Identity as Memory: Persistent Subject Modeling for Long-Form Video Generation

Marcus Elvain, Sofia Karnel

2025-12-30

Abstract

Long-form video generation remains fundamentally limited by identity drift, where subjects gradually lose appearance, geometry, or semantic consistency over time. While recent video diffusion models achieve strong short-term realism, they lack mechanisms for persistent identity modeling. In this paper, we argue that identity should be treated as a first-class temporal signal, analogous to memory in sequential decision-making systems. We analyze the causes of identity degradation, review memory-, geometry-, and alignment-based solutions, and propose a unifying perspective that frames identity preservation as a systems-level problem spanning representation, optimization, and control.

1 Introduction

Recent advances in diffusion-based video generation have enabled impressive visual fidelity and prompt adherence [1, 3, 5, 8, 15]. Large-scale models such as Sora demonstrate emergent physical reasoning and long-range motion synthesis [4]. However, long-form generation remains brittle: characters drift in facial structure, clothing mutates, and identities collapse across shots [20, 21].

These failures are not merely aesthetic artifacts; they indicate a structural deficiency in current generative pipelines. Identity is implicitly encoded in appearance features and local attention, rather than explicitly modeled as a persistent state. In contrast, research on world models and sequential decision-making shows that stable long-horizon behavior requires explicit latent memory [6, 7].

Motivated by this gap, we argue that identity preservation should be reframed as a memory modeling problem. We synthesize recent progress in adaptive memory [16], geometry-grounded generation [10], and alignment-based optimization [13], and propose a unified view of identity as a persistent temporal variable.

2 Failure Modes of Identity Drift

Identity drift manifests in several recurring failure modes. First, *appearance drift* causes gradual changes in facial features, texture, or clothing over time. Second, *structural drift* alters body proportions or spatial relationships, particularly under large pose changes. Third, *semantic collapse* occurs when multiple subjects merge or swap identities in multi-entity scenes [3, 12].

These issues are amplified by long temporal horizons, dynamic motion, and camera changes [21]. Existing video diffusion models rely on short-range temporal attention and frame-wise conditioning, which lack the capacity to enforce global identity constraints across extended sequences.

3 Identity as Persistent State

World models offer a useful conceptual analogy. In sequential environments, latent state variables summarize past observations to support long-horizon prediction and planning [6, 7]. Applying this paradigm to video generation suggests that identity should be represented as a slowly evolving latent variable that constrains per-frame synthesis.

Recent work explicitly adopts this principle. Temporal-ID introduces adaptive memory banks that store identity representations and re-inject them during generation, significantly improving long-form consistency [16]. These approaches decouple identity from instantaneous appearance, enabling robustness to pose, motion, and viewpoint variation.

4 Geometry as an Identity Anchor

Memory alone is insufficient if identity representations lack structural grounding. Geometry provides a complementary anchor by stabilizing shape and spatial relationships across time. Neural Radiance Fields and 3D Gaussian Splatting offer explicit, differentiable representations with strong multi-view consistency [11, 14].

Geometry-grounded video pipelines leverage this property to reduce identity ambiguity. DreamWM integrates world-model latent state with 3D scene representations, enabling consistent subject identity across narrative video segments [10]. Interactive editing frameworks further demonstrate that geometry-consistent attention constrains identity during localized edits [9].

These results suggest that identity is jointly supported by memory and geometry: memory preserves semantic continuity, while geometry stabilizes physical structure.

5 Optimization and Alignment

Beyond representation, optimization objectives play a critical role. Reinforcement-learning-based alignment methods introduce identity-specific reward models, directly optimizing generation policies for identity consistency [13]. This is particularly important in multi-human scenarios where appearance similarity alone is insufficient.

Physics-aware constraints further reduce identity drift under dynamic motion by enforcing physically plausible trajectories and interactions [18]. Together, these methods highlight that identity preservation requires coordinated advances in representation, training objectives, and control mechanisms.

6 Applications

Robust identity modeling enables several key applications:

- **Long-Form Narrative Video**: Persistent characters across scenes, viewpoints, and narrative arcs [10].
- **Multi-Human Interaction**: Stable identities in crowded or interactive settings [13].
- **Augmented Reality**: Identity-consistent avatars under real-time constraints [2, 17].

7 Discussion and Open Problems

Despite progress, open challenges remain. Identity representations must balance stability and adaptability; overly rigid constraints may limit generative diversity. Additionally, identity is multi-scale, encompassing geometry, texture, motion style, and semantics. Future work should explore hierarchical identity models that separate core attributes from transient appearance.

Integrating identity-aware memory into large-scale video foundation models remains an open research direction [19]. Such integration is essential for scalable, long-horizon generative systems.

8 Conclusion

We presented a systems-level perspective on identity preservation in long-form video generation. By framing identity as a persistent memory signal supported by geometry and alignment, we argue for a principled path toward more stable, interpretable, and controllable generative video 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] Danijar Hafner et al. Learning latent dynamics for planning from pixels. In *ICML*, 2019.

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

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

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

[10] 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.

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

[12] Willi Menapace et al. Snap: Spatio-temporal attention for video diffusion. *arXiv preprint arXiv:2401.06714*, 2024.

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

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

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

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

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

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

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

[20] Wenhai Wang et al. Internvid: A large-scale video dataset for video foundation models. *arXiv preprint arXiv:2307.06942*, 2023.

[21] Xinyu Yang et al. Cogvideox: Text-to-video diffusion models with 3d vae. *arXiv preprint arXiv:2403.09337*, 2024.