POSITION: NOT EVERY ACTION IS AN INTERVENTION – TOWARD CAUSALLY GROUNDED WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World models (WMs) have become increasingly capable of predicting and generating how environments evolve, yet predictive accuracy alone does not guarantee reliable reasoning about how the world responds when its variables or mechanisms are deliberately changed. We argue that addressing this gap requires treating causal interventions as a first-class consideration in world modeling. Importantly, actions and interventions should not be conflated: actions can have valid interventional semantics under appropriate assumptions, but the intervention space relevant to understanding a world may differ from and extend beyond the agent's action space. We propose viewing next-generation WMs as *intervention engines*: systems that represent where interventions enter a causal system, how their effects propagate through its mechanisms, and how informative interventions can reduce uncertainty about those mechanisms. This perspective shifts the emphasis from predicting trajectories under observed conditions toward reasoning about the consequences of deliberate changes. We develop a conceptual framework and identify open problems spanning causal representation learning, intervention-driven exploration, mechanism uncertainty, counterfactual reasoning, and interventional evaluation. Our goal is not to prescribe a particular causal architecture, but to argue for a broader formulation of world modeling centered on meaningful intervention spaces. Such a formulation provides a path toward WMs that combine predictive capability with more reliable reasoning under interventions, with implications for planning, embodied AI, and scientific discovery.

1 INTRODUCTION

World models (WMs) have become increasingly prominent in modern AI, particularly in model-based reinforcement learning frameworks such as Dreamer (Hafner et al., 2020a; 2025), as well as in emerging generative and interactive systems including video models like Sora (Brooks et al., 2024; OpenAI, 2024) and simulation approaches such as Genie (Bruce et al., 2024). These systems can generate coherent trajectories, capture temporal structure, and support increasingly sophisticated interactions. Yet current WMs often struggle when the conditions governing the world change, including under interventions, counterfactual queries, and distribution shifts. This raises a fundamental question: **does a model that accurately predicts how the world evolves also understand how it responds when deliberately changed?** WMs serve two intertwined functions (Ding et al., 2025). First, they support *prediction*: forecasting what will happen next for planning and control. Second, they support *understanding*: reasoning about why outcomes occur and how they would change under alternative conditions. Accurate prediction alone does not guarantee the ability to answer questions such as: what happens if I intervene, which mechanisms produced the resulting change, and what would have happened otherwise? Moving from prediction toward understanding therefore requires **reasoning about how the world responds to deliberate changes.**

Existing WMs exhibit important limitations outside their training conditions. Although Sora and related text-to-video systems generate physically plausible videos, they can fail out-of-distribution physics tests such as T2VPhysBench (Guo et al., 2025). Model-based policies can likewise degrade when environmental dynamics are perturbed (Pitis et al., 2020; Hansen et al., 2023), while LLM-based WMs face analogous challenges in causal reasoning (Bender et al., 2021; Jin et al., 2023; Pearl & Mackenzie, 2019; Kiciman et al., 2023; Zhou et al., 2025). Such failures do not by themselves establish an absence of causal knowledge, as support mismatch, insufficient coverage, or accumulated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

prediction error can produce similar behavior. Rather, they highlight that predictive success alone does not establish whether a model will respond correctly when variables or mechanisms are deliberately changed. These explanations may be nearly indistinguishable under previously observed trajectories while implying very different responses to future actions. More generally, distinct causal mechanisms can induce the same or similar observational distributions, making mechanism identification impossible without additional assumptions, inductive biases, or informative environmental variation (Pearl, 2009; Schölkopf et al., 2021).

Scaling can improve predictive accuracy and may allow useful causal structure to emerge implicitly. However, predictive capacity alone does not resolve causal identification. Multiple causal mechanisms can be compatible with the same observational distribution while implying different responses to interventions (Pearl, 2009; Schölkopf et al., 2021; Guo et al., 2020). The causal challenge is therefore not merely one of scale, but also of what changes to the world the model is expected to reason about.

Our Position. We argue that next-generation WMs need **explicit** causal interventions. This requires carefully distinguishing *actions* from *interventions*. In a fully observed and causally sufficient controlled system, an action that directly sets a control variable may have valid interventional semantics. Thus, action-conditioned prediction is not inherently restricted to Level 1 of Pearl's causal hierarchy (Pearl, 2009; Pearl & Mackenzie, 2019).

However, *not every action is an intervention* in the causal representation of interest. Consider a robot instructed to "push a ball to the right." The command does not directly set the ball's velocity; it acts through the robot's control interface, contact forces, and physical dynamics. Treating the action as a direct intervention on an arbitrary downstream latent variable can therefore assign it incorrect causal semantics. What matters is where the action enters the causal system and how its effects propagate.

Conversely, *not every intervention is an action*. Understanding the same system may require asking what happens if the ball becomes heavier, the surface becomes more slippery, gravity changes, or the actuator becomes weaker. These interventions modify variables or mechanisms of the world and may lie outside the agent's ordinary action space. The intervention space relevant to a WM can therefore be both different from and broader than its action space: actions characterize what an agent can do, whereas interventions characterize how the world can be deliberately changed.

We therefore argue that causal WMs should explicitly consider the intervention space over which they are expected to reason. They should represent where interventions enter the causal system, how their effects propagate through otherwise invariant mechanisms, and how uncertainty can be resolved through informative interventions. Level 3 counterfactual reasoning further asks what would have happened in the same realized situation under an alternative intervention and requires additional structural assumptions (Pearl, 2009; Pearl & Mackenzie, 2019). Importantly, we do not require an explicit causal graph or claim that existing WMs cannot learn causal structure implicitly. The central criterion is whether a WM supports reliable reasoning over relevant interventions, rather than whether it contains a particular causal architecture.

Contribution. This paper makes three contributions. (1) We clarify the relationship between actions and causal interventions in WMs, arguing that actions can have interventional semantics but need not coincide with the broader intervention space required for understanding a world. (2) We develop a constructive framework for intervention-centric WMs that incorporates causal variables, intervention targets, mechanism propagation, and uncertainty over causal mechanisms. (3) We identify major open problems including causal representation learning, intervention-driven training, counterfactual reasoning, interventional evaluation, and benchmark design.

2 Actions vs. Causal Interventions

Actions, interventions, and conditioning can appear superficially similar in a predictive model, but they have different causal semantics. Conditioning restricts what is observed, an action specifies a decision made through an agent's control interface, and an intervention modifies a variable or mechanism in the underlying causal system. Clarifying these distinctions is important because action-conditioned prediction can, under appropriate assumptions, already have an interventional interpretation, while the intervention space relevant to a WM may extend beyond its action space.

We describe the world as a structural causal model (SCM) $\mathcal{M}$ over the latent state $z_t = (z_t^1, \ldots, z_t^n)$, where each component is generated by a structural mechanism, $z_{t+1}^i = f_i\big(z^{\mathrm{Pa}_i}, \epsilon_{t+1}^i\big)$, where f_i is

the structural function governing z^i, Pa_i denotes its causal parents, z^{Pa_i} their values, and ϵ^i_{t+1} is an exogenous noise variable capturing influences not represented by those parents.

Action. An action a_t is a variable through which an agent interacts with the world. Its value is selected according to a policy, $a_t := \pi(\cdot, \epsilon^a_t)$, based on the information available to the agent. Its children are the variables through which the action enters the environment—for example, actuator or control variables in a robotic system. The action space $\mathcal{A}$ specifies the controls available to the agent.

Intervention. An intervention $I_t \in \mathcal{I}$ is an operation that maps the original causal model $\mathcal{M}$ to an intervened model $\mathcal{M}^{I_t}$ by modifying the mechanism of its target while leaving non-targeted mechanisms unchanged (Pearl, 2009; Schölkopf et al., 2021). A hard intervention may replace the mechanism of z^j_t with a specified value, $do(z^j_t := \tilde{z}^j_t)$, whereas a soft intervention modifies the mechanism itself, $f_j \to f_j^{I_t}$, while potentially retaining dependence on its original parents. For example, setting the floor friction coefficient to a specified value is naturally represented as a hard intervention, whereas weakening an actuator can be represented as a soft intervention on the mechanism mapping control commands to applied forces.

Conditioning. Conditioning, by contrast, leaves the causal model unchanged and restricts attention to cases in which a variable takes a particular value. Consequently, observing that a ball has high velocity can provide information about its ordinary causes, such as a strong applied force or a slippery surface. Under $do(v^{\mathrm{ball}}_t := v')$, however, the usual mechanism generating velocity is replaced, so the intervened value does not provide the same evidence about its ordinary causes. This distinction between observing and intervening is central to causal inference (Pearl, 2009).

When Actions are Interventions. An action-conditioned WM learns $p(z_{t+1} \mid z_t, a_t = a)$, whereas an interventional query asks for $p(z_{t+1} \mid z_t, do(a_t := a))$. These quantities coincide when the state available to the WM is sufficient to block relevant backdoor paths between the action and its outcome:

$$p(z_{t+1} \mid z_t, do(a_t := a)) = p(z_{t+1} \mid z_t, a_t = a). \tag{1}$$

This is a special case of the backdoor criterion (Pearl, 2009). In many controlled reinforcement learning (RL) settings, where actions are selected and executed by the agent from the state or history available to it, action-conditioned transitions can therefore have valid interventional semantics under appropriate causal sufficiency assumptions. The equivalence need not hold when the behavior that generated the data depends on information unavailable to the WM and that information also affects future outcomes. For example, in logged driving data, a driver may brake because of a pedestrian outside the model's field of view. Braking may then be statistically associated with a pedestrian appearing, even though braking does not cause the pedestrian to appear. In such settings, conditioning on the observed action and intervening to choose that action can imply different predictions.

Action Space vs. Intervention Space. Even when actions have valid interventional semantics, $\mathcal{A}$ need not coincide with $\mathcal{I}$. Consider a robot that pushes a ball. The action may intervene on the robot's control interface and reach the ball through actuator forces, contact dynamics, and subsequent physical mechanisms; it is not equivalent to directly setting the ball's velocity. Conversely, changing the floor friction, ball mass, or gravity may constitute meaningful interventions on the world that lie outside the robot's ordinary action space. This distinction exposes an important limitation of action data alone. A robot may learn accurately what every available push does on a particular floor, while having little information about what would happen if the floor became more slippery or the ball became heavier. Thus, a WM can correctly model the consequences of every action in $\mathcal{A}$ without being sufficient for the broader set of interventions in $\mathcal{I}$. Our position is not that action-conditioned WMs are inherently non-causal, but that next-generation WMs should reason over a meaningful intervention space that may extend beyond the actions available to the agent.

Table 1: This table summarizes representative WMs across major methodological categories and examines them from a causal intervention perspective. Specifically, we compare how each model interacts with or controls the modeled world, the space in which such control operates, and whether the model explicitly represents causal graphs or supports causal-based counterfactual prediction. Conditioning refers to providing contextual inputs (e.g., prompts, observations, or latent states) that guide prediction or generation; action refers to agent or physical actions that affect subsequent world states; and intervention refers to explicit manipulation of modeled variables or mechanisms. The perturbation space column specifies the type of variable being manipulated or conditioned on.

System	Cond. / Act. / Interv.	Graph	CF	Pertub. space
Model-based RL				
PlaNet 2019	Action	No	No	Action
Dreamer series 2020a; 2020b; 2025	Action	No	No	Action
TD-MPC2 2023	Action	No	No	Action
Self-supervised / latent				
JEPA/V-JEPA/V-JEPA 2 2022; 2024; 2025	Conditioning / action	No	No	Latent / action
Causal-JEPA 2026	Intervention	Implicit	CF-like	Latent / action
Video generation				
Sora / Wan / CogVideoX 2024; 2025; 2024a	Conditioning	No	No	Prompt
Genie 1 2024	Action	No	No	Latent
Genie 2 2024	Action	No	No	Action
Genie 3 2025	Action / environment	No	No	Action / prompt / world event
GAIA-1 / DriveDreamer 2023; 2024	Action / conditioning	No	No	Vehicle action / state / text
Physics-grounded				
Genesis 2024	Action / Intervention	Implicit (physics = SCM)	Physics simulator	Action / State
PhysGen / Cosmos 2024; 2025	Action / Intervention	Physics priors	Physics simulator	Action / State
Embodied / robotic				
DayDreamer 2023	Action	No	No	Action
Roboscape 2025	Action	No	No	Action / state
PlayWorld 2026	Action	No	No	Action
LLM-based				
LLM WMs 2024; 2024; 2024	Conditioning / hypothetical action	Usually No	Usually No	Prompt / language action and state
Social simulacra 2023; 2025; 2024b; 2023	Action / conditioning	Usually No	No	Action / prompt / state

3 From Predictive WM to Intervention Engines

A common formulation of WMs learns a latent representation z_t together with action-conditioned dynamics, $z_{t+1} = f(z_t, a_t)$. As discussed in Section 2, such dynamics can have valid interventional semantics for the agent's actions under appropriate assumptions. However, accurately modeling the consequences of actions does not necessarily identify how the world would respond to interventions on other variables or mechanisms. More generally, observational and action-conditioned data may be compatible with multiple causal mechanisms that make similar predictions under the observed conditions but disagree under interventions. Predictive accuracy over observed transitions therefore does not, by itself, guarantee reliable prediction over the broader intervention space. This gap motivates our positions below.

The central idea is that the same transition rule $f(\cdot)$ should remain valid across environments, even when the distribution of observations changes. While this view has led to important progress, we argue that it is fundamentally incomplete. A model that merely predicts the next latent state, even in a causally meaningful representation, is not yet a full causal WM. A true causal WM should answer not only "What happens next?" but also "*What can I change? Which mechanism would become active? What should I do next to distinguish competing explanations?*" We therefore propose a new view:

A causal WM is an *intervention engine*: a learned dynamical system that explicitly represents interventions, competing mechanisms, and the hierarchy of causal processes across scales.

Table 1 summarizes the intervention dimension in different types of WMs. First, no current system fully implements the do-operator (Pearl, 2009): intervention spaces are almost universally implicit. Second, causal graphs are nearly absent from learned WMs. Few exceptions also rely on limited causal structures, e.g., Genesis hard-codes a physics engine rather than learning causal structure from data. Third, counterfactual reasoning is largely missing across all categories, including the most capable video generation and embodied systems. The table reveals that current WMs have made uneven progress: strong on predictive capability and scale, but systematically weak on causal dimensions. Therefore, we propose the following position statements:

S1. Causal WM must represent explicit and meaningful intervention spaces. Learning which features are stably predictive across environments is a proxy for causality, not causality itself. A WM that only learns invariant features cannot answer the question that matters for acting agents: what happens when I intervene? This requires an explicit representation of which variables can be set externally, their valid ranges, and how effects propagate through causal graphs. Otherwise, the model cannot well distinguish the effect of an action from observed associations, and cannot generalize to different distributions. The transition from $p(z_{t+1}|z_t, a_t)$ to $p(z_{t+1}|z_t, do(a_t))$ is needed for a WM that genuinely supports decision-making. Open questions include: (1) **Which variables are actionable?** In a latent space, it is not obvious which dimensions correspond to variables an agent can physically set versus variables that are merely observable. How do we identify the actionable subset without exhaustive physical experimentation? (2) **What are the valid intervention ranges?** Physical constraints limit which values a variable can be forced to take. How should a WM represent and respect these constraints, especially when the latent space is continuous and unbounded? (3) **How does the intervention space change over time?** In non-stationary environments, which variables are actionable may itself change. How should a WM track and update its representation of the intervention space? (4) **How do we represent interventions that span multiple levels of abstraction?** A high-level action such as "pick up the object" corresponds to a complex sequence of low-level interventions on joint torques. How does the intervention space at one level of the causal hierarchy map onto the intervention space at another, and when is this mapping invertible?

S2: Causal WM must support interventional and counterfactual reasoning. Answering interventional and counterfactual queries distinguishes a WM that understands from one that merely predicts. At its most demanding, this involves counterfactual inference ("what would have happened under a different action?"), corresponding to Level 3 of Pearl's hierarchy. This capability is largely absent in current WM architectures. Forward rollout is not causal reasoning; it generates new trajectories rather than evaluating alternatives along the same one. True counterfactual reasoning requires shared exogenous noise between factual and counterfactual trajectories, ensuring differences arise solely from the intervention. More broadly, interventional and counterfactual reasoning demands that representations be causally structured from the outset, which should be a core design objective, not an emergent byproduct of predictive training. Open questions include: (1) **How do we validate counterfactual predictions when ground truth is unavailable?** By definition, the counterfactual world is unobserved, which is the fundamental challenge in causal reasoning. What proxy evaluations or experimental designs allow us to test whether a model's counterfactuals are accurate? (2) **How do we learn causal structure from heterogeneous modalities?** Causal relationships in the real world span vision, language, proprioception, and audio, often crossing modality boundaries. How do we identify causal structure over variables that do not correspond to any single modality? (3) **How do we reason causally when standard assumptions fail?** Classical causal inference assumes no hidden confounders, known intervention targets, and fully observed variables, conditions routinely violated in practice. When these assumptions hold only partially, how should a causal WM degrade gracefully, producing calibrated uncertainty over its conclusions rather than silent failures? (4) **How do we scale causal reasoning to real-world complexity?** Most causal inference methods are developed for small, well-specified graphs. Real-world WMs involve thousands of latent variables and long temporal horizons. What make counterfactual reasoning tractable at this scale?

S3. Causal WMs need to reason over multiple competing causal mechanisms with calibrated uncertainty. Most approaches treat the causal graph as a fixed structure to discover and commit to. However, in many real-world domains (e.g., multi-agent systems and non-stationary environments),

multiple causal mechanisms can govern different regimes. Fixing a single graph in such cases leads to systematic misattribution and overconfident errors, especially where robust WM are most needed. Instead, one should maintain a belief over competing causal graphs and update it with interventional evidence. Crucially, uncertainty varies across mechanisms: some edges are well-identified, others remain ambiguous. A causal WM must represent multiple mechanisms and track their uncertainty from different angles, including *aleatory uncertainty* that concerns stochastic variation in state transitions given a known graph, and *epistemic uncertainty* that concerns ignorance about graph structure itself. Open questions include: (1) **How do we represent a distribution over competing graphs tractably?** The number of causal graphs grows super-exponentially in the number of variables. What structured approximations (e.g., factored graph distributions, skeleton constraints) preserve the core benefits without combinatorial cost? (2) **How do we assign and update uncertainty per mechanism?** Different causal edges may be identified with very different confidence. What inference procedures correctly maintain and update uncertainty as new interventional evidence arrives? (3) **Can mechanisms be shared across contexts?** If two regimes share some but not all causal edges, how should the WM represent and exploit this partial overlap?

S4. The field needs benchmarks that test interventional and counterfactual correctness. Most current benchmarks, e.g., VBench (Huang et al., 2024), PhysBench (Chow et al., 2025), T2VPhysBench (Guo et al., 2025), EWMBench (Yue et al., 2025), measure predictive quality, physical realism, and task performance. These are necessary but insufficient: a model can achieve state-of-the-art scores on all of them while operating entirely at Level 1 of the causal hierarchy. We call for three categories of new benchmarks: interventional benchmarks that test whether a model correctly predicts outcomes of novel do-operations unseen in training; counterfactual benchmarks that test "what would have happened" accuracy against known ground truth; and mechanism uncertainty benchmarks that test whether confidence in causal attributions is calibrated to available evidence. Open questions include: (1) **How do we construct ground-truth counterfactuals for evaluation?** Real environments do not permit the same trajectory to be replayed under a different action. What experimental setups (e.g., simulated environments and synthetic data) can provide reliable counterfactual ground truth at scale? (2) **What is the right unit of evaluation for interventional correctness?** Interventional correctness requires comparing predicted post-intervention distributions against actual ones, but the choice of divergence measure, granularity, and test intervention set all affect results significantly. What conventions should the field standardize around? (3) **How do we prevent benchmark overfitting to spurious causal structure?** A model trained on a benchmark's specific intervention set may learn to pass it without genuine causal understanding. How do we design benchmarks that test transferable causal knowledge rather than memorized intervention-outcome pairs? (4) **How do we benchmark mechanism uncertainty calibration?** Unlike predictive calibration, there is no frequentist ground truth for graph uncertainty. What formal criteria (consistency, sharpness, or decision-theoretic performance) should define a well-calibrated causal WM?

4 A Possible Formulation: Intervention WM

We emphasize that the formulation presented in this section is one *possible* instantiation of an intervention-based WM. It is not intended to resolve all challenges outlined in Section 3 (e.g., learning intervention spaces, active causal exploration, and counterfactual reasoning), but rather to provide a concrete starting point for thinking about how such models might be architected in practice. We argue that an intervention WM should explicitly separate two questions:

1. **RQ1**. Where does an intervention enter the causal system, and what variable or mechanism does it modify?
2. **RQ2**. How does that change propagate through the remaining mechanisms of the world?

Concretely, we propose an intervention WM consisting of five modules (Fig. 1).

Causal Encoder. The first component is a *causal encoder*. Standard latent WMs typically learn a representation $z_t = E_\phi(x_t)$ optimized for compression and future prediction. The resulting representation may be highly entangled: a single latent dimension can simultaneously encode object position, lighting, background appearance, and other factors. Such representations may be sufficient for prediction but need not provide variables on which interventions have well-defined semantics. Instead, the causal encoder aims to recover a structured latent state $z_t = \left(z_t^{(1)}, z_t^{(2)}, \ldots, z_t^{(n)}\right)$,

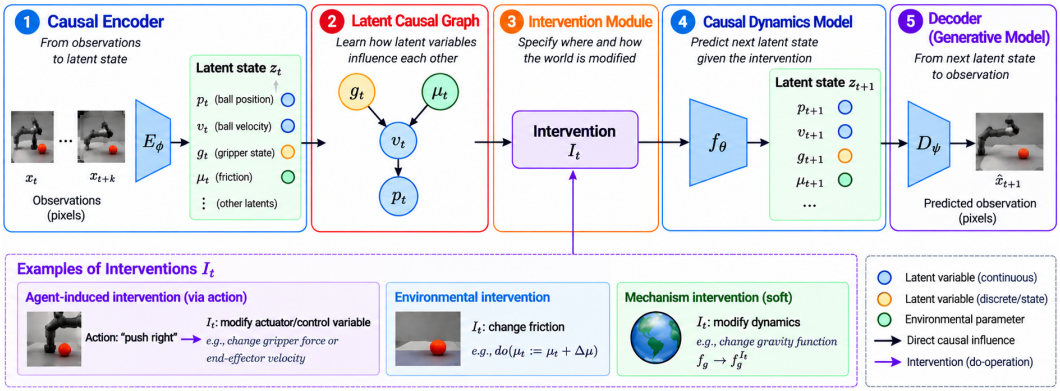

Figure 1: Architecture of an intervention WM. Rather than treating an action as simply another input, the model represents where an intervention enters the causal system and reasons about how its effects propagate through the world. Interventions may arise from agent actions or directly modify environmental variables or mechanisms.

whose components correspond to causally relevant properties of the environment. For example, in a robotic manipulation task, $z_t = \left(p_t^{\text{ball}}, v_t^{\text{ball}}, g_t^{\text{gripper}}, \mu_t^{\text{friction}}\right)$, where p_t^{ball} and v_t^{ball} denote ball position and velocity, g_t^{gripper} denotes the gripper state, and μ_t^{friction} represents floor friction.

The goal is not merely interpretability, but to learn variables on which interventions are meaningful. For example, changing the surface friction can be represented as an intervention on μ_t^{friction}. By contrast, a high-level action such as "push right" should not automatically be interpreted as directly setting the ball velocity; it may instead enter through the robot's control interface and affect the ball through downstream physical mechanisms. The causal encoder can therefore be viewed as a form of causal representation learning (Ahuja et al., 2023; Zhang et al., 2024; Schölkopf et al., 2021; Fan et al., 2026): rather than compressing observations into an arbitrary latent vector, it aims to recover variables sufficient for reasoning about interventions. We emphasize that recovering such representations is itself a difficult identification problem and generally requires appropriate assumptions, supervision, or interventional variation.

Latent Causal Graph. Once the latent variables have been identified, the model represents how they influence one another through a latent causal graph G. In the robot-and-ball example, the graph may capture how control inputs affect physical variables and how environmental properties such as friction modify the resulting dynamics.

More generally, the transition for each variable can be represented by a structural equation

$$z_{t+1}^i = f_i\left(z_t^{\text{Pa}_i^-}, z_{t+1}^{\text{Pa}_i^0}, \epsilon_{t+1}^i\right), \tag{2}$$

where Pa_i^- denotes temporal parents from the previous state, Pa_i^0 denotes contemporaneous parents within the new state, and ϵ_{t+1}^i denotes exogenous noise. The contemporaneous term can be omitted when the assumed dynamics contain only time-lagged edges. The graph provides an explicit representation of how local changes can propagate through the system. We emphasize, however, that an explicit graph is one possible realization of this idea; the essential requirement is that the learned representation preserve the causal dependencies needed for interventional reasoning.

Intervention Module. The intervention module represents where and how the causal system is modified. Given an intervention $I_t \in \mathcal{I}$, it identifies the targeted variable or mechanism and specifies the corresponding modification. This may take the form of a hard or soft intervention as in Section 2. When the intervention arises from an agent action, the module represents how a_t enters the causal system and induces I_t through the relevant control interface. The resulting intervention is then passed to the dynamics model to predict its downstream consequences.

Causal Dynamics Model. After an intervention is applied, the causal dynamics model predicts how its effects propagate through the rest of the world. The intervention directly modifies only its target

variable or mechanism; subsequent changes arise through the causal dependencies represented by the model, while unaffected mechanisms are preserved.

More generally, the intervention WM models

$$p^{I_t}(z_{t+1} \mid z_t) = \prod_{i=1}^{N} p^{I_t}\left(z_{t+1}^{i} \mid z_t^{\mathrm{Pa}_i^-}, z_{t+1}^{\mathrm{Pa}_i^0}\right), \tag{3}$$

where p^{I_t} denotes the distribution induced by intervention I_t. For a hard intervention, this corresponds to replacing the targeted structural equation; for a soft intervention, the targeted mechanism is modified instead. Mechanisms not targeted by I_t remain invariant. For example, changing the floor friction modifies the mechanism governing the ball's acceleration and velocity, which subsequently changes its position and potential collisions. Similarly, a robot action entering through an actuator can propagate through contact forces to the ball's motion without being treated as a direct intervention on the ball's velocity. The complete intervention WM therefore follows the conceptual pipeline

$$x_t \to z_t \to G, \qquad I_t \to \mathcal{M}^{I_t} \to z_{t+1} \to x_{t+1}, \tag{4}$$

where $\mathcal{M}^{I_t}$ denotes the causal dynamics modified by intervention I_t. When the intervention is induced by an agent action, the pipeline additionally includes $a_t \to I_t$.

This differs from standard action-conditioned prediction not because conditioning on actions is inherently non-causal, but because the intervention WM explicitly represents the *causal locus and semantics of change*. It distinguishes the intervention itself from its downstream consequences and allows the intervention space to extend beyond the agent's ordinary action space. By separating intervention from propagation, this formulation provides one concrete route toward WMs that can reason about how the world responds when its variables or mechanisms are deliberately changed.

5 Implications

The intervention-centric formulation provides a broader view of WMs: rather than only predicting how a system evolves under observed conditions, a WM should also reason about how the system responds when its variables or mechanisms are deliberately changed. This perspective has broad implications for how WMs are evaluated, trained, and deployed.

Rethinking Evaluation. Current WM evaluation largely emphasizes predictive accuracy on held-out trajectories. Such evaluation remains essential, but predictive performance alone does not establish whether a model has learned mechanisms that remain valid under intervention. Evaluation should therefore additionally measure *interventional fidelity*: whether a model correctly predicts the consequences of specified changes to variables or mechanisms. Relevant tests include generalization to unseen interventions, transfer of mechanisms across environments, discrimination among competing causal explanations, and calibrated uncertainty about interventional outcomes. Evaluation should also become more *interactive*. Rather than relying only on static test sets, benchmarks can assess whether a WM identifies informative interventions, predicts their consequences, observes the resulting outcomes, and updates its beliefs about the underlying mechanisms. Generic out-of-distribution failure is not by itself evidence of causal failure; support mismatch, insufficient coverage, and accumulated prediction error can produce similar behavior. Interventional evaluation should therefore specify what is changed and test whether the model responds correctly to that change.

Implicit Representation Models (Model-based RL). Model-based RL provides a natural setting for intervention-centric WMs because agent actions can often have well-defined interventional semantics. Under appropriate assumptions, an action-conditioned transition model may already support valid interventional prediction. The challenge is therefore not simply to replace action conditioning with a do-operator, but to learn where actions enter the causal system and which mechanisms govern their downstream effects. This perspective suggests objectives that encourage stable and reusable mechanisms rather than optimizing only one-step prediction on trajectories generated by a limited set of policies. It also changes the role of exploration. Actions can be selected not only to maximize reward or visit novel states, but also as experiments that distinguish competing models of the environment. A WM that represents uncertainty over these mechanisms can therefore couple planning with active causal discovery, particularly when reliable behavior requires generalization beyond previously observed trajectories.

Video Generation Models. Modern video models demonstrate increasingly powerful predictive models of visual dynamics, but accurate generation does not necessarily imply that the underlying mechanisms are identifiable or controllable. An intervention-centric perspective suggests evaluating these models not only by the realism of generated trajectories, but also by whether they respond consistently when particular properties of a scene or its dynamics are changed. A structured intervention interface could, for example, ask a model to generate the same scene under different gravity, friction, object mass, or initial conditions while preserving mechanisms unrelated to the intervention. Such capabilities would support controlled simulation, data generation, and mechanistic evaluation. Explicit causal graphs are not necessarily required; the central requirement is that the representation support well-defined and consistent changes to the modeled world.

Embodied AI. Embodied AI is particularly well aligned with the intervention perspective because agents can act on the physical world and observe the consequences. Actions can therefore serve both as controls for accomplishing a task and, when appropriately designed, as experiments for learning how the environment works. For manipulation, intervention-aware WMs could help distinguish changes caused by the robot from changes caused by properties such as object mass, friction, or contact dynamics, supporting diagnosis and adaptation when expected outcomes fail to occur. For navigation, they can distinguish controllable aspects of the environment from external constraints. For sim-to-real transfer, interventions on physical parameters and mechanisms provide a direct way to test which learned relationships remain invariant across simulated and physical environments. This brings world modeling, active experimentation, and planning into a common framework.

Scientific Discovery and Simulation. Scientific simulation provides another natural setting for intervention-centric WMs. Simulators already permit controlled changes that may be difficult, expensive, or impossible to realize experimentally. In molecular simulation, for example, one can perturb interactions, constrain molecular coordinates, or modify environmental conditions. Analogous interventions are possible in materials, climate, fluid, and physical-system simulations. An intervention-centric WM can use such simulations not only as sources of trajectories, but as environments for *active scientific experimentation*. The model can identify competing hypotheses about a mechanism, select simulations or perturbations most informative for distinguishing them, observe the resulting trajectories, and update its representation of the system. Interventional data can also expose regions of state and mechanism space that are poorly covered by passive trajectories, improving the data efficiency and transferability of learned dynamics. In this setting, the WM becomes more than a surrogate predictor for an expensive simulator: it can help determine *which simulation to perform next and why*, connecting world modeling to mechanistic discovery and experimental design.

6 Conclusion

WMs have evolved from a component of model-based RL into a central paradigm in AI. Despite impressive progress in predictive modeling, prediction alone does not guarantee reliable reasoning about how the world responds when its variables or mechanisms are deliberately changed. Failures under physical perturbations and distribution shifts highlight this gap, although such failures can also arise from limited coverage or predictive error. We argue that addressing this gap requires making causal interventions a first-class consideration in WM design. Importantly, actions and interventions should not be conflated: actions can have valid interventional semantics, but the intervention space relevant to understanding a world may differ from and extend beyond the agent's action space. Causal WMs should therefore represent where interventions enter the system, how their effects propagate through causal mechanisms, and how informative interventions can reduce uncertainty about those mechanisms. Our position does not require every WM to contain an explicit causal graph or symbolic *do*-operator. Rather, the central goal is to build WMs that support reliable reasoning over a meaningful intervention space. The framework developed in this paper highlights several steps toward this goal, including causal representations, intervention-driven exploration, uncertainty over competing mechanisms, counterfactual reasoning, and evaluation based on interventional correctness. Many of the necessary foundations already exist across causal inference, causal representation learning, and experimental design. Integrating these ideas with modern WMs offers a path from models that primarily *predict* how the world evolves toward models that can also reason about *how and why the world changes under intervention*.

AI DISCLOSURE

We used generative AI tools to assist with literature search, language editing and polishing, and the preparation of preliminary drafts of figures. All scientific ideas, analyses, views, interpretations, and final manuscript content were developed, verified, and approved by the authors.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pp. 372–407. PMLR, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Genesis Authors. Genesis: A generative and universal physics engine for robotics and beyond, December 2024. URL `https://github.com/Genesis-Embodied-AI/Genesis`.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? pp. 610–623, 2021.

Tim Brooks, William Peebles, Aleksander Holynski, Alexei A. Efros, and OpenAI. Video generation models as world simulators. *arXiv preprint arXiv:2405.07987*, 2024.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024.

Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. *arXiv preprint arXiv:2410.13232*, 2024.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Shicheng Fan, Kun Zhang, and Lu Cheng. Trace: Trajectory recovery for continuous mechanism evolution in causal representation learning. In *ICML*, 2026.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, et al. Is your llm secretly a world model of the internet? model-based planning for web agents. *arXiv preprint arXiv:2411.06559*, 2024.

Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020a.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020b.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065, 2023.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 2022.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-jepa: Learning world models through object-level latent interventions. *arXiv e-prints*, pp. arXiv–2602, 2026.

OpenAI. Sora: Creating video from text. OpenAI Technical Report, 2024. https://openai.com/sora.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, et al. Genie 2: A large-scale foundation world model. *URL: https://deepmind. google/discover/blog/genie-2-a-large-scale-foundation-world-model*, 2, 2024.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2019.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. 2025.

Silviu Pitis, Elliot Creager, and Animesh Garg. Counterfactual data augmentation using locally factored dynamics. *Advances in Neural Information Processing Systems*, 33:3976–3990, 2020.

Shuofei Qiao, Runnan Fang, Ningyu Zhang, Yuqi Zhu, Xiang Chen, Shumin Deng, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Agent planning with world knowledge model. *Advances in Neural Information Processing Systems*, 37:114843–114871, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. Roboscape: Physics-informed embodied world model. *arXiv preprint arXiv:2506.23135*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pp. 2226–2240. PMLR, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024a.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024b.

Tenny Yin, Zhiting Mei, Zhonghe Zheng, Miyu Yamane, David Wang, Jade Sceats, Samuel M Bateman, Lihan Zha, Apurva Badithela, Ola Shorinwa, et al. Playworld: Learning robot world models from autonomous play. *arXiv preprint arXiv:2603.09030*, 2026.

Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=Pte6iiXvpf`.

Xiaofan Zhou, Liangjie Huang, Pinyang Cheng, Wenpen Yin, Rui Zhang, Wenrui Hao, and Lu Cheng. Accelerating causal network discovery of alzheimer disease biomarkers via scientific literature-based retrieval augmented generation. In *IEEE BigData*, 2025.