From Generators to Worlds: Stateful Foundations for Persistent and Interactive Generative Systems

Alina Verroth, Ryan Kestrelon

2025-12-28

Abstract

Generative models have achieved striking perceptual fidelity for images and video, yet their behavior remains unstable when extended beyond short, offline generation. Long-horizon video, interactive editing, and physically grounded deployment consistently expose failures such as identity drift, geometric inconsistency, and loss of control under repeated interaction. This paper argues that these failures do not primarily stem from insufficient data or model scale, but from a deeper architectural limitation: most generative systems do not maintain persistent internal state.

We propose a reframing of generative media systems as *stateful world processes* rather than appearance-based generators. In this view, geometry, identity, physics, and interaction history are not auxiliary conditioning signals but persistent variables that evolve over time. We develop this argument through six systems—FaceSplat, Temporal-ID, DreamWM, Museum AR, VACE-PhysicsRL, and Inter-RoMaP—which together span reconstruction, memory, world modeling, deployment, control, and editing. Rather than surveying these works independently, we reinterpret them as partial solutions to a unified state management problem. This perspective yields new design criteria for generative systems that must remain coherent under time, viewpoint changes, and user intervention.

1 The Core Failure of Modern Generative Systems

Generative modeling has rapidly advanced across vision and graphics. Diffusion models [11, 22] produce high-quality images with remarkable diversity. Video generation has followed, incorporating temporal attention and spatiotemporal architectures [2, 4]. At scale, video prediction is increasingly framed as implicit world simulation [20]. In parallel, 3D representations have transitioned from implicit radiance fields [18] to explicit primitives optimized for real-time rendering [16].

Despite these advances, a consistent pattern emerges when these systems are pushed beyond single-shot generation. When outputs must persist across time, viewpoint, or interaction, models exhibit degradation that cannot be resolved by prompt engineering or larger networks. These issues appear across domains: cinematic video generation, AR content placement, avatar synthesis, and interactive editing.

We claim that these failures arise because most generative systems are *stateless*. They map inputs to outputs but do not maintain a durable representation of what exists in the world. Each frame, view, or edit effectively reconstructs the world from scratch, using pixels or latent features as a proxy for memory.

2 Generation Without State Is Fundamentally Ill-Posed

2.1 Temporal coherence without memory

Video diffusion models typically enforce coherence through short-range temporal attention [4]. While effective locally, such mechanisms cannot guarantee consistency over long horizons. Architectures that generate entire sequences jointly [2] improve global structure but remain brittle under mid-sequence edits or user interaction.

The underlying issue is not temporal modeling per se, but the absence of a persistent entity that carries information forward. When a system lacks memory, it must re-infer all relevant structure at every step, amplifying noise and compounding small errors.

2.2 Identity drift as repeated re-inference

Identity drift illustrates this problem clearly. In long-form generation, identity features become entangled with pose, lighting, and scene context. Because identity is not stored explicitly, it is repeatedly estimated from appearance, leading to gradual divergence.

Temporal-ID [25] addresses this failure by introducing adaptive memory banks that store identity representations over time. Importantly, this mechanism does not improve appearance modeling—it prevents identity from being regenerated at every step. This distinction is central to our argument: identity stability is a state retention problem, not a rendering problem.

2.3 Viewpoint changes without geometry

Similarly, view inconsistency arises when systems lack geometric state. NeRF enforces multi-view consistency by optimizing a shared volumetric representation [18], but its computational cost limits interactivity. Explicit representations such

as Gaussian splats [16] reduce this cost and enable incremental updates, revealing that geometry persistence—not just multi-view training—is the key enabler of stability.

3 Reframing Generative Media as World State Evolution

We propose treating generative systems as processes that evolve an internal world state:

$$\mathcal{S}_t = \{\mathbf{G}_t, \mathbf{I}_t, \mathbf{P}_t, \mathbf{H}_t\},$$

where $\mathbf{G}_t$ represents geometry, $\mathbf{I}_t$ identity, $\mathbf{P}_t$ physical constraints, and $\mathbf{H}_t$ interaction history. Outputs are observations rendered from $\mathcal{S}_t$, not direct predictions from input prompts.

This formulation aligns generative media with model-based reinforcement learning, where agents maintain latent dynamics models for planning [8, 9]. World models in RL support imagination and long-horizon reasoning [7], while planning-oriented systems such as MuZero learn to operate over internal state transitions [24]. Generative media systems face an analogous challenge: they must maintain and update a world under user actions rather than environment actions.

4 Geometry as a Stabilizing Prior

4.1 FaceSplat and partial geometric state

FaceSplat [12] demonstrates that even incomplete geometric structure dramatically reduces ambiguity in generation. By combining Gaussian splatting with facial priors, the method produces stable reconstructions from minimal input. The significance here is not reconstruction accuracy alone, but the introduction of a persistent spatial scaffold that downstream generation can rely on.

4.2 Geometry acquisition pipelines

Persistent geometry must be updated continuously in real systems. Offline SfM pipelines such as COLMAP [23] provide accurate reconstruction, while SLAM systems [6] estimate geometry and camera pose under real-time constraints. Learning-based motion estimation [28] complements these approaches in dynamic scenes. Together, these systems populate and update $\mathbf{G}_t$ rather than treating geometry as an implicit byproduct of rendering.

5 Identity as Explicit Memory

Temporal-ID [25] formalizes identity persistence as a retrieval problem. Rather than encoding identity implicitly in latent features, it maintains a memory bank that can be queried and updated. This mechanism decouples identity from transient visual factors and allows identity to survive edits, occlusions, and temporal gaps.

When combined with geometry-aware systems such as FaceSplat [12], identity becomes anchored both semantically and spatially. This dual anchoring is critical for avatars, characters, and recurring objects in generative worlds.

6 World Models Beyond Rendering

6.1 DreamWM and state-driven generation

DreamWM [15] explicitly models world state evolution for immersive narrative generation. Instead of synthesizing video frame-by-frame, DreamWM propagates a latent world representation that governs scene transitions and narrative logic. This approach mirrors world-model learning in RL, where state evolution enables long-horizon consistency [8].

6.2 Implicit simulators vs explicit state

Large video models are often described as implicit simulators [20]. While powerful, such simulators offer limited interpretability and control. Our position is not that implicit models are insufficient, but that interactive systems benefit from exposing and manipulating parts of the internal state explicitly.

7 Deployment Reveals Architectural Truths

7.1 Museum AR

Museum AR [26] situates generative models within the constraints of wearable AR: low latency, noisy sensors, and continuous interaction. In this setting, stateless generation fails immediately. Context—location, viewpoint, prior interaction—must persist across time to avoid perceptual instability.

Classic AR principles emphasize registration stability and latency budgets [1], while VR systems highlight human perceptual constraints [13]. Museum AR demonstrates that persistent state is not optional in deployed systems; it is a prerequisite.

8 Physics as a Control Interface

8.1 VACE-PhysicsRL

VACE-PhysicsRL [27] reframes controllable generation as a physics-aligned decision process. Physical laws provide interpretable constraints, while reinforcement learning aligns generative behavior with these constraints. This approach draws on decades of work in physically based simulation [3, 17, 19, 29].

The key insight is that physics is not merely for realism—it is a structured language for control. Unlike text prompts, physical parameters define admissible trajectories and interactions.

9 Editing as State Manipulation

9.1 Inter-RoMaP

Inter-RoMaP [14] treats editing as a localized state update rather than global regeneration. By leveraging geometry-consistent attention over Gaussian representations, it preserves global coherence while allowing targeted modification. This contrasts with instruction-based image editing [5] and NeRF-based editing pipelines [10], which often require costly re-optimization.

Text-to-3D methods such as DreamFusion [21] further highlight the tension between generative priors and persistent structure, reinforcing the need for explicit state.

10 Design Implications

From this reframing, several implications follow:

- Generative systems should maintain explicit state rather than relying on implicit re-inference.
- Geometry and identity should persist independently of appearance.
- Physics provides a structured, interpretable control space.
- Context and interaction history must be first-class variables.
- Evaluation should stress repeated interaction and long-horizon consistency [20].

11 Open Challenges

Scaling persistent memory remains unresolved [25]. Robust geometry estimation under real-world conditions is still difficult [6, 23, 28]. Balancing hard constraints with creative flexibility remains an open research question, as does deploying such systems efficiently on resource-limited devices [26].

12 Conclusion

This paper argues that the central limitation of modern generative media systems is not realism, but statelessness. By reinterpreting six complementary systems—FaceSplat [12], Temporal-ID [25], DreamWM [15], Museum AR [26], VACE-PhysicsRL [27], and Inter-RoMaP [14]—as components of a unified stateful architecture, we show how generative models can evolve into persistent, interactive world systems.

References

[1] Ronald T. Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385, 1997.

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[3] David Baraff. Fast contact force computation for nonpenetrating rigid bodies. *Computer Graphics (Proceedings of SIGGRAPH)*, 1994.

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2211.09800.

[6] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. Orb-slam3: An accurate open-source library for visual, visual-inertial and multi-map slam. *IEEE Transactions on Robotics*, 2021. arXiv:2007.11898.

[7] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[8] Danijar Hafner et al. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020. DreamerV2.

[9] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *International Conference on Machine Learning (ICML)*, 2019. arXiv:1811.04551.

[10] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2303.12789.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2006.11239.

[12] Sining Huang, Yixiao Kang, and Yukun Song. Facesplat: A lightweight, prior-guided framework for high-fidelity 3d face reconstruction from a single image.

[13] Jason Jerald. *The VR Book: Human-Centered Design for Virtual Reality*. ACM Books / Morgan & Claypool, 2015. ACM DOI:10.1145/2792790.

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

[15] 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.

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 2023. arXiv:2308.04079, ACM DOI:10.1145/3592433.

[17] Miles Macklin, Matthias Müller, Nuttapong Chentanez, and Tae-Yong Kim. Unified particle physics for real-time applications. *ACM SIGGRAPH*, 2014.

[18] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *European Conference on Computer Vision (ECCV)*, 2020. arXiv:2003.08934.

[19] Brian V. Mirtich. *Impulse-Based Dynamic Simulation of Rigid Body Systems*. PhD thesis, University of California, Berkeley, 1996.

[20] OpenAI. Video generation models as world simulators. OpenAI Blog, 2024. Published Feb 15, 2024.

[21] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.14988.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2112.10752.

[23] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 2020. Also arXiv:1911.08265.

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

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

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

[28] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, 2020. arXiv:2003.12039.

[29] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.