Interactive 3D Editing as a Consistency Problem: Geometry-Aware Generation in Gaussian Scene Representations

Tessa Norric, Daniel Iverton

2025-12-28

Abstract

Interactive editing of 3D scenes using generative models remains challenging due to multi-view inconsistency, structural drift, and temporal instability. While recent text- and instruction-based 3D editing systems enable intuitive interaction, they often lack guarantees of geometric coherence across views and time. In this paper, we argue that interactive 3D editing should be fundamentally framed as a consistency optimization problem. We review recent progress in geometry-aware representations, attention unprojection, and generative editing, and present a unified perspective that positions explicit 3D structure as the key enabler for stable, localized, and interactive 3D editing.

1 Introduction

Generative models have recently enabled intuitive interaction with visual content through natural language instructions [2]. Extensions to 3D content allow users to modify geometry, appearance, and semantics without manual modeling [4, 9]. However, interactive 3D editing remains fundamentally difficult: edits often fail to propagate consistently across viewpoints, resulting in structural artifacts and semantic ambiguity [12].

These limitations arise because most instruction-following models operate in image space or view-dependent latent spaces. Without explicit geometric grounding, edits are underconstrained across views. Recent work suggests that explicit 3D representations such as Gaussian splatting provide a promising substrate for consistent editing [7]. Building on this insight, we argue that interactive 3D editing should be reframed as a geometry-consistent optimization problem.

2 Background

2.1 Instruction-Based Editing

Instruction-based image editing models demonstrate strong semantic alignment but struggle with spatial consistency [2]. When extended to 3D via view-wise optimization, these methods often exhibit view conflict and geometry distortion [4]. Such issues worsen under complex edits or localized modifications.

2.2 Explicit 3D Scene Representations

Neural Radiance Fields provide continuous volumetric representations but are expensive to optimize for interactive editing [8]. In contrast, 3D Gaussian Splatting offers a compact, explicit, and differentiable scene representation, enabling real-time rendering and localized manipulation [7].

Recent text-to-3D methods leverage Gaussian splatting to accelerate generation while preserving multi-view consistency [3, 14]. These representations naturally support spatial indexing and region-level control.

3 Editing as a Consistency Optimization Problem

We argue that interactive editing can be formalized as minimizing inconsistency across three dimensions:

1. **View Consistency**: edits must remain coherent across camera viewpoints.
2. **Geometric Consistency**: modifications should respect underlying 3D structure.
3. **Temporal Consistency**: edits should persist across time and interaction steps.

Recent approaches explicitly incorporate geometry-aware attention to enforce such constraints. Inter-RoMaP introduces geometry-consistent attention priors to localize and propagate edits in Gaussian space [5]. Similar ideas appear in view-consistent Gaussian editing frameworks [12, 13].

4 Geometry-Aware Attention and Editing

Geometry-aware attention aligns 2D instruction signals with 3D structure by unprojecting attention maps into scene space. This mechanism enables precise localization and prevents semantic leakage across unrelated regions [5]. By grounding attention in geometry, edits remain spatially coherent under novel views and repeated interactions.

World-model-based approaches further integrate editing into a persistent latent state, enabling consistent scene evolution under multiple edits [6]. These systems treat editing as a state transition rather than an isolated operation.

5 Interactive and Real-Time Systems

For immersive applications, editing systems must operate under strict latency constraints. Real-time AR systems demonstrate that geometry-grounded generation is feasible on-device when representations are compact and explicit [10]. Gaussian-based pipelines are particularly well-suited for such scenarios due to their rendering efficiency [7].

Physics-aware constraints further improve realism and stability under interaction, especially when edits involve motion or deformation [11].

6 Applications

Consistency-aware 3D editing enables:

- **Creative Authoring**: intuitive scene modification without manual modeling.
- **Immersive AR/VR**: real-time world editing in spatial computing systems [1].
- **Narrative Generation**: persistent environments supporting interactive storytelling [6].

7 Discussion

Despite recent progress, challenges remain. Geometry estimation errors propagate through editing pipelines, and strong geometric constraints may limit creative flexibility. Future work should explore hybrid representations that combine explicit structure with learned semantic priors, and develop user-in-the-loop optimization strategies that balance control and expressiveness.

8 Conclusion

We presented a unifying perspective on interactive 3D editing, framing it as a geometry-consistency optimization problem. Explicit 3D representations, particularly Gaussian-based scenes, provide a principled foundation for stable, localized, and interactive generative editing across views and time.

References

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

[2] Tim Brooks et al. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

[3] Jiatao Gu et al. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024.

[4] Albert Haque et al. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *ICCV*, 2023.

[5] Yixiao Kang, Sining Huang, and Yukun Song. Robust and interactive localized 3d gaussian editing with geometry-consistent attention prior.

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

[9] Ben Poole et al. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

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

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

[12] Yifan Wang et al. View-consistent 3d editing with gaussian splatting. In *ECCV*, 2024.

[13] Jun Wu et al. Gaussctrl: Multi-view consistent text-driven 3d gaussian editing. In *ECCV*, 2024.

[14] Tianshuo Yuan et al. Gaussiandreamer: Fast text-to-3d generation via gaussian splatting. In *CVPR*, 2024.