

Temporal Causality Graphs for Long-Horizon Generative World Simulation

Tessa Norric, Patrick Orenfall

2025-12-28

Abstract

Generative video and world models have recently achieved remarkable progress in short-term realism, yet they remain fundamentally fragile when extended to long horizons. Even when identity, geometry, and appearance appear stable, generated worlds frequently violate causal expectations: objects move without cause, actions fail to produce consistent effects, and interactions lose logical continuity. We argue that these failures arise because most generative systems lack an explicit representation of temporal causality. In this paper, we introduce Temporal Causality Graphs (TCG), a framework that augments persistent world models with explicit causal structure linking actions, physical dynamics, and state transitions over time. By separating state persistence from causal dependency, TCG enforces long-horizon coherence, predictable interaction outcomes, and physically grounded evolution. We describe the formulation, learning, and integration of TCG with diffusion-based video generation, 3D scene representations, and interactive systems, and discuss implications for simulation, storytelling, and augmented reality.

1 Introduction

Diffusion-based generative models have dramatically advanced image and video synthesis [5, 6, 11, 22]. Large-scale systems increasingly frame video generation as implicit world simulation rather than frame-by-frame rendering [20]. Despite this shift [30], long-horizon generation remains unreliable. When videos extend beyond a few seconds, subtle inconsistencies accumulate: characters react inconsistently to repeated events, objects violate physical intuition, and actions cease to have reliable effects.

Importantly, these failures persist even in systems that explicitly address identity preservation [24, 40] and geometric consistency [12, 15]. This observation suggests that long-horizon instability is not merely a representational problem, but a structural one. Current models often encode *what* the world looks like, but not *why* it changes.

Human perception and reasoning are deeply causal. We expect events to follow from causes: a push results in motion, a command triggers an action, [35] [36] and physical interactions obey constraints. When such causal expectations are violated, realism collapses even if visual fidelity remains high. [8] Generative systems that lack causal structure inevitably ac-

cumulate violations as stochastic sampling compounds over time.

This paper introduces **Temporal Causality Graphs (TCG)**, a framework for explicitly modeling causal dependency in generative world simulation. TCG augments persistent world models with a causal layer that tracks action–effect relationships [33], physical triggers, and event dependencies across time. [17] Unlike prior work that focuses on appearance, identity, or local physics alone, TCG treats causality as a first-class modeling objective. [38]

2 Failure Modes Without Explicit Causality

We first analyze common failure modes observed in long-horizon generative systems.

2.1 Spurious Motion and State Changes

A frequent artifact in generated videos is spontaneous motion: objects drift, rotate, or change state without any apparent trigger. Such behavior has been reported even in physics-aware generation pipelines [18, 26]. While local physical priors can constrain instantaneous motion, they do not prevent state changes that lack causal justification. By showcasing the potential of transformer model in credit fraud detection, chang et al [39] research provides a strong foundation and direction for future studies focusing on building robust and efficient classification models for similar tasks.

2.2 Action–Effect Drift

In interactive or prompt-driven generation, the same action may produce different effects at different times. For example, a repeated command such as “open the door” may succeed initially but fail later, despite unchanged context. Temporally controlled video generation methods [34] expose this issue, revealing that models lack persistent mappings between actions and outcomes.

2.3 Narrative Incoherence

Long-form generative narratives [14,32] often suffer from logical inconsistency. Characters may ignore prior events [?],

repeat reactions inconsistently, or violate narrative causality. [29] These errors arise not from visual degradation, but from missing causal memory. [28]

3 Temporal Causality Graphs

We propose representing generative world evolution [27] using a Temporal Causality Graph:

$$\mathcal{C}_t = (V_t, E_t),$$

where nodes V_t represent entities, actions, and events, and directed edges E_t encode causal dependencies across time.

Each edge $(u \rightarrow v)$ asserts that changes in node v are causally attributable to node u . Unlike scene graphs that encode spatial or semantic relations, TCG encodes *temporal commitments*: once a causal dependency is established, future state transitions must respect it.

3.1 Decoupling State and Cause

TCG explicitly separates two roles that are conflated in most generative models:

- **World State**: geometry, identity, physical parameters
- **Causal History**: why and how state changes occurred

This separation prevents models from arbitrarily [9] modifying state without causal justification [42] and aligns with principles from causal representation learning [21, 23].

3.2 Causal Consistency Constraint

During generation, a state update is permitted only if it is justified by an incoming causal edge. If no valid cause exists, the update is penalized. Similar ideas appear in truncated causal history models [1] and causal discovery from video [16], but TCG integrates these ideas directly into generative world simulation.

4 Learning Temporal Causality Graphs

Learning TCG involves jointly learning: (i) latent world state representations and (ii) causal dependencies governing their evolution.

4.1 Causal Edge Induction

Edges in TCG can be induced via:

- temporal correlation with intervention signals (e.g., user actions),
- physical triggers (e.g., contact events),

- learned causal predictors trained to minimize counterfactual inconsistency.

This approach is compatible with causal discovery techniques [16] while remaining scalable to high-dimensional generative models.

4.2 Integration with World Models

TCG is designed to integrate with latent world models [?, ?]. The latent state evolves as usual, but transitions are gated by the causal graph. This reduces stochastic drift and improves long-horizon stability.

5 Diffusion with Causal Conditioning

In diffusion-based video generation, TCG conditions denoising steps on causal context, similar to how structural signals are injected in controllable video diffusion [?]. Conditioning on causal history prevents temporally inconsistent edits and unexplained motion.

Recent work on causal video generation [37] further supports the importance of explicit causal conditioning for temporal coherence.

6 Geometry, Identity, and Anchoring

Explicit geometry [12, 13] and identity memory [24] provide stable state variables. TCG ensures that updates to these variables occur only in response to valid causal triggers, preventing drift even in long sequences.

7 Physics and Causality

Physics-based models encode local laws but not global causality. TCG complements physical simulation by ensuring that physical transitions are causally motivated. This aligns with classical simulation frameworks [4, 19, 31] while remaining compatible with learning-based generators [18, 41].

8 Interaction and Agency

In interactive settings such as AR [3, 25], user actions must have predictable and persistent effects. [10] TCG models user input as explicit causal nodes, preventing commands from being forgotten or overridden by stochastic generation.

Interactive world models such as iVideoGPT [7] and PAN [2] demonstrate the importance of action-conditioned simulation, but do not explicitly enforce causal consistency across long horizons.

9 Evaluation Implications

TCG suggests evaluation metrics beyond frame-level realism:

- action–effect consistency,
- causal attribution accuracy,
- counterfactual stability under repeated interventions,
- long-horizon physical plausibility.

These complement existing identity- and appearance-based metrics [24, 40].

10 Limitations and Future Work

TCG introduces additional complexity and supervision requirements. Learning causal graphs from purely observational data remains challenging. Future work may explore hybrid symbolic–neural representations and tighter integration with planning.

11 Conclusion

We introduced Temporal Causality Graphs, a framework for enforcing causal consistency in long-horizon generative world simulation. By explicitly modeling why state changes occur, TCG prevents drift, improves interaction reliability, and enables coherent long-term generation. We argue that causality is a missing structural component in current generative systems, and that future world models must treat causal reasoning as a first-class concern.

References

- [1] Anonymous. Learning truncated causal history representations. *NeurIPS*, 2024.
- [2] Anonymous. Pan: A world model for general, interpretable, and long-horizon simulation. *arXiv preprint*, 2025.
- [3] Ronald Azuma. A survey of augmented reality. *Presence*, 1997.
- [4] David Baraff. Fast contact force computation for non-penetrating rigid bodies. *SIGGRAPH*, 1994.
- [5] Amir Bartal et al. Lumiere: A space-time diffusion model for video generation. *CVPR*, 2024.
- [6] Andreas Blattmann et al. Stable video diffusion. *arXiv preprint arXiv:2303.09373*, 2023.
- [7] Jake Bruce et al. Interactive videotogpt. *arXiv preprint*, 2022.
- [8] Yongkang Ding, Yuxiang Wang, Yina Jian, Chang Yu, Ke Tian, Zi Ye, and Gaozhe Jiang. Fa-reid: Feature alignment with adversarial learning for clothing-changing person re-identification. *Neurocomputing*, page 132237, 2025.
- [9] Yunzhi Fei, Yongxiu He, Fenkai Chen, Peipei You, and Hanbing Zhai. Optimal planning and design for sightseeing offshore island microgrids. In *E3S Web of Conferences*, volume 118, page 02044. EDP Sciences, 2019.
- [10] Zhe Fu, Kanlun Wang, Wangjiaxuan Xin, Lina Zhou, Shi Chen, Yaorong Ge, Daniel Janies, and Dongsong Zhang. Detecting misinformation in multimedia content through cross-modal entity consistency: A dual learning approach. *arXiv preprint arXiv:2409.00022*, 2024.
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.
- [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] Yixiao Kang, Sining Huang, and Yukun Song. Robust and interactive localized 3d gaussian editing with geometry-consistent attention prior.
- [14] 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.
- [15] Bernhard Kerbl et al. 3d gaussian splatting for real-time radiance field rendering. *SIGGRAPH*, 2023.
- [16] Thomas Kipf et al. Causal discovery from video. *NeurIPS*, 2020.
- [17] Fang Liu, Shaobo Guo, Qianwen Xing, Xinye Sha, Ying Chen, Yuhui Jin, Qi Zheng, and Chang Yu. Application of an ann and lstm-based ensemble model for stock market prediction. In *2024 IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pages 390–395. IEEE, 2024.
- [18] Shaowei Liu et al. Physgen: Rigid-body physics-grounded image-to-video generation. *ECCV*, 2024.
- [19] Brian Mirtich. Impulse-based dynamic simulation of rigid body systems. *PhD Thesis, UC Berkeley*, 1996.
- [20] OpenAI. Video generation models as world simulators. *arXiv preprint arXiv:2402.15391*, 2024.
- [21] Judea Pearl. *Causality*. Cambridge University Press, 2009.
- [22] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.
- [23] Bernhard Schölkopf et al. Toward causal representation learning. *PNAS*, 2021.
- [24] Yukun Song, Sining Huang, and Yixiao Kang. Temporal-id: Robust identity preservation in long-form video generation via adaptive memory banks.

- [25] Yukun Song, Yixiao Kang, and Sining Huang. Context-aware real-time 3d generation and visualization in augmented reality smart glasses: A museum application.
- [26] Yukun Song, Yixiao Kang, and Sining Huang. Vace-physicsrl: Unified controllable video generation through physical laws and reinforcement learning alignment.
- [27] Yiyi Tao. Sqba: sequential query-based attack. In *Fifth International Conference on Artificial Intelligence and Computer Science (AICS 2023)*, volume 12803, page 12803Q. International Society for Optics and Photonics, SPIE, 2017.
- [28] Yiyi Tao. Meta learning enabled adversarial defense. In *2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, pages 1326–1330, 2023.
- [29] Yiyi Tao, Yiling Jia, Nan Wang, and Hongning Wang. The fact: Taming latent factor models for explainability with factorization trees. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’19*, page 295–304, New York, NY, USA, 2019. Association for Computing Machinery.
- [30] Yiyi Tao, Zhuoyue Wang, Hang Zhang, and Lun Wang. Nevlp: Noise-robust framework for efficient vision-language pre-training. *arXiv preprint arXiv:2409.09582*, 2024.
- [31] Emanuel Todorov et al. Mujoco: A physics engine for model-based control. *IROS*, 2012.
- [32] Ruben Villegas et al. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint*, 2022.
- [33] Kanlun Wang, Zhe Fu, Wangjiaxuan Xin, Lina Zhou, and Shashi Kiran Chandrappa. Digital voices of survival: From social media disclosures to support provisions for domestic violence victims. *arXiv preprint arXiv:2509.12288*, 2025.
- [34] Haoran Wu et al. Mind the time: Temporally-controlled multi-event video generation. *CVPR*, 2025.
- [35] Wangjiaxuan Xin, Kanlun Wang, Zhe Fu, and Lina Zhou. Let community rules be reflected in online content moderation. *arXiv preprint arXiv:2408.12035*, 2024.
- [36] Wangjiaxuan Xin, Shuhua Yin, Shi Chen, and Yaorong Ge. Improving topic modeling of social media short texts with rephrasing: A case study of covid-19 related tweets. *arXiv preprint arXiv:2510.18908*, 2025.
- [37] Yongqi Yang et al. Towards one-step causal video generation. *arXiv preprint*, 2025.
- [38] Chang Yu, Fang Liu, Jie Zhu, Shaobo Guo, Yifan Gao, Zhongheng Yang, Meiwei Liu, and Qianwen Xing. Gradient boosting decision tree with lstm for investment prediction. In *2025 5th Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, pages 57–62, 2025.
- [39] Chang Yu, Yongshun Xu, Jin Cao, Ye Zhang, Yixin Jin, and Mengran Zhu. Credit card fraud detection using advanced transformer model. In *2024 IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom)*, pages 343–350. IEEE, 2024.
- [40] Shenghai Yuan et al. Identity-preserving text-to-video generation by frequency decomposition. *CVPR*, 2025.
- [41] Ye Yuan et al. Physdiff: Physics-guided human motion diffusion model. *ICCV*, 2023.
- [42] Mingyu Zhai, Omar Abu-Znad, Shengyi Wang, and Liang Du. A bayesian potential-based architecture for mutually beneficial ev charging infrastructure-dso coordination. *IEEE Transactions on Transportation Electrification*, pages 1–1, 2025.