Towards Persistent World Models for Generative Media: Geometry, Identity, Physics, and Interaction as First-Class Constraints

Julien Caldrex, Ryan Kestrelon

2025-12-28

Abstract

Diffusion models and large-scale generative systems have made remarkable progress for images and video, yet they remain brittle when deployed in long-horizon, interactive, or physically grounded settings. Common failure modes include temporal flicker, identity drift, geometry inconsistency under viewpoint change, and weak controllability beyond prompt-level guidance. This paper argues that these failures share a single underlying cause: most generative pipelines lack a persistent world model that explicitly represents stable state over time. We synthesize a unified perspective for generative world models in which geometry, identity, physics, and interaction history are treated as first-class constraints rather than auxiliary conditioning signals. Our discussion is anchored in six complementary systems from a single research repository—FaceSplat, Temporal-ID, DreamWM, Museum AR, VACE-PhysicsRL, and Inter-RoMaP—and connects them to established foundations in diffusion modeling, 3D reconstruction, SLAM, model-based reinforcement learning, and physically based simulation. We distill concrete design principles for building persistent generative systems and highlight open research directions for scalable memory, real-time deployment, and controllable editing.

1 Introduction

Generative modeling has experienced rapid advances across modalities. Diffusion probabilistic models [11] and latent diffusion models [22] enabled high-quality image synthesis and flexible conditional generation. Video generation similarly progressed via latent-space architectures and space-time modeling [2, 4], and recent large-scale efforts explore video generation systems as "world simulators" [20]. In parallel, 3D representations have shifted from implicit radiance fields [18] to explicit, real-time-friendly point-based or Gaussian representations [16]. These developments suggest a plausible future where generative models are not merely offline content creators but components of interactive systems: AR assistants that synthesize context-aware 3D overlays, immersive narrative engines that maintain story and scene continuity, or editing tools that allow localized, geometry-consistent modifications.

However, practitioners increasingly encounter a hard boundary between offline generation and interactive deployment. Long-form video generation exhibits temporal instability and identity drift even with strong per-frame realism; 3D generation suffers view inconsistency; and text-only control signals are too coarse for iterative user refinement. These deficiencies become critical in augmented reality and VR, where spatial anchoring, low latency, and perceptual stability are non-negotiable [1, 13]. More broadly, they matter whenever a system must maintain continuity across time, edits, or repeated interactions.

This paper advocates a systems-level hypothesis: *the primary bottleneck is the absence of persistent world modeling*. Most generative pipelines generate outputs that are *appearance-driven*—optimized to match local statistics or short windows of context—without maintaining an explicit state representation that persists, accumulates evidence, and supports consistent inference. In contrast, a persistent world model encodes latent state across time and interaction: geometry for spatial structure, identity for long-term consistency, physics for plausible dynamics, and memory for accumulated context. This hypothesis is both explanatory (it accounts for common failure modes) and constructive (it suggests how to build more robust systems).

We ground this perspective using six recent works in a single repository: (1) **FaceSplat** by Huang, Kang, and Song [12], (2) **Temporal-ID** by Song, Huang, and Kang [25], (3) **DreamWM** by Kang, Song, and Huang [15], (4) **Museum AR** by Song, Kang, and Huang [26], (5) **VACE-PhysicsRL** by Kang, Huang, and Song [27], and (6) **Inter-RoMaP** by Kang, Huang, and Song [14]. Each contributes a distinct constraint or systems insight; together they define a coherent research program for persistent generative world models.

Contributions. This paper provides: (i) a unified problem formulation for persistent generative media, (ii) an analysis of failure modes that arise without persistent state, (iii) a synthesis of six complementary repository systems and their roles, (iv) concrete design principles spanning representation, memory, and control, and (v) open problems and evaluation strategies for next-generation systems.

2 Why Appearance-Driven Generation Breaks

We first analyze why modern generative pipelines degrade in long-horizon or interactive regimes.

2.1 Temporal instability and error accumulation

Diffusion-based video models often combine spatial denoising with temporal layers or attention [4]. While these architectures improve short clips, long-horizon generation remains difficult because errors accumulate and the model lacks an explicit mechanism to maintain global state. Space-time U-Nets that generate entire videos in one pass can improve global temporal coherence [2], but the fundamental challenge persists when interaction or editing introduces off-distribution perturbations.

When users request changes mid-sequence—e.g., modify an object while preserving identity and motion—the model must maintain a consistent representation of what remains unchanged. In an appearance-driven system, "what remains unchanged" is not represented; it must be re-inferred from pixels. This leads to drift, flicker, and semantic inconsistency.

2.2 Identity drift as a memory failure

Identity preservation is central for long-form video generation, avatars, and narrative media. In practice, identity drift arises because identity is entangled with pose, lighting, and scene context in pixel-level features. Without persistent memory, a model re-estimates identity at each step, and small errors compound.

Temporal-ID by Song et al. [25] directly targets this issue. Their core contribution—adaptive memory banks—treats identity as a persistent embedding that is retrieved and reinjected during generation. In the context of this paper, Temporal-ID supports the claim that identity preservation is fundamentally a memory modeling problem: stable identity requires *explicit persistence* beyond local temporal attention.

2.3 View inconsistency and missing geometry

Without geometry, generative systems struggle to ensure view consistency. A classic example is multi-view rendering: generating a plausible image from one viewpoint says little about the scene from another viewpoint. NeRF introduced an implicit representation that is view-consistent by construction [18], but NeRF optimization and rendering costs make it challenging for interactive systems.

Recent explicit representations such as 3D Gaussian Splatting achieve real-time rendering and efficient optimization [16]. This shift is more than an engineering improvement—it enables *interactive* and *localized* updates, which are essential for persistent world models.

2.4 Control limitations of prompt-only interfaces

Text prompts provide high-level semantic control, but interactive systems demand finer control: specifying motion trajectories, constraints, physical plausibility, and localized edits. Prompt-only control lacks an interpretable parameterization for dynamics and interaction. This motivates constraint-based control signals (e.g., pose, depth, physics parameters) and learning objectives that align generation with these constraints.

VACE-PhysicsRL by Kang et al. [27] contributes here by unifying controllable video generation with physical laws and reinforcement learning alignment. In our synthesis, VACE-PhysicsRL shows that controllability becomes more reliable when expressed in a physically grounded control space and optimized with explicit alignment objectives.

3 A Persistent World Model Perspective

We define a persistent world model for generative media as a tuple:

$$\mathcal{W}_t = \{\mathbf{G}_t, \mathbf{I}_t, \mathbf{P}_t, \mathbf{M}_t\},$$

where $\mathbf{G}_t$ denotes geometry/state of the environment, $\mathbf{I}_t$ denotes persistent identities (people, objects, entities), $\mathbf{P}_t$ denotes physical constraints/dynamics, and $\mathbf{M}_t$ denotes memory of interaction and context. The generative model produces observations (images/video/3D renders) conditioned on $\mathcal{W}_t$, and updates $\mathcal{W}_t$ based on new observations and user interaction.

This perspective aligns with model-based reinforcement learning, where agents learn latent dynamics for planning [8,9] and world models for imagination [7]. Planning-oriented systems learn compact state representations that support multi-step prediction and action selection. For generative media, the analogous goal is multi-step *content* prediction under constraints and edits, with user actions replacing environment actions.

Relation to "world simulators" in video generation. Large-scale video generation efforts position models as "world simulators" [20], emphasizing that video prediction implicitly captures physical and semantic structure. Our perspective is complementary: we argue that interactive and controllable systems benefit from making portions of this structure explicit via geometry, identity memory, and physics constraints, rather than relying solely on implicit representations.

4 Geometry as the Structural Backbone

4.1 FaceSplat: Prior-guided geometry for stable identity

FaceSplat by Huang et al. [12] introduces a lightweight, prior-guided framework for high-fidelity 3D face reconstruction from a single image, leveraging Gaussian splatting and face priors. The key contribution for this paper is the demonstration that *partial but structured geometry* substantially stabilizes downstream generation and editing. Even when a single view is insufficient to recover perfect geometry, introducing a

geometric scaffold reduces ambiguity and constrains plausible solutions.

In persistent world models, faces and human identity are especially sensitive to drift. FaceSplat provides a mechanism to initialize and maintain a geometry-anchored identity representation, enabling stable rendering across viewpoint and consistent conditioning for video generation.

4.2 From NeRF to Gaussians: interactive representations

NeRF provides a powerful implicit representation for view synthesis [18]. Yet NeRF-based pipelines often require expensive optimization and are difficult to edit locally. 3D Gaussian Splatting addresses these challenges with explicit primitives optimized for fast rendering [16]. Explicit representations are particularly valuable for persistent systems because they: (i) support incremental updates, (ii) allow localized edits without re-optimizing the entire scene, (iii) provide direct handles for constraints (e.g., geometry alignment), and (iv) integrate naturally with real-time rendering engines.

4.3 Geometry acquisition in real systems: SfM and SLAM

Persistent world models depend on robust geometry acquisition. Classical SfM pipelines such as COLMAP [23] provide accurate offline reconstruction. For real-time AR/VR, SLAM systems (e.g., ORB-SLAM3 [6]) estimate camera trajectories and sparse maps under latency constraints. These perception modules provide the geometric state $\mathbf{G}_t$ upon which generative modules operate.

In interactive settings, motion and viewpoint changes occur continuously; thus, geometry must be updated online. Optical flow systems such as RAFT [28] support motion estimation and can complement SLAM in dynamic scenes.

5 Identity as Persistent Memory

5.1 Temporal-ID: adaptive memory banks for long-form generation

Temporal-ID by Song et al. [25] proposes adaptive memory banks to preserve identity in long-form video generation. The conceptual contribution to this paper is that identity should be treated as a *retrieval problem*: as generation proceeds, the model must retrieve stable identity features rather than re-derive them from evolving pixels.

This supports a broader design principle: persistent world models should modularize identity into an explicit component $\mathbf{I}_t$, distinct from local appearance. This separation reduces entanglement with pose, lighting, and background, and enables consistent re-conditioning after edits.

5.2 Identity-geometry synergy

Identity and geometry are complementary. Geometry anchors shape and viewpoint consistency; identity memory anchors semantic and appearance traits. In systems that generate humans or recurring objects, maintaining both $\mathbf{G}_t$ and $\mathbf{I}_t$ is necessary for stability. FaceSplat [12] provides geometry priors for faces, while Temporal-ID [25] provides memory mechanisms for identity persistence. Together, they imply that persistent identity is best represented as a *structured state* rather than a latent that must be re-inferred each frame.

6 World Models for Generative Media

6.1 DreamWM: state evolution for narrative generation

DreamWM by Kang et al. [15] introduces a world-model-guided 3D-to-video framework for immersive narrative generation in VR. DreamWM's contribution to this paper is methodological: it reframes generation as state evolution rather than per-frame synthesis. By maintaining a world-model latent state, DreamWM can better preserve narrative coherence and reduce logic violations across multi-segment sequences.

From our viewpoint, DreamWM operationalizes the persistent world model concept: it explicitly maintains and updates state across time and supports generation conditioned on that state. This aligns with RL world model approaches such as PlaNet [9] and DreamerV2 [8], where latent state supports multi-step prediction and planning.

6.2 Model-based planning parallels

World model literature provides a language for understanding why persistence matters. Ha and Schmidhuber [7] emphasize that an agent can learn a compact generative model and then plan in "dreamed" rollouts. DreamerV2 demonstrates that learning in latent space can yield strong control policies [8]. MuZero further shows how planning with a learned model can achieve strong performance without explicit environment rules [24].

For generative media, the analog is that coherent generation and editing require planning over latent state: choosing how to evolve scenes while preserving constraints. DreamWM [15] provides a concrete instantiation of this idea in a VR narrative setting.

7 Context-Aware Deployment in AR

7.1 Museum AR: constraints from real-time systems

Museum AR by Song et al. [26] introduces context-aware real-time 3D generation and visualization on AR smart glasses in a museum application. The contribution to this paper is a

systems perspective: persistent world models must be deployable under real-world constraints, including low latency, limited compute, noisy sensors, and continuous user interaction.

Museum AR highlights a critical point: *context is not just a prompt*. Context includes location in the environment, the user's viewpoint, object affordances, and interaction history. This corresponds directly to $\mathbf{M}_t$ in our formulation.

7.2 AR/VR stability requirements

Augmented reality requires stable registration and real-time integration of virtual content with the physical world [1]. VR/AR design principles emphasize latency budgets and human-centered constraints [13]. These requirements motivate explicit representations and incremental updates rather than slow, global recomputation.

Museum AR [26] thus supports a key thesis: persistent world models are not an abstract ideal; they are necessary for robust AR deployment.

8 Physics as an Interpretable Control Language

8.1 VACE-PhysicsRL: aligning generation with physical laws

VACE-PhysicsRL by Kang et al. [27] introduces unified controllable video generation through physical laws and reinforcement learning alignment. Its central contribution to this paper is showing that physics-based constraints provide a natural, interpretable control interface for generative dynamics. Where prompt-only control fails to specify detailed motion or interactions, physics parameters and constraints can.

This is consistent with decades of physically based simulation research. Rigid body contact and collision remain fundamental problems [3, 19]. Modern physics engines for control (e.g., MuJoCo) emphasize stable, efficient simulation for learning and optimization [29]. Constraint-based particle frameworks show how diverse materials can be simulated in unified systems [17]. These references collectively motivate why physics constraints are powerful control primitives.

8.2 Why RL alignment matters

Even with constraints, generative models can deviate from desired behavior due to optimization mismatch. RL alignment can operationalize preferences such as "obey physical constraints" or "maintain stable interaction outcomes" by rewarding constraint satisfaction. This connects VACE-PhysicsRL [27] to broader reinforcement learning methodology, where learned objectives can shape behavior beyond supervised loss.

9 Interactive Editing and Geometry-Consistent Manipulation

9.1 Inter-RoMaP: localized editing with geometry-consistent priors

Inter-RoMaP by Kang et al. [14] proposes robust and interactive localized 3D Gaussian editing with geometry-consistent attention priors. Its contribution to this paper is demonstrating that editing is fundamentally a consistency problem: localized changes must preserve global coherence, and edits must remain stable under viewpoint changes.

This aligns with recent instruction-based editing methods. InstructPix2Pix enables instruction-following image editing [5]. Instruct-NeRF2NeRF extends instruction-based editing to 3D scenes by iteratively editing input images and re-optimizing the NeRF [10]. However, NeRF editing can be costly; Gaussian-based representations and geometry-consistent priors, as in Inter-RoMaP [14], are particularly suitable for interactive pipelines.

9.2 Text-to-3D and the role of constraints

Text-to-3D methods such as DreamFusion demonstrate how 2D diffusion priors can guide 3D optimization [21]. Yet these approaches often struggle with consistency and controllability when used interactively. This reinforces our central claim: persistent representations and explicit constraints are crucial for robust interactive generation and editing.

10 Design Principles for Persistent Generative Systems

We distill practical design principles, each supported by the synthesis above.

10.1 Principle 1: Separate state from appearance

Treat stable aspects of the world as state variables (geometry, identity, physics) rather than repeatedly inferred from pixels. Temporal-ID [25] exemplifies this for identity memory; FaceSplat [12] exemplifies this for geometry priors. DreamWM [15] shows state evolution as a foundation for long-horizon generation.

10.2 Principle 2: Prefer explicit representations for interaction

Explicit geometry (e.g., Gaussian splats) enables localized edits and real-time rendering [16]. Inter-RoMaP [14] leverages this to support interactive localized editing with geometry-consistent attention. In contrast, purely implicit representations can be powerful but may be less practical for low-latency interaction [18].

10.3 Principle 3: Make controllability interpretable

Use physically meaningful controls (e.g., constraints, motion primitives, dynamics parameters) rather than only text prompts. VACE-PhysicsRL [27] demonstrates how physical laws and RL alignment yield more reliable control.

10.4 Principle 4: Context is multi-source and persistent

Museum AR [26] demonstrates that context includes environment semantics, user viewpoint, and interaction history. Persistent systems should explicitly model $\mathbf{M}_t$ and support retrieval from memory, not rely on short context windows.

10.5 Principle 5: Evaluation must stress persistence

Standard short-clip metrics may miss failure modes that appear only after long horizons or multiple edits. Persistent systems require benchmarks for identity drift, view consistency, constraint satisfaction, and interactive stability. This aligns with broader discussions of evaluating video generation as world simulation [20].

11 Open Problems

Scalable memory. Long-horizon generation requires memory mechanisms that scale with time and scene complexity. Memory banks (Temporal-ID [25]) are a step, but future systems must handle large environments and multiple entities.

Robust geometry in the wild. Geometry acquisition remains challenging under occlusion, dynamic scenes, and limited sensors. Combining SLAM [6] with learning-based motion estimation [28] and reconstruction pipelines [23] is promising, but robust integration remains open.

Balancing constraints and creativity. Physics and geometry constraints can limit creative outputs if over-enforced. Designing soft constraints and adaptive control remains a key research direction.

Real-time deployment. Museum AR [26] emphasizes real-time constraints. Persistent world models must be architected for latency, energy, and robustness on wearable devices, requiring co-design across model architecture and systems engineering.

12 Conclusion

This paper argued that persistent world modeling is the missing ingredient for reliable long-horizon, interactive, and physically grounded generative media systems. We synthesized six complementary repository works and showed how each contributes to a coherent architecture: Huang et al. (FaceSplat) [12] provide geometry priors for stable identity reconstruction; Song et al. (Temporal-ID) [25] provide adaptive memory banks for identity preservation; Kang et al. (DreamWM) [15] provide world-model-guided state evolution for narrative generation; Song et al. (Museum AR) [26] provide a real-time, context-aware deployment perspective; Kang et al. (VACE-PhysicsRL) [27] provide physics- and RL-aligned controllable generation; and Kang et al. (Inter-RoMaP) [14] provide interactive, localized, geometry-consistent editing.

Collectively, these works support a unified thesis: making geometry, identity, physics, and interaction history explicit transforms generative models from offline content creators into deployable, controllable, and persistent 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.