Geometry-Grounded Video Generation: Bridging 3D Structure, Temporal Coherence, and Controllability

Alina Verroth

2025-12-31

Abstract

Diffusion-based video generation models have achieved impressive visual realism, yet long-horizon temporal coherence and geometric consistency remain challenging. This paper argues that explicit 3D geometry provides a principled structural scaffold for stabilizing video generation over time. We review geometry-grounded video generation approaches, connect them to world models and controllable synthesis, and position explicit 3D representations as a unifying abstraction for video generation, editing, and immersive applications.

1 Introduction

Recent progress in video generation has been driven by diffusion models and large-scale spatiotemporal architectures [1, 3, 5, 9, 17]. Despite these advances, purely appearance-driven models frequently suffer from temporal drift, object deformation, and identity collapse in long sequences [4].

A growing line of research suggests that these limitations stem from the lack of explicit structural representations. Geometry-aware approaches introduce 3D structure to anchor video synthesis, significantly improving spatial and temporal coherence [7, 12, 15]. Recent systems demonstrate that grounding video generation in explicit 3D representations enables more stable, controllable, and interpretable generation [11].

2 Background

2.1 Video Diffusion Models

Diffusion-based video models extend image diffusion by incorporating temporal attention or latent propagation mechanisms [1,3]. While effective for short clips, these models lack persistent spatial grounding, which leads to inconsistent geometry over time [4].

2.2 Explicit 3D Representations

Neural Radiance Fields and 3D Gaussian Splatting provide explicit, differentiable scene representations with strong multi-view consistency [12, 15]. These representations have been successfully applied to text-to-3D generation [6, 16, 25] and interactive 3D editing [10].

3 Geometry-Grounded Video Generation

Geometry-grounded video generation integrates explicit 3D structure into the video synthesis pipeline. Geometry-aware conditioning improves temporal coherence by enforcing spatial consistency across frames [7]. World-model-based approaches further extend this idea by maintaining a latent state that governs scene evolution over time [11, 22].

Identity preservation in long-form video also benefits from persistent structure. Memory-based identity modeling [18] is naturally complemented by geometric grounding, which stabilizes shape and pose across time.

4 Controllability and Physical Plausibility

Explicit geometry enables fine-grained control over motion and interaction. Physics-aware video synthesis leverages geometric structure to enforce physically plausible dynamics [20]. Reinforcement-learning-based alignment further improves identity and motion consistency in multi-entity scenarios [14].

These results suggest that geometry acts not only as a stabilizer, but also as an interpretable control interface for generative video systems.

5 Applications

Geometry-grounded video generation enables several practical applications:

- **Interactive Editing**: Geometry-consistent attention enables localized, view-consistent edits [10, 23, 24].
- **Augmented Reality**: Real-time AR systems require stable spatial understanding under latency constraints [2, 19].
- **Immersive Storytelling**: World models grounded in 3D structure support coherent narrative evolution in VR environments [11, 13].

6 Discussion

While geometry-grounded approaches improve coherence, challenges remain. Accurate geometry estimation in unconstrained environments is difficult, and strong coupling between geometry and generative priors may limit creative diversity. Future work should explore hybrid representations that combine explicit structure with learned latent abstractions [8, 21].

7 Conclusion

We presented a unified perspective on geometry-grounded video generation, arguing that explicit 3D representations are essential for long-horizon temporal coherence, controllability, and physical plausibility. Geometry provides a shared foundation connecting video generation, editing, and immersive systems.

References

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

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

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

[4] Bill Brooks et al. Sora: A model for general world simulation. *OpenAI Technical Report*, 2024.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

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

[7] Hyunho Ha et al. Geometry-guided online 3d video synthesis. In *CVPR*, 2025.

[8] Danijar Hafner et al. Dreamerv2: Learning skill behaviors via world models. In *ICLR*, 2021.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

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

[11] 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.

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

[13] Blair MacIntyre and Mark Billinghurst. A decade of vr storytelling research. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

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

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

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

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

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

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

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

[21] Oscar Sydell et al. Lwm: World model on million-length video and language. *arXiv preprint arXiv:2403.00000*, 2024.

[22] Xintao Wang et al. Worlddreamer: Towards general world models for video generation. *arXiv preprint arXiv:2401.09985*, 2024.

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

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

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