CausalVAE as a Plug-in for World Models: Towards Reliable Counterfactual Dynamics

Ziyi Ding[1,3], Xianxin Lai[2], Weiyu Chen[2], Xiao-Ping Zhang[1]*, and Jiayu Chen[2,3]*

[1] Shenzhen Key Laboratory of Ubiquitous Data Enabling, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
xpzhang@ieee.org
[2] The University of Hong Kong, Hong Kong SAR, China
jiayuc@hku.hk
[3] INFIFORCE Intelligent Technology, China

Abstract. Latent world models roll observations forward accurately on factual transitions, but their predictive latents are typically entangled and weakly aligned with the underlying causal factors; as a result, they degrade under interventions and on counterfactual queries. We address this with CausalVAE, a plug-in structural module that attaches to diverse encoder–transition backbones and organizes their latents through a learned directed acyclic graph (DAG) over causal factors, without altering the base prediction architecture. Across four benchmarks and eight backbones, the plug-in largely preserves factual retrieval while improving intervention-aware counterfactual retrieval on most backbone–benchmark pairs; the gains are benchmark- and backbone-dependent rather than universal. Improvements are largest on the Physics benchmark: for a graph-network backbone trained with a negative-log-likelihood objective, counterfactual H@1 (CF-H@1) rises from **11.0** to **41.0**, with several other backbones gaining +10 to +30 CF-H@1 points. A causal analysis further shows that the learned structure recovers meaningful first-order physical interaction trends, supporting the interpretability of the latent causal structure.

Keywords: World Models · Causal Representation Learning · Counterfactual Reasoning · Structured Latent Dynamics

1 Introduction

Generalization under distribution shifts, interventions, and mechanism changes remains a central challenge for visual model-based learning [18]. World models achieve strong predictive performance by compressing observations into latent states and rolling them forward under actions [5, 6, 8], but predictive latents are often entangled and weakly aligned with underlying causal factors, limiting out-of-distribution robustness.

* Corresponding authors.

Recent work on causal representation learning argues that discovering high-level causal variables from low-level sensory inputs is essential for robust generalization and transfer [18]. In parallel, object-centric representation learning provides a compositional inductive bias by decomposing scenes into sets of object slots, improving systematic generalization [3, 4, 14]. However, neither standard world models nor generic object-centric models explicitly enforce that latent factors obey a directed acyclic graph (DAG) causal structure. This limitation is critical: without an explicit structural causal model, latent interventions are not identifiable, so counterfactual rollouts can deviate from physically valid alternative trajectories.

To address this, we integrate a structured causal disentanglement module into a world-model pipeline. Our structural branch adopts the CausalVAE causal layer to map independent exogenous factors into causally related endogenous variables, while learning a DAG over the latent factors [20]. We encourage acyclicity via a differentiable DAG constraint inspired by continuous optimization approaches for structure learning [21]. Because our setting is sequential and action-conditioned, existing static causal representation methods (e.g., CausalVAE in i.i.d. settings) cannot be directly applied for stable multi-step dynamics. Therefore, we introduce a staged training strategy that first learns predictive dynamics and then progressively activates structural regularization, while using alignment-only weak supervision to anchor latent coordinates (instead of the auxiliary-variable conditional prior used in identifiable VAE analyses [10]). This design yields an interpretable causal state space that supports intervention-style reasoning and improves robustness in model-based prediction.

Our contributions are: (i) a plug-in causal structural module for latent world models, which can be combined with diverse representative world-modeling baselines used in our experiments; (ii) a latent world-model formulation that explicitly captures causal relationships among latent elements, rather than treating latent dimensions as purely predictive features; (iii) a staged optimization scheme for stable sequential training, together with an alignment-anchored identifiability analysis that characterizes what becomes identifiable under this training setup, and empirical gains in generalization and counterfactual consistency over non-causal world-modelling baselines.

2 Related Work

2.1 Structured World Models and Latent Dynamics.

World models learn compact latent states to support multi-step prediction and visual planning [5, 7]. To improve compositionality in physical domains, structured variants introduce object- or relation-centric inductive biases, utilizing graph networks (e.g., NRI, GNS) [11, 17] or object-centric visual slots (e.g., C-SWM, Slot Attention) [12, 14]. While these models, including standard GNN-based dynamics, achieve strong factual rollout accuracy, their latent representations are optimized purely as predictive features. Without learning an explicit

Structural Causal Model (SCM) [16], the model does not capture the causal relationships among variables. As a result, it may perform well on factual prediction, but can break down under interventions or counterfactual tasks.

2.2 Causal Representations and Counterfactual Dynamics.

Causal representation learning grounds latents in SCMs to enable interpretable interventions. Methods such as CausalVAE [20] are mainly developed for static i.i.d. observations and are trained end-to-end with a causal layer tied to a specific encoder, using supervised attribute labels u as the anchor for identifiability; neither the end-to-end coupling nor the labels u transfer to action-conditioned pixel rollouts. In multi-step visual dynamics, this mismatch can manifest as compounding one-step errors and latent drift under temporal distribution shift. Our method therefore does not directly apply a static causal model: the causal branch is a *plug-in* that attaches to any pretrained encoder and is reused unchanged across all backbones in our experiments, and we replace the attribute supervision u with an alignment loss $\mathcal{L}_{\text{align}}$ to a low-dimensional state v_t (Sec. 3.3), stabilized by staged optimization. Conversely, recent counterfactual dynamics models (e.g., CWM, Causal-JEPA) [13, 15, 19] emphasize learning and evaluation under counterfactual/interventional settings, but many rely on masking/prompting heuristics: a masked slot carries no semantic label and can only reproduce factor combinations seen during training, so it cannot encode a genuinely new *do*-intervention. Empirically, replacing our learned causal layer with feature masking lowers CF-H@1 by 52 points on average (supplementary material), whereas the learned DAG matrix A routes interventions through identified causal edges.

2.3 Causal evaluation in dynamical systems.

Prior work stresses that predictive accuracy is insufficient to evaluate causal correctness from pixels, requiring rigorous intervention-centric metrics [9]. Counterfactual dynamics models further argue that counterfactual capability should be treated as a first-class objective beyond factual prediction [13, 19]. Motivated by these findings, we evaluate world models using intervention-centric protocols that probe interventional faithfulness and multi-step counterfactual consistency in addition to factual forecasting.

3 Method

3.1 Problem Setup

We study world modelling from visual observations under both factual and counterfactual settings. Following structured world-model formulations [9, 12], we adopt an object-centric latent dynamics pipeline and augment it with a causal structural branch inspired by CausalVAE [20].

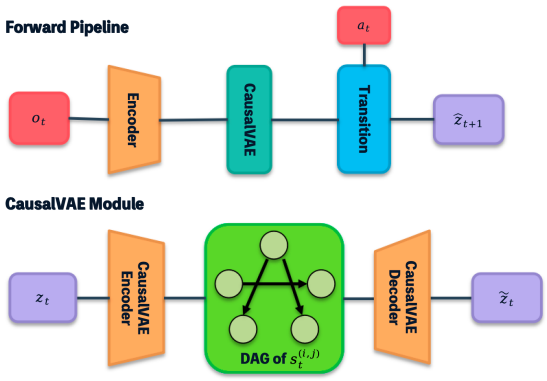

Fig. 1: Overview of our framework. Top: $o_t \to z_t \to \tilde{z}_t \xrightarrow{a_t} \hat{z}_{t+1}$. Bottom: the CausalVAE branch imposes DAG-structured causal constraints in latent space before decoding back to $\tilde{z}_t$.

Given. We are given transition tuples $\mathcal{D} = \{(o_t, a_t, o_{t+1})\}_{t=1}^{T}$, with optional simulator state $s_t \in \mathbb{R}^{d_s}$ used only for training-time alignment when available. Counterfactual queries are specified by intervention metadata (i, j, Δ) and implemented as $\mathrm{do}(s_t^{(i,j)}) := s_t^{(i,j)} + \Delta$, producing counterfactual targets such as o_{t+1}^{cf} (or latent counterparts) [16].

Unknowns and learning goal. We learn three coupled functions (with parameters (θ, ψ, ϕ)): encoder E_θ (observation $\to$ object-centric latent), causal branch C_ψ (latent $\to$ causally regularized latent), and action-conditioned transition F_ϕ (latent/action $\to$ next latent). The learning goal is to obtain representations and dynamics that are accurate on factual transitions and remain reliable under interventions, i.e., strong counterfactual performance under do-queries with causal selectivity rather than diffuse latent changes.

Evaluation target. We evaluate with retrieval metrics H@1, MRR, CF-H@1, and CF-MRR (formal definitions are given in Sec. 4.3); factual H@1/MRR are reported at 1/5/10-step horizons.

3.2 Method Overview

Figure 1 shows the architecture: an observation o_t is encoded to object-centric latents, refined by a causal structural branch, and rolled forward under the action a_t. The encoder produces

$$z_t = E_\theta(o_t), \quad z_t \in \mathbb{R}^{K \times d}, \tag{1}$$

with K object slots and per-slot dimension d; a causal branch then yields a structurally regularized latent

$$\tilde{z}_t = C_\psi(z_t), \tag{2}$$

where $C_\psi(\cdot)$ is the CausalVAE-inspired transformation of Fig. 1; and an action-conditioned transition predicts the next latent

$$\hat{z}_{t+1} = F_\phi(\tilde{z}_t, a_t). \tag{3}$$

The concrete instantiation of $\tilde{z}_t$ fed to the transition model is specified in the training-strategy section.

Figure 1 shows two complementary views: the end-to-end inference pipeline $o_t \to z_t \to \tilde{z}_t \xrightarrow{a_t} \hat{z}_{t+1}$ used for both factual and counterfactual prediction, and an expansion of the CausalVAE branch that encodes latents into a structured causal space and decodes them back under structural regularization. The design thus inherits action-conditioned latent transition modeling from structured world models [9, 12] and structural causal regularization from CausalVAE-style modeling [20], jointly targeting factual quality and counterfactual robustness.

3.3 CausalVAE Branch

Given the encoder latent z_t (Eq. (1)), our CausalVAE branch produces a structurally regularized latent $\tilde{z}_t$ (Eq. (2)), which is then used by the transition model in Eq. (3). The overall forward relation is consistent with Fig. 1:

$$o_t \to z_t \xrightarrow{\text{CausalVAE}} \tilde{z}_t \xrightarrow{a_t} \hat{z}_{t+1}. \tag{4}$$

The branch follows the CausalVAE principle [20]: latent factors are organized through a structural causal transformation parameterized by a DAG matrix A, and then mapped back to latent dynamics space. In our implementation, DAG-style structural constraints are applied inside the branch to encourage causally meaningful factorization before decoding to $\tilde{z}_t$. Importantly, this causal branch is a plug-in module: it operates on latent representations and can be attached to different latent world-model backbones (e.g., C-SWM-style, GNN-style, or other encoder–transition pipelines) without changing their base transition architecture.

The training objective of this branch is

$$\mathcal{L}_{\text{stage2}} = \lambda_1 \mathcal{L}_{\text{rec}} + \lambda_2 \mathcal{L}_{\text{KL}} + \lambda_3 \mathcal{L}_{\text{align}} + \lambda_4 \mathcal{L}_{\text{DAG}}, \tag{5}$$

where $\mathcal{L}_{\text{rec}}$ reconstructs latent representation, $\mathcal{L}_{\text{KL}}$ regularizes the variational posterior, $\mathcal{L}_{\text{align}}$ aligns to available state supervision (when s_t is available), and $\mathcal{L}_{\text{DAG}}$ encourages acyclicity of A.

When simulator states are available, we impose an auxiliary alignment objective to map latent dynamics to physically meaningful state variables. Let $g_{\text{align}}(\cdot)$ denote the alignment head and s_t the ground-truth state. Given the CausalVAE output latent $\tilde{z}_t$, we define

$$\mathcal{L}_{\text{align}} = \|g_{\text{align}}(\tilde{z}_t) - s_t\|_2^2. \tag{6}$$

This term is used only during training and is weighted by λ_3 in Eq. (5); at inference time the model does not require s_t. Because simulator state s_t is available

in all benchmarks reported here, the alignment term is kept active with λ_3=0.1. Unless otherwise noted, the remaining weights are $(\lambda_1, \lambda_2, \lambda_4)$=$(1, 0.1, 0.01)$ for $(\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{KL}}, \mathcal{L}_{\text{DAG}})$ with α_{KL}=0.3; a full sensitivity sweep over these weights is provided in the supplementary material, confirming that each term is load-bearing.

For the DAG constraint we use the standard smooth acyclicity penalty:

$$\mathcal{L}_{\text{DAG}} = \text{tr}(\exp(A \odot A)) - d, \tag{7}$$

with d the structural latent dimension [21].

Following the CausalVAE formulation [20], we instantiate the remaining terms as:

$$\mathcal{L}_{\text{rec}} = \|\tilde{z}_t - z_t\|_2^2, \quad \tilde{z}_t := C_\psi(z_t), \tag{8}$$

$$\mathcal{L}_{\text{KL}} = \mathbb{E}\left[\alpha_{\text{KL}}\, \text{KL}(q_\psi(\cdot \mid z_t) \,\|\, \mathcal{N}(0, I)) + \sum_{i=1}^{d} \text{KL}(q_\psi(\cdot \mid z_t, A)_i \,\|\, p(\cdot \mid s_t)_i)\right], \tag{9}$$

where $q_\psi(\cdot \mid z_t)$ is the approximate posterior induced by the CausalVAE branch from encoder latent z_t, $q_\psi(\cdot \mid z_t, A)_i$ denotes its i-th causal component after the DAG transformation, and $p(\cdot \mid s_t)_i$ is the corresponding state-conditioned prior component (when s_t is available). Here A is the learned DAG matrix, d is the structural latent dimension (consistent with Eq. (7)), and $\text{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence. Following CausalVAE [20], we set $\alpha_{\text{KL}} = 0.3$ in our implementation.

Identifiability of the Causal Branch. We analyze identifiability for the Stage-2 structural learner, where the world-model backbone (encoder/transition) is frozen and the CausalVAE branch is optimized with the same Stage-2 objective as in Eq. (5), i.e., $\min_{\psi, A, g_{\text{align}}} \mathcal{L}_{\text{stage2}}$ with $\tilde{z}_t = C_\psi(z_t)$ and $\mathcal{L}_{\text{DAG}}$ encouraging acyclicity [20].

Assumptions. (I1) *Stage-2 decoupling:* during Stage-2, transition/backbone parameters are frozen, so (5) is the only objective governing $(\psi, A, g_{\text{align}})$. (I2) *Latent invertibility (local):* on the low-dimensional manifold supervised by $\mathcal{L}_{\text{align}}$, there exists a *locally* invertible map g with $z_t = g(v_t)$ for the underlying causal state $v_t \in \mathbb{R}^d$; we do *not* assume the full image encoder is globally invertible, so KL compression of off-manifold detail is permitted. (I3) *Realizability + DAG:* there exists a ground-truth DAG adjacency $A^\star$ (acyclic) and parameters $(\psi^\star, g^\star_{\text{align}})$ attaining the population-risk minimum of (5), and $\mathcal{L}_{\text{DAG}}(A) = 0$ iff A is acyclic. (I4) *Alignment anchoring + scale normalization:* at the population optimum,

$$g_{\text{align}}(\tilde{z}_t) = s_t \quad \text{a.s.}, \tag{10}$$

and the coordinate system of $\tilde{z}_t$ is fixed by scale normalization (e.g., per-dimension zero-mean/unit-variance). Moreover, anchoring is axis-fixing: if an invertible $T : \mathbb{R}^d \to \mathbb{R}^d$ satisfies $g_{\text{align}}(T(\tilde{z}_t)) = g_{\text{align}}(\tilde{z}_t)$ a.s. under the same normalization, then T must be the identity. (I5) *Population/global optimum:* infinite-sample

and exact optimization. Why these assumptions are reasonable in our setting (and when they may fail) is discussed in the supplementary material.

Theorem 1 (Alignment-anchored identifiability). *Under (I1)–(I5), the Stage-2 optimizer identifies an adjacency $\hat{A}$ that matches $A^\star$ in the alignment-anchored coordinate system, i.e., $\hat{A} = A^\star$ in that coordinate frame.*

This statement is conditional on (I1)–(I5): a population-level characterization, not a finite-sample optimization guarantee.

Proof. See the supplementary material.

Relation to CausalVAE identifiability theorem in the original paper [20]. The identifiability proof in CausalVAE relies on an auxiliary variable u through a conditional prior $p(z \mid u)$, leading to $\sim$-identifiability of the generative parameters under that training setting. In our model, we do *not* assume access to u nor optimize a conditional prior; instead, we use only the alignment loss $\mathcal{L}_{\text{align}} = \|g_{\text{align}}(\tilde{z}_t) - s_t\|_2^2$ as weak supervision to anchor the latent coordinates. Therefore, our identifiability argument does not directly invoke CausalVAE's Theorem 1; it follows a different route in which alignment eliminates the permutation/scale (and more generally reparameterization) ambiguities, after which the DAG-constrained structural learner identifies the adjacency in the anchored coordinate system. We cite CausalVAE for (i) the formal ambiguity class captured by $\sim$-identifiability and (ii) its discussion on identifiability of the causal graph in the causal layer [20].

3.4 Transition Modeling

The transition module models action-conditioned dynamics in latent space. Following Sec. 3.2, the encoder latent and causal-refined latent are defined in Eq. (1) and Eq. (2), respectively. Next-state prediction then follows Eq. (3).

The supervision target is the latent encoding of the next-step observation:

$$z_{t+1} = E_\theta(o_{t+1}). \tag{11}$$

To stabilize long-horizon prediction, we use residual dynamics parameterization:

$$\Delta z_t = f_\phi(\tilde{z}_t, a_t), \qquad \hat{z}_{t+1} = \tilde{z}_t + \Delta z_t. \tag{12}$$

For multi-step rollout, predictions are generated recursively:

$$\hat{z}_{t+k+1} = F_\phi(\hat{z}_{t+k}, a_{t+k}), \quad k \geq 0, \tag{13}$$

with $\hat{z}_t \leftarrow \tilde{z}_t$, and targets

$$z_{t+k} = E_\theta(o_{t+k}). \tag{14}$$

The rollout objective is

$$\mathcal{L}_{\text{multistep}} = \frac{1}{H} \sum_{k=1}^{H} \ell(\hat{z}_{t+k}, z_{t+k}), \tag{15}$$

where H is the rollout horizon, and $\ell(\cdot,\cdot)$ is instantiated as either a contrastive objective (as in C-SWM-style latent energy matching [12]) or a latent negative log-likelihood objective (as in likelihood-based world models [5,6]). For the contrastive case, let z^-_{t+k} denote a negative latent target at step $t+k$ (sampled from non-matching futures in the candidate pool/batch), and let $m > 0$ be the margin. We use

$$\ell_{\text{con}}(\hat{z}, z) = \|\hat{z} - z\|_2^2 + \max\Big(0,\ m - \big\|\hat{z} - z^-\big\|_2^2\Big)\,. \tag{16}$$

For the NLL case, assuming an isotropic Gaussian decoder in latent space,

$$\ell_{\text{NLL}}(\hat{z}, z) = -\log p_\phi(z \mid \hat{z})\,, \quad p_\phi(z \mid \hat{z}) = \mathcal{N}\big(z; \hat{z}, \sigma^2 I\big)\,, \tag{17}$$

which is equivalent (up to constants) to a scaled squared-error term.

3.5 Training Strategy

We adopt a three-stage training pipeline (Fig. 2) to decouple (i) backbone dynamics pretraining, (ii) causal structure learning, and (iii) long-horizon transition refinement.

Stage 1 (Backbone pretraining). We optimize the encoder-transition backbone using one-step supervision defined in Sec. 3.2, i.e., $z_t = E_\theta(o_t)$ and $\hat{z}_{t+1} = F_\phi(z_t, a_t)$.

Stage 2 (Causal branch training). With backbone modules frozen, we train the CausalVAE branch in Sec. 3.3 using Eq. (5), including reconstruction, KL, alignment (when supervision is available), and DAG terms. Details on how supervision is obtained and how objective slots are selected for each benchmark are provided in the supplementary material.

Stage 3 (Transition Refinement with Alpha-Gated Fusion). We freeze the encoder and CausalVAE, and fine-tune the transition model with

$$z_t^{\text{gate}} = (1 - \alpha_t) z_t + \alpha_t \tilde{z}_t, \qquad \alpha_t \in [0, 1]. \tag{18}$$

In implementation, we use an exponentially decayed mixing coefficient ($\alpha_t = \alpha_0 \exp(-k_\alpha t)$), where α_0 is the initial fusion weight and k_α controls decay speed. We optionally apply an additional data-dependent gate with $\alpha_t^{\text{eff}} = g_t \alpha_t$ and $g_t = \sigma((\tau - \delta_t)\,\gamma)$, where $\delta_t = \|\tilde{z}_t - z_t\|_2$ and (τ, γ) are gate hyperparameters. The transition update is $\hat{z}_{t+1} = F_\phi(z_t^{\text{gate}}, a_t)$, and multi-step training follows Sec. 3.4 (Eq. (13) and Eq. (15)). We report an ablation over gate on/off in Sec. 4.6 (Tab. 3). The rationale for using a three-stage schedule is also empirically validated in Sec. 4.6 via explicit w/3-stage vs. w/o-3-stage comparisons.

4 Experiments

We evaluate whether injecting an explicit causal factorization via CausalVAE improves *counterfactual reliability* while preserving factual predictive performance.

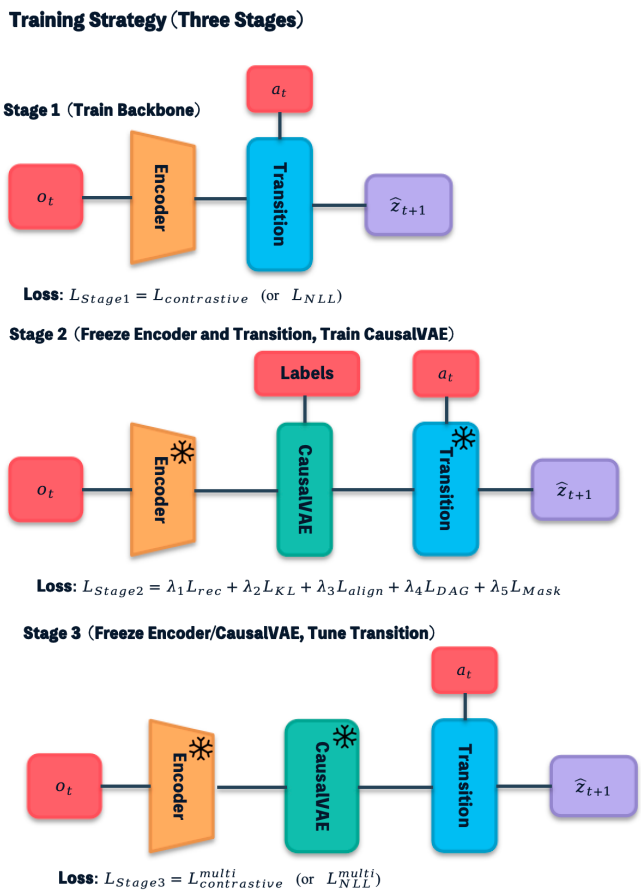

Fig. 2: Three-stage training strategy.

4.1 Benchmarks

We use four domains from the causal-discovery evaluation framework [9]: Physics (3-body gravitation), 2D Shapes, 3D Cubes, and Chemistry, with state-level interventions (e.g., position/velocity) on Physics, object-level (pose/push) interventions on Shapes/Cubes, and mechanism-level interventions on Chemistry. As illustrated in Fig. 3, counterfactual tasks start from factual tuples (o_t, a_t, o_{t+1}), apply a *do*-intervention at time t, re-simulate to obtain o^{cf}_{t+1}, and evaluate retrieval on paired factual/counterfactual futures; full construction details (query-group counts, intervention variables, magnitudes, and candidate sets) are in the supplementary material.

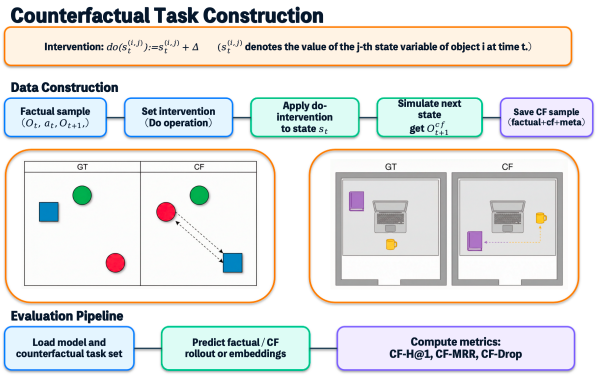

Fig. 3: Counterfactual task construction and evaluation pipeline. From factual tuples (o_t, a_t, o_{t+1}), we apply interventions $do(s_t^{(i,j)}):=s_t^{(i,j)}+\Delta$, re-simulate to obtain o_{t+1}^{cf}, and evaluate retrieval on paired factual/counterfactual futures using H@1, MRR, CF-H@1, and CF-MRR (factual H@1/MRR at 1/5/10-step horizons).

4.2 Baselines

Eight baselines (paired with +CausalVAE). We compare eight standard world-model baselines against matched **Baseline + CausalVAE** counterparts:

AE-based: AE_NLL, AE_Contrastive.
VAE-based: VAE_NLL, VAE_Contrastive.
Structured latent dynamics: Modular_NLL, Modular_Contrastive.
Graph dynamics: GNN_NLL, GNN_Contrastive. Detailed baseline network architectures and objective definitions (NLL/contrastive) are provided in the supplementary material.

Our model. Starting from each baseline backbone, we insert a latent CausalVAE structural layer and train with three stages (S1/S2/S3) for factual dynamics, structural constraint learning, and counterfactual refinement. This S1/S2/S3 schedule applies to **Baseline+CausalVAE** models; pure baselines follow the baseline objective settings (NLL, contrastive) described in the supplementary material.

Protocol. All paired comparisons (Baseline vs. Baseline+CausalVAE) use a unified protocol—identical splits, seeds, optimizer family, and rollout settings—and counterfactual reasoning is evaluated by controlled interventions at time t_0 with retrieval over factual/counterfactual candidates. Full AE/VAE/GNN/Modular architectures, hyperparameters, and intervention-construction details are provided in the supplementary material.

4.3 Metrics

We report four retrieval metrics, following the latent ranking-based evaluation of structured world models [12] and the intervention-centric protocol of

Ke *et al.* [9]: **H@1** (Hits@1) and **MRR** (mean reciprocal rank) for factual retrieval, together with their counterfactual counterparts **CF-H@1** and **CF-MRR**. For query i with factual rank r_i and counterfactual rank r_i^{cf}, we use $\text{H@1} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}[r_i = 1]$, $\text{MRR} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{r_i}$, $\text{CF-H@1} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}[r_i^{cf} = 1]$, and $\text{CF-MRR} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{r_i^{cf}}$. Factual H@1/MRR are reported at 1/5/10-step horizons using the same definitions on each horizon-specific retrieval task. CF-H@1/CF-MRR are reported for the counterfactual step.

4.4 Benchmarking Results

Factual retrieval comparison. Tab. 1 reports the factual retrieval performance (H@1 and MRR) of all baseline families and their +CausalVAE variants across all benchmarks. The key reading of this table is that the plug-in *preserves* factual prediction: across backbones, benchmarks, and the 1/5/10-step horizons, each +CausalVAE variant stays close to—and often matches or exceeds—its paired baseline, so attaching the causal branch does not trade away factual accuracy. This is the property that lets it serve as a drop-in module.

Table 1: Comparison on factual retrieval metrics. Each cell reports Baseline/+CausalVAE. Bold indicates the best value in each column.

		1 Step		5 Steps		10 Steps	
Benchmark	Method	H@1 ↑	MRR ↑	H@1 ↑	MRR ↑	H@1 ↑	MRR ↑
Physics 3-body	AE_Contrastive	89.00/98.00	94.25/99.00	10.00/9.00	27.91/25.72	1.00/3.00	6.59/9.31
	AE_NLL	92.00/94.00	95.67/96.83	4.00/4.00	16.58/16.39	0.00/1.00	8.93/9.04
	VAE_Contrastive	70.00/91.00	83.09/95.33	1.00/3.00	4.96/10.44	2.00/2.00	5.74/5.73
	VAE_NLL	92.00/94.00	95.83/97.00	1.00/2.00	10.19/9.46	0.00/0.00	3.61/3.64
	Modular_Contrastive	90.00/86.00	94.42/92.33	20.00/8.00	36.31/21.23	3.00/2.00	10.44/6.37
	Modular_NLL	96.00/95.00	97.83/97.50	11.00/11.00	30.28/28.36	1.00/3.00	9.38/9.99
	GNN_Contrastive	88.00/**99.00**	93.37/**99.50**	25.00/**36.00**	37.77/**54.58**	3.00/**7.00**	11.89/**17.00**
	GNN_NLL	98.00/99.00	99.00/99.50	9.00/17.00	29.36/33.32	**7.00**/5.00	15.27/12.57
Chemistry	AE_Contrastive	99.91/99.90	99.95/99.95	99.84/99.81	99.92/99.91	**99.92**/99.33	**99.96**/99.66
	AE_NLL	99.86/99.89	99.93/99.94	98.17/96.77	98.98/98.07	93.24/85.73	95.70/90.13
	VAE_Contrastive	99.95/99.96	99.98/99.98	99.94/99.73	99.97/99.87	**99.92**/98.60	**99.96**/99.30
	VAE_NLL	96.03/97.47	97.65/98.65	88.50/74.59	92.69/82.31	78.20/25.32	85.05/36.50
	Modular_Contrastive	99.50/16.03	99.75/33.95	77.97/6.48	88.06/17.97	30.50/3.76	53.24/11.83
	Modular_NLL	99.97/**99.99**	99.98/**99.99**	99.43/98.63	99.69/99.24	95.74/67.20	97.18/76.73
	GNN_Contrastive	99.95/99.95	99.98/99.98	**99.96**/99.74	**99.98**/99.87	99.86/99.48	99.93/99.74
	GNN_NLL	99.95/99.91	99.98/99.95	99.94/99.38	99.97/99.66	99.91/89.28	99.95/92.74
2D Shapes	AE_Contrastive	94.01/94.03	96.58/96.75	58.84/58.16	71.95/71.43	28.34/27.44	43.42/42.79
	AE_NLL	99.42/98.83	99.60/99.25	81.31/77.28	86.45/83.26	42.57/40.37	52.23/50.06
	VAE_Contrastive	76.51/87.70	84.81/93.35	20.30/1.75	34.95/7.11	5.58/0.29	13.67/1.37
	VAE_NLL	59.45/60.37	66.39/67.29	9.84/10.82	14.11/14.93	2.00/2.14	3.55/3.72
	Modular_Contrastive	3.25/2.16	14.23/9.06	0.73/0.19	4.42/1.63	0.33/0.08	2.48/0.80
	Modular_NLL	99.80/99.62	99.89/99.78	85.69/83.88	90.43/88.25	43.04/47.10	53.76/56.21
	GNN_Contrastive	**99.91**/99.90	**99.95**/99.95	**98.89**/97.60	**99.40**/98.60	**96.11**/86.02	**97.52**/90.72
	GNN_NLL	94.42/95.73	96.90/97.73	43.77/48.58	54.75/60.43	17.14/20.79	25.67/30.64
3D Cubes	AE_Contrastive	68.37/69.69	76.65/78.64	19.49/18.79	31.29/31.00	7.35/5.72	**15.50**/13.61
	AE_NLL	78.90/**79.64**	85.18/**85.79**	26.45/**27.42**	36.38/**37.44**	7.54/**8.23**	13.04/13.82
	VAE_Contrastive	65.25/66.09	74.08/74.73	14.72/17.49	25.15/28.54	4.36/6.71	9.88/13.87
	VAE_NLL	55.44/52.58	61.86/58.36	10.69/9.36	15.21/12.97	2.27/1.45	4.19/2.85
	Modular_Contrastive	3.18/11.32	9.06/23.37	0.39/0.42	1.78/2.12	0.21/0.13	1.09/0.78
	Modular_NLL	64.77/65.09	73.26/73.52	14.94/15.88	23.17/24.33	2.95/3.49	6.52/7.76
	GNN_Contrastive	60.88/59.38	69.17/68.34	15.71/2.67	24.89/7.67	5.45/0.13	11.09/0.86
	GNN_NLL	59.61/60.13	67.00/67.44	12.23/13.34	19.04/20.39	2.66/3.06	5.71/6.59

Counterfactual retrieval comparison. Tab. 2 reports the counterfactual retrieval performance (CF-H@1 and CF-MRR) of all baseline families and their +CausalVAE variants across all benchmarks.

Table 2: Comparison on counterfactual retrieval metrics. Each cell reports Baseline/+CausalVAE; Δ is the change of +CausalVAE relative to the paired baseline (positive = gain), and "–" marks a non-positive change.

Benchmark	Method	CF-H@1 ↑	ΔH@1 ↑	CF-MRR ↑	ΔMRR ↑
Physics 3-body	AE_Contrastive	10.00/24.00	+14.00	49.33/51.77	+2.44
	AE_NLL	10.50/41.00	+30.50	52.25/66.92	+14.67
	VAE_Contrastive	8.50/8.50	–	46.00/45.40	–
	VAE_NLL	10.50/13.00	+2.50	51.75/51.00	–
	Modular_Contrastive	22.50/25.00	+2.50	59.83/50.62	–
	Modular_NLL	14.50/22.00	+7.50	55.83/57.50	+1.67
	GNN_Contrastive	11.50/15.00	+3.50	52.96/53.17	+0.21
	GNN_NLL	11.00/41.00	+30.00	53.50/68.88	+15.38
Chemistry	AE_Contrastive	52.61/35.87	–	59.47/55.81	–
	AE_NLL	35.91/34.34	–	55.82/54.12	–
	VAE_Contrastive	15.38/24.61	+9.23	33.53/48.24	+14.71
	VAE_NLL	15.50/25.26	+9.76	33.62/45.80	+12.18
	Modular_Contrastive	13.05/27.49	+14.44	33.22/48.67	+15.45
	Modular_NLL	28.30/29.56	+1.26	50.46/51.09	+0.63
	GNN_Contrastive	13.54/25.76	+12.22	32.96/47.66	+14.70
	GNN_NLL	25.96/23.70	–	48.44/45.20	–
2D Shapes	AE_Contrastive	53.10/52.53	–	70.97/70.86	–
	AE_NLL	65.75/67.82	+2.07	80.22/81.64	+1.42
	VAE_Contrastive	46.97/45.31	–	66.15/65.74	–
	VAE_NLL	19.70/19.40	–	41.76/43.11	+1.35
	Modular_Contrastive	13.85/8.55	–	32.39/22.92	–
	Modular_NLL	62.80/68.35	+5.55	78.46/81.88	+3.42
	GNN_Contrastive	67.06/68.50	+1.44	81.11/81.94	+0.83
	GNN_NLL	33.00/54.75	+21.75	58.81/74.00	+15.19
3D Cubes	AE_Contrastive	29.90/33.20	+3.30	53.02/55.40	+2.38
	AE_NLL	39.90/41.65	+1.75	60.41/63.11	+2.70
	VAE_Contrastive	27.15/27.65	+0.50	50.08/51.01	+0.93
	VAE_NLL	22.15/20.35	–	38.32/41.01	+2.69
	Modular_Contrastive	7.90/16.75	+8.85	19.07/33.80	+14.73
	Modular_NLL	30.60/30.90	+0.30	53.74/54.83	+1.09
	GNN_Contrastive	20.05/18.50	–	44.28/42.57	–
	GNN_NLL	28.45/25.90	–	48.15/48.73	+0.58

Results analysis. Across benchmarks (Tabs. 1 and 2), CausalVAE usually preserves factual retrieval while improving counterfactual retrieval. The strongest evidence is on Physics: CF-H@1 jumps from 11.0 to 41.0 for GNN_NLL, and AE_NLL gains +30.5. On Chemistry and 2D/3D, multiple backbones (VAE, Modular, GNN_Contrastive) still gain roughly +9 to +21 CF-H@1 points.

The pattern is therefore not a small average effect; it is a large improvement on intervention-focused metrics in several settings. At the same time, gains are not universal: some factual and counterfactual cells drop (e.g. Chemistry Modular_Contrastive in factual metrics), showing a clear backbone-domain interaction. This sharpens the takeaway: the plug-in is most valuable when baseline dynamics are intervention-fragile, but it should be selected per backbone rather than assumed to dominate uniformly.

When does the plug-in help? Whether the plug-in helps is governed by the capacity of the host backbone's latent to carry the underlying factors. This is clearest on Chemistry, where the largest drops occur: each scene has 5^5=3125 discrete global color configurations, yet the AE/VAE backbones compress it into a single 5-dimensional code, so the causal branch's encode–decode path cannot retain the class-discriminative directions and collapses them—accounting for the AE_C drop (−16.7 CF-H@1) and the VAE_NLL degradation (78.2→25.3). Widening the backbone latent removes this bottleneck: a latent-capacity study on Chemistry (supplementary material) shows the plug-in gain rising monotonically with the per-object dimension and turning positive once the code is wide enough to represent the discrete factors. The effect is therefore a capacity limitation of the host backbone rather than a fundamental incompatibility of the causal module—the plug-in helps when the latent can carry the factors, and hurts only when it is too narrow.

Comparison to published counterfactual-dynamics methods. Beyond the paired-backbone study, the supplementary material compares our plug-in against three published counterfactual-dynamics methods (CFWM, CoPhy, and Causal-JEPA) on Physics-3body under a matched protocol that we re-implement (since none reports on this benchmark); there, our ∼1M-parameter plug-in on GNN_C attains the best step-1 CF-H@1 and CF-MRR, ahead of substantially larger models.

4.5 Causal Analysis (Causal Discovery Results)

Beyond accuracy, we ask whether the learned structure is interpretable, comparing the learned matrix A to a physics-derived first-order template on Physics. Linearizing the Euler-discretized dynamics $s_{t+1} \approx s_t + \Delta t\, f(s_t)$ around a reference state gives a local Jacobian J, and we take $A_{\mathrm{GT}} := |J - I|$ as the interaction-strength template (full derivation in the supplementary material). At the mechanism-trend level—which factors influence which, and with what relative strength (Fig. 4)—A_{Learned} matches A_{GT}, indicating that the model recovers an interpretable first-order physical law rather than merely fitting predictions; the correspondence is local, so it is a first-order approximation, not a full nonlinear law.

4.6 Ablation Results

As summarized in Tab. 3, we ablate the components of our staged design on Physics: **three-stage** vs. **joint training**, the **GNN_Contrastive** backbone baseline, and one-at-a-time removal of the **CausalVAE branch**, the Stage-2 **alignment loss**, **multi-step rollout** supervision, and the **contrastive objective**, together with the **α-gated fusion** gate and sensitivity to stage-split and rollout-policy schedules.

Note. Each ablation changes one factor relative to **three-stage training (ours)** (stage1=20, stage2=80 epochs, late-mixed rollout, contrastive loss, gate on). Abbreviations: w/o = without; s1/s2 = stage1/stage2 epochs (e.g. s1_8_s2_40); CF-H@1/CF-MRR defined in Sec. 4.3.

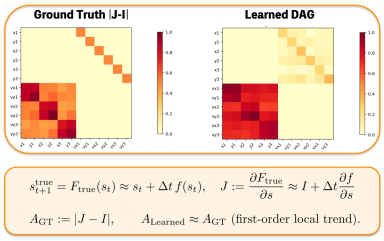

Fig. 4: Comparison between the learned structure A_{Learned} and the first-order physical template $A_{\text{GT}} = |J - I|$ obtained by local linearization. The learned structure recovers physically meaningful interaction trends in latent space, while remaining a first-order approximation rather than an exact nonlinear law.

Table 3: Ablations of training strategy, module removal, and hyperparameter sensitivity in the Physics benchmark.

Variant	H@1 ↑	MRR ↑	CF-H@1 ↑	CF-MRR ↑
three-stage training (ours)	**91.00**	**95.16**	31.00	52.15
joint training (w/o stage split)	90.33	94.75	28.00	51.05
Baseline (GNN_Contrastive)	88.00	93.37	11.50	**52.96**
w/o CausalVAE branch	80.00	89.55	27.67	50.06
w/o state align loss	9.00	28.02	26.67	50.57
single-step rollout	90.67	94.91	30.67	52.31
w/o contrastive loss	89.33	94.36	29.67	51.79
gate off	1.33	5.71	13.67	28.09
stage split: s1_8_s2_40	**91.00**	95.08	28.67	51.45
stage split: s1_12_s2_48	90.67	94.90	27.33	50.71
rollout policy: curriculum	**91.00**	95.06	**31.33**	52.32
rollout policy: mixed	90.00	94.75	29.67	51.88

Generality across backbones. To confirm these conclusions are not specific to GNN_C, we repeat the component ablation on all four backbones (VAE_C, AE_C, Modular_C, GNN_C) under the same protocol (full table in the supplementary material): the ranking full > noBranch > noAlign ≈ noStage3 ≫ noGate holds on every backbone, and the without-Stage-3 column gives a mean drop of −3.2, matching the 31→28 effect for GNN_C in Tab. 3.

5 Conclusion

We presented a plug-in CausalVAE layer that improves intervention-aware counterfactual behavior while keeping factual retrieval competitive across matched backbones, indicating that the gains come from added structural constraints rather than architecture replacement. This supports a modular path to robust world modeling—causal structure attached to existing predictive systems without redesigning the backbone—while the learned structural alignment adds interpretability without sacrificing forecasting. The added cost is training-time only (1.17× the GNN_C parameters; inference unchanged), and limitations remain in strongly nonlinear regimes and long-horizon drift; extending to larger JEPA-style architectures is natural future work [1, 2, 15].

Acknowledgements. This research is supported by Shenzhen Ubiquitous Data Enabling Key Lab under grant ZDSYS20220527171406015, and by Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Endowed Professorship Scheme of Shenzhen Pengrui Foundation. The computational resources for this work were provided by INFIFORCE Intelligent Technology.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15619–15629 (2023). https://doi.org/10.1109/cvpr52729.2023.01499
2. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=QaCCuDfBk2, last accessed 2026-06-26
3. Burgess, C.P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., Lerchner, A.: Monet: Unsupervised scene decomposition and representation. arXiv preprint arXiv:1901.11390 (2019)
4. Greff, K., Kaufman, R.L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., Lerchner, A.: Multi-object representation learning with iterative variational inference. In: International conference on machine learning. pp. 2424–2433. PMLR (2019)
5. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)
6. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603 (2019)
7. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering atari with discrete world models. In: International Conference on Learning Representations (ICLR) (2021), arXiv:2010.02193
8. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104 (2023)
9. Ke, N.R., Didolkar, A., Mittal, S., Goyal, A., Lajoie, G., Bauer, S., Rezende, D., Bengio, Y., Mozer, M., Pal, C.: Systematic evaluation of causal discovery in visual model based reinforcement learning. arXiv preprint arXiv:2107.00848 (2021)
10. Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A.: Variational autoencoders and nonlinear ica: A unifying framework. In: Chiappa, S., Calandra, R. (eds.) Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 108, pp. 2207–2217. PMLR (2020), https://proceedings.mlr.press/v108/khemakhem20a.html, last accessed 2026-06-26
11. Kipf, T., Fetaya, E., Wang, K.C., Welling, M., Zemel, R.: Neural relational inference for interacting systems. In: International conference on machine learning. pp. 2688–2697. Pmlr (2018)

12. Kipf, T., van der Pol, E., Welling, M.: Contrastive learning of structured world models. In: International Conference on Learning Representations (ICLR) (2020), arXiv:1911.12247
13. Li, M., Yang, M., Liu, F., Chen, X., Chen, Z., Wang, J.: Causal world models by unsupervised deconfounding of physical dynamics. arXiv preprint arXiv:2012.14228 (2020)
14. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in neural information processing systems **33**, 11525–11538 (2020)
15. Nam, H., Le Lidec, Q., Maes, L., LeCun, Y., Balestriero, R.: Causal-jepa: Learning world models through object-level latent masking. arXiv preprint arXiv:2602.11389 (2026)
16. Pearl, J.: Causality. Cambridge university press (2009). https://doi.org/10.1017/cbo9780511803161
17. Sánchez-González, Á., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., Battaglia, P.: Learning to simulate complex physics with graph networks. In: International conference on machine learning. pp. 8459–8468. PMLR (2020)
18. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021). https://doi.org/10.1109/jproc.2021.3058954
19. Venkatesh, R., Chen, H., Feigelis, K., Bear, D.M., Jedoui, K., Kotar, K., Binder, F., Lee, W., Liu, S., Smith, K.A., et al.: Understanding physical dynamics with counterfactual world modeling. In: European Conference on Computer Vision. pp. 368–387. Springer (2024). https://doi.org/10.1007/978-3-031-72691-0_21
20. Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., Wang, J.: Causalvae: Disentangled representation learning via neural structural causal models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9593–9602 (2021). https://doi.org/10.1109/cvpr46437.2021.00947
21. Zheng, X., Aragam, B., Ravikumar, P.K., Xing, E.P.: Dags with no tears: Continuous optimization for structure learning. Advances in neural information processing systems **31** (2018)