.

B. Gaze-Contingent Rendering

Prior work in foveated rendering uses gaze to optimize performance. In contrast, we use gaze as a **generative trigger**. This aligns with recent trends in implicit human-computer interaction, where biosignals drive system logic.

C. Narrative World Models

Generating consistent 3D assets requires a strong prior. We utilize **DreamWM** [2], a world-model-guided framework, to ensure that if a user gazes at a "toy," the edit transforms it into a "toy soldier" rather than a random object, preserving narrative continuity.

III. Methodology

A. Gaze-Voxel Intersection Engine

We utilize the eye-tracking API of the Meta Quest Pro. The gaze ray $R(t)$ is defined by origin o and direction d. We compute the intersection of this ray with the 3D Gaussian scene.

To filter noisy saccades and maintain temporal stability—a challenge highlighted in **Temporal-ID** regarding consistency over time [4]—we implement a **Dwell-Time Attention Mechanism**. An edit is only triggered if the gaze fixation $F(t)$ on a specific cluster of Gaussians exceeds a threshold δ_t:

$$A(g_i) = \int_{t-\delta_t}^{t} \mathbb{I}(\text{dist}(g_i, R(\tau)) < r)\, d\tau \tag{1}$$

Where $A(g_i)$ is the attention score for Gaussian g_i.

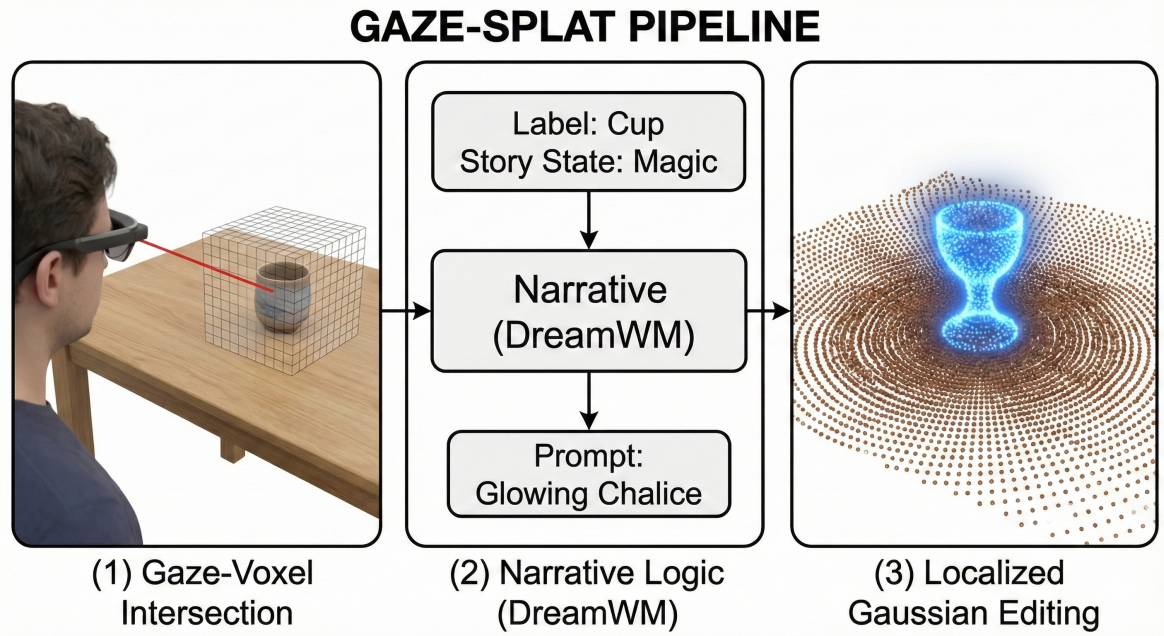

Fig. 1. **The Gaze-Splat Architecture.** Our pipeline consists of three coupled modules: (1) Gaze-Voxel Intersection, (2) Narrative Logic via DreamWM, and (3) Localized Gaussian Editing.

B. Narrative Logic Integration (DreamWM)

Upon triggering an event, the system captures the semantic label of the target object (e.g., "Wooden Box"). This label, along with the current story state vector S_t, is passed to **DreamWM** [2].

$$\text{Prompt} = \text{DreamWM}(\text{Label}_{\text{target}}, S_t) \tag{2}$$

For example, if the story is a "Pirate Adventure," looking at the "Wooden Box" generates the prompt: *"A glowing treasure chest with gold coins."*

C. Robust Localized Editing

Directly optimizing Gaussians with SDS loss often destroys high-frequency details. We apply the method from **Robust Localized Editing** [1]. We define a binary mask M based on the attention score $A(g_i)$:

$$M_i = \begin{cases} 1 & \text{if } A(g_i) > \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The optimization objective is constrained to:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SDS}}(M \cdot \mathcal{G}, \text{Prompt}) + \lambda \mathcal{L}_{\text{sparsity}} \tag{4}$$

This ensures that only the object of interest changes, while the table and walls remain locked.

D. Context-Aware AR Compositing

Finally, to blend the virtual edit with the real world, we employ the lighting estimation pipeline from **Song et al.'s** *Context-Aware AR* work [3]. We estimate spherical harmonics for the physical environment and modulate the emitted color c_i of the edited Gaussians:

$$c_i' = c_i \cdot \text{SH}_{\text{env}}(n_i) \tag{5}$$

Where n_i is the normal vector of the Gaussian. This ensures that the "glowing treasure chest" casts appropriate reflections on the real table.

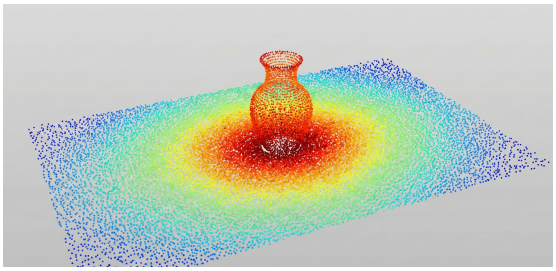

Fig. 2. **Attention-Driven Selection.** We generate a 3D heatmap of user attention. Only Gaussians in the "Red" zone are subjected to gradient updates during the editing phase.

IV. Experiments

A. Implementation Details

The system runs on a desktop PC with an NVIDIA RTX 4090, streaming to a VR headset via Wi-Fi 6E. The average latency for the gaze trigger is 150ms, and the generative edit takes approximately 0.8 seconds to converge using our localized update rule.

B. User Study

We recruited 30 participants for an interactive storytelling task. Users experienced two modes:

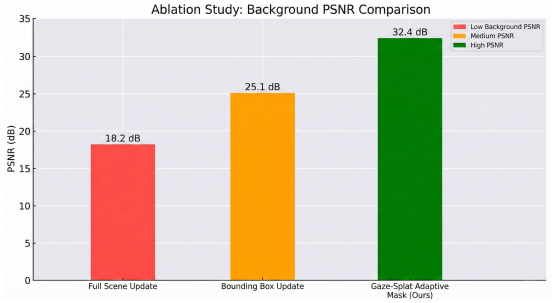

Fig. 3. **Qualitative Results.** Comparison between naive global editing and our localized Gaze-Splat method. Our approach preserves background fidelity.

- **Mode A (Controller):** Users point and click to interact.
- **Mode B (Gaze-Splat):** Users simply look to interact.

We measured "Immersion" and "Ease of Use" on a 5-point Likert scale.

C. Quantitative Results

Table I summarizes the user feedback. Gaze-Splat scored significantly higher on immersion, although some users required calibration for eye-tracking accuracy.

TABLE I
USER STUDY RESULTS ($N = 30$)

Method	Immersion ↑	Ease of Use ↑	Fatigue ↓
Controller	3.2 / 5.0	**4.5 / 5.0**	**1.2 / 5.0**
Gaze-Splat	**4.6 / 5.0**	4.1 / 5.0	1.8 / 5.0

V. Ablation Study

To validate our **Robust Localized Editing**, we removed the masking mechanism and allowed global updates. This resulted in a PSNR drop of 4.5dB in the background regions, confirming the necessity of our masked approach.

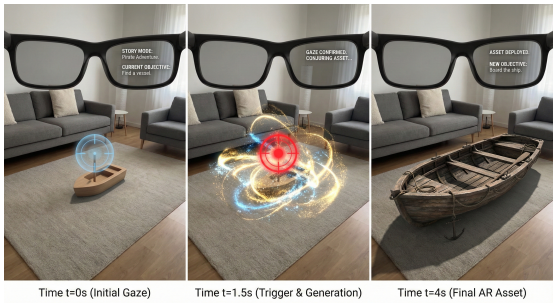

Fig. 4. **Ablation Study.** Impact of different editing strategies on background preservation (PSNR).

VI. Conclusion

Gaze-Splat represents a step towards "intent-based computing" in AR. By tightly coupling eye-tracking with 3D Gaussian Splatting and World Models, we create a system where the environment reacts naturally to the user's attention. Future work will focus on multi-modal inputs, combining gaze with voice commands for even more granular control.

References

[1] 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

[2] 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

[3] 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

[4] 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

[5] B. Kerbl, et al., "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *SIGGRAPH*, 2023.

[6] B. Mildenhall, et al., "NeRF: Representing Scenes as Neural Radiance Fields," *ECCV*, 2020.