CAUSAL REPRESENTATION LEARNING WITH INSTANTANEOUS AND LAGGED RELATIONS VIA NONSTATIONARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal representation learning for time-series data aims to identify latent states and their causal relations from observations. In this setting, an important challenge is to model both lagged causal relations across observation intervals and faster causal effects that appear as instantaneous relations within an interval, while accounting for nonstationarity in time-series data. However, methods that jointly handle these causal relations and nonstationarity remain limited. To address this gap, we establish sufficient conditions for identifying latent states up to component permutation and component-wise invertible transformations, and their instantaneous and lagged causal structures up to the same permutation, using an observed auxiliary variable, such as time or a condition label, associated with changes in transition-noise distributions. Based on these results, we propose iCReN, a framework that uses contrastive learning with discrete or continuous auxiliary variables to learn latent representations and estimate their instantaneous and lagged causal structures. Experiments demonstrate accurate recovery of latent states and both instantaneous and lagged causal structures on synthetic data and the utility of the learned representations for downstream forecasting on real-world data.

1 INTRODUCTION

Time-series analysis is used for a wide range of tasks (Yue et al., 2022; Wu et al., 2023), including weather forecasting (Bi et al., 2023; Lam et al., 2023) and human motion analysis (Bütepage et al., 2017; Martinez et al., 2017). Many time-series methods model dependencies among observed variables, while learning the factors underlying the observations and their causal relations may additionally support generalization under distribution shifts (Schölkopf et al., 2021; Shirahmad Gale Bagi et al., 2023; Song et al., 2026). When these factors are directly observed, causal discovery methods can be used to infer their causal structure (Shimizu et al., 2006; Zheng et al., 2018; Runge, 2020), whereas in many applications, the relevant factors are not directly observed but are instead entangled in observations such as sensor measurements or images. Causal representation learning (CRL) therefore seeks to identify these factors and their causal relations in learned representations (Komanduri et al., 2024). However, different latent representations can produce the same distribution of observations, so identifying the underlying factors requires additional conditions or information (Hyvärinen & Pajunen, 1999; Khemakhem et al., 2020). Consequently, a central question in CRL is under what conditions and with what additional information the underlying factors and their causal relations become identifiable (Moran & Aragam, 2026).

Two important aspects of CRL for time-series data are modeling the underlying factors as latent states and their causal relations within and across time steps, and accounting for nonstationarity of the data distribution. First, causal interactions occurring within an observation interval are represented as relations among variables at the same time step, termed instantaneous causal relations, whereas interactions across time steps are represented as lagged causal relations. While earlier CRL work for time-series data largely focused on lagged causal relations (Yao et al., 2022b;a), recent work explicitly models both instantaneous and lagged relations (Lippe et al., 2023; Li et al., 2025b;a; Song et al., 2025). These relations can be represented using structural causal models (SCMs) (Pearl, 2009), where each component of the latent state is modeled as a function of its instantaneous and lagged parents and exogenous noise (Hyvärinen et al., 2010; Peters et al., 2013). Second, the distri-

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.

bution of time-series data can change across seasons, times of day, operating conditions, or experimental conditions; accounting for such nonstationarity is important for time-series modeling (Liu et al., 2022; Ditzler et al., 2015; Song et al., 2023). Within this formulation, one way to model nonstationarity is to allow the distribution of the exogenous noise—which we call transition noise—to vary over time or across conditions. Prior work has used such transition-noise variation (Yao et al., 2022b) or, more generally, nonstationarity in the transitions between latent states (Yao et al., 2022a; Song et al., 2023; 2024) to identify latent states and their temporal relations. However, methods that explicitly model such nonstationarity and use it to identify latent states and their instantaneous and lagged causal structures remain underexplored.

To address this gap, we propose iCReN, a framework for identifying latent states and their instantaneous and lagged causal relations using nonstationarity of transition-noise distributions. We model nonstationarity through variation in transition-noise distributions associated with an observed auxiliary variable, while the structural functions governing transitions between latent states and the mapping from latent states to observations remain invariant. Under this setting, our main contributions are as follows. First, we establish sufficient conditions for identifying latent states and their instantaneous and lagged causal structures. Second, we develop iCReN-T and iCReN-G, which instantiate iCReN using time-contrastive learning (TCL) and generalized contrastive learning (GCL), respectively. Third, we demonstrate accurate recovery of latent states and causal structures on synthetic data and predictive utility on real-world data.

2 Related Work

CRL methods for time-series data differ in what causal relations they identify and what information they use to do so. While causal discovery methods for time-series data identify instantaneous and lagged causal relations among directly observed variables under assumptions based on non-Gaussianity, conditional independence, or regime-switching nonstationarity (Hyvärinen et al., 2010; Runge, 2020; Hulsman et al., 2026), CRL must identify both the latent states underlying the observations and their causal relations, requiring further assumptions or information. Several CRL approaches identify latent states and often their lagged causal relations using nonstationarity in transition-noise or latent transition distributions (Yao et al., 2022b;a; Song et al., 2023; 2024; Chen et al., 2024) or component-level interventions (Lippe et al., 2022). In particular, similar to our approach, LEAP (Yao et al., 2022b) exploits variation in transition-noise distributions while assuming invariant structural functions and an invariant observation process, but excludes instantaneous parents. Among methods that allow instantaneous causal relations, iCITRIS (Lippe et al., 2023) assumes known intervention targets and interventions that remove each targeted variable's dependence on its instantaneous causal parents. G-CaRL (Morioka & Hyvärinen, 2024) can also be applied to time-series data with both instantaneous and lagged causal relations; when applied in this temporal setting, however, it identifies only lagged causal relations. IDOL (Li et al., 2025b) identifies latent states and their instantaneous and lagged causal relations from stationary time-series data without auxiliary information, using sufficiently diverse dependencies on past states together with sparsity of the causal structure. More broadly, prior work has studied how sparse dependencies can support identification of latent variables (Lachapelle et al., 2022). Our work identifies latent states and both instantaneous and lagged causal structures when an observed auxiliary variable is associated with changes in the transition-noise distributions.

The methods above also differ in how representations are learned. LEAP and IDOL use variational inference with variational autoencoders (VAEs), while iCITRIS also offers a normalizing-flow variant. In contrast, our identification framework uses contrastive learning without requiring a variational posterior or decoder. Our approach builds on independent innovation analysis (IIA) (Morioka et al., 2021), which uses distributional changes associated with an auxiliary variable to identify transition noise up to a component permutation and component-wise invertible transformations. Its TCL and GCL variants accommodate discrete and general auxiliary variables, respectively, and identify the transition-noise components by exploiting sufficient variability in the associated distributional changes. However, IIA alone identifies neither the latent states nor their causal structure. Under additional structural constraints on the latent data-generating process and restrictions on the learned model class, we extend its transition-noise identification result to the latent states and their instantaneous and lagged causal structures. Table 1 summarizes the differences among the most closely related methods.

Table 1: Comparison of closely related CRL methods for time-series data by the causal relations they identify, the information they use for identification, and how they learn representations. For iCITRIS (*), the distribution changes are induced by interventions with known targets.

	Instantaneous causal structure identification	Nonstationarity	No intervention-target labels required	No variational inference required
LEAP (Yao et al., 2022b)	×	✓	✓	×
iCITRIS (Lippe et al., 2023)	✓	✓*	×	✓
G-CaRL (Morioka & Hyvärinen, 2024)	×	×	✓	✓
IDOL (Li et al., 2025b)	✓	×	✓	×
iCReN	✓	✓	✓	✓

3 Problem Setup

We consider a generative model for time-series observations with both instantaneous and lagged causal relations among latent states. As shown in Figure 1, the latent states evolve according to structural functions and are mapped to observations through an observation function. In our setting, the transition-noise distributions may vary depending on an auxiliary variable, whereas the structural functions and observation function remain invariant.

Observation model. At each time step t, let $\boldsymbol{x}_t \in \mathcal{X}$ denote an observation, with a corresponding continuous latent state $\boldsymbol{z}_t \in \mathbb{R}^{d_z}$. The observation is generated as $\boldsymbol{x}_t = g(\boldsymbol{z}_t)$, where g is an observation function that represents the process through which the latent states generate observations.

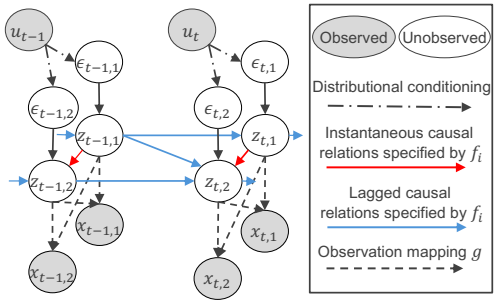


Figure 1: Generative model with $d_z = 2$ and $L = 1$. Gray dash-dotted, solid black, red, blue, and dashed black arrows denote conditioning of the transition-noise distributions on the auxiliary variable, transition-noise inputs to the structural functions, instantaneous causal relations, lagged causal relations, and the observation mapping, respectively.

Latent dynamics. We use structural equations (Pearl, 2009) to model the temporal evolution of the latent states, with instantaneous causal relations within each time step and lagged causal relations across time steps. We consider a finite-lag model with maximum lag $L \geq 1$, in which the current latent state may receive direct effects from latent states up to L time steps in the past. For each latent component $z_{t,i}$, we let f_i denote its structural function and define the same-time and past latent variables on which f_i depends nontrivially as its instantaneous and lagged parents, respectively. These parent sets satisfy

$$\mathrm{pa}_{\mathrm{inst}}(z_{t,i}) \subseteq \{z_{t,j} : j \neq i\}, \qquad \mathrm{pa}_{\mathrm{lags}}(z_{t,i}) \subseteq \{z_{t-\ell,j} : 1 \leq \ell \leq L,\ 1 \leq j \leq d_z\}. \tag{1}$$

Each latent component is generated as

$$z_{t,i} = f_i\big(\mathrm{pa}_{\mathrm{inst}}(z_{t,i}), \mathrm{pa}_{\mathrm{lags}}(z_{t,i}), \epsilon_{t,i}\big). \tag{2}$$

Here, $\epsilon_{t,i}$ is the transition noise for component i. This formulation includes, as special cases, models in which the instantaneous parent sets are empty and models in which transition noise enters additively. Each parent relation corresponds to a directed causal edge from the parent variable to $z_{t,i}$. The instantaneous causal structure is assumed to form a directed acyclic graph (DAG). We further assume that there are no unmodeled common causes of the latent components.

Nonstationary transition noise. To model nonstationarity, we allow each conditional transition-noise distribution $p_i(\epsilon_{t,i} \mid u_t)$ to vary depending on an observed auxiliary variable u_t. The auxiliary variable may be continuous, such as time, or discrete, such as a label indicating an environment, experimental condition, phase, season, or time of day. Conditioned on the auxiliary-variable sequence

$u_{1:T}$, the transition-noise components are independent across components and time steps, and $\boldsymbol{\epsilon}_t$ is independent of the latent states in the past L time steps given u_t:

$$p(\boldsymbol{\epsilon}_{1:T} \mid u_{1:T}) = \prod_{t=1}^{T} \prod_{i=1}^{d_z} p_i(\epsilon_{t,i} \mid u_t), \qquad \boldsymbol{\epsilon}_t \perp (\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-L}) \mid u_t. \tag{3}$$

Under this model, we seek to identify the latent states up to a component permutation and component-wise invertible transformations, and their instantaneous and lagged causal structures up to the same permutation, from the joint distribution of the observation sequence $(\boldsymbol{x}_t)_{t=1}^{T}$ and the auxiliary-variable sequence $(u_t)_{t=1}^{T}$.

4 Identifiability Analysis

In this section, we establish sufficient conditions for identifying the latent states and for identifying their instantaneous and lagged causal structures. For simplicity, we set $L = 1$ in this section and extend the results to any finite maximum lag L in Appendix A.5.

4.1 Identification of Latent Components

We first establish conditions for identifying the components of the latent state. Let ϕ be an encoder such that $\hat{\boldsymbol{z}}_t = \phi(\boldsymbol{x}_t)$. Because $\boldsymbol{x}_t = g(\boldsymbol{z}_t)$, the learned representation satisfies $\hat{\boldsymbol{z}}_t = (\phi \circ g)(\boldsymbol{z}_t)$. When $\phi \circ g$ is invertible, we denote its inverse by $H = (\phi \circ g)^{-1}$, so that $\boldsymbol{z}_t = H(\hat{\boldsymbol{z}}_t)$. We next define the conditional-dependence structures. Let $\mathcal{M}_z$ be the pairwise Markov network representing conditional-dependence relations among the components of $\boldsymbol{z}_t$ given $\boldsymbol{z}_{t-1}$ and u_t, and let $\mathcal{E}_{\text{inst}}$ denote its edge set. We further let $\mathcal{E}_{\text{lag}}$ denote the set of conditional-dependence edges between the components of $\boldsymbol{z}_{t-1}$ and $\boldsymbol{z}_t$ in the pairwise Markov network for $(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t)$ given u_t. For distinct i, j and any a, the absence of an edge is characterized by

$$\{z_{t,i}, z_{t,j}\} \notin \mathcal{E}_{\text{inst}} \iff z_{t,i} \perp z_{t,j} \mid \boldsymbol{z}_{t,\setminus i,j}, \boldsymbol{z}_{t-1}, u_t, \tag{4}$$

$$\{z_{t-1,a}, z_{t,i}\} \notin \mathcal{E}_{\text{lag}} \iff z_{t-1,a} \perp z_{t,i} \mid \boldsymbol{z}_{t-1,\setminus a}, \boldsymbol{z}_{t,\setminus i}, u_t. \tag{5}$$

Here, $\boldsymbol{z}_{s,\setminus S}$ denotes the vector obtained from $\boldsymbol{z}_s$ by excluding the components indexed by S. We denote the corresponding quantities for the learned representation by $\mathcal{M}_{\hat{z}}$, $\widehat{\mathcal{E}}_{\text{inst}}$, and $\widehat{\mathcal{E}}_{\text{lag}}$. Because these edge sets are defined by conditional dependence, their edges may correspond to direct causal edges or connect variables with no direct causal edge. For example, in $z_{t,i} \to z_{t,k} \leftarrow z_{t,j}$, the conditioning set in equation 4 includes the common child $z_{t,k}$, and conditioning on it can induce dependence between $z_{t,i}$ and $z_{t,j}$, even though no direct causal edge connects them.

We next introduce quantities that characterize how differences in the conditional log transition density across auxiliary-variable values depend on the current and past latent states. Fix a value $u_t^{(0)}$ of the auxiliary variable as a reference, and let $u_t^{(1)}, \ldots, u_t^{(N_{\text{aux}})}$ denote the comparison values. For each $n = 1, \ldots, N_{\text{aux}}$, define

$$\Delta_n(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) := \log p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, u_t^{(n)}) - \log p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, u_t^{(0)}), \tag{6}$$

and define $\widehat{\Delta}_n(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})$ in the same way for the learned representation. We then define the instantaneous and cross-time variation vectors

$$\boldsymbol{w}_n(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) := \left[\left(\frac{\partial^2 \Delta_n}{\partial z_{t,i}^2} \right)_{i=1}^{d_z}, \left(\frac{\partial^2 \Delta_n}{\partial z_{t,i} \partial z_{t,j}} \right)_{\{z_{t,i}, z_{t,j}\} \in \mathcal{E}_{\text{inst}}}, \left(\frac{\partial \Delta_n}{\partial z_{t,i}} \right)_{i=1}^{d_z} \right]^{\top}, \tag{7}$$

$$\boldsymbol{v}_{ia}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) := \left(\frac{\partial^2 \Delta_n}{\partial z_{t,i} \partial z_{t-1,a}} \right)_{n=1}^{N_{\text{aux}}}. \tag{8}$$

The corresponding learned quantity $\hat{\boldsymbol{v}}_{ia}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})$ is obtained from $\widehat{\Delta}_n(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})$ in the same manner. To examine cross-time relations between $\boldsymbol{z}_t$ and $\boldsymbol{z}_{t+1}$, we also use $\boldsymbol{v}_{bi}(\boldsymbol{z}_{t+1}, \boldsymbol{z}_t)$, with $\hat{\boldsymbol{v}}_{bi}(\hat{\boldsymbol{z}}_{t+1}, \hat{\boldsymbol{z}}_t)$ denoting its learned counterpart. Because H does not depend on the value of the auxiliary variable, the Jacobian term introduced by the change of variables cancels in the log-density

ratio, giving $\widehat{\Delta}_n(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}) = \Delta_n(H(\hat{\boldsymbol{z}}_t), H(\hat{\boldsymbol{z}}_{t-1}))$. For fixed learned-component indices j and a, differentiating this equation expresses $\hat{\boldsymbol{v}}_{ja}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})$ as a linear combination of the true variation vectors $\boldsymbol{v}_{km}(H(\hat{\boldsymbol{z}}_t), H(\hat{\boldsymbol{z}}_{t-1}))$ over current and past component indices k and m, where the coefficients are products of Jacobian entries of H. In Theorem 1, the instantaneous variation vector $\boldsymbol{w}_n$ and the cross-time variation vector $\boldsymbol{v}_{ia}$ play complementary roles in establishing component-wise identification. First, the condition involving $\boldsymbol{w}_n$ restricts possible mixing between latent components to a specific type of neighboring pair in $\mathcal{M}_z$. Following Zhang et al. (2024), we call $z_{t,j}$ an intimate neighbor of $z_{t,i}$ if $z_{t,j}$ is adjacent to $z_{t,i}$ and to every other neighbor of $z_{t,i}$; the formal definition is given in Appendix A.2. The condition involving $\boldsymbol{v}_{ia}$ then eliminates this remaining mixing, yielding component-wise identification.

Theorem 1 (Identification of latent components) *Suppose that the following conditions hold.*

(i) Regularity. *The inverse $H = (\phi \circ g)^{-1}$ exists and is a C^2 diffeomorphism independent of time and the value of the auxiliary variable. The transition densities are positive, and their log densities are twice continuously differentiable.*

(ii) Instantaneous variation. *Let $N_{\text{inst}} := 2d_z + |\mathcal{E}_{\text{inst}}|$. For every $\boldsymbol{z}_t$, there exists a set of N_{inst} pairs $(n_q, \bar{\boldsymbol{z}}_{t-1}^{(q)}) \in \{1, \dots, N_{\text{aux}}\} \times \mathbb{R}^{d_z}$, $q = 1, \dots, N_{\text{inst}}$, such that*

$$\boldsymbol{W}(\boldsymbol{z}_t) := \left[\boldsymbol{w}_{n_q}(\boldsymbol{z}_t, \bar{\boldsymbol{z}}_{t-1}^{(q)})\right]_{q=1}^{N_{\text{inst}}} \quad \textit{has rank } N_{\text{inst}}. \tag{9}$$

Here, the indices n_q need not be distinct from one another.

(iii) Cross-time variation and separation. *For each learned variation vector $\hat{\boldsymbol{v}}_{ja}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})$, the vectors $\boldsymbol{v}_{km}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$ indexed by all pairs (k, m) that can contribute to its linear expansion above are linearly independent almost everywhere. Moreover, for every intimate neighbor $z_{t,j}$ of $z_{t,i}$ in $\mathcal{M}_z$, there exists a past or future component with a conditional-dependence edge to $z_{t,i}$ and no such edge to $z_{t,j}$. The precise index sets and past and future conditions are given in Appendix A.2.*

(iv) Bounds on the learned edge sets. *$|\widehat{\mathcal{E}}_{\text{inst}}| \le |\mathcal{E}_{\text{inst}}|$ and $|\widehat{\mathcal{E}}_{\text{lag}}| \le |\mathcal{E}_{\text{lag}}|$ hold.*

Under these conditions, there exist a permutation π_z and univariate C^2 diffeomorphisms $\xi_1, \dots, \xi_{d_z}$ such that $\hat{z}_{t,\pi_z(i)} = \xi_i(z_{t,i})$ for every i and t. The same permutation π_z also defines an isomorphism between $\mathcal{M}_z$ and $\mathcal{M}_{\hat{z}}$.

The following discussion clarifies the scope of the conditions in Theorem 1 and the role of Condition (iii) when intimate neighbors are present. Compared with IDOL, whose sufficient-variability condition uses only variation across past states, Condition (ii) additionally exploits changes in transition-noise distributions across auxiliary-variable values, including changes in their mean or variance. Consequently, our condition may hold for additive Gaussian transition noise in cases where IDOL's condition does not (Li et al., 2025b). However, as in IDOL, Condition (ii) cannot hold when each transition-noise component can be expressed as the sum of a function of the current state and a function of the past state in the presence of instantaneous edges, as shown in Remark 1. Condition (iii) requires linear independence to prevent cancellation among the relevant cross-time variation vectors and uses differences in past or future conditional-dependence relations to eliminate mixing between intimate neighbors. Complete proofs are given in Appendices A.1 and A.2.

4.2 Identification of the Causal Graph

We then establish conditions for identifying the causal graph. For the learned representation, let $\hat{r}$ be a learned inverse transition function that estimates transition noise as $\hat{\boldsymbol{\epsilon}}_t = \hat{r}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})$. To express direct causal edges through derivatives, we regard each f_i as a function of all candidate instantaneous and lagged variables and characterize its parent sets by the variables whose corresponding partial derivatives are not identically zero. Theorem 2 builds on the latent-component identification in Theorem 1 and component-wise identification of transition noise to establish a correspondence between Jacobian entries of $\hat{r}$ that are not identically zero and direct causal edges.

Theorem 2 (Identification of the causal graph) *Suppose that Conditions (i)–(iv) of Theorem 1 and the following additional conditions hold, and let π_z denote the resulting permutation between the true and learned latent components.*

(v) Transition-noise identification. *The conditional transition-noise distribution factorizes across components, with each factor belonging to an exponential family with at least two twice continuously differentiable and nonredundant sufficient statistics, such as $\epsilon_{t,i}$ and $\epsilon_{t,i}^2$ for a Gaussian distribution. The natural parameters vary sufficiently across auxiliary-variable values to satisfy the full-rank condition required by IIA (Morioka et al., 2021). The remaining regularity conditions on the true and learned transformations, together with the estimation conditions for contrastive learning, are given in Assumption 3.*

(vi) Differentiability of the structural functions. *Each f_i is continuously differentiable.*

Under the above conditions, there exist a permutation π_ϵ and univariate C^2 diffeomorphisms q_i such that $\hat{\epsilon}_{t,\pi_\epsilon(i)} = q_i(\epsilon_{t,i})$ for every i and t. Moreover, because the instantaneous causal structure is a DAG, the instantaneous Jacobian support of $\hat{r}$ uniquely determines the correspondence between the indices $\pi_\epsilon(i)$ and $\pi_z(i)$. Under this correspondence, the following equivalence holds for $\ell \in \{0, 1\}$:

$$\frac{\partial \hat{r}_{\pi_\epsilon(i)}}{\partial \hat{z}_{t-\ell,\pi_z(j)}} \not\equiv 0 \quad \Longleftrightarrow \quad \frac{\partial f_i}{\partial z_{t-\ell,j}} \not\equiv 0. \tag{10}$$

Here, for $\ell = 0$, the case $i = j$ is excluded because instantaneous causal edges are defined only between distinct components. Consequently, the true instantaneous and lagged causal edges are identifiable up to the component permutation π_z from the corresponding Jacobian supports of $\hat{r}$.

We discuss the scope of the conditions and theoretical setting in Theorem 2. Condition (v), inherited from the IIA identification result (Morioka et al., 2021), requires the conditional transition-noise distributions to vary sufficiently across auxiliary-variable values and therefore does not cover data for which these distributions are invariant across those values. The smoothness in Conditions (v) and (vi) is used to derive the equivalence in equation 10. Using this equivalence to identify causal edges from the Jacobian support of $\hat{r}$ applies the approach of Reizinger et al. (2023) to latent transition dynamics. Appendix A.3 establishes transition-noise identification under the sufficient setting $\mathcal{X} = \mathbb{R}^{d_z}$ used in the IIA argument, and Appendix A.4 provides the complete causal-graph identification proof. The latent-component result relies on the C^2 diffeomorphism H in Condition (i), whereas the causal-graph result additionally uses the transition-noise correspondence; neither identification argument otherwise depends on the dimension of the observations. Related invertibility assumptions between learned and true latent states are also used in CRL analyses (Lippe et al., 2023; Morioka & Hyvärinen, 2024; Li et al., 2025b). For higher-dimensional observations, the same arguments would apply if the learned representation admitted such an H and allowed the transition noise to be identified; however, we do not establish conditions under which such a representation can be learned.

5 Learning Framework

This section describes the learning framework, summarized in Figure 2, and how the causal graph is subsequently estimated from the learned model. We first apply the shared encoder ϕ separately to each observation $\boldsymbol{x}_{t-\ell}$ for $\ell = 0, \ldots, L$ and then apply the learned inverse transition function $\hat{r}$ to the resulting representations. For contrastive learning, we also use a history feature map η to obtain a history feature $\boldsymbol{h}_t$ from the past learned representations:

$$\hat{\boldsymbol{z}}_{t-\ell} = \phi(\boldsymbol{x}_{t-\ell}), \qquad \hat{\boldsymbol{\epsilon}}_t = \hat{r}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}, \ldots, \hat{\boldsymbol{z}}_{t-L}), \qquad \boldsymbol{h}_t = \eta(\hat{\boldsymbol{z}}_{t-1}, \ldots, \hat{\boldsymbol{z}}_{t-L}). \tag{11}$$

Further parameterization details are provided in Appendix B.1.

Contrastive learning with auxiliary variables. Based on IIA (Morioka et al., 2021), we consider two variants of contrastive learning: TCL for discrete auxiliary variables, such as labels indicating an environment, experimental condition, motion phase, season, or time of day, and GCL for general auxiliary variables, including continuous values such as time. TCL and GCL use the following regression functions, respectively:

$$\begin{aligned} T_u^{\mathrm{TCL}}(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t) &= \sum_{i=1}^{d_z} \sum_{k=1}^{K} w_{uik} \psi_{ik}(\hat{\epsilon}_{t,i}) + \varphi(\boldsymbol{h}_t, u) + b_u, \\ T^{\mathrm{GCL}}(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t, u_t) &= \sum_{i=1}^{d_z} \sum_{k=1}^{K} \psi_{ik}(\hat{\epsilon}_{t,i}) \mu_{ik}(u_t) + \varphi(\boldsymbol{h}_t, u_t) + B(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t). \end{aligned} \tag{12}$$

Here, K is the number of feature functions for each transition-noise component, and ψ_{ik} denotes the k-th scalar-valued feature function applied to $\hat{\epsilon}_{t,i}$. The feature functions may be learned jointly with the remaining model components or, when the conditional noise family is specified in advance, fixed to include its sufficient statistics. For TCL, u indexes a class, T_u^{TCL} is its logit, and w_{uik} and b_u are learned separately for each class. For GCL, the learned function μ_{ik} determines the coefficient of the k-th feature of component i from the auxiliary variable. The learned function φ models the relation between the history feature and the auxiliary variable, while B collects terms independent of the auxiliary variable. In both regression functions, the terms that depend jointly on transition noise and the auxiliary variable decompose additively across components, as required by IIA.

The TCL logits over all possible values of the discrete auxiliary variable define a multinomial logistic regression model, which is trained using softmax cross-entropy with the observed u_t as the target. In GCL, T^{GCL} serves as the logit of a binary logistic regression model trained using binary cross-entropy, with matched tuples $(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t, u_t)$ as positive examples and mismatched tuples $(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t, u'_t)$ as negative examples; u'_t is obtained by permuting auxiliary-variable values within each minibatch. We write $\mathcal{L}_{\mathrm{IIA}}$ for the selected TCL or GCL loss; see Appendix B.2 for details.

Structural regularization and training objective. Motivated by the bounds on learned edge sets in Condition (iv), we impose sparsity regularization on the instantaneous and lagged input dependencies of $\hat{r}$. Following IDOL's use of sparsity regularization (Li et al., 2025b), we use ℓ_1-based regularizers $\mathcal{R}_{\mathrm{inst}}$ and $\mathcal{R}_{\mathrm{lag}}$ to encourage sparse instantaneous and lagged input dependencies of $\hat{r}$, respectively. We denote their sum by $\mathcal{R}_{\mathrm{sparsity}}$. Because latent components are identifiable only up to permutation, any acyclic instantaneous structure can be represented under a fixed triangular mask after an appropriate permutation of the latent coordinates. However, this requires the encoder to align the learned components with the fixed order encoded by the mask. We therefore also consider a variant using the NOTEARS acyclicity regularizer $\mathcal{R}_{\mathrm{DAG}}$ (Zheng et al., 2018), which does not fix an order during training. We optimize all learnable parameters using the objective $\mathcal{L}$:

$$\mathcal{L} := \mathcal{L}_{\mathrm{IIA}} + \lambda_{\mathrm{sparsity}}\mathcal{R}_{\mathrm{sparsity}} + \lambda_{\mathrm{DAG}}\mathcal{R}_{\mathrm{DAG}}. \tag{13}$$

Here, $\lambda_{\mathrm{sparsity}}$ and λ_{DAG} are nonnegative coefficients controlling the strengths of the sparsity and acyclicity regularization, respectively. When the fixed triangular mask is used, the $\mathcal{R}_{\mathrm{DAG}}$ term can be omitted. Further details of the regularizers and training objective are provided in Appendix B.3.

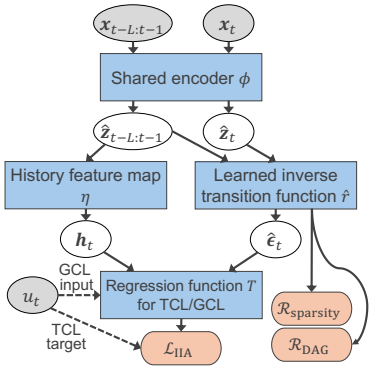


Figure 2: Learning framework. Gray nodes denote observed variables; white nodes denote learned representations, estimated transition noise, and history features. Dashed arrows denote auxiliary-variable inputs to TCL/GCL. The encoder ϕ is shared across time, and blue modules are trained jointly using the orange objectives.

Causal-graph estimation. After learning, we estimate a causal graph over the learned latent coordinates from the Jacobian support of $\hat{r}$; under the conditions of Theorem 2, the estimated edges correspond to the true causal edges up to a component permutation. For the case $L = 1$ considered in the theory, we estimate that an instantaneous edge $\hat{z}_{t,j} \to \hat{z}_{t,i}$ is present for $i \neq j$ when $\partial \hat{r}_i / \partial \hat{z}_{t,j} \not\equiv 0$, and that a lagged edge $\hat{z}_{t-1,j} \to \hat{z}_{t,i}$ is present when $\partial \hat{r}_i / \partial \hat{z}_{t-1,j} \not\equiv 0$. The same rule can be applied to each lag $\ell = 1, \ldots, L$.

6 Experiments

We evaluate latent-state and causal-graph recovery on synthetic data and the utility of learned representations for downstream forecasting on real-world data. The evaluated CRL baselines are IDOL (Li et al., 2025b), G-CaRL (Morioka & Hyvärinen, 2024), CtrlNS (Song et al., 2024), and LEAP (Yao et al., 2022b). iCITRIS (Lippe et al., 2023) is not evaluated because the experimental settings provide neither its required intervention targets nor interventions that remove incoming instantaneous edges. In addition, iVAE (Khemakhem et al., 2020) and β-VAE (Higgins et al., 2017) are included as representation-learning baselines that use and do not use the auxiliary variable, respectively. iVAE is evaluated only on synthetic data because, unlike the other methods, its inference model conditions on u when encoding observations at downstream inference.

Table 2: Comparison of the proposed methods with baselines on synthetic data. The left and right blocks report results for discrete-variation data and continuous-variation data, respectively. Each entry is the mean ± standard deviation over 10 seeds; boldface indicates the best value in each column, and −− indicates a metric not applicable to the method.

	Discrete variation			Continuous variation		
Method	MCC	Inst. F1	Lagged F1	MCC	Inst. F1	Lagged F1
iCReN-T	$\mathbf{0.951 \pm 0.011}$	$\mathbf{0.950 \pm 0.158}$	$\mathbf{0.980 \pm 0.063}$	$\mathbf{0.959 \pm 0.010}$	$\mathbf{0.975 \pm 0.079}$	$\mathbf{0.970 \pm 0.095}$
iCReN-G	0.940 ± 0.010	0.700 ± 0.158	$\mathbf{0.980 \pm 0.042}$	0.945 ± 0.013	$\mathbf{0.975 \pm 0.079}$	0.950 ± 0.071
IDOL	0.468 ± 0.046	0.225 ± 0.142	0.480 ± 0.079	0.534 ± 0.090	0.225 ± 0.142	0.520 ± 0.123
G-CaRL	0.764 ± 0.065	−−	0.440 ± 0.117	0.762 ± 0.047	−−	0.340 ± 0.084
CtrlNS	0.432 ± 0.071	−−	0.600 ± 0.000	0.398 ± 0.102	−−	0.600 ± 0.000
LEAP	0.373 ± 0.107	−−	0.380 ± 0.092	0.365 ± 0.086	−−	0.380 ± 0.063
iVAE	0.230 ± 0.106	−−	−−	0.373 ± 0.110	−−	−−
β-VAE	0.482 ± 0.067	−−	−−	0.473 ± 0.075	−−	−−

6.1 Experiments on Synthetic Data

We generate synthetic data with nonlinear structural and observation functions defined by three-layer neural networks that remain invariant across auxiliary-variable values, while Gaussian transition-noise distributions vary discretely across environments or continuously over time. The main comparison uses $\boldsymbol{x}_t, \boldsymbol{z}_t \in \mathbb{R}^5$ and $L = 1$, with 2^{16} samples from 256 environments for discrete-variation data and a sequence of length 2^{18} with time as the auxiliary variable for continuous-variation data. We measure latent-state recovery by the mean correlation coefficient (MCC) and graph recovery by instantaneous and lagged F1 scores. Each ranges from 0 to 1, with higher values indicating better recovery. Details of data generation and the evaluation protocol are provided in Appendix C.1.

We evaluate iCReN-T and iCReN-G under both transition-noise variation settings. For discrete-variation data, both use the observed environment labels; for continuous-variation data, iCReN-G uses continuous time and iCReN-T uses its discretization into 256 bins. The results in Table 2 show that iCReN-T attains the highest mean score for every metric in both settings. iCReN-G attains comparable MCC and lagged F1 scores, but its instantaneous F1 is lower on discrete-variation data. The sample-size analysis in Appendix C.4 indicates that iCReN-G is more sensitive to sample size than iCReN-T and that its lower instantaneous F1 on discrete-variation data is partly a finite-sample effect. Reducing the depth of the observation function from three layers to one increases mean MCC for every evaluated method, with the largest increases for IDOL and LEAP, which remain below both proposed methods, as shown in Appendix C.5. Appendix C additionally reports sensitivity analyses of latent causal structure and latent dimension, along with ablations of the objective components.

6.2 Experiments on Real-world Data

Because ground-truth latent states and causal structures are unavailable for the real-world datasets, we assess the utility of learned representations through two downstream forecasting tasks under held-out conditions: future-motion prediction at a walking speed and weather forecasting in a later period. For each task, we construct discrete auxiliary labels and use iCReN-T with an observation reconstruction loss during pretraining; we then freeze each encoder and train a forecasting head.

Treadmill motion forecasting. Using the treadmill walking data of Riglet et al. (2024), we evaluate future-motion forecasting at the Comfortable speed after pretraining and model selection on the Slow and Fast speeds. Participants are split among the training, validation, and test sets without overlap; during pretraining, iCReN and LEAP use gait phase divided into 12 classes as the auxiliary variable. With each encoder frozen, a forecasting head is trained and selected on the Slow and Fast data to predict the following 24 frames (1.92 seconds) from the latent representations of the preceding 12 frames (0.96 seconds). The head is then evaluated without updates on the Comfortable condition, which is excluded from training and model selection. We report mean marker error, defined as the three-dimensional Euclidean distance between predicted and true marker positions averaged over test windows and markers. The results in Table 3 show that iCReN achieves the lowest error at every reported horizon, suggesting that the learned representation supports forecasting at a walking speed excluded from training and model selection. Details of the experimental setup and results with standard deviations are provided in Appendix D.1.

Table 3: Forecasting results on real-world data. Entries are means across runs; lower is better. Treadmill reports mean marker error in mm, and weather reports standardized MSE for three-hour-ahead relative-humidity forecasting. Weather columns are southeastern Australia (AU), Canada Great Lakes–St. Lawrence (CAN), China (CN), France (FR), Germany (GER), Japan (JP), South Korea (KR), and US Atlantic (US). Boldface marks the best result in each column.

	Treadmill: marker error [mm] ↓				Weather: standardized MSE ↓							
Method	0.48 s	0.96 s	1.44 s	1.92 s	AU	CAN	CN	FR	GER	JP	KR	US
iCReN	**42.060**	**58.650**	**82.034**	**107.309**	**0.352**	0.481	**0.546**	**0.422**	0.474	**0.414**	**0.291**	0.427
IDOL	44.076	62.400	86.791	115.146	0.359	0.484	0.555	0.438	**0.469**	0.440	0.304	0.440
G-CaRL	52.197	75.790	103.133	135.169	0.649	0.673	0.664	0.750	0.754	0.615	0.535	0.554
CtrlNS	43.593	62.337	86.040	117.059	0.385	0.522	0.618	0.459	0.504	0.501	0.382	0.462
LEAP*	44.220	63.431	87.338	113.210	0.396	0.524	0.591	0.486	0.566	0.482	0.327	0.460
β-VAE	43.391	63.305	91.089	123.960	0.357	**0.475**	0.561	0.445	0.484	0.426	0.328	**0.416**

*For LEAP, pretraining on treadmill data was unstable for two seeds; results are based on eight runs.

Weather forecasting. Using the ERA5 global reanalysis (Hersbach et al., 2020), we examine whether representations pretrained on historical weather observations support forecasting of relative humidity in a later period across eight regions. We construct three-hourly observations from five cities in each region, using 2-m temperature, mean sea-level pressure, 10-m zonal and meridional wind components, and precipitation over the preceding three hours. Relative humidity is excluded from representation pretraining and serves as the downstream prediction target. For iCReN and LEAP, the auxiliary variable combines 12 months, four periods within each month, and eight UTC times at three-hour intervals, yielding 384 classes. Each method is pretrained and selected using data from 1981–2013; with the selected encoder frozen, a forecasting head is trained on 2014–2016, selected using 2017 validation data, and evaluated on 2018–2020. The head predicts relative humidity three hours ahead from the latent representations at the preceding eight time points. Performance is evaluated using standardized mean squared error (MSE). Among the compared representation-learning methods, iCReN achieves the lowest mean standardized MSE in five regions and the second-lowest in the remaining three, suggesting that the learned representations support forecasting in a later evaluation period across diverse regions. Details of the experimental setup and results with standard deviations are provided in Appendix D.2.

7 Discussion and Conclusion

The following discussion compares our framework with related methods and considers the limitations of our analysis and experiments. Our framework requires an observed auxiliary variable associated with sufficient variation in transition-noise distributions, but not the intervention-target labels required by iCITRIS (Lippe et al., 2023). IIA-HMM (Morioka et al., 2021) may offer a route to relaxing the requirement for an observed auxiliary variable by modeling nonstationarity with a hidden Markov model. For latent-component identification, our analysis draws on the identification argument of IDOL (Li et al., 2025b) by combining variation across auxiliary-variable values and past states; transition-noise identification builds on IIA (Morioka et al., 2021). The full-rank requirement in Condition (ii) does not cover models with instantaneous relations when the transition noise can be expressed as a sum of functions of the current and past latent states, including linear transition models. Consistent with this limitation, Appendix C.5 shows lower instantaneous F1 for linear transitions. Finally, because ground-truth latent states and causal structures are unavailable for the real-world datasets, we evaluate predictive utility rather than recovery of these quantities. Validating the learned states and causal relations on real-world data would require additional evidence, such as domain knowledge or interventional information.

We established sufficient conditions for identifying latent states and their instantaneous and lagged causal structures from time-series observations, using an observed auxiliary variable associated with changes in transition-noise distributions. Based on these results, we developed iCReN, a contrastive learning framework that accommodates discrete and continuous auxiliary variables. Experiments demonstrated accurate recovery of latent states and their causal structures on synthetic data and the utility of the learned representations for downstream forecasting on real-world data. Our results show how transition-noise nonstationarity associated with observed auxiliary variables can support CRL with both instantaneous and lagged causal relations.

AI USE STATEMENT

In this work, we used generative AI tools to assist with (i) theoretical work, including derivations and proof checking, as well as finding potentially relevant literature; (ii) experimental work and implementation, including experimental design, implementation and review of models and code for synthetic-data generation and real-world-data preprocessing, and interpretation of results; and (iii) manuscript preparation, including drafting, translation, and editing for clarity and readability. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

ETHICS STATEMENT

This work involved no new recruitment of human participants or collection of human-subject data. The motion-prediction experiment used the previously published, publicly available motion-capture dataset of Riglet et al. (2024). We used only the released kinematic measurements and pseudonymous subject identifiers and did not attempt to identify participants.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs underlying our identifiability results are provided in Appendix A. Appendix B specifies the model parameterizations, training objectives, structural regularization, and joint learning procedure. Appendices C and D document the data generation and preprocessing, optimization settings, data splits, auxiliary variables, method-specific inputs, model selection, evaluation protocols, sensitivity analyses, and ablations. We provide an anonymized supplementary code package containing the implementation and configurations of the proposed method, synthetic-data generation code, and instructions for preprocessing and running the proposed method on the real-world datasets.

REFERENCES

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023.

David Bolton. The computation of equivalent potential temperature. *Monthly Weather Review*, 108 (7):1046–1053, 1980.

Judith Bütepage, Michael Black, Danica Kragic, and Hedvig Kjellström. Deep representation learning for human motion prediction and classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1591–1599, 2017.

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. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7236–7259, 2024.

Copernicus Climate Change Service. ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2023. Accessed on 2 September 2026.

Gregory Ditzler, Manuel Roveri, Cesare Alippi, and Robi Polikar. Learning in nonstationary environments: A survey. *IEEE Computational Intelligence Magazine*, 10(4):12–25, 2015.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

Roel Hulsman, Carles Balsells-Rodas, and Sara Magliacane. Identifiable Markov switching models with instantaneous effects and exponential families. In *International Conference on Machine Learning (ICML)*, 2026.

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

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11: 1709–1731, 2010.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217, 2020.

Aneesh Komanduri, Xintao Wu, Yongkai Wu, and Feng Chen. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Sébastien Lachapelle, Pau Rodríguez, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning (CLeaR)*, volume 177 of *Proceedings of Machine Learning Research*, pp. 428–484, 2022.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Zijian Li, Minghao Fu, Junxian Huang, Yifan Shen, Ruichu Cai, Yuewen Sun, Guangyi Chen, and Kun Zhang. Towards identifiability of hierarchical temporal causal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 19680–19719, 2025a.

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

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13557–13603, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *International Conference on Learning Representations (ICLR)*, 2023.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 9881–9893, 2022.

Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4674–4683, 2017.

Gemma Moran and Bryon Aragam. Toward interpretable deep generative models via causal representation learning. *Journal of the American Statistical Association*, 121(553):259–275, 2026.

Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36249–36293, 2024.

Hiroshi Morioka, Hermanni Hälvä, and Aapo Hyvärinen. Independent innovation analysis for nonlinear vector autoregressive process. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1549–1557, 2021.

Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, 2013.

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*, 2023.

Louis Riglet, Corentin Delphin, Lauranne Claquesin, Baptiste Orliac, Paul Ornetti, Davy Laroche, and Mathieu Gueugnon. 3D motion analysis dataset of healthy young adult volunteers walking and running on overground and treadmill. *Scientific Data*, 11(1):556, 2024. URL `https://doi.org/10.1038/s41597-024-03420-y`.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1388–1397, 2020.

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.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

Shayan Shirahmad Gale Bagi, Zahra Gharaee, Oliver Schulte, and Mark Crowley. Generative causal representation learning for out-of-distribution motion forecasting. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31596–31612, 2023.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric P. Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 8092–8113, 2023.

Xiangchen Song, Zijian Li, Guangyi Chen, Yujia Zheng, Yewen Fan, Xinshuai Dong, and Kun Zhang. Causal temporal representation learning with nonstationary sparse transition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 77098–77131, 2024.

Xiangchen Song, Jiaqi Sun, Zijian Li, Yujia Zheng, and Kun Zhang. LLM interpretability with identifiable temporal-instantaneous representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 103963–104000, 2025.

Xinxin Song, Yuxiao Cheng, Tingxiong Xiao, and Jinli Suo. COGS: A causal representation learning framework for out-of-distribution generalization in time series. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 25572–25580, 2026.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2023.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 26492–26503, 2022a.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations (ICLR)*, 2022b.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 8980–8987, 2022.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60057–60075, 2024.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 9472–9483, 2018.

Appendix Contents

A Proofs **14**

A.1 Identification of the Instantaneous Markov Network 14

A.2 Identification of Latent Components 17

A.3 Transition-Noise Identification by IIA 26

A.4 Causal-Graph Identification from the Jacobian of the Inverse Transition Function . 29

A.5 Extension to Multiple Lags 31

B Learning Framework Details **33**

B.1 Model Parameterization 33

B.2 Contrastive Learning with Auxiliary Variables 33

B.3 Structural Regularization and Training Objective 34

C Synthetic-data Experimental Details and Additional Results **35**

C.1 Common Experimental Settings 35

C.2 Sensitivity to Latent Causal Structure 38

C.3 Sensitivity to Latent Dimension 39

C.4 Sensitivity to Total Sample Size 39

C.5 Sensitivity to Data-generating Function Complexity 41

C.6 Ablation of Objective Components 41

D Real-world Experimental Details **42**

D.1 Treadmill Motion Forecasting 42

D.2 Weather Forecasting 44

D.3 Data Sources and Availability 45

A PROOFS

A.1 IDENTIFICATION OF THE INSTANTANEOUS MARKOV NETWORK

This section establishes an isomorphism between the true and learned instantaneous Markov networks. The argument draws on the Markov-network identification step of IDOL (Li et al., 2025b), using a variability condition that incorporates variation across both past states and values of the auxiliary variable. Suppose that their latent representations are related by a C^2 diffeomorphism H that is independent of time and the value of the auxiliary variable:

$$\boldsymbol{z}_t = H(\hat{\boldsymbol{z}}_t), \qquad \boldsymbol{z}_{t-1} = H(\hat{\boldsymbol{z}}_{t-1}).$$

We use nonedges in the learned instantaneous network to derive constraints on J_H and combine them with instantaneous edge-count minimality to establish an isomorphism between the true and learned instantaneous Markov networks. Let $\mathcal{M}_z$ and $\mathcal{M}_{\hat{z}}$ denote the pairwise Markov networks over the components of $\boldsymbol{z}_t$ and $\hat{\boldsymbol{z}}_t$, respectively, defined by the instantaneous conditional-independence relations in the true and learned transition distributions, and write their edge sets as

$$\mathcal{E}_{\text{inst}} := \mathcal{E}(\mathcal{M}_z), \qquad \widehat{\mathcal{E}}_{\text{inst}} := \mathcal{E}(\mathcal{M}_{\hat{z}}).$$

For $i \neq j$ and $k \neq l$, their nonedges are characterized by

$$\begin{aligned} \{z_{t,i}, z_{t,j}\} \notin \mathcal{E}_{\text{inst}} \quad &\Longleftrightarrow \quad z_{t,i} \perp z_{t,j} \mid \boldsymbol{z}_{t,\setminus\{i,j\}}, \boldsymbol{z}_{t-1}, u_t, \\ \{\hat{z}_{t,k}, \hat{z}_{t,l}\} \notin \widehat{\mathcal{E}}_{\text{inst}} \quad &\Longleftrightarrow \quad \hat{z}_{t,k} \perp \hat{z}_{t,l} \mid \hat{\boldsymbol{z}}_{t,\setminus\{k,l\}}, \hat{\boldsymbol{z}}_{t-1}, u_t. \end{aligned}$$

Here and below, we write each edge as an unordered pair of its latent-variable nodes.

A nonedge in the learned instantaneous network implies that the corresponding mixed derivative of the log-density ratio in the learned coordinates is zero. After expansion through H, however, this condition is a linear combination with Jacobian products as coefficients. These products need not vanish individually because nonzero terms may cancel. To rule out such cancellation and additional learned edges, we impose sufficient variability and instantaneous edge-count minimality, respectively.

Assumption 1 (Sufficient variability and instantaneous edge-count minimality) *The following conditions hold.*

(i) Regularity. *For every auxiliary-variable value $u_t^{(n)}$, the density $p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, u_t^{(n)})$ is positive on $\mathbb{R}^{d_z} \times \mathbb{R}^{d_z}$, and its logarithm is twice continuously differentiable in $(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$.*

(ii) Sufficient variability across auxiliary-variable values and past states. *Choose a reference value $u_t^{(0)}$, let N_{aux} be the number of comparison values, and define, for $n \in \{1, \dots, N_{\text{aux}}\}$,*

$$\Delta_n(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) := \log p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, u_t^{(n)}) - \log p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, u_t^{(0)}).$$

Set $N_{\text{inst}} := 2d_z + |\mathcal{E}_{\text{inst}}|$ and define

$$\boldsymbol{w}_n(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) := \left(\frac{\partial^2 \Delta_n}{\partial z_{t,i}^2}\right)_{i=1}^{d_z} \oplus \left(\frac{\partial^2 \Delta_n}{\partial z_{t,i} \partial z_{t,j}}\right)_{\{z_{t,i}, z_{t,j}\} \in \mathcal{E}_{\text{inst}}} \oplus \left(\frac{\partial \Delta_n}{\partial z_{t,i}}\right)_{i=1}^{d_z} \in \mathbb{R}^{N_{\text{inst}}}.$$

For every $\boldsymbol{z}_t$, suppose that there exist pairs of auxiliary-variable indices and past-state values

$$\left(n_1, \bar{\boldsymbol{z}}_{t-1}^{(1)}\right), \dots, \left(n_{N_{\text{inst}}}, \bar{\boldsymbol{z}}_{t-1}^{(N_{\text{inst}})}\right) \in \{1, \dots, N_{\text{aux}}\} \times \mathbb{R}^{d_z}.$$

Here, the overbar indicates past-state evaluation values selected with $\boldsymbol{z}_t$ fixed, rather than the random variable $\boldsymbol{z}_{t-1}$; the indices $n_1, \dots, n_{N_{\text{inst}}}$ need not be distinct. Using these pairs, define

$$\boldsymbol{W}(\boldsymbol{z}_t) := \left[\boldsymbol{w}_{n_1}(\boldsymbol{z}_t, \bar{\boldsymbol{z}}_{t-1}^{(1)}), \dots, \boldsymbol{w}_{n_{N_{\text{inst}}}}(\boldsymbol{z}_t, \bar{\boldsymbol{z}}_{t-1}^{(N_{\text{inst}})})\right] \in \mathbb{R}^{N_{\text{inst}} \times N_{\text{inst}}},$$

and assume that $\boldsymbol{W}(\boldsymbol{z}_t)$ has full rank.

(iii) Instantaneous edge-count minimality. *The learned instantaneous edge set satisfies*

$$|\widehat{\mathcal{E}}_{\text{inst}}| \leq |\mathcal{E}_{\text{inst}}|.$$

We first fix two nonadjacent components in $\mathcal{M}_{\hat{z}}$ and derive Jacobian conditions that prevent a true component, or the endpoints of a true instantaneous edge, from depending simultaneously on these two learned components.

Lemma 1 (Jacobian conditions implied by a learned instantaneous nonedge) *Consider the learned representation above satisfying $\boldsymbol{z}_t = H(\hat{\boldsymbol{z}}_t)$, and suppose that Assumption 1(i)–(ii) holds. Fix distinct k, l such that $\{\hat{z}_{t,k}, \hat{z}_{t,l}\} \notin \widehat{\mathcal{E}}_{\text{inst}}$. Then, at every $\hat{\boldsymbol{z}}_t$ in the domain of H,*

$$\forall i, \quad \frac{\partial z_{t,i}}{\partial \hat{z}_{t,k}} \frac{\partial z_{t,i}}{\partial \hat{z}_{t,l}} = 0, \tag{14}$$

$$\forall \{z_{t,i}, z_{t,j}\} \in \mathcal{E}_{\text{inst}}, \quad \frac{\partial z_{t,i}}{\partial \hat{z}_{t,k}} \frac{\partial z_{t,j}}{\partial \hat{z}_{t,l}} = 0. \tag{15}$$

The relation in equation 14 states that a true component cannot depend simultaneously on two nonadjacent components of $\mathcal{M}_{\hat{z}}$. The relation in equation 15 states that, for a true instantaneous edge, its two endpoints cannot separately depend on the two nonadjacent learned components.

Proof. Define the learned log-density ratio by

$$\widehat{\Delta}_n(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}) := \log p(\hat{\boldsymbol{z}}_t \mid \hat{\boldsymbol{z}}_{t-1}, u_t^{(n)}) - \log p(\hat{\boldsymbol{z}}_t \mid \hat{\boldsymbol{z}}_{t-1}, u_t^{(0)}).$$

The density transformation induced by $\boldsymbol{z}_t = H(\hat{\boldsymbol{z}}_t)$ and $\boldsymbol{z}_{t-1} = H(\hat{\boldsymbol{z}}_{t-1})$ gives

$$\begin{aligned} p(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}, u_t) &= p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, u_t) \left|\det J_H(\hat{\boldsymbol{z}}_t)\right| \left|\det J_H(\hat{\boldsymbol{z}}_{t-1})\right|, \\ p(\hat{\boldsymbol{z}}_{t-1}, u_t) &= p(\boldsymbol{z}_{t-1}, u_t) \left|\det J_H(\hat{\boldsymbol{z}}_{t-1})\right|. \end{aligned}$$

Dividing the first equality by the second gives the conditional-density transformation. Taking logarithms and subtracting the reference expression from each comparison expression cancels the Jacobian term because it does not depend on the auxiliary variable, yielding

$$\widehat{\Delta}_n(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}) = \Delta_n\big(H(\hat{\boldsymbol{z}}_t), H(\hat{\boldsymbol{z}}_{t-1})\big). \tag{16}$$

Because $\{\hat{z}_{t,k}, \hat{z}_{t,l}\} \notin \widehat{\mathcal{E}}_{\text{inst}}$, the pairwise Markov property gives

$$\begin{aligned} \log p(\hat{\boldsymbol{z}}_t \mid \hat{\boldsymbol{z}}_{t-1}, u_t) = {} & \log p(\hat{z}_{t,k} \mid \hat{\boldsymbol{z}}_{t,\setminus\{k,l\}}, \hat{\boldsymbol{z}}_{t-1}, u_t) \\ & + \log p(\hat{z}_{t,l} \mid \hat{\boldsymbol{z}}_{t,\setminus\{k,l\}}, \hat{\boldsymbol{z}}_{t-1}, u_t) + \log p(\hat{\boldsymbol{z}}_{t,\setminus\{k,l\}} \mid \hat{\boldsymbol{z}}_{t-1}, u_t). \end{aligned}$$

No term on the right-hand side depends on both $\hat{z}_{t,k}$ and $\hat{z}_{t,l}$, so its mixed derivative with respect to these variables is zero for every auxiliary-variable value. Taking the difference between the comparison and reference values therefore gives $\partial^2 \widehat{\Delta}_n / \partial \hat{z}_{t,k} \partial \hat{z}_{t,l} = 0$. Substituting this result into the left-hand side of equation 16 and applying the chain rule to the right-hand side gives

$$\begin{aligned} 0 = {} & \sum_{a=1}^{d_z} \frac{\partial^2 \Delta_n}{\partial z_{t,a}^2} \frac{\partial z_{t,a}}{\partial \hat{z}_{t,k}} \frac{\partial z_{t,a}}{\partial \hat{z}_{t,l}} + \sum_{\{z_{t,a}, z_{t,b}\} \in \mathcal{E}_{\text{inst}}} \frac{\partial^2 \Delta_n}{\partial z_{t,a} \partial z_{t,b}} \left(\frac{\partial z_{t,a}}{\partial \hat{z}_{t,k}} \frac{\partial z_{t,b}}{\partial \hat{z}_{t,l}} + \frac{\partial z_{t,b}}{\partial \hat{z}_{t,k}} \frac{\partial z_{t,a}}{\partial \hat{z}_{t,l}} \right) \\ & + \sum_{a=1}^{d_z} \frac{\partial \Delta_n}{\partial z_{t,a}} \frac{\partial^2 z_{t,a}}{\partial \hat{z}_{t,k} \partial \hat{z}_{t,l}}. \end{aligned}$$

Here, all derivatives of Δ_n are evaluated at $(H(\hat{\boldsymbol{z}}_t), H(\hat{\boldsymbol{z}}_{t-1}))$. In the second sum, only pairs in $\mathcal{E}_{\text{inst}}$ remain because, by the same pairwise Markov property, the mixed derivative is zero for every true instantaneous nonedge.

Denote the coefficients of the derivatives of Δ_n in this expansion by

$$\kappa_a^{\text{diag}} := \frac{\partial z_{t,a}}{\partial \hat{z}_{t,k}} \frac{\partial z_{t,a}}{\partial \hat{z}_{t,l}}, \qquad \kappa_{\{a,b\}}^{\text{edge}} := \frac{\partial z_{t,a}}{\partial \hat{z}_{t,k}} \frac{\partial z_{t,b}}{\partial \hat{z}_{t,l}} + \frac{\partial z_{t,b}}{\partial \hat{z}_{t,k}} \frac{\partial z_{t,a}}{\partial \hat{z}_{t,l}}, \qquad \kappa_a^{\text{grad}} := \frac{\partial^2 z_{t,a}}{\partial \hat{z}_{t,k} \partial \hat{z}_{t,l}}.$$

Stack them in the same order as $\boldsymbol{w}_n$ into $\kappa \in \mathbb{R}^{N_{\text{inst}}}$. Then $\boldsymbol{w}_n(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})^\top \kappa = 0$ for all n and $\boldsymbol{z}_{t-1}$. For fixed $\boldsymbol{z}_t$, κ is independent of the past state and auxiliary-variable value. Stacking the equations for the pairs selected in Assumption 1(ii) gives

$$\boldsymbol{W}(\boldsymbol{z}_t)^\top \kappa = 0, \qquad \operatorname{rank} \boldsymbol{W}(\boldsymbol{z}_t) = N_{\text{inst}} \quad \Longrightarrow \quad \kappa = 0.$$

In particular, $\kappa_i^{\text{diag}} = 0$, which proves equation 14.

For $\{z_{t,i}, z_{t,j}\} \in \mathcal{E}_{\text{inst}}$, let

$$A_{ij} := \frac{\partial z_{t,i}}{\partial \hat{z}_{t,k}} \frac{\partial z_{t,j}}{\partial \hat{z}_{t,l}}, \qquad B_{ij} := \frac{\partial z_{t,j}}{\partial \hat{z}_{t,k}} \frac{\partial z_{t,i}}{\partial \hat{z}_{t,l}}.$$

The equalities $\kappa_{\{i,j\}}^{\text{edge}} = 0$ and equation 14 imply

$$A_{ij} + B_{ij} = 0, \qquad A_{ij} B_{ij} = 0.$$

Thus, $A_{ij} = B_{ij} = 0$, proving equation 15. □

We next use invertibility of the Jacobian of H to match every true latent component locally to a distinct learned coordinate.

Lemma 2 (Local component matching from an invertible Jacobian) *For every $\hat{\boldsymbol{z}}_t$, there exists a permutation $\pi_{\hat{\boldsymbol{z}}_t}$ such that*

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}} \neq 0 \qquad (\forall i).$$

Proof. Since H is a diffeomorphism, $\det J_H(\hat{\boldsymbol{z}}_t) \neq 0$ at every $\hat{\boldsymbol{z}}_t$. The determinant expands as

$$\det J_H(\hat{\boldsymbol{z}}_t) = \sum_{\pi \in \mathfrak{S}_{d_z}} \operatorname{sgn}(\pi) \prod_{i=1}^{d_z} \frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi(i)}},$$

where $\mathfrak{S}_{d_z}$ denotes the set of all permutations of $\{1, \ldots, d_z\}$. Since the determinant is nonzero, at least one product in the sum is nonzero. Choosing its associated permutation as $\pi_{\hat{\boldsymbol{z}}_t}$ proves the claim. □

The Jacobian conditions and local component matching now show that every true instantaneous edge is preserved in the learned representation; instantaneous edge-count minimality then yields network isomorphism.

Theorem 3 (Identification of the instantaneous Markov network) *Fix $\hat{\boldsymbol{z}}_t$ and take a local permutation $\pi_{\hat{\boldsymbol{z}}_t}$ from Lemma 2. Under Assumption 1, for every $i \neq j$,*

$$\{z_{t,i}, z_{t,j}\} \in \mathcal{E}_{\text{inst}} \quad \Longleftrightarrow \quad \{\hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}, \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}\} \in \widehat{\mathcal{E}}_{\text{inst}}.$$

Thus, $\mathcal{M}_z$ and $\mathcal{M}_{\hat{z}}$ are isomorphic under $\pi_{\hat{\boldsymbol{z}}_t}$. At this stage, $\pi_{\hat{\boldsymbol{z}}_t}$ may depend on the fixed point $\hat{\boldsymbol{z}}_t$.

Proof. Take any $\{z_{t,i}, z_{t,j}\} \in \mathcal{E}_{\text{inst}}$. If its image were a learned nonedge, Lemma 1 would give

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}} \frac{\partial z_{t,j}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} = 0,$$

contradicting Lemma 2. Thus, every true instantaneous edge maps injectively to a learned instantaneous edge, and

$$|\mathcal{E}_{\text{inst}}| \leq |\widehat{\mathcal{E}}_{\text{inst}}|, \qquad |\widehat{\mathcal{E}}_{\text{inst}}| \leq |\mathcal{E}_{\text{inst}}|.$$

The first inequality follows from edge preservation and injectivity of $\pi_{\hat{\boldsymbol{z}}_t}$, and the second from Assumption 1(iii). The edge sets therefore have equal cardinality, and the images of the true edges exhaust the learned edge set, proving the isomorphism. □

The following remark examines the full-rank condition in Assumption 1(ii) for additively separable inverse transitions.

Remark 1 (Rank bound for an additively separable inverse transition function) *Suppose that the inverse transition function satisfies $\epsilon_{t,i} = r_i(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) = a_i(\boldsymbol{z}_t) + b_i(\boldsymbol{z}_{t-1})$, where a_i depends only on $\boldsymbol{z}_t$ and b_i only on $\boldsymbol{z}_{t-1}$. Define the component-wise log-density ratio of the transition noise by*

$$\delta_{i,n}(\epsilon_{t,i}) := \log p_i(\epsilon_{t,i} \mid u_t^{(n)}) - \log p_i(\epsilon_{t,i} \mid u_t^{(0)}).$$

Conditioning on $\boldsymbol{z}_{t-1}$ and u_t, component-wise independence of the transition noise and the change of variables $\boldsymbol{\epsilon}_t = r(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$ give

$$p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, u_t) = \prod_{i=1}^{d_z} p_i\big(r_i(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) \mid u_t\big) \left| \det \frac{\partial r(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})}{\partial \boldsymbol{z}_t} \right|.$$

Because r is independent of the auxiliary-variable value, its Jacobian term cancels in the log-density ratio, yielding

$$\Delta_n(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) = \sum_{i=1}^{d_z} \delta_{i,n}\big(r_i(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})\big).$$

For all $j, k \in \{1, \dots, d_z\}$, evaluating the derivatives of $\delta_{i,n}$ at $\epsilon_{t,i} = r_i(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$ gives

$$\frac{\partial \Delta_n}{\partial z_{t,j}} = \sum_{i=1}^{d_z} \frac{\partial \delta_{i,n}}{\partial \epsilon_{t,i}} \frac{\partial a_i}{\partial z_{t,j}},$$

$$\frac{\partial^2 \Delta_n}{\partial z_{t,j} \partial z_{t,k}} = \sum_{i=1}^{d_z} \left\{ \frac{\partial^2 \delta_{i,n}}{\partial \epsilon_{t,i}^2} \frac{\partial a_i}{\partial z_{t,j}} \frac{\partial a_i}{\partial z_{t,k}} + \frac{\partial \delta_{i,n}}{\partial \epsilon_{t,i}} \frac{\partial^2 a_i}{\partial z_{t,j} \partial z_{t,k}} \right\}.$$

For fixed $\boldsymbol{z}_t$, all derivatives of a_i are fixed, and variation with the auxiliary-variable value and past state appears only through the two scalar coefficients $\partial \delta_{i,n} / \partial \epsilon_{t,i}$ and $\partial^2 \delta_{i,n} / \partial \epsilon_{t,i}^2$ for each i. Thus, every $\boldsymbol{w}_n$ is a linear combination of at most $2d_z$ fixed directions, and

$$\operatorname{rank} \boldsymbol{W}(\boldsymbol{z}_t) \leq 2d_z.$$

If $|\mathcal{E}_{\text{inst}}| > 0$, then $N_{\text{inst}} = 2d_z + |\mathcal{E}_{\text{inst}}| > 2d_z$, so the required full-rank condition cannot hold. A linear inverse transition function is a special case of this additively separable form, so the theorem does not cover such models when instantaneous edges are present.

Mixed derivatives with respect to a current and past component, however, may retain terms such as

$$\frac{\partial^2 \delta_{i,n}}{\partial \epsilon_{t,i}^2} \frac{\partial a_i}{\partial z_{t,j}} \frac{\partial b_i}{\partial z_{t-1,m}}.$$

Therefore, when the transition-noise variance changes with the auxiliary-variable value, cross-time sufficient variability may hold even for an additively separable model.

A.2 Identification of Latent Components

This section builds on the local isomorphism between the instantaneous Markov networks established in Appendix A.1 to identify the latent components. We first show that any remaining component mixing is restricted to intimate neighbors in the instantaneous Markov network. When such neighbors exist, we eliminate the remaining mixing using cross-time relations and their variation across auxiliary-variable values. Finally, we show that the resulting local permutations reduce to a single permutation independent of state and time, yielding component-wise identification. Our argument draws on the intimate-neighbor proof strategy of IDOL (Li et al., 2025b), while treating the instantaneous and cross-time relations separately and using changes across values of the auxiliary variable for the latter.

Throughout this section, suppose that the true and learned representations are related by the same C^2 diffeomorphism H, independent of time and the value of the auxiliary variable, and that Assumption 1 holds.

A.2.1 Definitions and Notation

To describe component mixing that cannot be ruled out by the isomorphism of the instantaneous Markov networks alone, we define neighborhoods and intimate neighbors in $\mathcal{M}_z$. The neighborhood of $z_{t,i}$ in $\mathcal{M}_z$ is

$$\mathcal{N}_{\mathcal{M}_z}(z_{t,i}) := \{z_{t,j} \mid \{z_{t,i}, z_{t,j}\} \in \mathcal{E}_{\text{inst}}\}.$$

Definition 1 (Intimate neighbor (Zhang et al., 2024)) *In the pairwise Markov network $\mathcal{M}_z$, the intimate-neighbor set of each node $z_{t,i}$ is*

$$\Psi_{\mathcal{M}_z}(z_{t,i}) := \{z_{t,j} \mid z_{t,j} \in \mathcal{N}_{\mathcal{M}_z}(z_{t,i}) \text{ and } \mathcal{N}_{\mathcal{M}_z}(z_{t,i}) \setminus \{z_{t,j}\} \subseteq \mathcal{N}_{\mathcal{M}_z}(z_{t,j}) \setminus \{z_{t,i}\}\}.$$

Thus, $z_{t,j}$ is an intimate neighbor of $z_{t,i}$ if they are adjacent in $\mathcal{M}_z$ and every neighbor of $z_{t,i}$ other than $z_{t,j}$ is also a neighbor of $z_{t,j}$ other than $z_{t,i}$. This relation is generally asymmetric: $z_{t,j} \in \Psi_{\mathcal{M}_z}(z_{t,i})$ does not imply $z_{t,i} \in \Psi_{\mathcal{M}_z}(z_{t,j})$.

Even after the instantaneous Markov networks have been shown to be isomorphic, mixing may remain between components corresponding to intimate neighbors. We express this possibility using off-diagonal entries of the Jacobian of the mapping $H : \hat{\boldsymbol{z}}_t \mapsto \boldsymbol{z}_t$.

Definition 2 (Residual mixing) *Fix a point $\hat{\boldsymbol{z}}_t$ and choose the local permutation $\pi_{\hat{\boldsymbol{z}}_t}$ given by Lemma 2. For $i \neq j$, we say that residual mixing remains between $z_{t,i}$ and $\hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}$ under H if*

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} \neq 0.$$

The following lemma shows that residual mixing is restricted to components corresponding to intimate neighbors.

Lemma 3 (Residual mixing is restricted to intimate neighbors) *Fix a point $\hat{\boldsymbol{z}}_t$ and choose the local permutation $\pi_{\hat{\boldsymbol{z}}_t}$ given by Lemma 2. For every $i \neq j$,*

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} \neq 0 \quad \Longrightarrow \quad z_{t,j} \in \Psi_{\mathcal{M}_z}(z_{t,i}).$$

Proof. Suppose, for contradiction, that residual mixing occurs, so that

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} \neq 0, \tag{17}$$

while $z_{t,j} \notin \Psi_{\mathcal{M}_z}(z_{t,i})$.

First consider the case $\{z_{t,i}, z_{t,j}\} \notin \mathcal{E}_{\text{inst}}$. Theorem 3 implies

$$\{\hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}, \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}\} \notin \widehat{\mathcal{E}}_{\text{inst}}.$$

The relation in equation 14 from Lemma 1 therefore gives

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}} \frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} = 0.$$

The two factors are nonzero by Lemma 2 and equation 17, respectively, yielding a contradiction.

Next consider $\{z_{t,i}, z_{t,j}\} \in \mathcal{E}_{\text{inst}}$. Because $z_{t,j}$ is not an intimate neighbor of $z_{t,i}$, there exists an index k such that

$$\{z_{t,i}, z_{t,k}\} \in \mathcal{E}_{\text{inst}}, \qquad \{z_{t,j}, z_{t,k}\} \notin \mathcal{E}_{\text{inst}}.$$

Theorem 3 then implies

$$\{\hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}, \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(k)}\} \notin \widehat{\mathcal{E}}_{\text{inst}}.$$

Applying equation 15 in Lemma 1 to the true edge $\{z_{t,i}, z_{t,k}\}$ yields

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} \frac{\partial z_{t,k}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(k)}} = 0.$$

Again, the two factors are nonzero by equation 17 and Lemma 2, respectively, which is a contradiction. Thus, whether or not $z_{t,i}$ and $z_{t,j}$ are adjacent in $\mathcal{M}_z$, assuming $z_{t,j} \notin \Psi_{\mathcal{M}_z}(z_{t,i})$ leads to a contradiction. Hence $z_{t,j} \in \Psi_{\mathcal{M}_z}(z_{t,i})$. □

A.2.2 EMPTY INTIMATE-NEIGHBOR SETS

When every intimate-neighbor set is empty, Lemma 3 immediately rules out all residual mixing. Specifically, suppose

$$\Psi_{\mathcal{M}_z}(z_{t,i}) = \emptyset \qquad (\forall i),$$

and fix any point $\hat{\boldsymbol{z}}_t$ and the local permutation $\pi_{\hat{\boldsymbol{z}}_t}$ given by Lemma 2. By Lemma 3, a nonzero off-diagonal entry of the locally permuted Jacobian would imply the existence of an intimate neighbor, contradicting the assumption. Therefore,

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} = 0 \qquad (\forall i,\ \forall j \neq i).$$

After its columns are arranged according to $\pi_{\hat{\boldsymbol{z}}_t}$, the local Jacobian is thus diagonal with nonzero diagonal entries.

A.2.3 NONEMPTY INTIMATE-NEIGHBOR SETS: USING CROSS-TIME RELATIONS

When intimate neighbors remain as candidates for residual mixing after considering the local structure of $\mathcal{M}_z$, we use conditional-dependence relations with the past and future time steps to eliminate residual mixing. We first define the cross-time relations and their changes across auxiliary-variable values, derive Jacobian conditions from learned cross-time nonedges, and then combine these conditions with cross-time edge-count minimality to match the true and learned cross-time relations.

Definition 3 (Cross-time conditional-dependence relations) *Within this section, combine the latent variables at two time steps as*

$$\boldsymbol{c}_t := (\boldsymbol{z}_{t-1}, \boldsymbol{z}_t), \qquad \hat{\boldsymbol{c}}_t := (\hat{\boldsymbol{z}}_{t-1}, \hat{\boldsymbol{z}}_t).$$

Let $\mathcal{M}_{\boldsymbol{c}_t}$ be the pairwise Markov network for the conditional joint distribution $p(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t \mid u_t)$, and define $\mathcal{M}_{\hat{\boldsymbol{c}}_t}$ analogously for $p(\hat{\boldsymbol{z}}_{t-1}, \hat{\boldsymbol{z}}_t \mid u_t)$. The current–current part of $\mathcal{M}_{\boldsymbol{c}_t}$ represents the instantaneous conditional-independence relations given the past state and therefore coincides with $\mathcal{M}_z$ and its edge set $\mathcal{E}_{\mathrm{inst}}$ defined in Appendix A.1. Here, we use only the past–current part of these Markov networks.

Define the relations between the past and current time steps by

$$\mathcal{E}^-_{\mathrm{lag}} := \left\{ \{z_{t-1,a}, z_{t,i}\} \;\middle|\; z_{t-1,a} \not\perp\!\!\!\perp z_{t,i} \mid \boldsymbol{z}_{t-1,\setminus\{a\}}, \boldsymbol{z}_{t,\setminus\{i\}}, u_t \right\}.$$

For the learned representation, define

$$\widehat{\mathcal{E}}^-_{\mathrm{lag}} := \left\{ \{\hat{z}_{t-1,a}, \hat{z}_{t,i}\} \;\middle|\; \hat{z}_{t-1,a} \not\perp\!\!\!\perp \hat{z}_{t,i} \mid \hat{\boldsymbol{z}}_{t-1,\setminus\{a\}}, \hat{\boldsymbol{z}}_{t,\setminus\{i\}}, u_t \right\}.$$

Thus, $\mathcal{E}^-_{\mathrm{lag}}$ and $\widehat{\mathcal{E}}^-_{\mathrm{lag}}$ are the past–current parts of $\mathcal{M}_{\boldsymbol{c}_t}$ and $\mathcal{M}_{\hat{\boldsymbol{c}}_t}$, respectively. They correspond to $\mathcal{E}_{\mathrm{lag}}$ and $\widehat{\mathcal{E}}_{\mathrm{lag}}$ in the main text. The future-directed sets $\mathcal{E}^+_{\mathrm{lag}}$ and $\widehat{\mathcal{E}}^+_{\mathrm{lag}}$ are obtained by applying the same definitions to $\boldsymbol{c}_{t+1} = (\boldsymbol{z}_t, \boldsymbol{z}_{t+1})$ and $\hat{\boldsymbol{c}}_{t+1} = (\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t+1})$.

For a reference value $u_t^{(0)}$ and comparison value $u_t^{(n)}$, define the log-density ratio between the corresponding transition densities by

$$\Delta^-_n(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) := \log p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, u_t^{(n)}) - \log p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, u_t^{(0)}).$$

For each current–past component pair (i, a), collect these mixed derivatives over the comparison values as

$$\boldsymbol{v}^-_{ia}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) := \left(\frac{\partial^2 \Delta^-_n(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})}{\partial z_{t,i} \partial z_{t-1,a}} \right)_{n=1}^{N_{\mathrm{aux}}} \in \mathbb{R}^{N_{\mathrm{aux}}}.$$

The quantities Δ_n and $\boldsymbol{v}_{ia}$ in the main text correspond to Δ^-_n and $\boldsymbol{v}^-_{ia}$, respectively. Define $\widehat{\Delta}^-_n$ and $\hat{\boldsymbol{v}}^-_{ia}$ analogously for the learned representation. Analogously, the future-directed quantities Δ^+_n, $\boldsymbol{v}^+_{bi}$, and their learned counterparts are obtained by shifting these definitions forward by one time step. The following lemma relates cross-time nonedges to zeros of the corresponding derivative vectors.

Lemma 4 (A cross-time nonedge implies zero mixed derivatives of the log-density ratio) *For the true representation, the past- and future-directed statements are, respectively,*

$$\{z_{t-1,a}, z_{t,i}\} \notin \mathcal{E}^{-}_{\text{lag}} \implies \boldsymbol{v}^{-}_{ia}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) = 0,$$
$$\{z_{t,i}, z_{t+1,b}\} \notin \mathcal{E}^{+}_{\text{lag}} \implies \boldsymbol{v}^{+}_{bi}(\boldsymbol{z}_{t+1}, \boldsymbol{z}_t) = 0.$$

For the learned representation, the corresponding statements are

$$\{\hat{z}_{t-1,m}, \hat{z}_{t,k}\} \notin \widehat{\mathcal{E}}^{-}_{\text{lag}} \implies \hat{\boldsymbol{v}}^{-}_{km}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}) = 0,$$
$$\{\hat{z}_{t,k}, \hat{z}_{t+1,m}\} \notin \widehat{\mathcal{E}}^{+}_{\text{lag}} \implies \hat{\boldsymbol{v}}^{+}_{mk}(\hat{\boldsymbol{z}}_{t+1}, \hat{\boldsymbol{z}}_t) = 0.$$

Proof. We first consider the true distribution in the past case. Suppose $\{z_{t-1,a}, z_{t,i}\} \notin \mathcal{E}^{-}_{\text{lag}}$. By the definition of $\mathcal{E}^{-}_{\text{lag}}$, for every auxiliary-variable value u_t,

$$z_{t-1,a} \perp z_{t,i} \mid \boldsymbol{z}_{t-1,\setminus\{a\}}, \boldsymbol{z}_{t,\setminus\{i\}}, u_t.$$

The transition density therefore factorizes as

$$p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, u_t) = p\big(z_{t,i} \mid \boldsymbol{z}_{t,\setminus\{i\}}, \boldsymbol{z}_{t-1,\setminus\{a\}}, u_t\big) p(\boldsymbol{z}_{t,\setminus\{i\}} \mid \boldsymbol{z}_{t-1}, u_t).$$

The first factor does not depend on $z_{t-1,a}$, and the second does not depend on $z_{t,i}$. Because the transition density is positive, taking its logarithm gives

$$\frac{\partial^2}{\partial z_{t,i} \partial z_{t-1,a}} \log p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, u_t) = 0$$

for every auxiliary-variable value. Taking the difference between each comparison value and the reference value therefore makes every entry of $\boldsymbol{v}^{-}_{ia}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$ zero, and hence

$$\boldsymbol{v}^{-}_{ia}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) = 0.$$

For the learned representation, the same conclusion follows directly from the definition of a nonedge in the learned Markov network. In particular, if

$$\{\hat{z}_{t-1,m}, \hat{z}_{t,k}\} \notin \widehat{\mathcal{E}}^{-}_{\text{lag}},$$

then, for every u_t,

$$\hat{z}_{t-1,m} \perp \hat{z}_{t,k} \mid \hat{\boldsymbol{z}}_{t-1,\setminus\{m\}}, \hat{\boldsymbol{z}}_{t,\setminus\{k\}}, u_t.$$

The same factorization argument gives

$$\frac{\partial^2}{\partial \hat{z}_{t,k} \partial \hat{z}_{t-1,m}} \log p(\hat{\boldsymbol{z}}_t \mid \hat{\boldsymbol{z}}_{t-1}, u_t) = 0,$$

and hence $\hat{\boldsymbol{v}}^{-}_{km}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}) = 0$. Shifting time forward by one step proves the statements for the future case for both representations. □

The Jacobian conditions in Appendix A.1 restrict which true components can have nonzero derivatives with respect to each learned coordinate according to the local structure of $\mathcal{M}_z$. To express this restriction at each time step, let $[d_z] := \{1, \ldots, d_z\}$ and, for each time s and index $a \in [d_z]$, define the candidate component index set

$$\mathcal{C}_s(a) := \{i \in [d_z] \mid i = a \text{ or } z_{s,a} \in \Psi_{\mathcal{M}_z}(z_{s,i})\}.$$

By Lemma 3,

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(a)}} \neq 0 \quad \implies \quad i \in \mathcal{C}_t(a),$$

and the analogous implication holds after shifting time by one step. Thus, for the learned current and past coordinates indexed by j and a, respectively, only true components indexed by $\mathcal{C}_t(j)$ and $\mathcal{C}_{t-1}(a)$ can have nonzero derivatives.

Among the index pairs in $\mathcal{C}_t(j) \times \mathcal{C}_{t-1}(a)$, only those corresponding to true cross-time relations need to be retained, because Lemma 4 makes the derivative vector zero for every cross-time nonedge. We therefore define the past-directed candidate block as the following set of index pairs:

$$\mathcal{B}^{-}(j,a) := \left\{(k,m) \in [d_z]^2 \;\middle|\; k \in \mathcal{C}_t(j),\ m \in \mathcal{C}_{t-1}(a),\ \{z_{t-1,m}, z_{t,k}\} \in \mathcal{E}^{-}_{\text{lag}}\right\}.$$

In the future direction, define

$$\mathcal{B}^+(j,b) := \left\{ (k,m) \in [d_z]^2 \;\middle|\; k \in \mathcal{C}_t(j),\ m \in \mathcal{C}_{t+1}(b),\ \{z_{t,k}, z_{t+1,m}\} \in \mathcal{E}^+_{\text{lag}} \right\}.$$

Here, $\mathcal{B}^-(j,a)$ contains index pairs, whereas edges in $\mathcal{E}^-_{\text{lag}}$ are written as unordered pairs of the corresponding latent-variable nodes. Using these candidate blocks, we next express each learned cross-time derivative vector as a linear combination of the corresponding true derivative vectors.

Lemma 5 (Transformation of cross-time derivatives over local candidate blocks) *Fix $(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})$ and choose the corresponding local permutations $\pi_{\hat{\boldsymbol{z}}_t}$ and $\pi_{\hat{\boldsymbol{z}}_{t-1}}$ from Lemma 2. For any $j, a \in [d_z]$,*

$$\hat{\boldsymbol{v}}^-_{\pi_{\hat{\boldsymbol{z}}_t}(j), \pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}) = \sum_{(k,m)\in\mathcal{B}^-(j,a)} \chi_{km} \boldsymbol{v}^-_{km}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}),$$

where $\boldsymbol{z}_t = H(\hat{\boldsymbol{z}}_t)$, $\boldsymbol{z}_{t-1} = H(\hat{\boldsymbol{z}}_{t-1})$, and

$$\chi_{km} := \frac{\partial z_{t,k}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} \frac{\partial z_{t-1,m}}{\partial \hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}}.$$

The analogous equality holds in the future direction after shifting time by one step.

Proof. We prove the past case. Transforming the conditional density under $\boldsymbol{z}_s = H(\hat{\boldsymbol{z}}_s)$ gives

$$p(\hat{\boldsymbol{z}}_t \mid \hat{\boldsymbol{z}}_{t-1}, u_t) = p\big(H(\hat{\boldsymbol{z}}_t) \mid H(\hat{\boldsymbol{z}}_{t-1}), u_t\big) \left|\det J_H(\hat{\boldsymbol{z}}_t)\right|.$$

Applying this equality under the reference and comparison values and taking their difference cancels $\log|\det J_H(\hat{\boldsymbol{z}}_t)|$ because H is independent of the auxiliary variable. Thus, for every n,

$$\widehat{\Delta}^-_n(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}) = \Delta^-_n\big(H(\hat{\boldsymbol{z}}_t), H(\hat{\boldsymbol{z}}_{t-1})\big).$$

Differentiate this equality once with respect to the learned current coordinate $\hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}$ and once with respect to the learned past coordinate $\hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}$. Since the transformation acts separately at each time step, the chain rule leaves only products of first-order Jacobian entries:

$$\frac{\partial^2 \widehat{\Delta}^-_n(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)} \partial \hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}} = \sum_{k=1}^{d_z} \sum_{m=1}^{d_z} \chi_{km} \frac{\partial^2 \Delta^-_n(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})}{\partial z_{t,k} \partial z_{t-1,m}}.$$

Collecting this equality over $n = 1, \ldots, N_{\text{aux}}$ gives the corresponding equality for $\boldsymbol{v}^-$. By Lemma 3, χ_{km} can be nonzero only if $k \in \mathcal{C}_t(j)$ and $m \in \mathcal{C}_{t-1}(a)$. Moreover, Lemma 4 gives $\boldsymbol{v}^-_{km} = 0$ whenever $\{z_{t-1,m}, z_{t,k}\} \notin \mathcal{E}^-_{\text{lag}}$. The sum can therefore be restricted to $\mathcal{B}^-(j,a)$. The equality for the future case follows by shifting time forward by one step. □

Figure 3 illustrates how differences in cross-time relations can resolve intimate-neighbor ambiguities that remain after considering the instantaneous Markov network alone. Based on the preceding results, we impose the following three conditions to eliminate the remaining residual mixing. The first is a local linear-independence condition that prevents cancellation among the cross-time derivative vectors in the preceding transformation. The second is a cross-time edge-count minimality condition that rules out additional learned cross-time edges. The third requires each intimate-neighbor ambiguity left by the instantaneous Markov network to be resolved by a difference in the cross-time relations with some past or future component.

Assumption 2 (Cross-time variation, separation, and edge-count minimality) *For every time t, the following conditions hold.*

(i) Local cross-time variability. *For every $j, a \in [d_z]$, at almost every $(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$,*

$$\operatorname{rank}\left[\boldsymbol{v}^-_{km}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})\right]_{(k,m)\in\mathcal{B}^-(j,a)} = |\mathcal{B}^-(j,a)|,$$

where the vectors $\boldsymbol{v}^-_{km}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$ form the columns of the matrix.

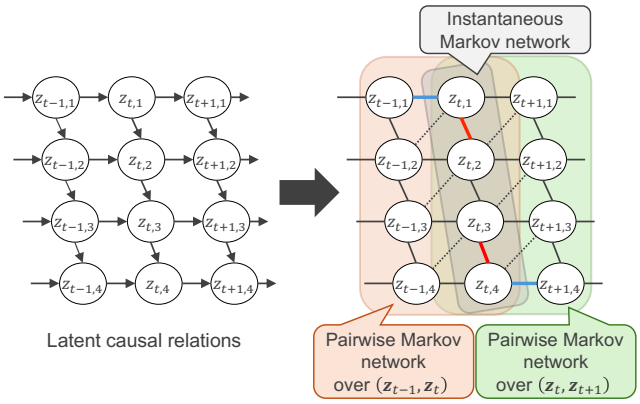


Figure 3: Illustration of how cross-time relations resolve intimate-neighbor ambiguities. The left panel shows a four-component latent causal model, and the right panel shows the corresponding instantaneous and cross-time pairwise Markov networks. In the instantaneous network, $z_{t,2} \in \Psi_{\mathcal{M}_z}(z_{t,1})$ and $z_{t,3} \in \Psi_{\mathcal{M}_z}(z_{t,4})$, as indicated by the red edges. The blue cross-time relations resolve the ambiguity of the first pair through $z_{t-1,1}$ and that of the second through $z_{t+1,4}$, illustrating that either a past or a future relation can suffice for each intimate-neighbor pair. Dotted edges denote conditional-dependence relations that are not direct causal edges. All Markov-network relations are defined conditionally on the auxiliary variable.

(ii) Cross-time edge-count minimality. *The learned cross-time edge set has no greater cardinality than the true one:*

$$|\widehat{\mathcal{E}}^{-}_{\text{lag}}| \leq |\mathcal{E}^{-}_{\text{lag}}|.$$

(iii) Intimate-neighbor separation. *For every $z_{t,j} \in \Psi_{\mathcal{M}_z}(z_{t,i})$, at least one of the following conditions holds:*

1. *There exists a past component $z_{t-1,a}$ such that*

$$\{z_{t-1,a}, z_{t,i}\} \in \mathcal{E}^{-}_{\text{lag}}, \qquad \{z_{t-1,a}, z_{t,j}\} \notin \mathcal{E}^{-}_{\text{lag}}.$$

2. *There exists a future component $z_{t+1,b}$ such that*

$$\{z_{t,i}, z_{t+1,b}\} \in \mathcal{E}^{+}_{\text{lag}}, \qquad \{z_{t,j}, z_{t+1,b}\} \notin \mathcal{E}^{+}_{\text{lag}}.$$

We first note a dimensional requirement for the local cross-time variability condition in Assumption 2(i). A necessary condition is

$$N_{\text{aux}} \geq \max_{j,a} |\mathcal{B}^{-}(j,a)|.$$

This dimensional bound is necessary but does not by itself guarantee linear independence. Because Assumption 2 holds at every time step, Conditions (i) and (ii) also apply in the future direction after shifting time forward by one step. Thus, no separate variability, dimensional, or edge-count condition is required for the future case. We next combine the transformation in Lemma 5 with Condition (i) to derive Jacobian constraints from learned cross-time nonedges.

Lemma 6 (Jacobian condition implied by a learned cross-time nonedge) *Fix $(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})$ and choose the corresponding local permutations $\pi_{\hat{\boldsymbol{z}}_t}$ and $\pi_{\hat{\boldsymbol{z}}_{t-1}}$ from Lemma 2. Suppose that $\{z_{t-1,a}, z_{t,i}\} \in \mathcal{E}^{-}_{\text{lag}}$ and $i \in \mathcal{C}_t(j)$. Then*

$$\{\hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}, \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}\} \notin \widehat{\mathcal{E}}^{-}_{\text{lag}} \qquad \Longrightarrow \frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} \frac{\partial z_{t-1,a}}{\partial \hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}} = 0.$$

The future-directed result follows by applying the same argument after shifting time forward by one step.

Proof. We prove the past case. The assumed learned nonedge and Lemma 4 imply

$$\hat{\boldsymbol{v}}^{-}_{\pi_{\hat{\boldsymbol{z}}_t}(j),\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}) = 0.$$

Applying Lemma 5 gives

$$0 = \sum_{(k,m)\in\mathcal{B}^-(j,a)} \chi_{km} \boldsymbol{v}^{-}_{km}(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}).$$

Since $a \in \mathcal{C}_{t-1}(a)$ by definition, the conditions $i \in \mathcal{C}_t(j)$ and $\{z_{t-1,a}, z_{t,i}\} \in \mathcal{E}^{-}_{\text{lag}}$ imply $(i,a) \in \mathcal{B}^-(j,a)$. By the linear independence in Assumption 2(i), all coefficients in this linear combination must be zero. In particular, $\chi_{ia} = 0$, which gives the claimed Jacobian condition.

The rank condition in Assumption 2(i) holds almost everywhere, so it remains to consider points where this condition may fail. At such a point, the Jacobian entries associated with the chosen local permutations are nonzero by Lemma 2. By continuity, these entries remain nonzero in a sufficiently small neighborhood, so the same local permutations can be used there. Since the rank condition holds almost everywhere, every sufficiently small neighborhood contains points where it holds. At such points, the preceding argument gives $\chi_{ia} = 0$, and continuity then gives $\chi_{ia} = 0$ at the original point as well. The future-directed result follows by applying the same argument after shifting time forward by one step. □

Lemma 6 shows that every true cross-time edge is preserved under the local permutations. Combined with cross-time edge-count minimality, this yields the local correspondence between the true and learned cross-time edge sets stated next.

Lemma 7 (Local isomorphism of cross-time conditional-dependence relations) *Fix $(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})$ and choose the corresponding local permutations $\pi_{\hat{\boldsymbol{z}}_t}$ and $\pi_{\hat{\boldsymbol{z}}_{t-1}}$ from Lemma 2. Under Assumptions 1 and 2, for every $i, a \in [d_z]$,*

$$\{z_{t-1,a}, z_{t,i}\} \in \mathcal{E}^{-}_{\text{lag}} \quad \Longleftrightarrow \quad \{\hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}, \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}\} \in \widehat{\mathcal{E}}^{-}_{\text{lag}}.$$

The future-directed result follows by applying the same argument after shifting time forward by one step.

Proof. We first show that every true cross-time relation is preserved under the local permutations. Take $\{z_{t-1,a}, z_{t,i}\} \in \mathcal{E}^{-}_{\text{lag}}$. Suppose, for contradiction, that the corresponding learned pair is a nonedge:

$$\{\hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}, \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}\} \notin \widehat{\mathcal{E}}^{-}_{\text{lag}}.$$

Applying Lemma 6 with $j = i$ then gives

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}} \frac{\partial z_{t-1,a}}{\partial \hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}} = 0.$$

Both factors are nonzero by Lemma 2, yielding a contradiction. Therefore,

$$\{z_{t-1,a}, z_{t,i}\} \in \mathcal{E}^{-}_{\text{lag}} \quad \Longrightarrow \quad \{\hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}, \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}\} \in \widehat{\mathcal{E}}^{-}_{\text{lag}}.$$

Because $\pi_{\hat{\boldsymbol{z}}_t}$ and $\pi_{\hat{\boldsymbol{z}}_{t-1}}$ are permutations, this edge-preserving map is injective, and hence

$$|\mathcal{E}^{-}_{\text{lag}}| \leq |\widehat{\mathcal{E}}^{-}_{\text{lag}}|.$$

Assumption 2(ii) gives the reverse inequality, so the two edge sets have equal cardinality. Since every true cross-time relation is already mapped to a learned one, the learned edge set contains no additional relations, proving the equivalence. The future-directed result follows by applying the same argument after shifting time forward by one step. □

Lemma 3 leaves only intimate neighbors as candidates for residual mixing. Assumption 2(iii) requires each such pair to differ in its cross-time relation with at least one past or future component. The local cross-time correspondence established in Lemma 7 then allows this difference to rule out the remaining residual mixing, as shown next.

Lemma 8 (Elimination of residual mixing associated with intimate neighbors) *Under Assumptions 1 and 2, for every $z_{t,j} \in \Psi_{\mathcal{M}_z}(z_{t,i})$,*

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} = 0.$$

Proof. Fix $z_{t,j} \in \Psi_{\mathcal{M}_z}(z_{t,i})$ and suppose, for contradiction, that

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} \neq 0.$$

By the definition of $\mathcal{C}_t(j)$, $z_{t,j} \in \Psi_{\mathcal{M}_z}(z_{t,i})$ implies $i \in \mathcal{C}_t(j)$.

First suppose that the past case in Assumption 2(iii) holds for some component $z_{t-1,a}$:

$$\{z_{t-1,a}, z_{t,i}\} \in \mathcal{E}^-_{\text{lag}}, \qquad \{z_{t-1,a}, z_{t,j}\} \notin \mathcal{E}^-_{\text{lag}}.$$

By Lemma 7, the second relation implies

$$\{\hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}, \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}\} \notin \widehat{\mathcal{E}}^-_{\text{lag}}.$$

Together with $\{z_{t-1,a}, z_{t,i}\} \in \mathcal{E}^-_{\text{lag}}$ and $i \in \mathcal{C}_t(j)$, Lemma 6 then gives

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} \frac{\partial z_{t-1,a}}{\partial \hat{z}_{t-1,\pi_{\hat{\boldsymbol{z}}_{t-1}}(a)}} = 0.$$

The second factor is nonzero by Lemma 2, so

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} = 0,$$

contradicting the assumed residual mixing.

If instead the future case in Assumption 2(iii) holds for some $z_{t+1,b}$, then

$$\{z_{t,i}, z_{t+1,b}\} \in \mathcal{E}^+_{\text{lag}}, \qquad \{z_{t,j}, z_{t+1,b}\} \notin \mathcal{E}^+_{\text{lag}}.$$

Applying Lemma 7 and the future-directed form of Lemma 6 gives, by the same argument,

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} = 0,$$

again contradicting the assumed residual mixing. Thus, residual mixing is ruled out for every intimate-neighbor pair. □

A.2.4 From Local Jacobian Conditions to Component-Wise Transformations

With residual mixing eliminated, we now use the resulting local Jacobian structure to obtain a global component-wise transformation. Lemma 3 restricts potentially nonzero off-diagonal Jacobian entries to intimate-neighbor pairs, while Lemma 8 shows that these entries are also zero. Indeed, fix $\hat{\boldsymbol{z}}_t$ and its local permutation $\pi_{\hat{\boldsymbol{z}}_t}$. If, for some $i \neq j$,

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} \neq 0,$$

then Lemma 3 gives $z_{t,j} \in \Psi_{\mathcal{M}_z}(z_{t,i})$, whereas Lemma 8 forces the same derivative to be zero. Therefore,

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} = 0 \qquad (i \neq j).$$

The local Jacobian is thus diagonal with nonzero diagonal entries after its columns are arranged according to $\pi_{\hat{\boldsymbol{z}}_t}$. The next lemma makes this pointwise permutation independent of the state.

Lemma 9 (Globalizing the local latent-component permutation) *Let H be a C^2 diffeomorphism from the range of the learned representation onto $\mathbb{R}^{d_z}$. Suppose that for every $\hat{\boldsymbol{z}}_t$ in the domain of H there exists a permutation $\pi_{\hat{\boldsymbol{z}}_t}$ such that, for every i,*

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,k}} = 0 \quad (k \neq \pi_{\hat{\boldsymbol{z}}_t}(i)), \qquad \frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}} \neq 0.$$

Then there exist a permutation π_z independent of $\hat{\boldsymbol{z}}_t$ and univariate C^2 diffeomorphisms ξ_i such that

$$\hat{z}_{t,\pi_z(i)} = \xi_i(z_{t,i}) \qquad (\forall i).$$

Proof. For each $\hat{\boldsymbol{z}}_t$ in the domain of H, the assumed derivative pattern uniquely determines $\pi_{\hat{\boldsymbol{z}}_t}$:

$$\frac{\partial z_{t,i}}{\partial \hat{z}_{t,k}} = 0 \quad \text{if } k \neq \pi_{\hat{\boldsymbol{z}}_t}(i), \qquad \frac{\partial z_{t,i}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(i)}} \neq 0.$$

By continuity of J_H, the d_z selected nonzero entries at any fixed point remain nonzero in a common neighborhood. Since the assumption allows only one nonzero entry in each row at every point, the same permutation holds throughout this neighborhood. Thus, $\pi_{\hat{\boldsymbol{z}}_t}$ is locally constant.

The domain of H is connected because H is a diffeomorphism from that domain onto the connected space $\mathbb{R}^{d_z}$. Therefore, the permutation is constant over the domain; denote it by π_z.

Differentiating the identity $H(H^{-1}(\boldsymbol{z})) = \boldsymbol{z}$ gives

$$J_{H^{-1}}(\boldsymbol{z}_t) = J_H\big(H^{-1}(\boldsymbol{z}_t)\big)^{-1}.$$

Since J_H has exactly one nonzero entry in each row and column under the fixed permutation π_z, its inverse has the corresponding pattern. Consequently, writing $(H^{-1})_k$ for the k-th component of H^{-1}, for every i and every $\boldsymbol{z}_t$,

$$\frac{\partial (H^{-1})_{\pi_z(i)}(\boldsymbol{z}_t)}{\partial z_{t,j}} = 0 \quad (j \neq i), \qquad \frac{\partial (H^{-1})_{\pi_z(i)}(\boldsymbol{z}_t)}{\partial z_{t,i}} \neq 0.$$

Since the domain of H^{-1} is the connected set $\mathbb{R}^{d_z}$, the zero cross-component derivatives imply that $(H^{-1})_{\pi_z(i)}$ depends only on z_i. Denote the resulting univariate function by ξ_i. Because H^{-1} is bijective and has this component-wise form, each ξ_i is bijective between the corresponding coordinate ranges. Since H^{-1} is C^2 and the derivative of each ξ_i is everywhere nonzero, each ξ_i is a univariate C^2 diffeomorphism. Thus,

$$\hat{z}_{t,\pi_z(i)} = \xi_i(z_{t,i}) \qquad (i = 1, \ldots, d_z).$$

□

The preceding results yield the following latent-component identification theorem. The first case shows that no cross-time conditions are needed when all intimate-neighbor sets are empty, whereas the second case gives the general result used in the main theorem.

Theorem 4 (Identification of latent components) *Consider a learned representation related to the true representation by a C^2 diffeomorphism H, independent of time and the value of the auxiliary variable, from the range of the learned representation onto $\mathbb{R}^{d_z}$, with $\boldsymbol{z}_t = H(\hat{\boldsymbol{z}}_t)$, and suppose Assumption 1 holds. Suppose further that either*

1. *$\Psi_{\mathcal{M}_z}(z_{t,i}) = \emptyset$ for every i; or*
2. *Assumption 2 holds.*

Then there exists a permutation π_z, independent of time and the values of the latent variables, under which $\mathcal{M}_z$ and $\mathcal{M}_{\hat{z}}$ are isomorphic via the correspondence $z_{t,i} \mapsto \hat{z}_{t,\pi_z(i)}$. Moreover, there exist univariate C^2 diffeomorphisms $\xi_1, \ldots, \xi_{d_z}$ such that

$$\hat{z}_{t,\pi_z(i)} = \xi_i(z_{t,i}) \qquad (i = 1, \ldots, d_z).$$

Proof. In the first case, the argument in Section A.2.2 shows that, at every point, arranging the local Jacobian columns according to the local permutation produces a diagonal matrix with nonzero diagonal entries. In the second case, Lemma 3 restricts all off-diagonal mixing to intimate-neighbor pairs, and Lemma 8 eliminates this remaining mixing. Thus, in either case, the local Jacobian becomes diagonal with nonzero diagonal entries at every point after its columns are arranged according to some local permutation $\pi_{\hat{\boldsymbol{z}}_t}$.

Lemma 9 gives a permutation π_z independent of $\hat{\boldsymbol{z}}_t$ and univariate C^2 diffeomorphisms ξ_i such that

$$\hat{z}_{t,\pi_z(i)} = \xi_i(z_{t,i}) \qquad (\forall i).$$

Substituting $\pi_{\hat{\boldsymbol{z}}_t} = \pi_z$ into the local isomorphism of the instantaneous Markov networks established in Theorem 3 shows that $\mathcal{M}_z$ and $\mathcal{M}_{\hat{z}}$ are isomorphic under the same fixed permutation. Finally, because the coordinate transformation H is common to all time steps, π_z and the functions ξ_i are also independent of time. □

A.3 Transition-Noise Identification by IIA

This appendix adapts the transition-noise identification argument of independent innovation analysis (IIA) (Morioka et al., 2021) to the encoder and learned inverse transition function used in our framework. Here, transition noise plays the role of the innovations in IIA. In the sufficient setting proved in this appendix, we take $\mathcal{X} = \mathbb{R}^{d_z}$, as in the IIA argument. We consider two contrastive-learning approaches used in IIA (Morioka et al., 2021): time-contrastive learning (TCL) and generalized contrastive learning (GCL). TCL predicts a discrete auxiliary-variable value from the estimated transition noise and a learned feature of the past observations, whereas GCL distinguishes matched tuples from tuples in which only the auxiliary-variable value is permuted. Let ϕ be the encoder and define the maps $\rho_{\mathrm{aug}}, \hat{\rho}_{\mathrm{aug}} : \mathbb{R}^{2d_z} \to \mathbb{R}^{2d_z}$ by

$$\rho_{\mathrm{aug}}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) := (\boldsymbol{\epsilon}_t, \boldsymbol{x}_{t-1}), \qquad \hat{\rho}_{\mathrm{aug}}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) := \big(\hat{r}\big(\phi(\boldsymbol{x}_t), \phi(\boldsymbol{x}_{t-1})\big), \boldsymbol{x}_{t-1}\big)\,.$$

For the theoretical setting with $L = 1$, the history feature $\boldsymbol{h}_t = \eta(\phi(\boldsymbol{x}_{t-1}))$ is a deterministic function of $\boldsymbol{x}_{t-1}$, so we write the regression functions below in terms of $\boldsymbol{x}_{t-1}$. Accordingly, in the analysis below, we regard the implemented regression function $T(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t, u_t)$ as a function of $(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{x}_{t-1}, u_t)$. For either TCL or GCL, the regression function decomposes over the learned transition-noise components as

$$T(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{x}_{t-1}, u_t) = \sum_{i=1}^{d_z} \sum_{k=1}^{K} \psi_{ik}(\hat{\epsilon}_{t,i}) \mu_{ik}(u_t) + \varphi(\boldsymbol{x}_{t-1}, u_t) + B(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{x}_{t-1}). \tag{18}$$

For TCL, $\mu_{ik}(u) = w_{uik}$, and b_u is absorbed into φ. The auxiliary-variable-independent term B can be omitted because it cancels when the class logits are compared. Here, ψ_{ik} is a scalar-valued function of the estimated transition-noise component $\hat{\epsilon}_{t,i}$, μ_{ik} modulates its coefficient with the auxiliary variable, and the remaining terms account for dependence involving the past observation and terms independent of the auxiliary variable. The corresponding finite-sample TCL and GCL objectives are given in Appendix B. A regression function is population-optimal when it minimizes the expected multiclass cross-entropy for TCL or the expected binary cross-entropy for GCL under the corresponding population distributions. Let $u_t^{(0)}$ denote a reference auxiliary-variable value and let $u_t^{(n)}$ denote the n-th comparison value. For both TCL and GCL, comparing the reference value with each comparison value gives

$$T^\star(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{x}_{t-1}, u_t^{(n)}) - T^\star(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{x}_{t-1}, u_t^{(0)}) = \log \frac{p(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{x}_{t-1} \mid u_t^{(n)})}{p(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{x}_{t-1} \mid u_t^{(0)})} + c_n. \tag{19}$$

Assumption 3 (Conditions for transition-noise identification by IIA) *The following conditions hold.*

(i) Invertibility and regularity. *The learned inverse transition function $\hat{r}$ is C^1, and ρ_{aug}, $\hat{\rho}_{\mathrm{aug}}$, and the observation function g are C^2 diffeomorphisms on their respective domains.*

(ii) Exponential-family modulation. *There exist $K \geq 2$, base measures Q_i, normalizing terms Z_i, twice continuously differentiable sufficient statistics s_{ik}, and natural parameters λ_{ik} depending on*

the auxiliary variable such that

$$p(\boldsymbol{\epsilon}_t \mid u_t) = \prod_{i=1}^{d_z} \frac{Q_i(\epsilon_{t,i})}{Z_i(u_t)} \exp\left[\sum_{k=1}^{K} s_{ik}(\epsilon_{t,i})\lambda_{ik}(u_t)\right].$$

For each i, the sufficient statistics are nonredundant up to a constant: if $\sum_{k=1}^{K} c_k s_{ik}(e)$ is constant in e, then $c_1 = \cdots = c_K = 0$. Let $\boldsymbol{\lambda}(u)$ collect all $d_z K$ natural parameters. For an arbitrarily selected reference value $u_t^{(0)}$ and distinct comparison values $u_t^{(n_1)}, \ldots, u_t^{(n_{d_z K})}$ satisfying $1 \le n_1 < \cdots < n_{d_z K} \le N_{\text{aux}}$, the matrix

$$\boldsymbol{\Lambda} := \left[\boldsymbol{\lambda}(u_t^{(n_1)}) - \boldsymbol{\lambda}(u_t^{(0)}), \ldots, \boldsymbol{\lambda}(u_t^{(n_{d_z K})}) - \boldsymbol{\lambda}(u_t^{(0)})\right]$$

has rank $d_z K$. For a finite-valued auxiliary variable, this condition requires at least $d_z K + 1$ distinct values, including the reference.

(iii) Population regression and learned-feature rank. *The learned regression function attains a population optimum satisfying equation 19. For the feature functions ψ_{ik} in this solution, define*

$$\boldsymbol{D}_{\psi,i}(e) := \left[\frac{\partial \psi_{ik}(e)}{\partial e}, \frac{\partial^2 \psi_{ik}(e)}{\partial e^2}\right]_{k=1}^{K} \in \mathbb{R}^{K\times 2},$$

and suppose that $\operatorname{rank} \boldsymbol{D}_{\psi,i}(e) = 2$ for every i and every e. This condition plays the role of the learned-feature nonredundancy condition in the IIA argument of Morioka et al. (2021); we state the required derivative-rank property directly for simplicity.

Under these conditions, the transition-noise components are identified up to a permutation and component-wise C^2 diffeomorphisms.

Theorem 5 (Transition-noise identification) *Under Assumption 3, there exist a permutation π_ϵ and component-wise C^2 diffeomorphisms q_i, all independent of time and the auxiliary-variable value, such that*

$$\hat{\epsilon}_{t,\pi_\epsilon(i)} = q_i(\epsilon_{t,i}) \qquad (i = 1, \ldots, d_z).$$

Proof. The assumptions $\boldsymbol{\epsilon}_t \perp \boldsymbol{z}_{t-1} \mid u_t$ from the problem setup and $\boldsymbol{x}_{t-1} = g(\boldsymbol{z}_{t-1})$ imply $\boldsymbol{\epsilon}_t \perp \boldsymbol{x}_{t-1} \mid u_t$. By Assumption 3(ii), the density ratio between a reference value and a comparison value in the true augmented coordinates $(\boldsymbol{\epsilon}_t, \boldsymbol{x}_{t-1})$ decomposes as

$$\log \frac{p(\boldsymbol{\epsilon}_t, \boldsymbol{x}_{t-1} \mid u_t^{(n)})}{p(\boldsymbol{\epsilon}_t, \boldsymbol{x}_{t-1} \mid u_t^{(0)})} = \sum_{i=1}^{d_z}\sum_{k=1}^{K} s_{ik}(\epsilon_{t,i}) \left[\lambda_{ik}(u_t^{(n)}) - \lambda_{ik}(u_t^{(0)})\right] + P_n(\boldsymbol{x}_{t-1}). \tag{20}$$

Here, $P_n(\boldsymbol{x}_{t-1})$ collects the density-ratio term for the past observations and the normalizing terms of the exponential-family distributions.

The transformation $\hat{\rho}_{\text{aug}} \circ \rho_{\text{aug}}^{-1}$ between the true and learned augmented coordinates is a diffeomorphism that does not depend on the auxiliary variable. Consequently, the Jacobian terms from the change of variables cancel in the density ratio between the reference and comparison values, and at corresponding points,

$$\frac{p(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{x}_{t-1} \mid u_t^{(n)})}{p(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{x}_{t-1} \mid u_t^{(0)})} = \frac{p(\boldsymbol{\epsilon}_t, \boldsymbol{x}_{t-1} \mid u_t^{(n)})}{p(\boldsymbol{\epsilon}_t, \boldsymbol{x}_{t-1} \mid u_t^{(0)})}. \tag{21}$$

Let $\boldsymbol{y} := (\boldsymbol{\epsilon}_t, \boldsymbol{x}_{t-1})$, and write the learned transition noise expressed in the true augmented coordinates as $\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{y})$. Stack the learned features, sufficient statistics, and learned coefficients into $\boldsymbol{\psi}(\hat{\boldsymbol{\epsilon}}_t)$, $\boldsymbol{s}(\boldsymbol{\epsilon}_t)$, and $\boldsymbol{\mu}(u_t)$, respectively, and define

$$\boldsymbol{M} := \left[\boldsymbol{\mu}(u_t^{(n_1)}) - \boldsymbol{\mu}(u_t^{(0)}), \ldots, \boldsymbol{\mu}(u_t^{(n_{d_z K})}) - \boldsymbol{\mu}(u_t^{(0)})\right].$$

Taking the difference in equation 19 eliminates the term B because it does not depend on the auxiliary variable. Using equation 20 and equation 21 and stacking the resulting equations over the comparison values gives

$$\boldsymbol{M}^\top \boldsymbol{\psi}\big(\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{y})\big) + \Phi(\boldsymbol{x}_{t-1}) = \boldsymbol{\Lambda}^\top \boldsymbol{s}(\boldsymbol{\epsilon}_t) + P(\boldsymbol{x}_{t-1}) + c, \tag{22}$$

where Φ and P depend only on the past observation and c stacks the constant terms.

By nonredundancy of the sufficient statistics, the space spanned by their derivative vectors has dimension K. Hence, there exist evaluation points $e_i^{(1)}, \ldots, e_i^{(K)}$ at which these derivative vectors are linearly independent. Define the matrices obtained by evaluating the derivatives of the true sufficient statistics and learned features at these points as

$$\boldsymbol{A}_i := \left[\frac{\partial s_{ik}(e_i^{(m)})}{\partial e} \right]_{k,m=1}^{K}, \qquad \boldsymbol{A} := \operatorname{diag}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{d_z}),$$
$$\widetilde{\boldsymbol{A}} := \left[\frac{\partial}{\partial \epsilon_{t,i}} \boldsymbol{\psi}\big(\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{y})\big)\Big|_{\epsilon_{t,i}=e_i^{(m)}} \right]_{(i,m)}.$$

Differentiating equation 22 with respect to each $\epsilon_{t,i}$ eliminates the terms depending only on $\boldsymbol{x}_{t-1}$ and the constants. Evaluation at the selected points gives $\boldsymbol{M}^\top \widetilde{\boldsymbol{A}} = \boldsymbol{\Lambda}^\top \boldsymbol{A}$. Since $\boldsymbol{\Lambda}$ and $\boldsymbol{A}$ are nonsingular, the equality above implies that $\boldsymbol{M}$ is also nonsingular.

For distinct j_1 and j_2, take the mixed derivative of equation 22 with respect to ϵ_{t,j_1} and ϵ_{t,j_2}. Since the right-hand side is a sum of univariate functions of the individual transition-noise components, its mixed derivative vanishes; the nonsingularity of $\boldsymbol{M}$ therefore gives $\partial^2 \boldsymbol{\psi}(\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{y}))/\partial \epsilon_{t,j_1} \partial \epsilon_{t,j_2} = 0$. Expanding the K entries corresponding to $\hat{\epsilon}_{t,i}$ by the chain rule gives

$$\boldsymbol{D}_{\psi,i}(\hat{\epsilon}_{t,i}) \begin{pmatrix} \dfrac{\partial^2 \hat{\epsilon}_{t,i}}{\partial \epsilon_{t,j_1} \partial \epsilon_{t,j_2}} \\ \dfrac{\partial \hat{\epsilon}_{t,i}}{\partial \epsilon_{t,j_1}} \dfrac{\partial \hat{\epsilon}_{t,i}}{\partial \epsilon_{t,j_2}} \end{pmatrix} = 0,$$
$$\frac{\partial \hat{\epsilon}_{t,i}}{\partial \epsilon_{t,j_1}} \frac{\partial \hat{\epsilon}_{t,i}}{\partial \epsilon_{t,j_2}} = 0, \tag{23}$$

where the second relation follows from the full column rank of $\boldsymbol{D}_{\psi,i}$. Similarly, taking the mixed derivative of equation 22 with respect to $\epsilon_{t,j}$ and $x_{t-1,a}$ and using the full column rank of $\boldsymbol{D}_{\psi,i}$ gives

$$\frac{\partial \hat{\epsilon}_{t,i}}{\partial \epsilon_{t,j}} \frac{\partial \hat{\epsilon}_{t,i}}{\partial x_{t-1,a}} = 0. \tag{24}$$

The composition $\hat{\rho}_{\mathrm{aug}} \circ \rho_{\mathrm{aug}}^{-1}$ has the form $(\boldsymbol{\epsilon}_t, \boldsymbol{x}_{t-1}) \mapsto (\hat{\boldsymbol{\epsilon}}_t(\boldsymbol{\epsilon}_t, \boldsymbol{x}_{t-1}), \boldsymbol{x}_{t-1})$, and its Jacobian is block triangular:

$$J_{\hat{\rho}_{\mathrm{aug}} \circ \rho_{\mathrm{aug}}^{-1}} = \begin{pmatrix} \dfrac{\partial \hat{\boldsymbol{\epsilon}}_t}{\partial \boldsymbol{\epsilon}_t} & \dfrac{\partial \hat{\boldsymbol{\epsilon}}_t}{\partial \boldsymbol{x}_{t-1}} \\ 0 & I \end{pmatrix}. \tag{25}$$

Since this map is a diffeomorphism, its Jacobian is nonsingular, and equation 25 therefore implies that $\partial \hat{\boldsymbol{\epsilon}}_t / \partial \boldsymbol{\epsilon}_t$ is nonsingular. Equation 23 implies that each row of this Jacobian has at most one nonzero entry, while nonsingularity implies exactly one nonzero entry in each row and column. Thus, its support defines a permutation of the transition-noise components at every point. Equation 24 then implies that each learned transition-noise component does not depend on the past observation. Continuity makes this permutation locally constant, and connectedness of $\mathbb{R}^{2d_z}$ therefore fixes it globally to a single permutation π_ϵ. For each i, the fixed permutation together with the product-zero conditions implies that $\hat{\epsilon}_{t,\pi_\epsilon(i)}$ has zero derivatives with respect to all coordinates other than $\epsilon_{t,i}$. Since the augmented-coordinate domain is $\mathbb{R}^{2d_z}$, it follows that $\hat{\epsilon}_{t,\pi_\epsilon(i)}$ is a function only of $\epsilon_{t,i}$, yielding

$$\hat{\epsilon}_{t,\pi_\epsilon(i)} = q_i(\epsilon_{t,i}) \qquad (i = 1, \ldots, d_z). \tag{26}$$

Since $\hat{\rho}_{\mathrm{aug}} \circ \rho_{\mathrm{aug}}^{-1}$ is a diffeomorphism that preserves the past observation, the map $\boldsymbol{\epsilon}_t \mapsto \hat{\boldsymbol{\epsilon}}_t$ is bijective for each fixed $\boldsymbol{x}_{t-1}$, and the corresponding derivatives are nonzero. Thus, every q_i is a C^2 diffeomorphism.

□

A.4 Causal-Graph Identification from the Jacobian of the Inverse Transition Function

This section combines transition-noise identification from Theorem 5 with latent-component identification from Theorem 4 to identify the true instantaneous and lagged causal edges from the Jacobian support of the learned inverse transition function $\hat{r}$. We first relate the Jacobian supports of the true and learned inverse transition functions. Acyclicity of the instantaneous DAG then determines the correspondence between learned output rows and current latent components. Finally, the Jacobian support of the true inverse transition function r is related to the direct causal edges defined by the structural functions f_i. Throughout this section, suppose that the conditions of Theorems 5 and 4 hold.

Let π_ϵ and π_z denote the permutations for the transition-noise and latent components, respectively. Using component-wise C^2 diffeomorphisms q_i and ξ_i, write

$$\hat{\epsilon}_{t,\pi_\epsilon(i)} = q_i(\epsilon_{t,i}), \qquad \hat{z}_{t,\pi_z(i)} = \xi_i(z_{t,i}). \tag{27}$$

Here, π_ϵ and π_z specify the orderings of the learned output and latent input components, respectively, and are not assumed to be identical. When describing the functional support of f_i, we regard it as a function of all candidate instantaneous and lagged variables, with derivatives with respect to nonparents understood to be identically zero. We write the true inverse transition function as $r = (r_1, \dots, r_{d_z})$, where $\epsilon_{t,i} = r_i(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$.

Lemma 10 (Jacobian-support correspondence for inverse transition functions) *The Jacobian supports of the true and learned inverse transition functions coincide up to the row permutation π_ϵ and column permutation π_z. Specifically, for $\ell \in \{0, 1\}$ and all i, j,*

$$\frac{\partial \hat{r}_{\pi_\epsilon(i)}}{\partial \hat{z}_{t-\ell,\pi_z(j)}} \not\equiv 0 \quad \Longleftrightarrow \quad \frac{\partial r_i}{\partial z_{t-\ell,j}} \not\equiv 0. \tag{28}$$

Proof. Substituting $\boldsymbol{\epsilon}_t = r(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$ and $\hat{\boldsymbol{\epsilon}}_t = \hat{r}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})$ into the transition-noise correspondence in equation 27 gives

$$\hat{r}_{\pi_\epsilon(i)}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1}) = q_i\big(r_i(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})\big).$$

The latent-component correspondence in equation 27 also gives $z_{t-\ell,j} = \xi_j^{-1}(\hat{z}_{t-\ell,\pi_z(j)})$. Differentiating the preceding identity with respect to $\hat{z}_{t-\ell,\pi_z(j)}$ and applying the chain rule yields

$$\frac{\partial \hat{r}_{\pi_\epsilon(i)}(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t-1})}{\partial \hat{z}_{t-\ell,\pi_z(j)}} = \frac{\partial q_i(\epsilon_{t,i})}{\partial \epsilon_{t,i}} \frac{\partial r_i(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})}{\partial z_{t-\ell,j}} \frac{\partial \xi_j^{-1}(\hat{z}_{t-\ell,\pi_z(j)})}{\partial \hat{z}_{t-\ell,\pi_z(j)}}. \tag{29}$$

Because q_i and ξ_j are univariate diffeomorphisms, the first and last factors on the right-hand side are nonzero, proving equation 28. □

We next use acyclicity of the instantaneous DAG to match the learned output rows ordered by π_ϵ uniquely to the current latent components ordered by π_z.

Lemma 11 (Matching transition-noise and latent components using the instantaneous DAG) *Suppose that the instantaneous causal structure is a DAG as in the problem setup and that every structural function f_i is continuously differentiable. Consider the bipartite graph whose left nodes are r_i, whose right nodes are $z_{t,j}$, and which contains an edge whenever $\partial r_i / \partial z_{t,j} \not\equiv 0$. A perfect matching pairs every r_i with exactly one $z_{t,j}$ and every $z_{t,j}$ with exactly one r_i. This graph has a unique perfect matching, consisting of the edges between r_i and $z_{t,i}$. Consequently, the learned Jacobian support uniquely determines the pairing between each output row $\hat{r}_{\pi_\epsilon(i)}$ and its corresponding current latent component $\hat{z}_{t,\pi_z(i)}$.*

Proof. We first show that the pairs $(r_i, z_{t,i})$ form a perfect matching and then prove that no other perfect matching exists. Assumption 3(i) ensures that the true inverse transition function r is continuously differentiable. Substituting $\epsilon_{t,i} = r_i(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$ into the structural equation gives

$$z_{t,i} = f_i\big(\mathrm{pa}_{\mathrm{inst}}(z_{t,i}), \mathrm{pa}_{\mathrm{lags}}(z_{t,i}), r_i(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})\big)\,.$$

Because $z_{t,i}$ is not its own instantaneous parent, differentiating this identity with respect to $z_{t,i}$ produces no direct parent term, leaving only the term through r_i. For $j \neq i$, differentiating the same identity with respect to $z_{t,j}$ gives zero on the left-hand side. Thus,

$$1 = \frac{\partial f_i}{\partial \epsilon_{t,i}} \frac{\partial r_i}{\partial z_{t,i}}, \qquad 0 = \frac{\partial f_i}{\partial z_{t,j}} + \frac{\partial f_i}{\partial \epsilon_{t,i}} \frac{\partial r_i}{\partial z_{t,j}}.$$

For each past component $z_{t-1,j}$, differentiating the same identity gives

$$0 = \frac{\partial f_i}{\partial z_{t-1,j}} + \frac{\partial f_i}{\partial \epsilon_{t,i}} \frac{\partial r_i}{\partial z_{t-1,j}}.$$

The first equality implies that both factors are nonzero throughout the domain. Therefore,

$$\frac{\partial r_i}{\partial z_{t,i}} = \left(\frac{\partial f_i}{\partial \epsilon_{t,i}} \right)^{-1} \neq 0, \qquad \frac{\partial r_i}{\partial z_{t,j}} = - \left(\frac{\partial f_i}{\partial \epsilon_{t,i}} \right)^{-1} \frac{\partial f_i}{\partial z_{t,j}}.$$

Similarly, for each past component,

$$\frac{\partial r_i}{\partial z_{t-1,j}} = - \left(\frac{\partial f_i}{\partial \epsilon_{t,i}} \right)^{-1} \frac{\partial f_i}{\partial z_{t-1,j}}.$$

Thus, the edges joining r_i to $z_{t,i}$ form a perfect matching, while the off-diagonal Jacobian support corresponds to the instantaneous causal edges. It remains to show that this perfect matching is unique. Suppose, for contradiction, that a different perfect matching exists. Relative to the matching $(r_i, z_{t,i})$, its associated permutation contains a nontrivial cycle $i_1, \ldots, i_s$. Writing $i_{s+1} = i_1$, this matching pairs each row r_{i_h} with the distinct current component $z_{t,i_{h+1}}$. Hence,

$$\frac{\partial r_{i_h}}{\partial z_{t,i_{h+1}}} \not\equiv 0 \qquad (h = 1, \ldots, s).$$

Because $i_h \neq i_{h+1}$, the relation derived above between the derivatives of r_i and f_i implies

$$\frac{\partial f_{i_h}}{\partial z_{t,i_{h+1}}} \not\equiv 0 \qquad (h = 1, \ldots, s).$$

These dependencies form the directed instantaneous cycle

$$z_{t,i_2} \to z_{t,i_1}, \quad z_{t,i_3} \to z_{t,i_2}, \quad \ldots, \quad z_{t,i_1} \to z_{t,i_s},$$

contradicting the instantaneous DAG assumption. Therefore, the matching between corresponding components is unique. By Lemma 10, the true and learned supports coincide up to π_ϵ and π_z; hence, the unique learned matching pairs $\hat{r}_{\pi_\epsilon(i)}$ with $\hat{z}_{t,\pi_z(i)}$. This argument uses only acyclicity of the instantaneous DAG and does not require a unique topological order. □

Together, the two lemmas match each output row of $\hat{r}$ to its corresponding current latent component. The Jacobian supports with respect to its current and past inputs can then be interpreted as direct causal edges.

Theorem 6 (Causal-graph identification from the learned inverse transition function) *Suppose that Theorems 5 and 4 hold. Under the instantaneous DAG assumed in the problem setup, also suppose that every structural function f_i is continuously differentiable. Define the direct causal edges induced by f_i by*

$$z_{t,j} \to z_{t,i} \iff \frac{\partial f_i}{\partial z_{t,j}} \not\equiv 0 \quad (j \neq i), \qquad z_{t-1,j} \to z_{t,i} \iff \frac{\partial f_i}{\partial z_{t-1,j}} \not\equiv 0.$$

Then, for all $i \neq j$,

$$\frac{\partial \hat{r}_{\pi_\epsilon(i)}}{\partial \hat{z}_{t,\pi_z(j)}} \not\equiv 0 \iff \frac{\partial f_i}{\partial z_{t,j}} \not\equiv 0,$$

and, for all i, j,

$$\frac{\partial \hat{r}_{\pi_\epsilon(i)}}{\partial \hat{z}_{t-1,\pi_z(j)}} \not\equiv 0 \iff \frac{\partial f_i}{\partial z_{t-1,j}} \not\equiv 0.$$

Consequently, the true instantaneous and one-step lagged causal edges are identifiable up to a single component permutation π_z from the off-diagonal instantaneous and one-step lagged Jacobian supports of $\hat{r}$, respectively.

Proof. For $\ell \in \{0, 1\}$, Lemma 10 gives

$$\frac{\partial \hat{r}_{\pi_\epsilon(i)}}{\partial \hat{z}_{t-\ell,\pi_z(j)}} \not\equiv 0 \quad \Longleftrightarrow \quad \frac{\partial r_i}{\partial z_{t-\ell,j}} \not\equiv 0.$$

The unique perfect matching in Lemma 11 assigns the learned output row $\hat{r}_{\pi_\epsilon(i)}$ uniquely to its child component $\hat{z}_{t,\pi_z(i)}$. For $\ell = 0$ with $j \neq i$, or for $\ell = 1$, a nonzero entry in this output row at the column $\hat{z}_{t-\ell,\pi_z(j)}$ therefore corresponds to the edge

$$\hat{z}_{t-\ell,\pi_z(j)} \to \hat{z}_{t,\pi_z(i)}.$$

This assignment does not assume $\pi_\epsilon = \pi_z$; the unique perfect matching associates each output row with its child component indexed according to π_z.

As derived in the proof of Lemma 11, differentiating the structural equation after substituting $\epsilon_{t,i} = r_i(\boldsymbol{z}_t, \boldsymbol{z}_{t-1})$ gives, for $\ell = 0$ with $j \neq i$, and for $\ell = 1$,

$$\frac{\partial r_i}{\partial z_{t-\ell,j}} = -\left(\frac{\partial f_i}{\partial \epsilon_{t,i}}\right)^{-1} \frac{\partial f_i}{\partial z_{t-\ell,j}}.$$

As shown in the proof of Lemma 11, $\partial f_i / \partial \epsilon_{t,i}$ is nonzero throughout the domain. Therefore, a nonzero off-diagonal instantaneous component or a nonzero one-step lagged component of the Jacobian of r_i is equivalent to the corresponding partial derivative of f_i being nonzero. Thus, under the single node permutation π_z, the off-diagonal instantaneous and one-step lagged Jacobian supports of $\hat{r}$ correspond exactly to the true instantaneous and one-step lagged causal edges, respectively. □

Remark 2 (Relation to Jacobian-based causal-graph identification) *The proof follows the idea of Reizinger et al. (2023): after identifying transition noise and latent states component-wise, the causal graph is identified from the Jacobian support of the inverse transition function. We apply this idea to latent transition dynamics, using the instantaneous Jacobian to match transition-noise components to current latent components and the instantaneous and lagged Jacobians to identify their respective causal edges.*

A.5 Extension to Multiple Lags

We extend the preceding identification results from $L = 1$ to any finite maximum lag L. The single past state is replaced by a finite history, while cross-time relations and variability are considered separately at each lag. Write the finite history over the past L time steps as $\boldsymbol{z}_t^- := (\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_{t-L})$. Let $\mathcal{M}_z$ denote the pairwise Markov network representing the instantaneous conditional-independence relations in $p(\boldsymbol{z}_t \mid \boldsymbol{z}_t^-, u_t)$, and let $\mathcal{E}_{\text{inst}}$ denote its edge set. For each $\ell = 1, \ldots, L$, let $\mathcal{E}_{\text{lag}}^{(\ell)}$ denote the past–current edges between $\boldsymbol{z}_{t-\ell}$ and $\boldsymbol{z}_t$ in the pairwise Markov network of the conditional joint distribution $p(\boldsymbol{z}_t^-, \boldsymbol{z}_t \mid u_t)$. Define the corresponding learned Markov networks analogously, with instantaneous edge set $\widehat{\mathcal{E}}_{\text{inst}}$ and lag-specific past–current edge sets $\widehat{\mathcal{E}}_{\text{lag}}^{(\ell)}$, and let

$$\Delta_n^{(L)}(\boldsymbol{z}_t, \boldsymbol{z}_t^-) := \log p(\boldsymbol{z}_t \mid \boldsymbol{z}_t^-, u_t^{(n)}) - \log p(\boldsymbol{z}_t \mid \boldsymbol{z}_t^-, u_t^{(0)}),$$
$$\boldsymbol{v}_{ia}^{(\ell)}(\boldsymbol{z}_t, \boldsymbol{z}_t^-) := \left(\frac{\partial^2 \Delta_n^{(L)}}{\partial z_{t,i} \partial z_{t-\ell,a}}\right)_{n=1}^{N_{\text{aux}}}.$$

Here, we retain the assumption on H in Condition (i) of Theorem 1 throughout this section and define $\widehat{\Delta}_n^{(L)}$ and $\hat{\boldsymbol{v}}_{ia}^{(\ell)}$ analogously for the learned representation. Define the candidate sets $\mathcal{C}_s$ as in Appendix A.2, using the instantaneous Markov network $\mathcal{M}_z$ defined above, and let

$$\mathcal{B}^{(\ell)}(j, a) := \left\{(k, m) \in [d_z]^2 \;\middle|\; k \in \mathcal{C}_t(j),\ m \in \mathcal{C}_{t-\ell}(a),\ \{z_{t-\ell,m}, z_{t,k}\} \in \mathcal{E}_{\text{lag}}^{(\ell)}\right\}.$$

The corresponding future-directed candidate blocks are defined by shifting these quantities forward by ℓ time steps.

Assumption 4 (Finite-lag extension conditions) *For every time t, the following conditions hold whenever the required past and future variables exist.*

(i) Finite-history transition-noise identification. *All conditions in Appendix A.3 hold after replacing the past observation $\boldsymbol{x}_{t-1}$ by the finite history $\boldsymbol{x}_t^- := (\boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-L})$ in the augmented-coordinate construction.*

(ii) Instantaneous variability and edge-count minimality. *The positivity and regularity condition in Assumption 1(i) holds with $\boldsymbol{z}_t^-$ in place of $\boldsymbol{z}_{t-1}$. Let $N_{\mathrm{inst}} := 2d_z + |\mathcal{E}_{\mathrm{inst}}|$ and define $\boldsymbol{w}_n^{(L)}(\boldsymbol{z}_t, \boldsymbol{z}_t^-) \in \mathbb{R}^{N_{\mathrm{inst}}}$ from $\Delta_n^{(L)}$ in the same way as $\boldsymbol{w}_n$ is defined in Assumption 1(ii), replacing $\boldsymbol{z}_{t-1}$ with $\boldsymbol{z}_t^-$. For every $\boldsymbol{z}_t$, suppose that there exist auxiliary-variable indices and finite-history evaluation values*

$$(n_\alpha, \bar{\boldsymbol{z}}_t^{-,(\alpha)}) \in \{1, \ldots, N_{\mathrm{aux}}\} \times \mathbb{R}^{Ld_z}, \qquad \alpha = 1, \ldots, N_{\mathrm{inst}},$$

such that

$$\boldsymbol{W}^{(L)}(\boldsymbol{z}_t) := \left[\boldsymbol{w}_{n_\alpha}^{(L)}(\boldsymbol{z}_t, \bar{\boldsymbol{z}}_t^{-,(\alpha)})\right]_{\alpha=1}^{N_{\mathrm{inst}}}$$

has rank N_{inst}. The indices n_α need not be distinct. The learned instantaneous edge set also satisfies

$$|\widehat{\mathcal{E}}_{\mathrm{inst}}| \le |\mathcal{E}_{\mathrm{inst}}|.$$

(iii) Per-lag cross-time variability and edge-count minimality. *For every $\ell = 1, \ldots, L$ and $j, a \in [d_z]$, at almost every $(\boldsymbol{z}_t, \boldsymbol{z}_t^-)$,*

$$\operatorname{rank}\left[\boldsymbol{v}_{km}^{(\ell)}(\boldsymbol{z}_t, \boldsymbol{z}_t^-)\right]_{(k,m)\in\mathcal{B}^{(\ell)}(j,a)} = |\mathcal{B}^{(\ell)}(j,a)|.$$

The learned cross-time edge sets satisfy

$$\sum_{\ell=1}^{L} |\widehat{\mathcal{E}}_{\mathrm{lag}}^{(\ell)}| \le \sum_{\ell=1}^{L} |\mathcal{E}_{\mathrm{lag}}^{(\ell)}|.$$

(iv) Intimate-neighbor separation across lags. *For every $z_{t,j} \in \Psi_{\mathcal{M}_z}(z_{t,i})$, there exists at least one lag $\ell \in \{1, \ldots, L\}$ and a past or future component at that lag whose cross-time relation is present with $z_{t,i}$ but absent with $z_{t,j}$, as in Assumption 2(iii).*

This yields the following extension of the $L = 1$ result.

Theorem 7 (Identification for a general finite maximum lag) *Suppose that Assumption 4 holds. Then there exist a time-independent permutation π_z and univariate C^2 diffeomorphisms ξ_i such that $\hat{z}_{t,\pi_z(i)} = \xi_i(z_{t,i})$. Furthermore, under the instantaneous DAG assumed in the problem setup, suppose that every structural function f_i is continuously differentiable. Let π_ϵ be the transition-noise permutation obtained by IIA. Then, for all $i \neq j$,*

$$\frac{\partial \hat{r}_{\pi_\epsilon(i)}}{\partial \hat{z}_{t,\pi_z(j)}} \not\equiv 0 \quad \Longleftrightarrow \quad \frac{\partial f_i}{\partial z_{t,j}} \not\equiv 0,$$

and, for every $\ell = 1, \ldots, L$ and all i, j,

$$\frac{\partial \hat{r}_{\pi_\epsilon(i)}}{\partial \hat{z}_{t-\ell,\pi_z(j)}} \not\equiv 0 \quad \Longleftrightarrow \quad \frac{\partial f_i}{\partial z_{t-\ell,j}} \not\equiv 0. \tag{30}$$

Consequently, the true instantaneous and lagged causal edges are identifiable up to a single component permutation π_z from the off-diagonal instantaneous and per-lag Jacobian supports of $\hat{r}$.

Proof. In Appendix A.3, replacing $\boldsymbol{x}_{t-1}$ with $\boldsymbol{x}_t^-$ preserves the block-triangular Jacobian of the augmented coordinate transformation, so the same mixed-derivative argument applies to every component of the finite history. In Appendix A.1, replacing past-state evaluation values with finite-history evaluation values does not change the fact that the unknown coefficients constructed from the coordinate transformation H depend only on $\boldsymbol{z}_t$. The full-rank condition on $\boldsymbol{W}^{(L)}(\boldsymbol{z}_t)$ and instantaneous edge-count minimality therefore imply an isomorphism of $\mathcal{M}_z$.

For each lag ℓ, taking the mixed derivative of $\widehat{\Delta}_n^{(L)}$ with respect to $\hat{\boldsymbol{z}}_t$ and $\hat{\boldsymbol{z}}_{t-\ell}$ introduces no terms from other lags because H acts separately at each time step, and gives

$$\hat{\boldsymbol{v}}^{(\ell)}_{\pi_{\hat{\boldsymbol{z}}_t}(j),\pi_{\hat{\boldsymbol{z}}_{t-\ell}}(a)} = \sum_{(k,m)\in\mathcal{B}^{(\ell)}(j,a)} \frac{\partial z_{t,k}}{\partial \hat{z}_{t,\pi_{\hat{\boldsymbol{z}}_t}(j)}} \frac{\partial z_{t-\ell,m}}{\partial \hat{z}_{t-\ell,\pi_{\hat{\boldsymbol{z}}_{t-\ell}}(a)}} \boldsymbol{v}^{(\ell)}_{km}.$$

The local full-rank condition preserves every true edge at each lag, so $|\mathcal{E}^{(\ell)}_{\text{lag}}| \leq |\widehat{\mathcal{E}}^{(\ell)}_{\text{lag}}|$. Thus, each difference $|\widehat{\mathcal{E}}^{(\ell)}_{\text{lag}}| - |\mathcal{E}^{(\ell)}_{\text{lag}}|$ is nonnegative, while the aggregate edge-count minimality requires their sum to be nonpositive. Every difference must therefore be zero, yielding edge-set correspondence at each lag. The argument of Appendix A.2 then eliminates the remaining mixing and yields π_z and ξ_i.

Finally, the chain-rule support correspondence in Appendix A.4 extends directly to every input lag, while the unique perfect matching between output rows and current latent components is determined from the instantaneous Jacobian as before. Thus, at every lag, the Jacobian supports of the true and learned inverse transition functions correspond under the same row permutation π_ϵ and column permutation π_z, proving the stated graph-identification result. □

B Learning Framework Details

This section provides details of the learning framework described in the main text, including its model parameterization, contrastive objectives, structural regularization, and optimization.

B.1 Model Parameterization

Shared encoder and learned inverse transition function. For maximum lag L, our method applies the shared encoder ϕ separately to each observation in $(\boldsymbol{x}_{t-L}, \ldots, \boldsymbol{x}_t)$, yielding $\hat{\boldsymbol{z}}_{t-\ell} = \phi(\boldsymbol{x}_{t-\ell})$ for $\ell = 0, \ldots, L$. Each transition-noise component is then estimated by a component-specific learned inverse transition function. Each component of $\hat{r} = (\hat{r}_1, \ldots, \hat{r}_{d_z})$ is parameterized as

$$\hat{\epsilon}_{t,i} = \hat{r}_i\left(\hat{z}_{t,i}, \boldsymbol{G}^{(0)}_{i,-i} \odot \hat{\boldsymbol{z}}_{t,-i}, \left\{\boldsymbol{G}^{(\ell)}_{i,:} \odot \hat{\boldsymbol{z}}_{t-\ell}\right\}_{\ell=1}^{L}\right). \tag{31}$$

Here, $\boldsymbol{G}^{(0)}, \boldsymbol{G}^{(\ell)} \in \mathbb{R}^{d_z \times d_z}$, $\ell = 1, \ldots, L$, are learnable weight matrices that gate individual current and lagged input components, respectively. For these matrices, row i corresponds to the target component $\hat{z}_{t,i}$ and column j to the source component $\hat{z}_{t-\ell,j}$. The notation $-i$ denotes all components except component i, whereas : denotes all components. Accordingly, $\boldsymbol{G}^{(0)}_{i,-i}$ gates the other current components and $\boldsymbol{G}^{(\ell)}_{i,:}$ gates all components at lag ℓ. Because the current component $\hat{z}_{t,i}$ is supplied directly to $\hat{r}_i$ without a gate, the diagonal entries of $\boldsymbol{G}^{(0)}$ are fixed to zero. For each i, the gated inputs are concatenated and processed jointly by a general nonlinear network parameterizing $\hat{r}_i$.

History feature. For the contrastive objectives below, the past learned representations are summarized by the history feature

$$\boldsymbol{h}_t = \eta(\hat{\boldsymbol{z}}_{t-1}, \ldots, \hat{\boldsymbol{z}}_{t-L}), \tag{32}$$

where the history feature map η is learned jointly with the other modules.

B.2 Contrastive Learning with Auxiliary Variables

We implement the IIA contrastive-learning framework (Morioka et al., 2021) in two forms. iCReN-T treats a discrete auxiliary variable as a finite set of classes, whereas iCReN-G accommodates a general observed auxiliary variable. We describe the corresponding regression functions and training losses below.

iCReN-T. For a discrete auxiliary variable, iCReN-T treats each possible value as a class and predicts it from the estimated transition noise and history feature. For a candidate value u, the logit

is

$$T_u^{\mathrm{TCL}}(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t) = \sum_{i=1}^{d_z} \sum_{k=1}^{K} w_{uik} \psi_{ik}(\hat{\epsilon}_{t,i}) + \varphi(\boldsymbol{h}_t, u) + b_u. \tag{33}$$

Here, K is the number of feature functions for each transition-noise component, and ψ_{ik} is the k-th scalar-valued feature function of $\hat{\epsilon}_{t,i}$. The feature functions may be learned or fixed; when the conditional noise family is specified, fixed functions can be chosen to include its sufficient statistics. The coefficients w_{uik} and biases b_u are learned separately for each class. The learned function $\varphi(\boldsymbol{h}_t, u)$ represents the relation between the history feature and the auxiliary variable. We omit an auxiliary-variable-independent common term because it cancels from the softmax probabilities. The softmax cross-entropy for the observed u_t defines $\mathcal{L}_{\mathrm{TCL}}$.

iCReN-G. iCReN-G uses changes in the transition-noise distributions associated with a general observed auxiliary variable through binary classification of matched and mismatched tuples. Its binary logistic regression function is

$$T^{\mathrm{GCL}}(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t, u_t) = \sum_{i=1}^{d_z} \sum_{k=1}^{K} \psi_{ik}(\hat{\epsilon}_{t,i}) \mu_{ik}(u_t) + \varphi(\boldsymbol{h}_t, u_t) + B(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t). \tag{34}$$

Here, μ_{ik} is a learned function of the auxiliary variable that gives the coefficient of the k-th feature of component i. The learned function φ represents the relation between the history feature and the auxiliary variable as well as terms that depend only on the auxiliary variable, whereas B collects terms independent of the auxiliary variable. The component-wise additive restriction applies only to the auxiliary-variable-dependent term; B may therefore depend jointly on all estimated transition-noise components and the history feature. For each minibatch, matched tuples $(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t, u_t)$ are positive examples. Mismatched tuples $(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t, u'_t)$ are constructed by permuting only the auxiliary-variable values within the minibatch, while keeping $\hat{\boldsymbol{\epsilon}}_t$ and $\boldsymbol{h}_t$ fixed. We use a nonzero cyclic shift for this permutation so that each sample receives the auxiliary-variable value from a different sample. The resulting binary cross-entropy defines $\mathcal{L}_{\mathrm{GCL}}$.

B.3 Structural Regularization and Training Objective

Sparsity regularization of input dependencies. Motivated by the edge-count minimality used in the identification analysis in Appendix A.2, we impose the following regularizers on the input gates in equation 31:

$$\mathcal{R}_{\mathrm{inst}} = \frac{1}{d_z^2} \sum_{i,j} \left| \boldsymbol{G}_{ij}^{(0)} \right|, \qquad \mathcal{R}_{\mathrm{lag}} = \frac{1}{L d_z^2} \sum_{\ell=1}^{L} \sum_{i,j} \left| \boldsymbol{G}_{ij}^{(\ell)} \right|. \tag{35}$$

For any gated current or lagged input $\hat{z}_{t-\ell,j}$ supplied to $\hat{r}_i$ through $\boldsymbol{G}_{ij}^{(\ell)}$, the chain rule gives

$$\frac{\partial \hat{r}_i}{\partial \hat{z}_{t-\ell,j}} = \boldsymbol{G}_{ij}^{(\ell)} \frac{\partial \hat{r}_i}{\partial \left(\boldsymbol{G}_{ij}^{(\ell)} \hat{z}_{t-\ell,j} \right)}. \tag{36}$$

Consequently, $\boldsymbol{G}_{ij}^{(\ell)} = 0$ makes the corresponding Jacobian component zero. Regularizing the gates therefore provides a computationally simple surrogate for encouraging sparse input dependencies, since it acts directly on the weight matrices without explicitly evaluating the Jacobian. IDOL instead applies an ℓ_1 penalty directly to the Jacobian (Li et al., 2025b), which can also be used here but requires explicit Jacobian evaluation during training. We combine the instantaneous and lagged penalties as

$$\mathcal{R}_{\mathrm{sparsity}} := \mathcal{R}_{\mathrm{inst}} + \mathcal{R}_{\mathrm{lag}}. \tag{37}$$

Instantaneous DAG regularization. When the instantaneous causal order is not fixed, we encourage acyclicity using the following NOTEARS-style regularizer (Zheng et al., 2018):

$$\mathcal{R}_{\mathrm{DAG}} := \mathrm{tr}\left[\exp\left(\boldsymbol{G}^{(0)} \odot \boldsymbol{G}^{(0)} \right) \right] - d_z. \tag{38}$$

Because this term is optimized as an ordinary loss penalty in our implementation, equation 38 is a soft penalty and does not strictly guarantee acyclicity. When a causal order is fixed in advance, a hard mask restricts the learnable entries to the triangular part consistent with that order, structurally guaranteeing an instantaneous DAG; in this case, $\mathcal{R}_{\mathrm{DAG}}$ can be omitted.

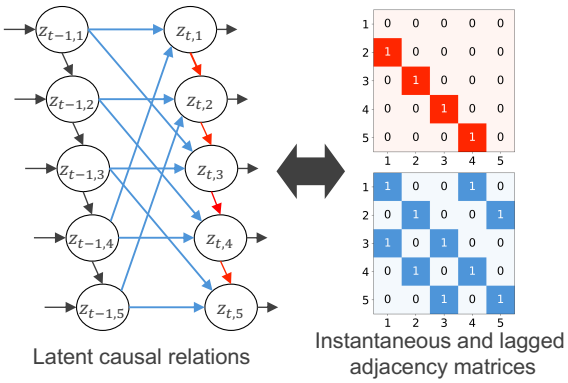


Figure 4: Instantaneous and lagged causal graphs in the default five-component setting and their adjacency matrices. Red and blue indicate instantaneous and lagged causal edges, respectively; rows and columns index targets and sources.

Joint optimization. The regularizers are combined with the selected TCL or GCL loss as in equation 13, and all learnable parameters are optimized jointly. A decoder and reconstruction loss may also be added, although neither is required by the learning framework or identification theory.

C Synthetic-data Experimental Details and Additional Results

The main text presents the primary comparison between iCReN and the baselines in the default setting. This section details data generation, training conditions, baselines, and evaluation, followed by analyses of latent causal structure, latent dimension, sample size, data-generating function complexity, and objective components. General definitions of the model, TCL/GCL losses, input-gate sparsity regularization, and DAG regularization are given in Appendix B.

C.1 Common Experimental Settings

C.1.1 Data Generation

Data-generating model. In the default setting, we generate sequences using nonlinear structural functions and a nonlinear observation function that remain invariant across auxiliary-variable values, while only the transition-noise distribution changes. We set $d_z = d_x = 5$ and $L = 1$ and generate data as

$$z_{t,i} = f_i\big(\mathrm{pa}_{\mathrm{inst}}(z_{t,i}), \mathrm{pa}_{\mathrm{lags}}(z_{t,i}), \epsilon_{t,i}\big)\,, \qquad \boldsymbol{x}_t = g(\boldsymbol{z}_t). \tag{39}$$

We use the additive-noise special case $f_i(a, b, e) = \widetilde{f}_i(a, b) + e$. In the default setting and dimension sweep, the instantaneous graph is the chain DAG $z_{t,i} \to z_{t,i+1}$ for $i = 1, \ldots, d_z - 1$. The lagged edges are $z_{t-1,i} \to z_{t,i}$ and $z_{t-1,i} \to z_{t,1+((i+1) \bmod d_z)}$ for $i = 1, \ldots, d_z$. Figure 4 illustrates these graphs for the default five-component setting, together with their adjacency matrices.

Function parameterization and data splits. Each $\widetilde{f}_i$ is a three-layer multilayer perceptron (MLP) that uses Leaky-ReLU activations and jointly takes all instantaneous and lagged parents specified by the true causal graph as input. The nonlinear observation function g consists of three successive layers, each applying a Leaky-ReLU activation followed by an orthogonal linear transformation. For each seed, the training, validation, and test splits use independent transition-noise samples but share the true graph, structural and observation functions, and transition-noise parameters.

Discrete variation in transition-noise distributions. For discrete-variation data, the transition-noise distribution switches across discrete environments. Each split contains 256 independent sequences of length 256. For every environment $u \in \{1, \ldots, 256\}$ and component i, we sample once and fix

$$\mu_{u,i} \sim \mathrm{Unif}[-0.7, 0.7], \qquad \sigma_{u,i} \sim \mathrm{Unif}[0.4, 1.1], \tag{40}$$

and define

$$p(\boldsymbol{\epsilon}_t \mid u_t = u) = \prod_{i=1}^{d_z} \mathcal{N}\big(\epsilon_{t,i}; \mu_{u,i}, \sigma_{u,i}^2\big)\,. \tag{41}$$

Transition noise is sampled independently from this distribution at each time. Thus, its distribution is constant within a sequence, while its component-wise means and standard deviations switch across sequences. Each split contains $256 \times 256 = 2^{16}$ samples.

Continuous variation in transition-noise distributions. For continuous-variation data, the transition-noise distribution varies smoothly with continuous time. Each split is one sequence of length $T = 2^{18}$, with $t = 0, \ldots, T-1$ used as a continuous auxiliary variable. For each component i, we construct two smooth time series from separate random linear combinations of the sine and cosine functions $\{\sin(2\pi kt/T), \cos(2\pi kt/T)\}_{k=1}^{64}$. We standardize each time series, apply the standard Gaussian cumulative distribution function, and linearly rescale the results to $[-0.7, 0.7]$ and $[0.4, 1.1]$ to obtain $\mu_i(t)$ and $\sigma_i(t)$, respectively. We then set

$$p(\boldsymbol{\epsilon}_t \mid t) = \prod_{i=1}^{d_z} \mathcal{N}\big(\epsilon_{t,i}; \mu_i(t), \sigma_i(t)^2\big)\,. \tag{42}$$

For methods requiring discrete labels, we divide the sequence into 256 equal-width bins and use the pseudo-label $\widetilde{u}_t = \lfloor 256t/T \rfloor \in \{0, \ldots, 255\}$, which we distinguish from the continuous auxiliary variable t. In both data types, the conditional transition-noise distribution is a component-wise Gaussian exponential family with sufficient statistics $(\epsilon_{t,i}, \epsilon_{t,i}^2)$ and auxiliary-variable-dependent natural parameters, matching the form assumed in the IIA analysis in Appendix A.3.

Windows used for training and evaluation. We remove windows containing extreme observations using a threshold determined solely from the training split. Specifically, we set

$$\tau = Q_{0.999}\Big(\big\{|x_{t,j}^{\text{train}}|\big\}_{t,j}\Big)\,, \tag{43}$$

and retain windows satisfying $\max_{\ell=0,\ldots,L} \max_j |x_{t-\ell,j}| \le \tau$. For window-based training and validation, we apply this threshold without using the true latent states. Because LEAP is trained on contiguous sequences, we do not remove individual windows from its training or validation sequences. At test time, however, we evaluate all methods only on windows satisfying the threshold.

C.1.2 Training Conditions and Baselines

Model architecture. We use the model and objective defined in Appendix B with the dimensions and auxiliary variables specified for these experiments. The encoder and learned inverse transition function are five-layer Leaky-ReLU MLPs, and the history feature map is a three-layer Leaky-ReLU MLP; all hidden widths are $4d_z$. The main comparison and all additional analyses except the reconstruction-loss ablation use no decoder or reconstruction loss. iCReN-T classifies 256 discrete environments using the Gaussian sufficient statistics $(\hat{\epsilon}_{t,i}, \hat{\epsilon}_{t,i}^2)$ for each estimated transition-noise component, together with the history-feature vector $\boldsymbol{h}_t$ and its element-wise square. iCReN-G applies the same element-wise transformations to the estimated transition noise and history feature. For discrete-variation data, each coefficient $\mu_{ik}(u)$, the coefficients multiplying the transformed history features in φ, and the offset in φ are represented by separate parameters for each of the 256 environments. For continuous-variation data, these coefficients and the offset are linear functions of a Fourier basis consisting of 64 pairs of sine and cosine functions. Its auxiliary-variable-independent term $B(\hat{\boldsymbol{\epsilon}}_t, \boldsymbol{h}_t)$ in equation 34 is implemented as the sum of two MLP outputs, one for $\hat{\boldsymbol{\epsilon}}_t$ and one for $\boldsymbol{h}_t$, each with hidden width 32. Although the theory assumes C^2 mappings, the experimental implementation uses Leaky-ReLU activations, which are only piecewise smooth. This mismatch at the activation level can be removed by replacing Leaky-ReLU with a smooth, strictly increasing activation applied element-wise.

Optimization. We clip the global gradient norm at 5 and set $\lambda_{\text{sparsity}} = 0.005$ and $\lambda_{\text{DAG}} = 10$. Because the implementation averages the instantaneous and lagged mean ℓ_1 quantities, this corresponds to an implementation-level sparsity coefficient of 0.01. We use AdamW with weight decay 10^{-5}, initial learning rate 10^{-2}, and minibatches of 256 windows, reducing the learning rate to

Table 4: Main optimization settings for the baselines in the synthetic experiments.

Method	Optimization
IDOL	Adam, initial learning rate 10^{-3}; multiplied by 0.1 at epoch 120
G-CaRL	SGD, momentum 0.9, initial learning rate 10^{-2}; multiplied by 0.1 after 80% of updates
CtrlNS	AdamW, initial learning rate 10^{-3}
iVAE / β-VAE	AdamW, initial learning rate 10^{-2}; multiplied by 0.1 at epoch 120
LEAP	AdamW (model/transition) and SGD (discriminator), initial learning rate 2×10^{-3}

Table 5: Auxiliary variables supplied to each method in the synthetic experiments.

Method	Discrete variation	Continuous variation
iCReN-T	Discrete environment u_t	256-bin pseudo-label $\widetilde{u}_t$
iCReN-G	Discrete environment u_t	Continuous time t
IDOL	None	None
G-CaRL	None	None
CtrlNS	None	None
LEAP	Discrete environment u_t	256-bin pseudo-label $\widetilde{u}_t$
iVAE	Discrete environment u_t	Continuous time t
β-VAE	None	None

0.001 from epoch 120. We selected these coefficients by inspecting validation-loss trajectories and avoiding values for which overly strong regularization produced degenerate solutions. The ablation in Table 10 shows that removing either or both penalties does not consistently reduce latent-state or causal-graph recovery in the evaluated synthetic settings.

Training and checkpoint selection. All methods are trained independently with the same 10 seeds, and checkpoints are selected using validation objectives that do not use the true latent states or causal graph. Methods with the same seed share the generated data and splits; data generation, model initialization, training, and evaluation are independent across seeds. Each method is trained for at most 150 epochs, and we select the checkpoint minimizing its corresponding validation objective. Neither MCC nor F1 is used for checkpoint selection.

Baselines and their inputs. We follow public implementations where possible, adapting network sizes to the present setting and making limited adjustments to optimization settings, including learning rates, based on validation behavior. We evaluate IDOL (Li et al., 2025b), G-CaRL (Morioka & Hyvärinen, 2024), CtrlNS (Song et al., 2024), LEAP (Yao et al., 2022b), iVAE (Khemakhem et al., 2020), and β-VAE (Higgins et al., 2017), with $\beta = 2$ for β-VAE. Tables 4 and 5 summarize the final optimization settings and auxiliary variables, respectively.

Method-specific sequence handling. For G-CaRL, CtrlNS, and LEAP, we adapt the inputs to the sequence structure of the default data. G-CaRL receives $(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})$ as current- and past-time groups through a shared encoder, and we report past-to-current lagged F1. CtrlNS receives no auxiliary variable; its gating network estimates at most 256 regimes from consecutive learned representations. The 256 time bins are used only for chunk sampling and are not provided as labels. For LEAP, we compute the total-correlation loss from the residual components of each transition separately, rather than jointly across an entire residual sequence. For continuous-variation data, we divide the single sequence into the 256 pseudo-label bins and present each bin to LEAP as a separate sequence; consequently, transitions across bin boundaries are excluded.

C.1.3 Evaluation Protocol

Latent-representation and causal-graph evaluation. We choose the permutation of the learned latent components that maximizes the mean absolute correlation with the true components and use

Table 6: Latent-state and causal-graph recovery by iCReN-T under different instantaneous and lagged causal structures on synthetic data with Gaussian transition-noise distributions that vary across discrete environments. Random DAGs are distinct across seeds. Values are the mean $\pm$ standard deviation over 10 seeds.

Instant. graph	Lagged graph	MCC	Inst. F1	Lagged F1
Default chain	Default	0.951 ± 0.011	0.950 ± 0.158	0.980 ± 0.063
Random DAGs	Default	0.960 ± 0.008	1.000 ± 0.000	0.970 ± 0.067
Default chain	Self-loops	0.955 ± 0.009	0.925 ± 0.169	0.940 ± 0.135
Random DAGs	Self-loops	0.948 ± 0.032	0.975 ± 0.079	0.960 ± 0.126

the same permutation to order the learned edge scores before computing F1. The resulting MCC is

$$\mathrm{MCC} = \max_{\pi \in \mathfrak{S}_{d_z}} \frac{1}{d_z} \sum_{i=1}^{d_z} \left| \mathrm{corr}\big(\boldsymbol{z}_{:,i}, \hat{\boldsymbol{z}}_{:,\pi(i)}\big) \right|, \tag{44}$$

where $\mathfrak{S}_{d_z}$ is the set of permutations of the d_z components, corr denotes the sample Pearson correlation, and $\boldsymbol{z}_{:,i}$ and $\hat{\boldsymbol{z}}_{:,i}$ collect the values of component i over the evaluated test samples. Let $\widehat{\mathbb{E}}_{\mathrm{test}}$ denote the empirical mean over the evaluated test windows. For iCReN, we define the empirical edge scores as

$$\begin{aligned} \boldsymbol{S}^{(0)}_{ij} &= \widehat{\mathbb{E}}_{\mathrm{test}} \left[\left| \frac{\partial \hat{r}_i}{\partial \hat{z}_{t,j}} \right| \right], \qquad i \neq j, \\ \boldsymbol{S}^{(\ell)}_{ij} &= \widehat{\mathbb{E}}_{\mathrm{test}} \left[\left| \frac{\partial \hat{r}_i}{\partial \hat{z}_{t-\ell,j}} \right| \right], \quad \ell = 1, \ldots, L, \end{aligned} \tag{45}$$

and use the edge scores produced by each baseline. Here, t denotes the target time of each evaluated window. Diagonal entries of $\boldsymbol{S}^{(0)}$ are excluded from instantaneous edges, and rows and columns index targets and sources, respectively. Using the row-to-target convention of the learned inverse transition function, we apply the MCC-based latent-component permutation to the target and source axes of the empirical graph scores. We then select for each target the same number of highest-scoring sources as its true number of parents. This evaluation depends on the ranking of the edge scores rather than their absolute scale, which is not invariant to component-wise invertible transformations. From this binary support and the true support, we compute

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{46}$$

Here, TP, FP, and FN denote the numbers of true-positive, false-positive, and false-negative edges, respectively. We report instantaneous F1 and lagged F1, abbreviating the former as Inst. F1 in tables and figures. True latent states and true parent counts are not used for training or checkpoint selection. For G-CaRL, we separately align the current- and past-time groups and apply the two permutations to the target and source axes of its lagged-edge scores. Tables and figures report the mean $\pm$ standard deviation over 10 seeds; $--$ marks metrics not applicable to the method. Figure 5 illustrates two seeds that recover the same support despite having different score scales.

C.2 Sensitivity to Latent Causal Structure

To examine sensitivity to the latent causal structure, we evaluate all four combinations of two instantaneous and two lagged graphs. For the instantaneous graph, we compare the default chain with ten distinct randomly generated DAGs, one for each seed. Each random DAG has five nodes and four edges, respects the fixed component order, is connected but not chain-shaped when edge directions are ignored, and contains at least one fork and one collider. For the lagged graph, we compare the default pattern, in which each component has two lagged parents, with a graph containing only the lagged self-loop $z_{t-1,i} \to z_{t,i}$ for each component i. Because this replacement reduces the number of lagged parents per component from two to one, we double the weight assigned to the retained self-loop input in each true structural function. Table 6 shows that iCReN-T attains high MCC and F1 scores across all four combinations.

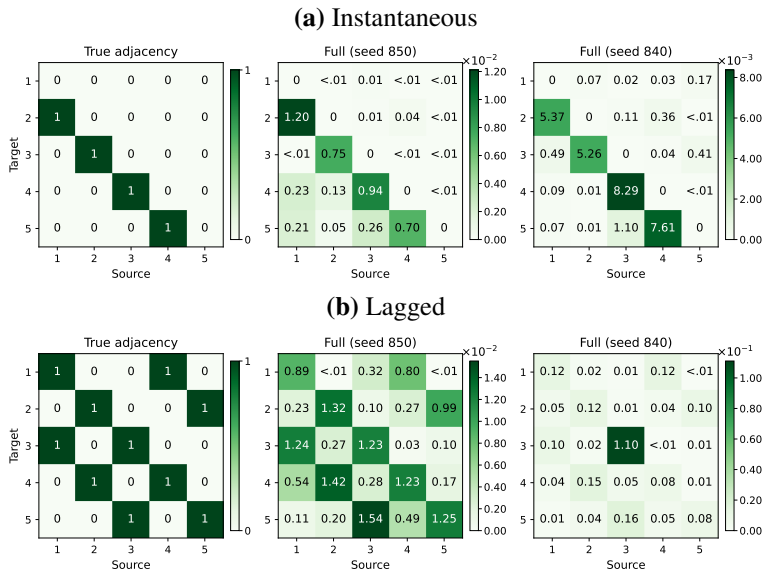


Figure 5: Jacobian edge scores in the default setting with Gaussian transition-noise distributions that vary across discrete environments: (a) instantaneous and (b) lagged. In each panel, the left matrix is the true support, and the middle and right matrices show seeds 850 and 840 trained with the full objective. Both seeds recover the same support, although their score scales differ under component-wise invertible transformations. Rows and columns index targets and sources, respectively.

Table 7: Latent-state and causal-graph recovery by iCReN-T as the latent dimension varies on synthetic data with Gaussian transition-noise distributions that vary across discrete environments. Values are the mean $\pm$ standard deviation over 10 seeds.

d_z	MCC	Inst. F1	Lagged F1
4	0.958 ± 0.009	1.000 ± 0.000	1.000 ± 0.000
5	0.951 ± 0.011	0.950 ± 0.158	0.980 ± 0.063
8	0.860 ± 0.021	0.743 ± 0.284	0.869 ± 0.127
12	0.886 ± 0.017	0.527 ± 0.200	0.812 ± 0.118
16	0.792 ± 0.056	0.140 ± 0.058	0.569 ± 0.136

C.3 Sensitivity to Latent Dimension

We vary the latent dimension of discrete-variation data over $d_x = d_z \in \{4, 5, 8, 12, 16\}$ while fixing the total sample size at 2^{16}. For $d_z = 4$, the lagged graph contains only self-loops; otherwise, we use the default instantaneous and lagged graphs. Table 7 shows that MCC and F1 scores generally decrease as d_z increases, with the most pronounced decrease in instantaneous F1.

C.4 Sensitivity to Total Sample Size

We vary the total sample size from 2^{13} to 2^{18} while retaining the respective discrete-variation and continuous-variation generation schemes. For discrete-variation data, we fix 256 environments and vary the sequence length per environment, whereas for continuous-variation data we vary the length of the single continuous sequence. Figure 6 shows that MCC and F1 generally improve with sample size for both variants under both data-generation schemes. For both data-generation schemes, iCReN-G is more sensitive to sample size than iCReN-T, particularly in instantaneous F1. At $N = 2^{16}$, its instantaneous F1 is 0.700 ± 0.158 for discrete variation and 0.700 ± 0.197 for continuous variation. Increasing N to 2^{18} raises both values to 0.975 ± 0.079, indicating that the lower discrete-variation result at the default sample size is partly a finite-sample effect.

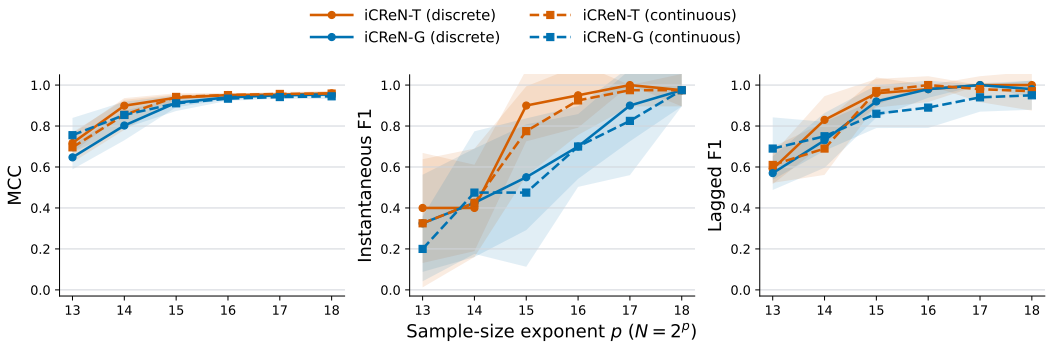


Figure 6: Latent-state and causal-graph recovery as total sample size varies on synthetic data with Gaussian transition-noise distributions. Orange and blue denote iCReN-T and iCReN-G, respectively; solid curves with circular markers and dashed curves with square markers denote discrete variation across environments and continuous variation over time, respectively. Shaded regions show mean $\pm$ standard deviation over 10 seeds.

Table 8: Latent-state and causal-graph recovery by iCReN-T as the complexity of the true data-generating functions varies on synthetic data with Gaussian transition-noise distributions that vary across discrete environments. We vary structural-function depth with a fixed three-layer observation function in (a), and observation-function depth with fixed three-layer structural functions in (b). Values are the mean $\pm$ standard deviation over 10 seeds.

(a) Structural-function depth

Depth	MCC	Inst. F1	Lagged F1
1	0.913 ± 0.012	0.275 ± 0.275	1.000 ± 0.000
2	0.953 ± 0.007	0.975 ± 0.079	1.000 ± 0.000
3	0.951 ± 0.011	0.950 ± 0.158	0.980 ± 0.063
4	0.952 ± 0.007	0.925 ± 0.237	0.950 ± 0.127
5	0.955 ± 0.006	1.000 ± 0.000	1.000 ± 0.000

(b) Observation-function depth

Depth	MCC	Inst. F1	Lagged F1
1	0.957 ± 0.008	1.000 ± 0.000	1.000 ± 0.000
2	0.953 ± 0.007	0.975 ± 0.079	1.000 ± 0.000
3	0.951 ± 0.011	0.950 ± 0.158	0.980 ± 0.063
4	0.946 ± 0.011	0.950 ± 0.105	1.000 ± 0.000
5	0.930 ± 0.008	0.875 ± 0.177	0.980 ± 0.063

Table 9: MCC of both proposed methods and the baselines with one- and three-layer observation functions on discrete-variation synthetic data. Each entry is the mean ± standard deviation over 10 seeds.

Method	One layer	Three layers
iCReN-T	0.957 ± 0.008	0.951 ± 0.011
iCReN-G	0.947 ± 0.010	0.940 ± 0.010
IDOL	0.704 ± 0.061	0.468 ± 0.046
G-CaRL	0.775 ± 0.060	0.764 ± 0.065
CtrlNS	0.503 ± 0.050	0.432 ± 0.071
LEAP	0.536 ± 0.063	0.373 ± 0.107
iVAE	0.264 ± 0.037	0.230 ± 0.106
β-VAE	0.582 ± 0.058	0.482 ± 0.067

Table 10: Ablation of structural regularization on synthetic data with Gaussian transition-noise distributions that vary across discrete environments, with $d_z = 5$ (left) and $d_z = 8$ (right). Values are the mean ± standard deviation over 10 seeds.

	$d_z = 5$			$d_z = 8$		
Objective	MCC	Inst. F1	Lagged F1	MCC	Inst. F1	Lagged F1
Full	0.951 ± 0.011	0.950 ± 0.158	0.980 ± 0.063	0.860 ± 0.021	0.743 ± 0.284	0.869 ± 0.127
No sparsity	0.955 ± 0.007	1.000 ± 0.000	1.000 ± 0.000	0.867 ± 0.016	0.900 ± 0.191	0.950 ± 0.097
No DAG	0.957 ± 0.004	0.950 ± 0.105	0.940 ± 0.135	0.861 ± 0.027	0.786 ± 0.205	0.894 ± 0.135
Neither	0.955 ± 0.008	0.925 ± 0.121	0.890 ± 0.191	0.861 ± 0.017	0.800 ± 0.204	0.800 ± 0.164

C.5 Sensitivity to Data-generating Function Complexity

We vary the depth of the true structural or observation function to examine its effect on latent-state and causal-graph recovery by iCReN-T. We use the default dimensions $d_x = d_z = 5$ and the default instantaneous and lagged graphs. First, we fix the observation function at three layers and vary the structural-function depth from one to five layers. One layer gives linear structural functions, whereas two to five layers are Leaky-ReLU MLPs with increasing numbers of hidden layers; all conditions use the same 10 seeds. As shown in Table 8(a), MCC and both F1 scores remain high for the nonlinear two- to five-layer structural functions. For the linear functions, instantaneous F1 decreases to 0.275 ± 0.275, whereas MCC and lagged F1 remain high. An additively separable linear transition cannot satisfy the sufficient-variability rank condition when instantaneous edges are present (Remark 1). The reduced instantaneous F1 is consistent with this limitation.

Next, we fix the structural functions at three layers and vary the observation-function depth from one to five layers. Table 8(b) shows that iCReN-T has mean MCC and instantaneous F1 that tend to decrease as the observation-function depth increases, whereas lagged F1 shows no clear trend. We also compare iCReN-T, iCReN-G, and the baselines using one- and three-layer observation functions while keeping the structural functions at three layers. Within each seed, the two settings share the same latent states, transition noise, environment labels, and time indices. Table 9 shows higher mean MCC at one layer for every evaluated method, with the largest increases for IDOL and LEAP; both proposed methods have higher mean MCC than all baselines at both depths.

C.6 Ablation of Objective Components

Structural regularization. We remove the sparsity and DAG regularizers separately or jointly to assess their empirical contributions to latent-state and causal-graph recovery. On discrete-variation data with the default $d_z = 5$ and with $d_z = 8$, where recovery begins to deteriorate, we compare the full objective with three ablations: No sparsity with $\lambda_{\text{sparsity}} = 0$, No DAG with $\lambda_{\text{DAG}} = 0$, and Neither with both coefficients set to zero. The results in Table 10 show that removing either or both regularizers does not consistently lower MCC, instantaneous F1, or lagged F1 in the evaluated settings, although the regularizers are motivated by the structural conditions used in the theory.

Reconstruction loss. We add a decoder and reconstruction loss to the default objective to assess how observation reconstruction affects latent-state and causal-graph recovery. For the default discrete-variation data, let $\hat{g}$ be a three-layer decoder from the learned representation to the observa-

Table 11: Effect of adding a decoder and reconstruction loss to the default objective on synthetic data with Gaussian transition-noise distributions that vary across discrete environments. Values are the mean $\pm$ standard deviation over 10 seeds.

Objective	MCC	Inst. F1	Lagged F1	Reconstruction loss
Full	0.951 ± 0.011	0.950 ± 0.158	0.980 ± 0.063	$--$
Full + reconstruction	0.957 ± 0.009	1.000 ± 0.000	0.970 ± 0.048	0.0022 ± 0.0005

Table 12: Main pretraining objectives and auxiliary information for the treadmill motion forecasting experiment. KL denotes Kullback–Leibler.

Method	Pretraining objective	Auxiliary information
iCReN	Reconstruction, TCL, and sparsity regularization	12-class gait phase
IDOL	Reconstruction, KL divergence, and sparsity regularization	None
G-CaRL	Group-contrastive objective	None
CtrlNS	Variational transition objective	No observed label; at most five inferred regimes
LEAP	Variational transition, total correlation, and sparsity regularization	12-class gait phase
β-VAE	Reconstruction and β-weighted KL divergence	None

tion space. For a minibatch $\mathcal{B}$ of observation windows, define

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \sum_{\ell=0}^{L} \left\| \boldsymbol{x}_{t-\ell}^{(b)} - \hat{g}\Big(\phi\Big(\boldsymbol{x}_{t-\ell}^{(b)}\Big)\Big) \right\|_2^2. \tag{47}$$

We compare the full objective with Full + reconstruction, which adds $\mathcal{L}_{\mathrm{rec}}$ with $\lambda_{\mathrm{rec}} = 1$. Table 11 shows that adding reconstruction yields similarly high MCC and F1 scores but no consistent improvement across the three metrics.

D Real-world Experimental Details

We use iCReN-T in both real-world experiments because it generally achieved higher mean recovery scores in the synthetic comparisons. Its classifier applies the element-wise map $a \mapsto (a, a^2, \log(1+a^2))$ to both the estimated transition noise and the history feature. Inputs are standardized using statistics computed only from the training data, and each method's pretraining checkpoint is selected using its method-specific validation objective. After pretraining, each encoder is frozen, and a forecasting head with the same architecture across methods is trained, selected using validation data, and evaluated on test data. Each random seed defines one complete run, including representation pretraining, training of the forecasting head, and evaluation.

D.1 Treadmill Motion Forecasting

Dataset and preprocessing. We use the treadmill walking data provided by Riglet et al. (2024), including sequences recorded at Slow, Comfortable, and Fast walking speeds. From the three-dimensional coordinates of markers on the left and right sides of the pelvis, thigh, knee, ankle, heel, and toe, we construct 36-dimensional sequences. We align the coordinate axes, convert the coordinates from millimeters to meters, center each frame by subtracting the midpoint of the left and right pelvic markers, and retain every eighth frame of the original 100-Hz trajectories, yielding a sampling rate of 12.5 Hz. We split the 30 participants into 19 for training, 6 for validation, and 5 for testing. We estimate the normalization statistics using only sequences recorded at the Slow and Fast walking speeds from the training participants and apply them unchanged to validation and test data. We use Slow and Fast for training and model selection and Comfortable only for final test evaluation.

Pretraining. We pretrain each method for at most 20,000 steps with latent dimension 5 and batch size 256. Following the motion-capture setting of IDOL (Li et al., 2025b), we set the maximum lag to

Table 13: Mean marker error (mm) at every third predicted frame for Comfortable test sequences, which are excluded from training and model selection. Values are the mean ± standard deviation across runs; boldface indicates the best value in each column. We report 10 runs for iCReN, IDOL, G-CaRL, CtrlNS, and β-VAE and the eight numerically stable runs for LEAP.

(a) Early prediction horizons

Prediction horizon [s]	0.24	0.48	0.72	0.96
iCReN	$\mathbf{35.518 \pm 1.236}$	$\mathbf{42.060 \pm 1.939}$	$\mathbf{47.863 \pm 1.255}$	$\mathbf{58.650 \pm 2.051}$
IDOL	36.654 ± 2.106	44.076 ± 2.304	50.717 ± 1.786	62.400 ± 2.119
G-CaRL	45.942 ± 4.276	52.197 ± 4.917	60.568 ± 3.640	75.790 ± 7.785
CtrlNS	37.152 ± 1.859	43.593 ± 2.267	48.861 ± 1.610	62.337 ± 1.978
LEAP	37.281 ± 1.326	44.220 ± 1.825	50.902 ± 1.568	63.431 ± 2.825
β-VAE	36.429 ± 0.877	43.391 ± 1.865	51.058 ± 2.257	63.305 ± 2.546

(b) Later prediction horizons

Prediction horizon [s]	1.20	1.44	1.68	1.92
iCReN	$\mathbf{67.673 \pm 2.831}$	$\mathbf{82.034 \pm 2.725}$	$\mathbf{95.912 \pm 2.611}$	$\mathbf{107.309 \pm 4.345}$
IDOL	71.989 ± 2.691	86.791 ± 2.547	102.094 ± 3.543	115.146 ± 3.822
G-CaRL	86.037 ± 6.819	103.133 ± 9.088	118.191 ± 10.616	135.169 ± 8.541
CtrlNS	70.025 ± 1.759	86.040 ± 4.300	101.372 ± 3.605	117.059 ± 7.501
LEAP	70.623 ± 2.281	87.338 ± 4.548	99.185 ± 3.018	113.210 ± 7.738
β-VAE	73.886 ± 2.671	91.089 ± 4.090	106.227 ± 4.279	123.960 ± 5.890

one for all dynamic methods. Using the right-foot-strike events provided with the dataset, we define each gait cycle as the interval between consecutive strikes, divide it uniformly into 12 phases, and use the phase index as the auxiliary variable for iCReN and LEAP. We fit each representation model on the Slow and Fast walking speeds from the 19 training participants and select its checkpoint on the same walking speeds from the six validation participants using the method-specific validation objective. For iCReN, a frame-wise decoder reconstructs the 36-dimensional marker coordinates from each latent representation, and the pretraining objective combines reconstruction and TCL classification losses with sparsity regularization on the input gates, using $\lambda_{\text{rec}} = 1$, $\lambda_{\text{IIA}} = 0.001$, and $\lambda_{\text{sparsity}} = 5 \times 10^{-5}$. Because we use a fixed lower-triangular mask for the instantaneous structure, no DAG regularization is applied in this experiment. Two LEAP runs produced non-finite losses or gradients during pretraining and were therefore excluded from downstream evaluation without replacement; we report the remaining eight runs. Table 12 summarizes the pretraining objectives and auxiliary information.

Downstream forecasting. After pretraining, we freeze each selected encoder and train the same two-layer forecasting head with hidden width 64 and Gaussian error linear unit (GELU) activations for all methods to directly predict the marker coordinates for the following 24 frames (1.92 seconds) from the latent representations of the preceding 12 frames (0.96 seconds). For downstream learning, we further divide the 19 training participants into 15 for fitting the forecasting head and 4 for selecting its checkpoint; both subsets contain only the Slow and Fast walking speeds. We train the head for at most 100,000 steps using Adam with learning rate 10^{-3} and batch size 256 and select the checkpoint with the lowest mean marker error averaged over all 24 prediction horizons on the four-participant selection subset. We then evaluate the selected encoder and forecasting head on the Comfortable walking speed from the five test participants without further parameter updates or model selection.

Evaluation metric. Before computing the error, we return the standardized predictions and targets to the physical coordinate scale using the training-set normalization statistics. Let $n = 1, \ldots, N$ index test windows, $h = 1, \ldots, H$ index prediction horizons, and $j = 1, \ldots, J$ index markers, where $N = 7{,}214$, $H = 24$, and $J = 12$. Let $\boldsymbol{y}_{n,h,j}$ and $\hat{\boldsymbol{y}}_{n,h,j}$ denote the true and predicted three-dimensional marker coordinates, respectively. We compute the mean marker error at horizon h as

$$\frac{1}{NJ} \sum_{n=1}^{N} \sum_{j=1}^{J} \left\| \boldsymbol{y}_{n,h,j} - \hat{\boldsymbol{y}}_{n,h,j} \right\|_2 . \tag{48}$$

We convert the resulting distances from meters to millimeters and report every third prediction horizon, from $h = 3$ (0.24 seconds) to $h = 24$ (1.92 seconds). As shown in Table 13, iCReN attains the lowest mean marker error at every reported horizon.

D.2 Weather Forecasting

Dataset and preprocessing. We construct regional weather time series from the ERA5 global climate reanalysis dataset (Hersbach et al., 2020; Copernicus Climate Change Service, 2023) for the eight regions listed in Table 14. For each of the five cities in a region, we use the nearest 0.25° grid point and sample the data at three-hour intervals. At each time point, the encoder input comprises 2-m temperature (`t2m`), mean sea-level pressure (`msl`), 10-m zonal and meridional wind components (`u10` and `v10`), and total precipitation over the preceding three hours (`tp`) for all five cities. We convert the three-hour precipitation totals to millimeters and apply a `log1p` transformation. We use only windows consisting of consecutive three-hour observations. Within each month of the 1981–2013 pretraining period, we divide the days into four periods (1–7, 8–14, 15–21, and 22 onward) and reserve the final day of each period for validation. The preceding L time points used as context for these validation windows are also excluded from training and from the computation of normalization statistics. We standardize each input variable separately for each city using only the remaining 1981–2013 training samples and apply the resulting statistics unchanged to the pretraining validation and later downstream data.

Pretraining. We use a calendar index as the auxiliary variable during pretraining. The index combines month (12 categories), the four within-month periods defined above, and UTC time (8 three-hour categories), yielding 384 classes. We exclude February 29 and do not include the year, so the same 384 classes recur across years. Only iCReN and LEAP use this observed auxiliary variable during pretraining; the other methods receive no calendar label. We pretrain each method for at most 10,000 steps on data from 1981–2013, use latent dimension 5, and set the maximum lag to 8 for the dynamic methods, covering the preceding 24 hours at the three-hour sampling interval. For iCReN, the pretraining objective combines reconstruction and TCL classification losses with sparsity regularization on the input gates, using $\lambda_{\text{rec}} = 1$, $\lambda_{\text{IIA}} = 0.005$, and $\lambda_{\text{sparsity}} = 5 \times 10^{-5}$. A fixed lower-triangular mask is used for the instantaneous structure, so no DAG penalty contributes to the objective. Table 15 summarizes the pretraining objectives and auxiliary information.

Downstream forecasting. After pretraining, we freeze each selected encoder and train the same two-layer forecasting head with hidden width 256, GELU activations, and dropout 0.1 for all methods. The head uses the latent representations at the preceding eight time points to predict the relative humidity of all five cities three hours ahead. We fit the head on data from 2014–2016 for at most 5,000 steps using Adam with learning rate 10^{-3} and select its checkpoint by standardized MSE on 2017 validation data. We then evaluate the selected encoder and head on data from 2018–2020 without further parameter updates or model selection. The calendar index is not provided to the forecasting head.

Target and evaluation metric. We derive relative humidity, which is not used during pretraining, from 2-m temperature and dew-point temperature and evaluate the three-hour-ahead forecasts using standardized MSE. Let T and T_d denote 2-m temperature and dew-point temperature in kelvin, and

Table 14: Cities used in the weather forecasting experiment. We use the ERA5 0.25° grid point nearest to each city center.

Region	Cities
Southeastern Australia	Brisbane, Sydney, Canberra, Melbourne, Hobart
Canada	Toronto, Kingston, Ottawa, Montreal, Quebec City
China	Harbin, Beijing, Shanghai, Wuhan, Guangzhou
France	Lille, Paris, Strasbourg, Bordeaux, Marseille
Germany	Hamburg, Berlin, Cologne, Frankfurt, Munich
Japan	Sapporo, Sendai, Tokyo, Osaka, Fukuoka
South Korea	Seoul, Daejeon, Daegu, Gwangju, Busan
US Atlantic	Boston, New York, Washington, D.C., Atlanta, Miami

Table 15: Main pretraining objectives and auxiliary information for the weather forecasting experiment.

Method	Pretraining objective	Auxiliary information
iCReN	Reconstruction, TCL, and sparsity regularization	384-class calendar index
IDOL	Reconstruction, KL divergence, and sparsity regularization	None
G-CaRL	Group-contrastive objective	None
CtrlNS	Variational transition objective	No observed label; at most five inferred regimes
LEAP	Variational transition, total correlation, and sparsity regularization	384-class calendar index
β-VAE	Reconstruction and β-weighted KL divergence	None

Table 16: Standardized MSE for three-hour-ahead relative-humidity forecasting in eight regions. Each region comprises five cities. Values for the learned methods are the mean $\pm$ standard deviation over 10 runs; boldface indicates the best learned method in each column. The two panels partition the columns only for readability. Persistence is a deterministic forecasting reference and is shown without a standard deviation.

(a) Southeastern Australia, Canada, China, and France

Method	AU	CAN	CN	FR
iCReN	$\mathbf{0.352 \pm 0.004}$	0.481 ± 0.011	$\mathbf{0.546 \pm 0.010}$	$\mathbf{0.422 \pm 0.015}$
IDOL	0.359 ± 0.004	0.484 ± 0.010	0.555 ± 0.006	0.438 ± 0.007
G-CaRL	0.649 ± 0.017	0.673 ± 0.013	0.664 ± 0.009	0.750 ± 0.019
CtrlNS	0.385 ± 0.021	0.522 ± 0.084	0.618 ± 0.028	0.459 ± 0.061
LEAP	0.396 ± 0.006	0.524 ± 0.009	0.591 ± 0.012	0.486 ± 0.014
β-VAE	0.357 ± 0.003	$\mathbf{0.475 \pm 0.011}$	0.561 ± 0.015	0.445 ± 0.011
Persistence (reference)	0.477	0.373	0.356	0.434

(b) Germany, Japan, South Korea, and US Atlantic

Method	GER	JP	KR	US
iCReN	0.474 ± 0.011	$\mathbf{0.414 \pm 0.013}$	$\mathbf{0.291 \pm 0.007}$	0.427 ± 0.009
IDOL	$\mathbf{0.469 \pm 0.010}$	0.440 ± 0.009	0.304 ± 0.007	0.440 ± 0.008
G-CaRL	0.754 ± 0.028	0.615 ± 0.026	0.535 ± 0.016	0.554 ± 0.020
CtrlNS	0.504 ± 0.043	0.501 ± 0.057	0.382 ± 0.048	0.462 ± 0.026
LEAP	0.566 ± 0.018	0.482 ± 0.015	0.327 ± 0.011	0.460 ± 0.012
β-VAE	0.484 ± 0.017	0.426 ± 0.009	0.328 ± 0.009	$\mathbf{0.416 \pm 0.008}$
Persistence (reference)	0.437	0.446	0.374	0.334

define their values in degrees Celsius as $\tau = T - 273.15$ and $\tau_d = T_d - 273.15$. Using Bolton's approximation of saturation vapor pressure (Bolton, 1980), we compute

$$\mathrm{RH}(T, T_d) = \mathrm{clip}_{[0,100]}\left[100 \exp\left(\frac{17.67\tau_d}{\tau_d + 243.5} - \frac{17.67\tau}{\tau + 243.5}\right)\right]. \tag{49}$$

Dew-point temperature is used only for this calculation and is not provided to the encoder. We standardize relative humidity separately for each city using statistics from the 1981–2013 pretraining period. Let $n = 1, \ldots, N$ index test time points and $c = 1, \ldots, C$ index cities, and denote the standardized true and predicted relative humidity by $y_{n,c}$ and $\hat{y}_{n,c}$, respectively. We compute

$$\mathrm{MSE}_{\mathrm{std}} = \frac{1}{NC} \sum_{n=1}^{N} \sum_{c=1}^{C} \left(y_{n,c} - \hat{y}_{n,c}\right)^2. \tag{50}$$

For each run, we evaluate the metric over all test time points and cities from 2018–2020. Table 16 reports the results. We also report persistence, which uses the most recently observed relative humidity as the forecast. Because the learned methods do not receive relative humidity as an input, we treat persistence as a reference and exclude it from the boldface comparison.

D.3 Data Sources and Availability

The ERA5 hourly data on single levels used in the weather experiments are publicly available from the Copernicus Climate Change Service Climate Data Store (Copernicus Climate Change Service,

2023) under DOI 10.24381/cds.adbb2d47; we retrieved them on September 2, 2026. We do not redistribute the data, and document their acquisition and preprocessing in the supplementary code package. The results contain modified Copernicus Climate Change Service information 2020. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

The treadmill experiment uses the publicly available motion-capture dataset of Riglet et al. (2024), distributed through Figshare under DOI 10.6084/m9.figshare.c.7056797.v1. We do not redistribute the source data; the supplementary package provides the split definition and preprocessing code.