Counterfactual World Models: Evaluating and Training Generative Systems via Intervention Consistency

Irene Solvane

2025-11-18

Abstract

Recent generative world models demonstrate impressive visual realism and short-term temporal coherence, yet they remain fragile under intervention. Actions applied at different times yield inconsistent outcomes, removing events does not undo their downstream effects, and repeated interactions accumulate unpredictable drift. We argue that these failures arise because most generative systems lack counterfactual reasoning. In this paper, we introduce Counterfactual World Models (CWM), a framework for evaluating and training generative systems based on their consistency under explicit counterfactual interventions. CWM formalizes intervention operators over persistent world state, defines counterfactual consistency objectives, and proposes evaluation protocols that test causal correctness rather than surface-level realism. We show how CWM complements persistent, editable, and causal world models, and argue that counterfactual correctness is a necessary condition for reliable long-horizon generative simulation.

1 Introduction

Diffusion-based generative models have fundamentally reshaped image and video synthesis, enabling high-fidelity outputs at unprecedented scale [10, 19]. Recent extensions to video generation further improve temporal continuity [2, 4], and large-scale systems increasingly describe video generators as implicit world simulators rather than frame-wise predictors [16].

In parallel, substantial progress has been made toward stabilizing individual aspects of generative worlds, including identity persistence [22], geometric consistency [11, 13], localized editing [12], physical plausibility [15, 24], and real-time interaction in augmented reality settings [23]. Despite these advances, generative world models remain unreliable when subjected to intervention.

If an action is applied earlier, later, or removed entirely, generated outcomes often violate causal expectations. These failures persist even in long-horizon models with memory [5, 26] and interaction support [17]. We argue that the root cause is structural: most generative systems learn correlations, not counterfactual structure.

2 Limitations of Correlational Generative Models

2.1 Correlation Without Commitment

Standard generative training encourages models to reproduce observed patterns, but does not require them to commit to underlying causal mechanisms. As a result, visually plausible trajectories may violate causal semantics when conditions change. This issue has been observed in long-form video generation [26] and temporally controlled synthesis.

2.2 Intervention Fragility

A defining failure mode of current world models is inconsistent action–effect mapping. Repeating an identical intervention often produces divergent outcomes, indicating that the model does not encode persistent causal relationships. Even physics-aware diffusion models [15, 28] cannot guarantee that removing a physical interaction removes its downstream effects.

3 Counterfactual World Models

We define a generative world model with latent state $\mathcal{S}_t$. A counterfactual world model must support interventions of the form:

$$\mathrm{do}(X = x),$$

where X may correspond to an action, a physical parameter, or a component of the world state [18].

3.1 Intervention Semantics

Given an intervention applied at time t_0, the model must generate a counterfactual trajectory $\mathcal{S}^{(cf)}_{t>t_0}$ such that:

- only variables causally downstream of X are affected,
- all unrelated aspects of the world remain unchanged.

This definition follows classical causal intervention semantics [18] but is adapted to high-dimensional generative simulation.

3.2 Counterfactual Consistency

We define counterfactual consistency as the degree to which observed and counterfactual trajectories diverge only where causally justified:

$$\mathcal{L}_{cf} = \sum_t d\big(\mathcal{S}_t^{(obs)}, \mathcal{S}_t^{(cf)}\big),$$

where $d(\cdot)$ penalizes divergence outside affected subspaces. Similar ideas have been explored in causal representation learning [20] and counterfactual analysis of generative models [3].

4 Training with Counterfactual Objectives

CWM augments standard generative training with counterfactual objectives that explicitly test intervention robustness.

4.1 Synthetic Interventions

Counterfactual supervision can be introduced by:

- removing, repeating, or reordering actions,
- modifying physical parameters or contact events,
- altering interaction inputs in AR or embodied settings.

Such interventions are naturally supported by editable world representations [12] and physics-aligned generation pipelines [15, 24].

4.2 Compatibility with Diffusion Models

In diffusion-based generators [19], counterfactual losses can be applied at the latent level, enforcing consistency across denoising trajectories conditioned on different intervention histories. Recent work on causal video generation [27] supports the necessity of explicit causal constraints for stable long-horizon synthesis.

5 Evaluation Protocols

CWM proposes evaluation beyond perceptual metrics such as FID. We introduce three complementary tests:

- **Repeatability**: identical interventions yield identical outcomes.
- **Reversibility**: removing an intervention removes its effects.
- **Temporal Shift**: applying an intervention earlier or later produces causally shifted outcomes.

These tests expose failure modes invisible to appearance-based metrics and complement identity-focused evaluations [22].

6 Relation to World Models and Planning

World models are central to planning and decision-making [7–9]. Planning fundamentally relies on counterfactual rollouts—evaluating hypothetical futures without executing actions. Classical systems such as MuZero demonstrate the power of counterfactual planning even without explicit supervision [21].

Recent long-horizon interactive world models [5, 17] demonstrate impressive scalability, yet still lack guarantees under counterfactual intervention. CWM provides a complementary training and evaluation layer that enforces such guarantees.

7 Physics, Geometry, and Persistent State

Explicit geometry and physical structure constrain the space of valid counterfactuals. Geometry-aware representations [11, 13] and localized editing mechanisms [12] allow interventions to be spatially precise. Classical physics formulations [1, 25] further motivate reversibility and action–effect consistency as fundamental correctness criteria.

8 Discussion

CWM reframes generative modeling as hypothesis testing rather than passive synthesis. A system that cannot answer "what would have happened if this action had not occurred?" lacks a meaningful internal model of the world. Counterfactual evaluation reveals structural deficiencies that remain hidden under standard benchmarks.

9 Limitations and Future Work

Counterfactual supervision introduces additional complexity and may require carefully designed interventions. Future work may integrate causal discovery from video [14], modular dynamics [6], and symbolic abstractions to scale counterfactual reasoning to open-world settings.

10 Conclusion

We introduced Counterfactual World Models, a framework for evaluating and training generative systems via intervention consistency. By requiring models to behave correctly under counterfactual changes, CWM enforces a stronger notion of world understanding than appearance-based realism. We argue that counterfactual correctness is a necessary criterion for trustworthy generative world models.

References

[1] David Baraff. Fast contact force computation for non-penetrating rigid bodies. *SIGGRAPH*, 1994.

[2] Amir Bartal et al. Lumiere: A space-time diffusion model for video generation. *CVPR*, 2024.

[3] Michel Besserve, Ruoyu Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *ICLR*, 2024.

[4] Andreas Blattmann et al. Stable video diffusion. *arXiv preprint arXiv:2303.09373*, 2023.

[5] Jake Bruce et al. Interactive videogpt. *arXiv preprint*, 2022.

[6] Anirudh Goyal et al. Recurrent independent mechanisms. *ICLR*, 2021.

[7] David Ha and Jürgen Schmidhuber. World models. *NeurIPS*, 2018.

[8] Danijar Hafner et al. Learning latent dynamics for planning from pixels. *ICML*, 2019.

[9] Danijar Hafner et al. Mastering atari with discrete world models. *ICLR*, 2021.

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

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

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

[13] Bernhard Kerbl et al. 3d gaussian splatting for real-time radiance field rendering. *SIGGRAPH*, 2023.

[14] Thomas Kipf et al. Causal discovery from video. *NeurIPS*, 2020.

[15] Shaowei Liu et al. Physgen: Rigid-body physics-grounded image-to-video generation. *ECCV*, 2024.

[16] OpenAI. Video generation models as world simulators. *arXiv preprint arXiv:2402.15391*, 2024.

[17] Jiahui Pan et al. Pan: A world model for general, interactable, and long-horizon simulation. *arXiv preprint arXiv:2511.09057*, 2025.

[18] Judea Pearl. *Causality*. Cambridge University Press, 2009.

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

[20] Bernhard Schölkopf et al. Toward causal representation learning. *PNAS*, 2021.

[21] Julian Schrittwieser et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 2020.

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

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

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

[25] Emanuel Todorov et al. Mujoco: A physics engine for model-based control. *IROS*, 2012.

[26] Ruben Villegas et al. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint*, 2022.

[27] Yongqi Yang et al. Towards one-step causal video generation. *arXiv preprint arXiv:2511.01419*, 2025.

[28] Ye Yuan et al. Physdiff: Physics-guided human motion diffusion model. *ICCV*, 2023.