IDENTIFIABLE WORLD MODELS FROM PRETRAINED DIFFUSION REPRESENTATIONS *

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based world models can generate and predict trajectories in high-dimensional dynamical systems, but predictive accuracy does not imply that their latent coordinates recover the underlying state variables or causal interactions. We ask whether a frozen pretrained diffusion model can be equipped with identifiable coordinates without retraining its generative backbone. We show that auxiliary-variable nonlinear ICA guarantees can be transferred to Contrastive Diffusion Alignment (ConDA), which learns only a lightweight alignment map on top of frozen diffusion latents. Under standard TCL/GCL assumptions, the aligned representation identifies latent dynamical states up to permutation and componentwise invertible transformations, preserves the latent dynamic structural causal model, and reduces lagged graph recovery to transition-Jacobian sparsity. We evaluate TCL-, GCL-, and CEBRA-based ConDA against TDRL, CaRiNG, IDOL, temporal SuaVE, and iVAE across physical and robotic video systems. TCL and GCL achieve near-perfect blockwise state recovery and competitive lagged graph recovery, including exact recovery in a simulated falling-body system. In a simulated bipedal robot, learned dynamics recover the sign and temporal structure of responses to held-out control perturbations. These results show that a frozen generative diffusion model can be equipped with coordinates that are identifiable, structurally interpretable, and useful for analyzing intervention-relevant dynamics.

1 INTRODUCTION

Learning meaningful representations of high-dimensional dynamical systems is a central problem in scientific machine learning. Many scientific, biomedical, and robotic systems are governed by low-dimensional latent dynamics but are observed through high-dimensional measurements. Diffusion models provide powerful generative representations of such observations (Ho et al., 2020; Rombach et al., 2022), but generative fidelity alone does not imply that their internal representations recover the variables or interactions governing the underlying system. Indeed, multiple latent parameterizations can explain the same observations (Hyvärinen & Pajunen, 1999; Locatello et al., 2019; Khemakhem et al., 2020), and predictive representations need not preserve the causal structure required for intervention, control, or scientific reasoning (Bengio et al., 2013; Schölkopf et al., 2021). Contrastive learning, in turn, can exploit temporal or auxiliary information to recover structured low-dimensional representations (Hyvarinen & Morioka, 2016; Oord et al., 2018; Chen et al., 2020; Schneider et al., 2023). This raises a natural question: can a pretrained diffusion model retain its generative capacity while supporting an identifiable and controllable latent coordinate system that preserves causal structure?

Nonlinear ICA provides one route to identifiability through auxiliary variables such as time, conditions, interventions, or control inputs (Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020), but its extension to high-dimensional dynamical systems represented through diffusion latents remains largely unexplored. Building on Contrastive Diffusion Alignment (ConDA) (Sandilya et al., 2026), we study identifiability and causal graph recovery in diffusion-based world models. We distinguish three representation levels: the true low-dimensional dynamical state s_t, the high-dimensional diffusion latent x_t, and the observed frame y_t. Generatively, the system follows $(s_t \to x_t \to y_t)$. For identification, we invert the observation into diffusion space and learn an

*Code is provided in the supplementary material and will be released under the MIT license upon publication.

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.

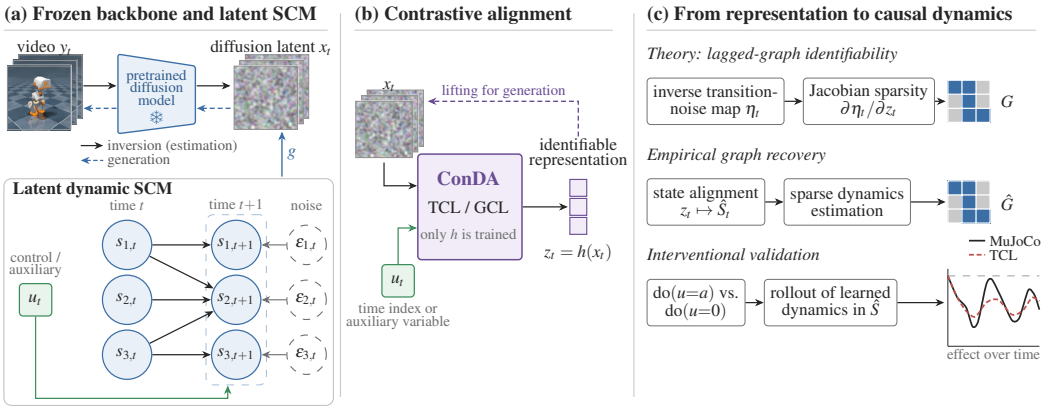

Figure 1: **Overview of the proposed framework.** (a) A frozen pretrained diffusion model serves as a generative world model: it embeds video frames y_t into diffusion latents x_t by inversion (solid arrows) and decodes latents back into frames (dashed arrows). The latents are in turn generated by a lagged dynamic SCM over the latent state s_t, driven by inputs u_t and independent noise, through $x_t = g(s_t)$. (b) TCL- or GCL-based ConDA trains only a lightweight map h, using temporal or auxiliary information, to obtain $z_t = h(x_t)$. Under nonlinear ICA assumptions, z_t identifies s_t up to permutation and componentwise invertible transformations, and can be lifted back to x_t for generation. (c) Lagged causal structure is characterized theoretically by inverse transition-noise Jacobian sparsity and estimated empirically from states aligned from z_t. Dynamics learned in these coordinates are validated by predicting responses to held-out control perturbations.

auxiliary-aligned embedding $z_t = h(x_t)$ ($y_t \to x_t \to z_t$). Identifiability then asks whether the learned coordinates z_t recover the underlying state s_t up to the admissible ambiguities of permutation and componentwise invertible transformations.

We use the standard density-ratio interpretation of supervised InfoNCE to connect auxiliary-aligned contrastive learning on diffusion latents with auxiliary-variable nonlinear ICA. Under the stated separability and identifiability assumptions, the learned embedding recovers the latent dynamical state up to permutation and componentwise invertible transformations. We further show that these coordinates preserve the factorization and parent structure of the latent dynamic SCM, and that the lagged causal graph is characterized by the sparsity of the inverse transition-noise Jacobian. In practice, under additive-noise dynamics, we estimate the corresponding graph using the sparsity of a learned forward-dynamics Jacobian. Empirically, we evaluate TCL-, GCL-, and CEBRA-based ConDA embeddings (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019; Schneider et al., 2023) on CausalVerse physical and robotic simulations with known ground-truth dynamics (Chen et al., 2025), as well as on Microduck, a controlled bipedal locomotion environment based on the official Pollen Robotics reinforcement-learning framework (Pollen Robotics, 2026). We assess whether these representations recover low-dimensional latent state structure and support lagged causal graph recovery; in Microduck, we additionally evaluate interventional prediction under held-out control perturbations. Overall, our framework evaluates world models beyond reconstruction and forecasting through latent identifiability, causal graph recovery, and preservation of intervention-relevant dynamics.

The role of diffusion in our framework is not to claim that diffusion latents are universally superior representations for state recovery. Rather, diffusion supplies a pretrained generative world model capable of synthesizing trajectories. Our question is whether this generative representation can be equipped, without retraining the backbone, with an identifiable coordinate system suitable for causal analysis and intervention. Accordingly, z_t is deliberately not a replacement generative representation: it is an identifiable coordinate chart on the frozen generative representation x_t.

Contributions. (1) We formulate identifiability and causal graph recovery directly in the latent space of a frozen pretrained diffusion model, separating the underlying dynamical state s_t, diffusion representation x_t, observed frame y_t, and auxiliary-aligned embedding $z_t = h(x_t)$. This decouples the generative world model from causal identification, as the diffusion backbone retains its capacity

to generate trajectories, while the learned alignment supplies identifiable coordinates for analyzing and controlling its dynamics. (2) We establish sufficient conditions under which TCL- or GCL-based alignment on frozen diffusion latents identifies the latent dynamical state up to permutation and componentwise invertible transformations, preserves the factorization and lagged parent structure of the latent dynamic SCM, and characterizes lagged causal edges through sparsity of the inverse transition-noise Jacobian. (3) We evaluate identifiability, lagged causal graph recovery, and interventional prediction across physical and robotic video dynamical systems. TCL- and GCL-based ConDA variants achieve strong, and in several settings near-perfect, blockwise state recovery and support recovery of lagged causal structure, including exact graph recovery in the simulated falling-body system. On controlled Microduck dynamics, we further evaluate whether the learned representations preserve responses to held-out control perturbations. Together, these experiments provide an evaluation of diffusion-based world models beyond reconstruction and forecasting, using state recovery, structural recovery, and intervention-relevant dynamics.

2 Related Work

World models and diffusion representations. World models learn latent representations that support prediction, planning, and control in high-dimensional environments (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2023). While effective for forecasting and decision making, these representations are not generally identifiable and need not recover the underlying physical state or causal dependencies required for mechanistic interpretation and intervention. Diffusion models provide powerful generative representations for high-dimensional data (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021), while latent diffusion models improve scalability by operating in compressed representation spaces (Rombach et al., 2022). However, diffusion representations are primarily optimized for generative fidelity rather than recovery of identifiable dynamical states or causal structure. We therefore study whether auxiliary-aligned embeddings of frozen pretrained diffusion representations can recover low-dimensional latent states and the lagged causal dependencies governing their dynamics.

Contrastive learning and nonlinear ICA. Contrastive learning learns low-dimensional representations by bringing positive pairs together while separating negative pairs (Oord et al., 2018; Chen et al., 2020; He et al., 2020). For dynamical systems, TCL exploits temporal nonstationarity to recover latent sources under nonlinear ICA assumptions (Hyvarinen & Morioka, 2016; 2017), while generalized contrastive learning (GCL) uses auxiliary variables to establish nonlinear ICA identifiability under broader forms of auxiliary information (Hyvarinen et al., 2019). CEBRA similarly learns low-dimensional embeddings aligned with behavioral, temporal, or neural variables (Schneider et al., 2023). More generally, nonlinear latent variables are not identifiable without additional assumptions (Hyvärinen & Pajunen, 1999; Locatello et al., 2019); auxiliary information such as time, labels, experimental conditions, or control inputs can provide the structure required for identifiability (Hyvarinen et al., 2019; Khemakhem et al., 2020). We build on these ideas by applying auxiliary-aligned contrastive objectives to frozen diffusion latents and using the density-ratio interpretation of contrastive learning to connect the learned embeddings to auxiliary-variable nonlinear ICA.

Causal representation learning and causal discovery. Causal representation learning seeks latent representations that support causal reasoning, intervention, and transfer (Bengio et al., 2013; Schölkopf et al., 2021). Standard learned representations may capture predictive or statistical structure without preserving the underlying causal mechanisms. Causal discovery, in turn, aims to recover directed dependencies from observational or interventional data (Spirtes et al., 2000; Pearl, 2009; Peters et al., 2017). In time-series and dynamical systems, temporal ordering and lagged dependencies provide additional structure for identifying directed interactions (Granger, 1969; Runge et al., 2019; Tank et al., 2021). However, many causal discovery methods assume access to meaningful state variables, whereas in high-dimensional systems the relevant dynamical variables may themselves be latent. We address this gap by studying when embeddings learned from diffusion representations recover identifiable latent states and preserve their lagged causal structure. Under our dynamic SCM assumptions, lagged causal edges are characterized by the sparsity pattern of the Jacobian of the inverse transition-noise map. This characterization is related to Jacobian-based causal discovery under nonlinear ICA (Reizinger et al., 2023), adapted here to auxiliary-aligned representations of frozen diffusion latents in dynamical systems.

Positioning of this work. Our work addresses a setting largely unexplored in causal representation learning: recovering identifiable and causally meaningful states from *frozen pretrained diffusion representations* of high-dimensional spatiotemporal systems. Rather than learning the encoder jointly with the temporal or causal model (Yao et al., 2022; Chen et al., 2024; Li et al., 2025), we keep the diffusion representation fixed and learn only a lightweight auxiliary-aligned mapping. Building on auxiliary-variable nonlinear ICA (Hyvarinen et al., 2019; Khemakhem et al., 2020), contrastive density-ratio estimation (Oord et al., 2018), and Jacobian-based causal discovery (Reizinger et al., 2023), we show that the resulting embeddings identify latent dynamical states up to permutation and componentwise invertible transformations, preserve the factorization and lagged parent structure of the dynamic SCM, and characterize lagged causal edges through inverse transition-noise Jacobian sparsity. Empirically, TCL- and GCL-based ConDA variants achieve strong, often near-perfect, blockwise identifiability and recover lagged causal structure across CausalVerse systems, while Microduck experiments show that the learned representations also preserve the direction and temporal structure of responses to held-out control perturbations. Together, these results position pretrained diffusion representations as a reusable substrate for identifiable, structurally faithful, and intervention-relevant world models of high-dimensional dynamical systems.

3 METHOD

This section develops the theoretical connection between identifiable representations learned from frozen diffusion latents, preservation of latent dynamical structure and lagged causal graph recovery. Proofs of the main theoretical statements are deferred to Appendix A.1-A.5.

Modeling Framework. We consider a latent dynamical system in which high-dimensional observations are generated from an underlying low-dimensional state:

$$s_{t+1} = F(s_t, u_t, \varepsilon_t), \qquad x_t = g(s_t), \qquad y_t = r(x_t), \tag{1}$$

where $s_t \in \mathbb{R}^d$ denotes the true dynamical state, u_t denotes observed auxiliary information or externally imposed controls, $x_t \in \mathbb{R}^p$ denotes the frozen diffusion representation, and y_t denotes the observed frame. Thus, the generative hierarchy is $s_t \to x_t \to y_t$. For identification and causal analysis, we learn a structured embedding $z_t = h(x_t)$, yielding the estimation pipeline $y_t \to x_t \to z_t$. The embedding z_t is not a generative layer; rather, it provides a low-dimensional coordinate system for identifiability and causal analysis. We assume that $g : \mathbb{R}^d \to \mathbb{R}^p$ is smooth and injective on the support of the latent state distribution, so that the diffusion representation preserves the information in s_t. The latent dynamics follow a dynamic structural causal model (SCM),

$$s_{i,t+1} = F_i\big(s_{\mathrm{Pa}(i),t}, s_{i,t}, u_t, \varepsilon_{i,t}\big)\,, \tag{2}$$

where $\mathrm{Pa}(i)$ denotes the lagged causal parents of $s_{i,t+1}$ and the noise variables $\varepsilon_{i,t}$ are mutually independent across coordinates conditional on u_t. Throughout, we include the self-lag in the parent set, $i \in \mathrm{Pa}(i)$.

Our objective is to learn $z_t = h(x_t)$ such that the underlying latent coordinates are identified up to permutation and componentwise invertible transformations, while preserving the dynamical structure required for graph recovery and intervention analysis.

Theorem 1 (Identifiability of nonlinear ICA with auxiliary variables). *Assume: (1) $x_t = g(s_t)$, where g is smooth and injective; (2) the latent process is first-order Markov conditional on u_t,*

$$p(s_{t+1} \mid s_{1:t}, u_{1:t}) = p(s_{t+1} \mid s_t, u_t); \tag{3}$$

(3) the transition density factorizes coordinatewise,

$$p(s_{t+1} \mid s_t, u_t) = \prod_{i=1}^{d} p_i(s_{i,t+1} \mid s_t, u_t); \tag{4}$$

(4) each conditional log-density has exponential-family form

$$\log p_i(s_{i,t+1} \mid s_t, u_t) = \sum_{k=1}^{K_i} \lambda_{i,k}(s_t, u_t) q_{i,k}(s_{i,t+1}) - \log Z_i(s_t, u_t), \tag{5}$$

with linearly independent sufficient statistics $q_{i,k}$; (5) the modulation coefficients $\{\lambda_{i,k}(s_t, u_t)\}_{i,k}$ vary with sufficient rank over the auxiliary contexts. Then the latent coordinates are identifiable from the diffusion representation up to permutation and componentwise invertible scalar transformations:

$$z_{i,t} = \phi_i(s_{\pi(i),t}), \qquad i = 1, \dots, d. \tag{6}$$

Contrastive Diffusion Alignment. We learn an encoder h on top of the frozen diffusion representation, $z_t = h(x_t)$, using temporal or auxiliary information. ConDA is therefore plug-and-play: the pretrained diffusion model remains fixed and only the alignment map h is learned. For a supervised contrastive formulation, positive pairs are selected using similarity in auxiliary context: $\mathcal{P}(i) = \{p \neq i : d(u_i, u_p) \leq \delta\}$ with negative samples corresponding to auxiliary contexts farther than δ. The supervised InfoNCE objective is

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\sum_{p\in\mathcal{P}(i)} \exp(\text{sim}(z_i, z_p)/\tau)}{\sum_{a\in\mathcal{A}(i)} \exp(\text{sim}(z_i, z_a)/\tau)}, \tag{7}$$

where $\tau > 0$ denotes the temperature parameter. The same framework accommodates different alignment objectives, including TCL-, GCL-, and CEBRA-based variants, which differ in how temporal or auxiliary structure defines positive and negative examples.

Lemma 1 (**Contrastive Density Ratio Estimation**). *Let positive pairs be sampled from matched auxiliary contexts and negative pairs from*

$$p_{\text{neg}}(x_t, x_t^- \mid u_t) = p(x_t \mid u_t)p(x_t^-). \tag{8}$$

If the score class is sufficiently expressive and the population optimum is attained, then

$$r^*(x_t, x_t^+, u_t) = \log \frac{p_{\text{pos}}(x_t, x_t^+ \mid u_t)}{p_{\text{neg}}(x_t, x_t^+ \mid u_t)} + C(u_t). \tag{9}$$

Lemma 1 provides the connection between auxiliary-aligned contrastive learning and auxiliary-variable nonlinear ICA. We next connect the density-ratio representation learned by contrastive alignment to the latent state coordinates.

Theorem 2 (**Identifiability Through Frozen Diffusion Representations**). *Assume that: (1) $x_t = g(s_t)$, where g is smooth and injective on the support of s_t, with a smooth inverse on its image; (2) the observed auxiliary variable, latent process, sampling scheme, and classifier satisfy the assumptions of the applicable TCL or GCL identifiability result (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019); and (3) any unmixing required by that result is performed. Then the specified TCL or GCL procedure applied to x_t identifies the latent state up to the equivalence class guaranteed by that result. If that result gives coordinatewise identification after any required unmixing, its output z_t satisfies*

$$z_{i,t} = \phi_i(s_{\pi(i),t}), \qquad i = 1, \dots, d, \tag{10}$$

for a permutation π and invertible scalar functions ϕ_i.

Interpretation. Theorem 2 is a sufficient-condition result. It does not imply that arbitrary contrastive training on diffusion representations automatically yields coordinate-wise identifiable variables. When the identification assumptions are only approximately satisfied, the learned representation may recover mixtures or blocks of underlying factors rather than individual coordinates. We therefore assess identifiability empirically using MCC and blockwise R^2, together with downstream graph recovery. A generic pairwise similarity objective such as Eq. (7), or the InfoNCE-based CEBRA variant, is not covered by this coordinatewise-identification claim.

Theorem 3 (**Preservation of dynamic SCM factorization**). *Assume the latent dynamics admit the structural factorization*

$$s_{i,t+1} = f_i\big(s_{\text{Pa}(i),t}, \varepsilon_{i,t}, u_t\big), \tag{11}$$

where $\text{Pa}(i)$ denotes the set of lagged parents of $s_{i,t+1}$, including i, and the noises $\varepsilon_{i,t}$ are mutually independent across i conditional on u_t. Suppose further that the output z_t of the procedure in Theorem 2, after any required unmixing, is coordinatewise identified, so that, after resolving the permutation ambiguity,

$$z_{i,t} = \phi_i(s_{i,t}), \tag{12}$$

for invertible scalar functions ϕ_i. Then the dynamics in the learned coordinates admit the factorization

$$z_{i,t+1} = \tilde{f}_i\big(z_{\mathrm{Pa}(i),t}, \varepsilon_{i,t}, u_t\big)\,, \tag{13}$$

where $\tilde{f}_i = \phi_i \circ f_i \circ \phi^{-1}_{\mathrm{Pa}(i)}$, with the inverse transformations applied componentwise to the parent coordinates. Consequently, the transformed dynamics preserve the same lagged parent sets and conditional factorization as the original dynamic SCM.

For the transformed structural model, let $\eta_t = T(z_t, z_{t+1}, u_t)$ denote the inverse transition-noise map.

Theorem 4 (**Lagged graph recovery**). *Assume that the representation is identifiable up to permutation and componentwise invertible transformations, each structural mechanism is smooth and bijective in its noise coordinate, the inverse map T is differentiable almost everywhere, and the model satisfies structural faithfulness. Then*

$$\frac{\partial \eta_{i,t}}{\partial z_{j,t}} = 0 \quad \Longleftrightarrow \quad j \notin \mathrm{Pa}(i) \tag{14}$$

almost everywhere. Hence, the sparsity pattern of the inverse transition-noise Jacobian recovers the lagged directed causal graph.

Forward versus inverse Jacobian. Our experiments estimate a forward transition model $\hat{f}_\theta(z_t, u_t) \approx \mathbb{E}[z_{t+1} \mid z_t, u_t]$. For additive-noise dynamics, $z_{i,t+1} = \tilde{f}_i(z_{\mathrm{Pa}(i),t}, z_{i,t}, u_t) + \eta_{i,t}$, we have

$$\frac{\partial \eta_{i,t}}{\partial z_{j,t}} = -\frac{\partial \tilde{f}_i}{\partial z_{j,t}}, \tag{15}$$

so the forward and inverse Jacobians have the same sparsity pattern. We therefore use forward-Jacobian magnitudes as empirical edge scores. For externally imposed controls, derivatives with respect to u_t quantify control-to-state dependencies separately from the endogenous state graph.

Corollary 1. *Under the assumptions of Theorems 2, 3, and 4, the learned representation identifies the latent coordinates up to permutation and componentwise invertible transformations, preserves the dynamic SCM factorization after permutation resolution, and reduces lagged graph recovery to recovery of the sparsity pattern of the inverse transition-noise Jacobian. For additive-noise dynamics, the same sparsity pattern is obtained from the forward transition Jacobian.*

4 Experiments

Experimental Setup. We evaluate identifiability and lag-1 causal graph recovery on video datasets from CausalVerse (Chen et al., 2025), which contain simulated physical and robotic systems with known ground-truth causal structure. Details of the corresponding reference graphs are provided in Appendices A.6, A.7, and A.8. We further evaluate identifiability, controlled dynamics graph recovery, and interventional prediction on videos generated in Microduck, a bipedal robotic locomotion environment based on the official Microduck RL framework (Pollen Robotics, 2026) and simulated in MuJoCo (Todorov et al., 2012). The Microduck dataset is generated using walking and standing policies and, unlike the CausalVerse Physical Fall and Robotics settings, explicitly includes externally imposed control commands. This enables us to assess whether learned representations recover not only the underlying physical state and its dynamical dependencies, but also the system's response to controlled interventions.

For CausalVerse Physical Fall and Microduck, we use 5000 videos, with 1000 held out for testing. For the CausalVerse Robotics Study, we use 2700 videos, with 540 held out for testing. We evaluate ConDA with three contrastive objectives: CEBRA (Schneider et al., 2023), GCL (Hyvarinen et al., 2019), and TCL (Hyvarinen & Morioka, 2016). We compare against five CRL baselines: IDOL (Li et al., 2025), CaRiNG (Chen et al., 2024), TDRL (Yao et al., 2022), temporal variant of SuaVE (Liu et al., 2026), and iVAE (Khemakhem et al., 2020). All methods are evaluated over five random seeds, and we report the mean and standard deviation across seeds.

Our pipeline consists of three stages. First, we train a conditional latent diffusion model (cLDM) on the video data. After training, the diffusion model is frozen and used only as a pretrained

representation model. Each video frame y_t is first encoded into the latent space of the diffusion model using the Stable Diffusion VAE (Rombach et al., 2022). We then apply Rex-RK4 (Blasingame & Liu, 2026) with 8 inversion steps to the frozen diffusion model to obtain the corresponding inverted diffusion latent x_t. Second, we learn ConDA representations $z_t = h(x_t)$ from these frozen diffusion latents. ConDA uses temporal or auxiliary information u_t through CEBRA, GCL, or TCL to recover structured latent coordinates. The resulting frame-level embeddings are then assembled into latent trajectories $\{z_t\}_{t=1}^{T}$ for downstream identifiability, dynamics, and causal analysis. Third, we evaluate latent-state recovery and lag-1 graph recovery from the aligned representations. For CausalVerse Physical Fall and Robotics, lagged dependencies are estimated using residual Lasso with trajectory-level bootstrap stability. For Microduck, we instead learn a controlled transition model $[\hat{S}_t, U_t] \mapsto \hat{S}_{t+1}$ and estimate state- and control-dependent edges from the average absolute transition Jacobian. Binary graphs are obtained using the reference parent count for each child, applied consistently across methods. This fixes graph density across methods and isolates recovery of the correct parent structure from the separate problem of choosing a sparsity threshold. We therefore interpret binary graph metrics as measuring edge recovery conditional on the true parent count, and additionally report AUPRC as a threshold-free measure of edge-ranking performance. Detailed graph-recovery procedures are given in Algorithms 1 and 2 (Appendix A.11).

Evaluation Metrics. We measure coordinate-wise identifiability by MCC, the mean absolute Pearson correlation between ground-truth state coordinates and learned embedding coordinates under optimal matching, and block-wise identifiability by R^2, the variance of ground-truth state blocks explained by nonlinear regression on z_t. For graph recovery, we report F1, accuracy, structural Hamming distance (SHD), AUPRC, and structural intervention distance (SID); formal definitions are provided in Appendix A.9. We evaluate interventional prediction in Microduck using 100 matched simulator trajectories. Starting from the same state at $t = 50$, we compare $do(u_x = \pm 0.25)$ and $do(u_y = \pm 0.17)$ with the zero-command baseline, holding the other commands at zero, and roll out each intervention for 25 steps (1 s at 25 Hz). The intervention values lie within the empirical training distribution (Fig. 4). For a command u_c and its target velocity v, we define

$$\mathrm{DIE}_v(k) = \mathbb{E}_i\big[v_{i,t+k}^{do(u_c=a)} - v_{i,t+k}^{do(u_c=0)}\big].$$

We compare the MuJoCo response with a true-state dynamics baseline trained on $[S_t, U_t]$ and a representation-based model trained on $[\hat{S}_t, U_t]$, where $\hat{S}_t$ is obtained from the learned representation; additional details are provided in Appendix A.10.

5 Results

Identifiability and Graph Recovery. Table 1 summarizes latent-state identifiability and lag-1 graph recovery across CausalVerse Physical Fall, CausalVerse Robotics Study, and Microduck. On Physical Fall, ConDA-TCL and ConDA-GCL achieve near-perfect block-wise recovery ($R^2 \approx 1$) and exact graph recovery (F1/AUPRC $= 1$, SHD/SID $= 0$), while iVAE attains the highest MCC (0.714). On the Robotics Study, ConDA-TCL and ConDA-GCL achieve the highest MCC among the ConDA variants (0.679 and 0.683) with $R^2 = 1.000$; iVAE performs slightly better on the graph-recovery metrics, whereas ConDA-TCL/GCL remain competitive. On Microduck, ConDA-GCL attains the highest MCC (0.629), while ConDA-TCL gives the strongest graph recovery among the reported methods in F1 (0.769), accuracy (0.888), SHD (0.111), and SID (0.076), with AUPRC comparable to iVAE. Overall, ConDA-TCL and ConDA-GCL provide consistently strong performance across both state recovery and lag-1 dependency recovery, although no single method dominates every metric and dataset.

Figure 2 visualizes the corresponding binary adjacency matrices. The exact recovery obtained by ConDA-TCL and ConDA-GCL on Physical Fall is clearly visible, while the Robotics and Microduck results show that the ConDA variants preserve much of the sparse reference structure despite several remaining false positive or missed edges.

Block-wise vs. Coordinate-wise Recovery. Across the three datasets, ConDA-TCL and ConDA-GCL exhibit substantially stronger block-wise R^2 than coordinate-wise MCC. This indicates that the learned representations retain nearly all information needed to reconstruct the underlying state, even when individual latent coordinates are not in one-to-one correspondence with physical state variables. The graph-recovery results further show that near-perfect block-wise recovery can be

Table 1: **Identifiability and lag-1 graph recovery across dynamical systems.** Mean ± standard deviation across seeds are reported for CausalVerse Physical Fall, CausalVerse Robotics Study, and Microduck. MCC and R^2 quantify latent-state recovery, while F1, accuracy, SHD, AUPRC, and SID measure recovery of the lag-1 dynamics graph. ConDA-TCL and ConDA-GCL show consistently strong performance across both identifiability and graph recovery, with exact graph recovery on Physical Fall and competitive performance on the Robotics Study and Microduck. Arrows indicate whether higher (↑) or lower (↓) values are preferred.

Method	MCC ↑	R^2 ↑	F1 ↑	Acc. ↑	SHD ↓	AUPRC ↑	SID ↓
CausalVerse Physical Fall							
CaRiNG	0.129 ± 0.013	0.497 ± 0.016	0.800 ± 0.093	0.900 ± 0.046	0.400 ± 0.185	0.743 ± 0.147	0.047 ± 0.021
IDOL	0.307 ± 0.014	0.418 ± 0.005	0.733 ± 0.061	0.866 ± 0.030	0.533 ± 0.121	0.618 ± 0.060	0.062 ± 0.014
TDRL	0.293 ± 0.045	0.480 ± 0.017	0.755 ± 0.049	0.877 ± 0.024	0.488 ± 0.099	0.676 ± 0.100	0.057 ± 0.011
SuaVE	0.438 ± 0.014	0.600 ± 0.020	0.711 ± 0.060	0.855 ± 0.030	0.577 ± 0.121	0.580 ± 0.072	0.068 ± 0.014
iVAE	**0.714 ± 0.004**	0.998 ± 0.000	0.977 ± 0.049	0.988 ± 0.024	0.044 ± 0.099	0.983 ± 0.037	0.005 ± 0.011
ConDA-CEBRA	0.538 ± 0.006	0.759 ± 0.001	0.800 ± 0.093	0.900 ± 0.046	0.400 ± 0.185	0.696 ± 0.123	0.047 ± 0.021
ConDA-TCL	0.694 ± 0.001	0.998 ± 0.000	**1.000 ± 0.000**	**1.000 ± 0.000**	**0.000 ± 0.000**	**1.000 ± 0.000**	**0.000 ± 0.000**
ConDA-GCL	0.700 ± 0.002	**0.999 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**0.000 ± 0.000**	**1.000 ± 0.000**	**0.000 ± 0.000**
CausalVerse Robotics Study							
CaRiNG	0.509 ± 0.005	0.749 ± 0.003	0.874 ± 0.006	0.988 ± 0.000	0.225 ± 0.014	0.786 ± 0.013	0.010 ± 0.000
IDOL	0.412 ± 0.001	0.650 ± 0.003	0.877 ± 0.000	0.988 ± 0.000	0.218 ± 0.000	0.792 ± 0.000	0.010 ± 0.000
TDRL	0.481 ± 0.018	0.740 ± 0.002	0.853 ± 0.012	0.986 ± 0.001	0.268 ± 0.028	0.751 ± 0.024	0.010 ± 0.000
SuaVE	0.574 ± 0.010	0.768 ± 0.002	0.858 ± 0.019	0.985 ± 0.002	0.275 ± 0.040	0.771 ± 0.023	0.007 ± 0.000
iVAE	0.678 ± 0.003	**1.000 ± 0.000**	**0.934 ± 0.016**	**0.993 ± 0.001**	**0.125 ± 0.031**	**0.898 ± 0.016**	**0.005 ± 0.000**
ConDA-CEBRA	0.587 ± 0.002	0.753 ± 0.003	0.889 ± 0.011	0.989 ± 0.000	0.206 ± 0.017	0.815 ± 0.023	0.007 ± 0.001
ConDA-TCL	0.679 ± 0.005	**1.000 ± 0.000**	0.917 ± 0.011	0.992 ± 0.001	0.156 ± 0.022	0.875 ± 0.013	0.006 ± 0.000
ConDA-GCL	**0.683 ± 0.002**	**1.000 ± 0.000**	0.911 ± 0.026	0.991 ± 0.002	0.168 ± 0.052	0.869 ± 0.016	0.006 ± 0.000
Microduck							
CaRiNG	0.075 ± 0.025	0.323 ± 0.010	0.569 ± 0.087	0.792 ± 0.042	0.207 ± 0.042	0.703 ± 0.017	0.099 ± 0.031
IDOL	0.374 ± 0.003	0.357 ± 0.004	0.615 ± 0.094	0.814 ± 0.045	0.185 ± 0.045	0.739 ± 0.075	0.096 ± 0.033
TDRL	0.377 ± 0.010	0.383 ± 0.003	0.584 ± 0.116	0.800 ± 0.056	0.200 ± 0.056	0.713 ± 0.033	0.097 ± 0.045
SuaVE	0.378 ± 0.029	0.390 ± 0.004	0.584 ± 0.042	0.800 ± 0.020	0.200 ± 0.020	0.714 ± 0.053	0.124 ± 0.008
iVAE	0.626 ± 0.002	**1.000 ± 0.000**	0.753 ± 0.034	0.881 ± 0.016	0.118 ± 0.016	0.798 ± 0.026	0.081 ± 0.009
ConDA-CEBRA	0.493 ± 0.001	0.695 ± 0.006	0.507 ± 0.042	0.763 ± 0.020	0.237 ± 0.020	0.682 ± 0.007	0.123 ± 0.010
ConDA-TCL	0.628 ± 0.001	**1.000 ± 0.000**	**0.769 ± 0.054**	**0.888 ± 0.026**	**0.111 ± 0.026**	**0.796 ± 0.011**	**0.076 ± 0.017**
ConDA-GCL	**0.629 ± 0.000**	**1.000 ± 0.000**	0.738 ± 0.042	0.874 ± 0.020	0.125 ± 0.020	0.786 ± 0.032	0.085 ± 0.011

sufficient for accurate recovery of lagged dependency structure, although exact coordinate alignment is not required in every setting.

Dynamic Intervention Effect. We further evaluated whether a physical-state readout of the learned representation predicts responses to held-out control perturbations. Figure 3 compares dynamic intervention effects predicted from the TCL-aligned representation with matched MuJoCo rollouts and a true-state dynamics baseline. For the backward-velocity intervention, $do(u_x = -0.25)$, TCL recovers the correct sign and temporal trend, although the response magnitude is attenuated relative to MuJoCo. For the lateral intervention, $do(u_y = -0.17)$, TCL follows the simulator response more closely over the rollout and captures the dominant negative effect on v_y. The true-state baseline generally tracks the transient MuJoCo response more closely, whereas the representation-based model underestimates the immediate effect. Overall, the learned representation preserves intervention direction and much of the longer-horizon dynamics, while exhibiting a smoother and attenuated short-term response. Additional intervention results are provided in Appendix A.10.1.

6 Conclusion and Limitations

We studied when frozen pretrained diffusion representations can support identifiable and causally meaningful world models for high-dimensional dynamical systems. Building on ConDA, we showed that, under stated assumptions, TCL- or GCL-based alignment can recover latent dynamical states up to permutation and componentwise invertible transformations, preserve the factorization and lagged parent structure of the dynamic SCM, and characterize lagged causal dependencies through inverse transition-noise Jacobian sparsity. Empirically, TCL- and GCL-based variants achieved

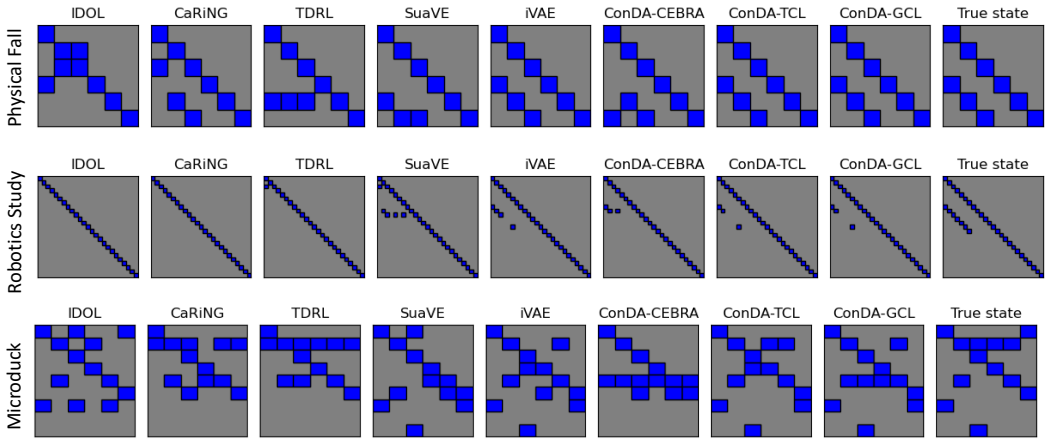

Figure 2: **Recovered lag-1 dynamics graphs across datasets and methods.** Binary adjacency matrices are shown for IDOL, CaRiNG, TDRL, temporal SuaVE, iVAE, ConDA-CEBRA, ConDA-TCL, ConDA-GCL, and the true-state reference. Blue entries indicate recovered edges and gray entries indicate absent edges. Rows correspond to the different dynamical systems considered in the study. The ConDA variants more closely reproduce the sparsity structure of the true-state dynamics, with TCL and GCL showing particularly strong agreement across settings.

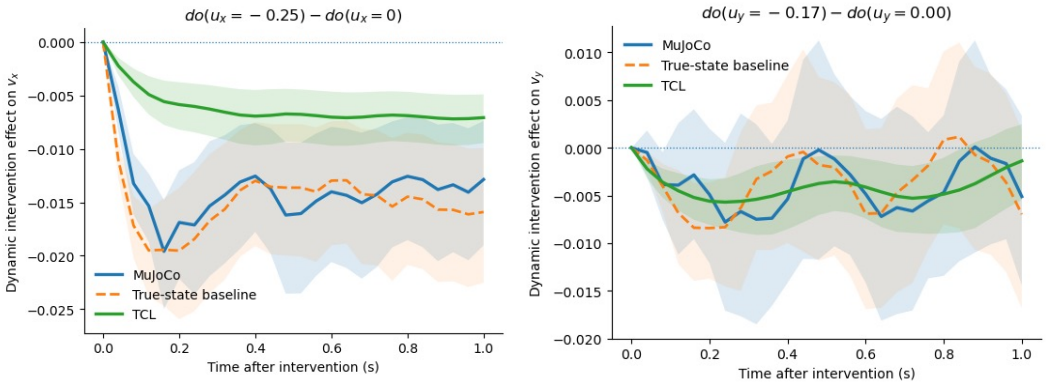

Figure 3: **Dynamic intervention effects under controlled Microduck perturbations.** Mean dynamic intervention effect over time for interventions on forward velocity command u_x (left) and lateral velocity command u_y (right). The curves compare the MuJoCo simulator response, the true-state dynamics baseline, and the TCL-aligned representation. Shaded regions denote 95% bootstrap confidence intervals across matched intervention trajectories. The left panel shows the effect of $do(u_x = -0.25)$ relative to $do(u_x = 0)$ on v_x, while the right panel shows the effect of $do(u_y = -0.17)$ relative to $do(u_y = 0)$ on v_y.

strong blockwise identifiability and supported lagged graph recovery across physical and robotic systems, while Microduck experiments showed preservation of intervention direction and temporal response structure under held-out controls. The theoretical guarantees are population-level and rely on informative auxiliary variables, separability, and regularity assumptions; empirical graph recovery also uses alignment to known state coordinates and, for forward-Jacobian estimation, additive-noise assumptions. Our intervention studies are limited to controlled simulation and open-loop perturbations. Extending the framework to partially observed, higher-order, nonstationary, and real-world scientific systems remains an important direction for future work.

AI USE STATEMENT

Generative AI tools were used to assist with code development and debugging, interpretation of empirical results, and preparation and editing of the manuscript. They were also used for literature-related queries, improving presentation and readability, and refining figures, captions, and LaTeX. AI tools were not used to generate the experimental datasets or to autonomously conduct the research. All theoretical claims, experimental designs, code, numerical results, citations, and AI-assisted text were reviewed and verified by the authors; generated code was tested against the corresponding experiments, and mathematical arguments were checked by the authors before inclusion. The authors take full responsibility for the final content of this work, including all text, claims, code, analyses, and artifacts produced with the assistance of generative AI.

REFERENCES

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Zander W. Blasingame and Chen Liu. Rex: A family of reversible exponential (stochastic) Runge-Kutta solvers. *Forty-third International Conference on Machine Learning (ICML)*, 2026.

Gustau Camps-Valls, Andreas Gerhardus, Urmi Ninad, Gherardo Varando, Georg Martius, Emili Balaguer-Ballester, Ricardo Vinuesa, Emiliano Diaz, Laure Zanna, and Jakob Runge. Discovering causal relations and equations from data. *Physics Reports*, 1044:1–68, 2023.

Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. CaRiNG: Learning temporal causal representation under non-invertible generation process. *arXiv preprint arXiv:2401.14535*, 2024.

Guangyi Chen, Yunlong Deng, Peiyuan Zhu, Yan Li, Yifan Shen, Zijian Li, and Kun Zhang. CausalVerse: Benchmarking causal representation learning with configurable high-fidelity simulations. In *NeurIPS*, 2025. URL `https://huggingface.co/CausalVerse`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, pp. 1597–1607. PmLR, 2020.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

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 (ICML)*, pp. 2555–2565. PMLR, 2019.

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

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9729–9738, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *NeurIPS*, 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Artificial intelligence and statistics*, pp. 460–469. PMLR, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868. PMLR, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the identification of temporally causal representation with instantaneous dependence. *International Conference on Learning Representations (ICLR)*, 2025.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifying weight-variant latent causal models. *Journal of Machine Learning Research*, 27(4):1–49, 2026.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning (ICML)*, pp. 4114–4124. PMLR, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Pollen Robotics. Microduck rl: Reinforcement learning environments for microduck. `https://github.com/pollen-robotics/microduck_rl`, 2026.

Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár, and Wieland Brendel. Jacobian-based causal discovery with nonlinear ICA. *Transactions on Machine Learning Research (TMLR)*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10684–10695, 2022.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5 (11):eaau4996, 2019.

Ruchi Sandilya, Sumaira Perez, Charles Lynch, Lindsay Victoria, Benjamin Zebley, Derrick Matthew Buchanan, Mahendra T. Bhati, Nolan Williams, Timothy J. Spellman, Faith M. Gunning, Conor Liston, and Logan Grosenick. Contrastive Diffusion Alignment: Learning structured latents for controllable generation. *Forty-third International Conference on Machine Learning (ICML)*, 2026.

Steffen Schneider, Jin H Lee, and Mackenzie W Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 2023. doi: https://doi.org/10.1038/s41586-023-06031-6.

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.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning (ICML)*, pp. 2256–2265. pmlr, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*, 2021.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, et al. Causal representation learning from multi-modal biomedical observations. *International Conference on Learning Representations (ICLR)*, pp. arXiv–2411, 2025.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:71705–71736, 2024.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *NeurIPS*, 35:26492–26503, 2022.

DATASETS AND CODE ACCESS INFORMATION

All code used to generate the results presented in this work will be released as open source under the MIT License. The benchmark dataset is publicly available through Hugging Face at `https://huggingface.co/CausalVerse`

A APPENDIX

A.1 IDENTIFIABILITY OF LATENT STATES FROM AUXILIARY-VARIABLE NONLINEAR ICA

Proof Sketch of Theorem 1. Under conditional independence and exponential-family assumptions, the latent variables are identifiable up to permutation and componentwise invertible transformations Hyvarinen & Morioka (2016); Hyvarinen et al. (2019); Khemakhem et al. (2020). □

A.2 CONTRASTIVE LEARNING AS DENSITY RATIO ESTIMATION

Proof of Lemma 1. At the population level, the supervised contrastive objective equation 7 can be interpreted as a classification problem that distinguishes matched pairs drawn from p_{pos} from mismatched pairs drawn from p_{neg}. For a sufficiently expressive score class and in the large-negative-sample limit, the Bayes-optimal score is the log-likelihood ratio between the positive and negative-pair distributions, up to an additive term depending only on the conditioning variable u_t. Therefore,

$$r^*(x_t, x_t^+, u_t) = \log \frac{p_{\text{pos}}(x_t, x_t^+ \mid u_t)}{p_{\text{neg}}(x_t, x_t^+ \mid u_t)} + C(u_t).$$

Substituting $p_{\text{neg}}(x_t, x_t^+ \mid u_t) = p(x_t \mid u_t)p(x_t^+)$ gives the stated expression. □

A.3 IDENTIFIABILITY THROUGH FROZEN DIFFUSION REPRESENTATIONS

Proof Sketch of Theorem 2. Because g has a smooth inverse on its image, observing $x_t = g(s_t)$ preserves the state information available to the specified TCL or GCL procedure. Under assumptions (2)–(3), the applicable identification result yields its stated equivalence class when applied to x_t. If that class is coordinatewise after any required unmixing, its output satisfies Eq. (10).

□

A.4 CAUSAL REPRESENTATION PRESERVATION

Proof Sketch of Theorem 3. By Theorem 2, after resolving the permutation ambiguity there exist invertible scalar functions ϕ_i such that

$$z_{i,t} = \phi_i(s_{i,t}), \qquad i = 1, \dots, d.$$

Substituting $s_{i,t} = \phi_i^{-1}(z_{i,t})$ into the latent SCM

$$s_{i,t+1} = f_i(s_{\text{Pa}(i),t}, s_{i,t}, \varepsilon_{i,t}, u_t)$$

yields

$$z_{i,t+1} = \phi_i\Big(f_i\big(\phi_{\text{Pa}(i)}^{-1}(z_{\text{Pa}(i),t}), \phi_i^{-1}(z_{i,t}), \varepsilon_{i,t}, u_t\big)\Big).$$

This defines a transformed mechanism $\tilde{f}_i$ depending on the same parent coordinates. Because the transformation is componentwise and invertible, it does not mix coordinates and therefore preserves the parent sets. Independence of the noise coordinates is preserved under componentwise reparameterization, yielding the stated transformed SCM. □

A.5 JACOBIAN-BASED GRAPH RECOVERY

Proof Sketch of Theorem 4. Because each mechanism $\tilde{f}_i$ is bijective in its own noise coordinate $\eta_{i,t}$, the structural equations can be inverted to express each $\eta_{i,t}$ as a function $T_i(z_t, z_{t+1}, u_t)$. By construction, T_i depends only on those coordinates of z_t that appear in the corresponding structural

equation for $z_{i,t+1}$, namely $z_{i,t}$ and $z_{j,t}$ for $j \in \mathrm{Pa}(i)$. Thus if $j \notin \mathrm{Pa}(i)$, then $\eta_{i,t}$ does not depend on $z_{j,t}$ and

$$\frac{\partial \eta_{i,t}}{\partial z_{j,t}} = 0.$$

Conversely, if $j \in \mathrm{Pa}(i)$, then by structural faithfulness this dependence does not cancel almost everywhere, so

$$\frac{\partial \eta_{i,t}}{\partial z_{j,t}} \neq 0$$

on a set of positive measure. This establishes the equivalence. □

A.6 Causal Graph for Physical Fall

We first evaluate our framework on the Physical Fall video dataset from CausalVerse (Chen et al., 2025). This dataset simulates objects undergoing free-fall motion under gravity. The system evolves in discrete time, and the underlying state variables describe the object's position and velocity.

For each spatial coordinate, the temporal transition dynamics are given by

$$v_{t+1} = v_t + g\Delta t,$$
$$h_{t+1} = h_t + v_t\Delta t + \tfrac{1}{2}g(\Delta t)^2,$$

where h_t denotes position or height at time t, v_t denotes velocity, g is gravitational acceleration, and Δt is the simulation time step. For the three-dimensional setting, the true latent state is

$$s_t = [h_{1,t}, h_{2,t}, h_{3,t}, v_{1,t}, v_{2,t}, v_{3,t}].$$

The corresponding dynamic causal graph contains three types of directed time-lagged dependencies:

$$h_t \to h_{t+1}, \quad v_t \to v_{t+1}, \quad v_t \to h_{t+1}.$$

Thus, each future position depends on its current position and velocity, while each future velocity depends on its current velocity. We use gravity and time as auxiliary variables. Because the transition mechanism is known analytically, the Physical Fall setting provides a controlled benchmark for evaluating both source identifiability, through alignment between the learned representations and the true latent states, and dynamic causal graph recovery, through comparison of the estimated lagged adjacency matrix with the known graph.

A.7 Physics-informed reference lag-1 dynamics graph for Robotics Study

We then evaluate our framework on the Robotics video dataset from CausalVerse (Chen et al., 2025), which consists of simulated robotic manipulation scenes in which a robot interacts with objects over time. Compared with the Physical Fall setting, this dataset involves a higher-dimensional state space and more complex temporal dependencies arising from joint motion, end-effector dynamics, and gripper motion.

For each video, we construct a 25-dimensional robot state comprising joint positions and velocities, end-effector position and orientation, and gripper position and velocity:

$$s_t = \left[q_t,\, \dot{q}_t,\, p_t^{\mathrm{eef}},\, Q_t^{\mathrm{eef}},\, g_t,\, \dot{g}_t\right],$$

where $q_t, \dot{q}_t \in \mathbb{R}^7$ denote the robot joint positions and velocities, $p_t^{\mathrm{eef}} \in \mathbb{R}^3$ denotes the end-effector position, $Q_t^{\mathrm{eef}} \in \mathbb{R}^4$ its quaternion orientation, and $g_t, \dot{g}_t \in \mathbb{R}^2$ denote the gripper position and velocity, respectively.

We define a physics-informed reference lag-1 dynamics graph in which every state variable has a self-transition and each joint velocity additionally influences the corresponding joint position at the next time step:

$$q_{k,t} \to q_{k,t+1}, \qquad \dot{q}_{k,t} \to q_{k,t+1}, \qquad \dot{q}_{k,t} \to \dot{q}_{k,t+1}, \quad k = 1, \ldots, 7,$$

together with

$$p_{j,t}^{\mathrm{eef}} \to p_{j,t+1}^{\mathrm{eef}}, \qquad Q_{j,t}^{\mathrm{eef}} \to Q_{j,t+1}^{\mathrm{eef}}, \qquad g_{j,t} \to g_{j,t+1}, \qquad \dot{g}_{j,t} \to \dot{g}_{j,t+1}.$$

This yields 25 self-lag edges and seven cross-lag edges, for a total of 32 reference dynamic edges. We use time, object position and object velocity as auxiliary information for representation learning.

A.8 Controlled Causal Dynamics for Microduck

We further evaluate our framework on Microduck, a bipedal robotic locomotion environment implemented using the official Microduck RL framework (Pollen Robotics, 2026) and simulated in MuJoCo (Todorov et al., 2012). Unlike the Physical Fall and CausalVerse Robotics settings, Microduck explicitly incorporates externally imposed control commands, allowing us to evaluate whether learned representations capture not only the underlying physical state but also its response to interventions.

For each trajectory, we define a six-dimensional physical state

$$s_t = [v_{x,t}, v_{y,t}, \omega_{z,t}, h_t, \phi_t, \theta_t],$$

where $v_{x,t}$ and $v_{y,t}$ denote the forward and lateral body velocities, $\omega_{z,t}$ is the body yaw rate, h_t is the trunk height, and ϕ_t and θ_t denote body roll and pitch, respectively. The robot is driven by a three-dimensional external command

$$u_t = [u_{x,t}, u_{y,t}, u_{\omega,t}],$$

corresponding to commanded forward velocity, lateral velocity, and yaw rate. The command u_t is applied before recording the state at time t and is held fixed while the simulator advances to $t+1$. We therefore model the controlled transition mechanism as $s_{t+1} = F(s_t, u_t)$, and recover lagged dependencies of the form $[s_t, u_t] \to s_{t+1}$. This controlled formulation is also the basis for comparing graph recovery in the true simulator state and in the learned representation.

The control variables have direct physical interpretations. In particular, commanded forward and lateral velocities are expected to influence the corresponding future body velocities, while the commanded yaw rate influences the future body yaw rate:

$$u_{x,t} \to v_{x,t+1}, \qquad u_{y,t} \to v_{y,t+1}, \qquad u_{\omega,t} \to \omega_{z,t+1}.$$

The remaining state variables interact through the robot's locomotion dynamics, including coupling between translational velocity, body attitude, and trunk height. Rather than imposing an arbitrary sparse causal graph, we determine the reference lag-1 structure from the simulator state dynamics and controlled perturbations, using graph recovery from the true state as an oracle benchmark.

A.9 Evaluation Metric Definitions

To evaluate identifiability and causal graph recovery, we use both component-level and block-level metrics. Let $s_t \in \mathbb{R}^d$ denote the ground-truth state at time t, and let $z_t \in \mathbb{R}^k$ denote the learned representation. For a trajectory of length T, we stack these quantities as $S = [s_1, \ldots, s_T]^\top$ and $Z = [z_1, \ldots, z_T]^\top$.

Mean Correlation Coefficient (MCC). We use MCC to measure component-wise identifiability between the ground-truth state dimensions and the learned representation dimensions. We first compute the absolute Pearson correlation matrix $R_{ij} = |\mathrm{corr}(S_i, Z_j)|$, then select an injective matching $\pi : \{1, \ldots, d\} \to \{1, \ldots, k\}$ maximizing the total matched correlation: $\pi^\star = \arg\max_\pi \sum_{i=1}^d R_{i,\pi(i)}$. Finally, the MCC value is defined as $\mathrm{MCC}(S, Z) = \frac{1}{d}\sum_{i=1}^d R_{i,\pi^\star(i)}$.

Coefficient of Determination (R^2). Block-wise R^2 measures the proportion of variance in a ground-truth state block s_b that can be explained by the learned representation z. Formally, $R^2 = 1 - \frac{\mathrm{Var}(s_b - \mathrm{Reg}(z))}{\mathrm{Var}(s_b)}$, where Reg is a regression function, either linear or nonlinear, trained to predict the ground-truth state block s_b from the learned representation z. A value of $R^2 = 1$ indicates perfect block-wise predictability, meaning that the state block can be fully recovered from the learned representation.

Causal Graph Recovery. For causal graph recovery, we compare the estimated adjacency matrix $\hat{B}$ with the ground-truth adjacency matrix B, where $B_{ij} = 1$ denotes a directed lagged edge from variable i at time t to variable j at time $t+1$. For threshold-based metrics, $\hat{B}$ is binarized and we compute true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) over all candidate edges. We report

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}, \qquad \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$

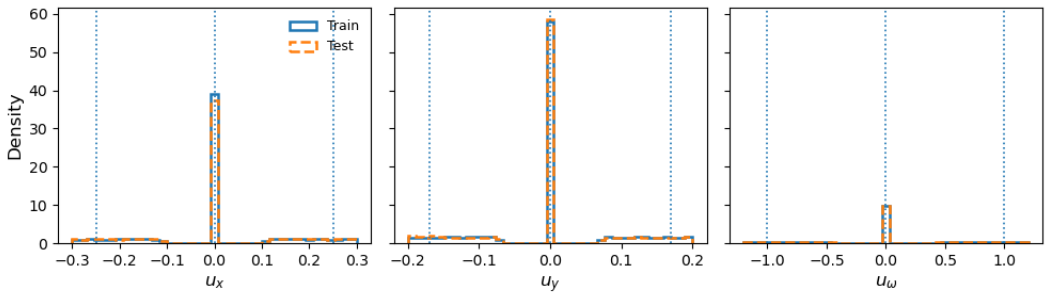

Figure 4: **Distribution of Microduck control commands.** Empirical distributions of the forward-velocity command u_x, lateral-velocity command u_y, and yaw-rate command u_ω in the training and test sets. The distributions closely overlap across the two splits. Dotted vertical lines indicate the negative, zero, and positive command values used in the intervention experiments.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{F1} = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}}.$$

To evaluate graph recovery independently of a single threshold, we additionally report the area under the precision–recall curve (AUPRC), computed from the continuous edge scores. We report structural Hamming distance (SHD), which quantifies structural disagreement between $\hat{B}$ and B through missing, extra, and incorrectly oriented edges; lower values indicate better recovery, with $\text{SHD} = 0$ corresponding to exact structural recovery. We also report structural intervention distance (SID), which measures disagreement between the estimated and ground-truth graphs in the interventional distributions implied by their graph structures; lower SID indicates greater agreement in causal effects, with $\text{SID} = 0$ indicating no structural intervention errors. We report SHD and SID normalized by the number of candidate edges. Figure 2 shows the binarized adjacency matrices, with recovered edges in blue.

A.10 Dynamic Intervention Effect

To assess whether the learned representation supports prediction under controlled interventions, we perform matched interventional rollouts in the Microduck environment. For each of 100 trajectories, the simulator is reconstructed to the same state at intervention time $t = 50$. We then compare $do(u_x = \pm 0.25)$ and $do(u_y = \pm 0.17)$ with the corresponding zero-command baseline, holding the remaining commands at zero. The intervention values lie within the empirical support of the training control distribution (Fig. 4).

Each intervention is propagated for 25 simulation steps, corresponding to 1 s at 25 Hz. For a command $u_c \in \{u_x, u_y\}$ and state component s_j, we define the dynamic intervention effect (DIE) at horizon k as

$$\text{DIE}_j(k; a, b) = \mathbb{E}_i \left[s_{i,t+k,j}^{do(u_c=a)} - s_{i,t+k,j}^{do(u_c=b)} \right], \tag{16}$$

where the expectation is over matched trajectories i. We evaluate the direct effects of u_x on v_x and of u_y on v_y.

We compare the MuJoCo simulator response with two learned dynamics models: (i) a true-state baseline trained directly on the physical state, $[S_t, U_t] \longrightarrow S_{t+1}$, and (ii) a representation-based model in which the learned latent representation is first aligned to the physical state, $Z_t \longrightarrow \hat{S}_t$, after which the controlled dynamics are modeled as $[\hat{S}_t, U_t] \longrightarrow \hat{S}_{t+1}$.

For a rollout horizon H and target velocity $v \in \{v_x, v_y\}$, we quantify intervention prediction error as

$$\text{RMSE}_H = \sqrt{\frac{1}{H} \sum_{k=1}^{H} \left(\widehat{\text{DIE}}_v(k) - \text{DIE}_v^{\text{sim}}(k) \right)^2}, \tag{17}$$

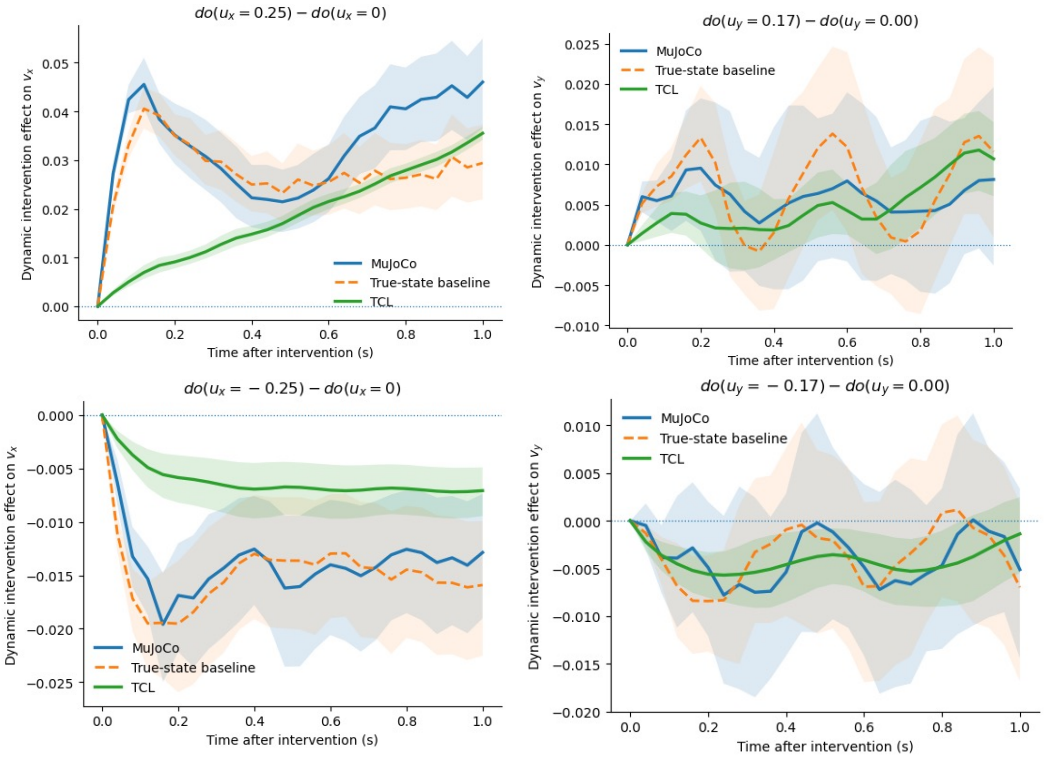

Figure 5: **Dynamic intervention effects under positive and negative control perturbations in Microduck.** Mean dynamic intervention effects over a 1s rollout horizon for interventions on the forward command u_x and lateral command u_y. The top row shows positive interventions, $do(u_x = 0.25)$ and $do(u_y = 0.17)$, relative to the corresponding zero-control baselines, while the bottom row shows negative interventions, $do(u_x = -0.25)$ and $do(u_y = -0.17)$. Responses are measured on v_x for u_x interventions and on v_y for u_y interventions. Curves compare the MuJoCo simulator, the true-state dynamics baseline, and the TCL-aligned representation. Shaded regions indicate 95% bootstrap confidence intervals across matched intervention trajectories.

and report the corresponding relative error

$$\mathrm{RelErr}_H = \frac{\mathrm{RMSE}_H}{\sqrt{\frac{1}{H}\sum_{k=1}^{H}\left(\mathrm{DIE}_v^{\mathrm{sim}}(k)\right)^2}}. \tag{18}$$

A.10.1 Results

To assess whether the learned representation preserves intervention-relevant dynamics, we evaluated controlled rollouts under matched positive and negative perturbations of the forward and lateral velocity commands. For each intervention, we compared the dynamic intervention effect predicted from the aligned TCL representation with the corresponding MuJoCo response and with a true-state dynamics baseline. The resulting intervention trajectories are shown in Fig. 5, and the corresponding horizon-wise quantitative results are summarized in Tables 2–5.

For interventions on the forward command u_x, the learned representation recovered the correct direction of the response for both positive and negative perturbations, but generally underestimated its magnitude (Fig. 5; Tables 2 and 4). Under $do(u_x = 0.25)$, the representation RMSE decreased from 0.0248 at $H = 1$ to 0.0199 at $H = 25$, while the relative error decreased from 0.910 to 0.571. The endpoint response at 1 s was 0.0281 for the representation compared with 0.0461 in MuJoCo. For the negative intervention $do(u_x = -0.25)$, the representation similarly preserved the sign of the effect but produced an attenuated response. At $H = 25$, the representation endpoint was -0.0042,

Table 2: **Dynamic intervention effect on forward velocity in Microduck.** Errors are computed relative to the matched MuJoCo response for $do(u_x = 0.25)$ versus $do(u_x = 0)$.

H	t (s)	RMSE ↓		Rel. ↓	Endpoint	
		Rep.	State	Rep.	Sim.	Rep.
1	0.04	0.0248	0.0065	0.910	0.0273	0.0025
2	0.08	0.0320	0.0092	0.897	0.0424	0.0046
5	0.20	0.0323	0.0072	0.845	0.0351	0.0085
10	0.40	0.0259	0.0059	0.772	0.0223	0.0130
25	1.00	0.0199	0.0110	0.571	0.0461	0.0281

Table 3: **Dynamic intervention effect on lateral velocity in Microduck.** Errors are computed relative to the matched MuJoCo response for $do(u_y = 0.17)$ versus $do(u_y = 0)$.

H	t (s)	RMSE ↓		Rel. ↓	Endpoint	
		Rep.	State	Rep.	Sim.	Rep.
1	0.04	0.0046	0.0009	0.765	0.0060	0.0014
2	0.08	0.0038	0.0014	0.660	0.0055	0.0027
5	0.20	0.0047	0.0023	0.626	0.0097	0.0028
10	0.40	0.0041	0.0028	0.630	0.0038	0.0019
25	1.00	0.0036	0.0035	0.572	0.0084	0.0107

Table 4: **Dynamic intervention effect on forward velocity in Microduck.** Errors are computed relative to the matched MuJoCo response for $do(u_x = -0.25)$ versus $do(u_x = 0)$.

H	t (s)	RMSE ↓		Rel. ↓	Endpoint	
		Rep.	State	Rep.	Sim.	Rep.
1	0.04	0.0042	0.0042	0.672	-0.0063	-0.0021
2	0.08	0.0075	0.0038	0.728	-0.0132	-0.0034
5	0.20	0.0109	0.0030	0.727	-0.0169	-0.0050
10	0.40	0.0104	0.0021	0.700	-0.0125	-0.0047
25	1.00	0.0102	0.0021	0.711	-0.0129	-0.0042

Table 5: **Dynamic intervention effect on lateral velocity in Microduck.** Errors are computed relative to the matched MuJoCo response for $do(u_y = -0.17)$ versus $do(u_y = 0)$.

H	t (s)	RMSE ↓		Rel. ↓	Endpoint	
		Rep.	State	Rep.	Sim.	Rep.
1	0.04	0.0017	0.0008	3.705	-0.0005	-0.0022
2	0.08	0.0012	0.0006	0.452	-0.0039	-0.0036
5	0.20	0.0014	0.0033	0.387	-0.0048	-0.0056
10	0.40	0.0016	0.0034	0.286	-0.0056	-0.0045
25	1.00	0.0019	0.0028	0.403	-0.0049	-0.0014

compared with -0.0129 in the simulator, with a relative error of 0.711. The true-state dynamics baseline was consistently more accurate for the forward-velocity interventions, indicating that part of the remaining error arises from representation recovery in addition to finite-horizon dynamics prediction.

The lateral interventions were recovered more accurately (Fig. 5; Tables 3 and 5). For $do(u_y = 0.17)$, the representation RMSE remained below 0.0048 across all evaluated horizons and reached 0.0036 at $H = 25$, with a relative error of 0.572. The endpoint effect at 1 s was 0.0107, compared with 0.0084 in MuJoCo. For $do(u_y = -0.17)$, the representation achieved low absolute error throughout the rollout, with RMSE values between 0.0012 and 0.0019 for $H \geq 2$. At $H = 10$, the relative error was 0.286, with a representation endpoint of -0.0045 compared with a simulator effect of -0.0056. The larger relative error at $H = 1$ for this case is due to the very small simulator effect magnitude at the first prediction step. For this intervention, the representation-based model also attains lower RMSE than the true-state baseline for $H \geq 5$ (Table 5), consistent with the small effect magnitude relative to rollout variability.

Overall, Fig. 5 and Tables 2–5 show that the aligned TCL representation preserves the qualitative direction and temporal structure of controlled responses across both intervention signs and both control dimensions. Recovery is strongest for lateral-velocity perturbations, while forward-velocity effects are systematically attenuated in magnitude. The comparison with the true-state baseline further separates representation-related error from error introduced by the learned rollout dynamics.

A.11 Graph-Recovery Algorithms

We use trajectory-level rather than frame-level resampling so that temporal dependence within each trajectory is preserved. The reference parent count k_j is used only for graph binarization and is applied identically to all methods, thereby controlling graph density and separating dependency ranking from threshold calibration. AUPRC is additionally reported from the continuous edge scores without fixing a binary threshold.

Algorithm 1 Residual-Lasso graph recovery with trajectory-bootstrap stability

Require: Aligned state trajectories $\{\hat{S}^{(n)}\}_{n=1}^{N}$, $\hat{S}^{(n)} \in \mathbb{R}^{T_n \times d}$; reference parent counts $\{k_j\}_{j=1}^{d}$; Lasso penalty λ; number of bootstrap samples B; trajectory sampling fraction ρ.

Ensure: Stability scores $\Pi \in [0,1]^{d \times d}$ and binary lag-1 graph $\hat{B} \in \{0,1\}^{d \times d}$.

Standardize the aligned states using statistics estimated from the training trajectories.
for $b = 1, \ldots, B$ **do**
    Resample a fraction ρ of complete trajectories.
    Construct lag-one predictors and residual targets

$$X_t = \hat{S}_t, \qquad R_t = \hat{S}_{t+1} - \hat{S}_t.$$

    for $j = 1, \ldots, d$ **do**
        Exclude $\hat{S}_{t,j}$ from the candidate cross-parent set and fit

$$\hat{\beta}_j^{(b)} = \arg\min_{\beta} \frac{1}{2M_b} \|R_{:,j} - X_{:,-j}\beta\|_2^2 + \lambda \|\beta\|_1.$$

        Let $r_j = k_j - 1$ denote the number of cross-lag parents for child j.
        Record the r_j predictors with the largest $|\hat{\beta}_{ij}^{(b)}|$.
    end for
end for
Compute the trajectory-bootstrap stability score

$$\Pi_{ij} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}[i \text{ is selected as a parent of } j] \,.$$

Initialize $\hat{B} = 0$ and include the reference self-lag edge $\hat{B}_{jj} = 1$ for each state variable.
for $j = 1, \ldots, d$ **do**
    Select the $r_j = k_j - 1$ cross-parents with the largest stability scores Π_{ij} and set the corresponding $\hat{B}_{ij} = 1$.
end for
return Π, $\hat{B}$.

For Microduck, $S_t = [v_x, v_y, \omega_z, h, \phi, \theta]$, $U_t = [u_x, u_y, u_\omega]$, so the transition model has nine candidate state/control inputs and six next-state outputs. The same controlled-dynamics procedure is applied to the true physical state to obtain the true-state dynamics baseline.

A.12 Detailed Related Work

World models and latent dynamical representations. World models learn latent representations that support prediction, planning, and control in high-dimensional environments (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2023). These models have achieved strong performance in reinforcement learning and sequential decision-making, where the primary objectives are typically reconstruction, forecasting, reward prediction, or control. However, predictive accuracy does not ensure that the learned latent state corresponds to the underlying physical or causal state of the system. This distinction is especially important in scientific settings, where intervention and mechanistic interpretation require representations that are not only predictive but also identifiable and structurally meaningful. Our work studies whether frozen diffusion representations can be transformed into latent states that recover the underlying dynamical variables and their directed dependencies.

Diffusion models and latent representations. Diffusion models are powerful generative models for high-dimensional data, including images and videos (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). Latent diffusion models improve scalability by performing generation in compressed representation spaces rather than directly in pixel space (Rombach et al., 2022). Although these representations often contain rich semantic and structural information, they are primarily optimized for generative fidelity rather than identifiability or causal interpretability. Their coordinates therefore need not correspond to the latent dynamical sources governing the observations. We build on frozen

Algorithm 2 Controlled lag-1 graph recovery in Microduck

Require: Aligned state trajectories $\{\hat{S}^{(n)}\}_{n=1}^{N}$, $\hat{S}_t \in \mathbb{R}^d$; control trajectories $\{U^{(n)}\}_{n=1}^{N}$, $U_t \in \mathbb{R}^q$; reference parent counts $\{k_j\}_{j=1}^{d}$.

Ensure: Dependency scores $J \in \mathbb{R}^{(d+q)\times d}$ and binary controlled graph $\hat{B} \in \{0,1\}^{(d+q)\times d}$.

1: Construct controlled lag-one pairs

$$X_t = [\hat{S}_t, U_t], \qquad Y_t = \hat{S}_{t+1}.$$

2: Standardize X_t and Y_t using training-set statistics.

3: Train a neural transition model $f_\theta : \mathbb{R}^{d+q} \to \mathbb{R}^d$ by minimizing

$$\mathcal{L}(\theta) = \frac{1}{M} \sum_t \|Y_t - f_\theta(X_t)\|_2^2 .$$

4: Compute the average absolute transition Jacobian

$$J_{ij} = \frac{1}{M} \sum_t \left| \frac{\partial f_{\theta,j}(X_t)}{\partial X_{t,i}} \right|, \qquad i = 1, \ldots, d+q, \quad j = 1, \ldots, d.$$

5: Initialize $\hat{B} = 0$.

6: **for** $j = 1, \ldots, d$ **do**

7: Select the k_j inputs with the largest values of J_{ij}:

$$\mathcal{P}_j = \mathrm{TopK}\left(\{J_{ij}\}_{i=1}^{d+q}, k_j\right).$$

8: **for** $i \in \mathcal{P}_j$ **do**

9: Set $\hat{B}_{ij} = 1$.

10: **end for**

11: **end for**

12: **return** $J, \hat{B}$.

diffusion representations and ask whether auxiliary-aligned low-dimensional embeddings can recover identifiable dynamical states and support structural causal analysis.

Contrastive representation learning. Contrastive learning learns structured representations by bringing positive pairs together while separating negative pairs (Oord et al., 2018; Chen et al., 2020; He et al., 2020). The information captured by the learned embedding depends strongly on how positive and negative pairs are constructed. In dynamical settings, time-contrastive learning (TCL) uses temporal nonstationarity to recover latent sources under nonlinear ICA assumptions (Hyvarinen & Morioka, 2016; 2017), while generalized contrastive learning (GCL) uses auxiliary variables to obtain identifiability under broader forms of side information (Hyvarinen et al., 2019). CEBRA similarly learns low-dimensional embeddings aligned with behavioral, temporal, or neural variables (Schneider et al., 2023). In our setting, these objectives serve as auxiliary-alignment mechanisms applied to frozen diffusion latents, and we evaluate whether the resulting embeddings recover identifiable states and support downstream causal structure recovery.

Nonlinear ICA and identifiability with auxiliary variables. Nonlinear latent-variable models are generally non-identifiable without additional assumptions (Hyvärinen & Pajunen, 1999; Locatello et al., 2019). Nonlinear ICA provides conditions under which latent sources can nevertheless be recovered from nonlinear observations. In particular, auxiliary information such as time, labels, experimental conditions, or context can provide sufficient structure for identifiability (Hyvarinen & Morioka, 2016; 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020). These results motivate our theoretical analysis. We connect auxiliary-aligned contrastive learning on diffusion latents to density-ratio estimation and auxiliary-variable nonlinear ICA, and use this connection to establish sufficient conditions for identifying latent dynamical states from frozen pretrained representations.

Causal representation learning. Causal representation learning seeks latent representations that support causal reasoning, intervention, and transfer (Bengio et al., 2013; Schölkopf et al., 2021). A central challenge is that representations may capture predictive or statistical regularities without

preserving the variables and mechanisms required for causal reasoning. Recent work therefore studies conditions under which learned coordinates correspond to causal variables or preserve structural mechanisms. Our analysis contributes to this direction by showing that componentwise identifiable dynamical coordinates preserve the factorization and lagged parent structure of the underlying dynamic SCM. This links latent identifiability directly to preservation of causal structure in time-dependent systems.

Causal discovery in dynamical systems. Causal discovery aims to recover directed dependencies from observational or interventional data (Spirtes et al., 2000; Pearl, 2009; Peters et al., 2017). In time-series and dynamical systems, temporal ordering and lagged dependencies provide additional structure for identifying directed interactions (Granger, 1969; Runge et al., 2019; Tank et al., 2021). Many such methods, however, assume that the observed variables already correspond to meaningful system states. In high-dimensional observations such as video, these variables may instead be latent, making causal discovery dependent on the quality of the learned representation. We address this issue by first establishing conditions for identifiable state recovery and then characterizing lagged causal edges through sparsity of the inverse transition-noise Jacobian. This connection is related to Jacobian-based causal discovery under nonlinear ICA (Reizinger et al., 2023), but is developed here for auxiliary-aligned representations of dynamical systems and time-lagged causal structure.

Causal representation learning for scientific dynamical systems. Recent work has increasingly connected causal representation learning with scientific dynamical modeling. Yao et al. (2024) study the integration of identifiable causal representations with dynamical-system models for scientific applications, while Sun et al. (2025) develop identification theory for causal representation learning from multimodal biomedical observations. More broadly, Camps-Valls et al. (2023) review methods for discovering causal relations, equations, and scientific laws from observational data. Our work is complementary, focusing specifically on frozen pretrained diffusion representations of high-dimensional dynamical systems and evaluating whether auxiliary alignment recovers latent state structure, lagged causal dependencies, and intervention-relevant dynamics.