LATENT-SPACE DENOISING FOR TEMPORAL CAUSAL REPRESENTATION LEARNING WITH MECHANISM SPARSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Observation noise poses a major challenge to causal representation learning. In temporal data, noise affects latent representations at successive time steps, making temporal dependencies harder to estimate and causal structures harder to recover. To address this challenge, we integrate our latent-space denoising module with an existing mechanism-sparsity method to form a joint learning framework, Latent Denoising with Mechanism Sparsity (LDMS). The framework uses free-energy-guided particle dynamics to refine the latent distribution and combines denoising regularization with variational reconstruction and sparse transition estimation to improve the representations used for temporal mechanism learning. Our theoretical analysis considers the effects of both latent states and auxiliary variables at the preceding time step on the current latent state. We establish stability bounds for mechanism strengths after denoising, sufficient conditions for preserving sparse dependencies, and consistency conditions for joint denoising and mechanism estimation. These results quantify how denoising error affects mechanism estimation and graph recovery, providing a theoretical basis for the joint framework. Experimental results show that LDMS improves latent factor alignment and graph recovery, demonstrating the benefits of latent-space denoising for existing temporal causal representation learning methods and its potential as a plug-and-play module.

1 INTRODUCTION

Representation learning seeks useful features from complex observations (Bengio et al., 2013; LeCun et al., 2015). Causal inference provides a framework for modeling mechanisms and interventions (Pearl, 2009; Peters et al., 2017). Causal representation learning connects these problems by seeking high-level causal variables from low-level observations (Schölkopf et al., 2021). Identifiability depends on suitable inductive biases (Locatello et al., 2019); auxiliary variables provide identifying information for nonlinear ICA and variational models (Hyvärinen et al., 2019; Khemakhem et al., 2020). Interventions and partially observed multiple views provide further routes to causal representation learning (Brehmer et al., 2022; Ahuja et al., 2023; Yao et al., 2024). For time-series data, the goal is to recover the latent variables that describe a system and the causal dependencies that govern their evolution. This requires more than reconstructing observations or predicting future measurements: the learned variables must also reflect the underlying causal structure, within the ambiguities allowed by the model.

Temporal causal representation learning uses successive observations to identify latent variables and their dependencies. Nonlinear ICA exploits temporal nonstationarity and dependence to identify latent sources (Hyvärinen & Morioka, 2016; 2017). Existing methods establish identification results using assumptions on temporal mechanisms and distribution changes (Yao et al., 2022b;a), or observed intervention targets (Lippe et al., 2022). Mechanism-sparsity learning provides another approach by restricting the number of past latent states or auxiliary inputs that influence each latent variable (Lachapelle et al., 2022; 2026). Under suitable structural and variability conditions, sparsity restricts the representations consistent with the observations. Some graph structures permit complete disentanglement; others allow mixing within groups of factors. Representation recovery and graph recovery must therefore be assessed together and in relation to the assumed model.

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.

Recovering latent variables and their temporal dependencies from noisy observations remains an estimation challenge, even when identification conditions are available. A nonlinear encoder can transform measurement noise into state-dependent errors in the latent representation. Errors in preceding representations affect the inputs used to estimate a transition, while errors in current representations affect its response. Denoising autoencoders and variational approaches learn from corrupted inputs through reconstruction objectives (Vincent et al., 2008; Im et al., 2017), and structured nonlinear ICA explicitly addresses additive observation noise (Hälvä et al., 2021). However, better reconstruction does not by itself establish accurate latent dependencies. Physical causal benchmarks also show the need to assess representation learning under measurement noise and nonlinear mixing (Gamella et al., 2025b;a).

For a mechanism-sparsity model, adding latent denoising raises a specific question: when does refinement preserve the sparse dependencies that the original method seeks to recover? Denoising changes the distributions used to estimate the mechanisms. A closer fit to the latent marginal distribution alone need not preserve the joint distribution of consecutive states or their dependence on auxiliary inputs. The analysis must therefore relate representation error to changes in mechanism strength and account for errors in the fitted mechanisms. This analysis specifies the representation quality needed to preserve the original sparse parent sets.

We introduce Latent Denoising with Mechanism Sparsity (LDMS) to combine latent distributional refinement with sparse temporal mechanism estimation. A free-energy objective fits a noise-convolved candidate latent density to the encoded observations, and particle updates provide a regularizer for the encoder. We combine this regularizer with variational reconstruction and sparse transition learning. Encoder outputs remain associated with their observations, preserving the temporal pairing needed to estimate dependencies. The module refines the representations while retaining the original model's temporal and auxiliary inputs.

Our contributions are threefold. First, we formulate a joint objective for latent-space denoising and sparse temporal mechanism estimation. Second, we derive perturbation bounds for latent and auxiliary mechanism strengths, sufficient conditions for sparse-support preservation, and consistency conditions for mechanism estimation. These results specify when sufficiently accurate representations and fitted mechanisms preserve the original sparse parent sets. Third, we evaluate recovery of latent variables and graph structure on nine synthetic datasets over five trials per method and dataset. The results show joint improvements on several datasets. We also evaluate LDMS on real hydraulic observations, where it improves reconstruction and the test variational objective across the examined training budgets.

2 Related Work

Causal representation learning from time series uses temporal structure to recover latent causal variables and their relations. SlowVAE uses sparse temporal changes to identify latent factors (Klindt et al., 2021), while independent innovation analysis identifies the innovations driving nonlinear autoregressive processes (Morioka et al., 2021). LEAP and TDRL study latent causal dynamics under assumptions on temporal mechanisms and distribution changes (Yao et al., 2022b;a). CtrlNS uses sparse transitions to learn causal representations across distribution shifts without observed domain labels (Song et al., 2024). IDOL studies both instantaneous and time-delayed latent influences (Li et al., 2025b). CaRiNG uses temporal context to handle non-invertible observation maps (Chen et al., 2024), while CHiLD studies hierarchical temporal causal representations (Li et al., 2025a). These methods address different latent dynamics and observation processes under distinct identification assumptions.

Intervention information and mechanism sparsity provide further ways to identify temporal causal representations. CITRIS uses observed intervention targets, and iCITRIS also allows instantaneous causal effects (Lippe et al., 2022; 2023b). BISCUIT uses observed interaction information to infer causal variables and unknown binary interaction variables (Lippe et al., 2023a). Mechanism-sparsity learning jointly estimates latent variables and sparse dependencies on past states or auxiliary inputs (Lachapelle et al., 2022). Its nonparametric extension characterizes partial disentanglement through graph-preserving transformations (Lachapelle et al., 2026).

3 Model and Training Method

LDMS combines our latent denoising module with the sparse temporal model of Lachapelle et al. (2026). The shared encoder supports both distributional refinement and mechanism estimation (Figure 1).



Figure 1: Joint LDMS training procedure. Noisy observations are generated for the benchmarks and encoded for latent-space learning. Particle-guided latent refinement and sparse conditional learning share a joint objective. The arrows indicate the training stages. Denoising regularizes the encoder, and gradients are not propagated through particle updates.

3.1 Temporal Model and Graph Convention

Let t index discrete time, and let $X^t \in \mathbb{R}^p$, $Z^t \in \mathbb{R}^d$, and $A^{t-1} \in \mathbb{R}^q$ denote the observation, latent state, and preceding auxiliary input, respectively; p, d, q are their dimensions. Uppercase letters denote random vectors and lowercase letters their realizations. Both Z^{t-1} and A^{t-1} influence Z^t. We consider the first-order temporal model

$$X^t = g_\star(Z^t) + \epsilon^t, \tag{1}$$

$$p_\star(z^t \mid z^{t-1}, a^{t-1}) = \prod_{i=1}^{d} p_{\star,i}\Big(z_i^t \mid z^{t-1}_{\mathrm{Pa}_z(i)}, a^{t-1}_{\mathrm{Pa}_a(i)}\Big). \tag{2}$$

Here $g_\star : \mathbb{R}^d \to \mathbb{R}^p$ is the injective mixing function, $\epsilon^t \in \mathbb{R}^p$ is observation noise independent of the clean latent sequence, and $p_\star$ and $p_{\star,i}$ are the true joint and coordinatewise transition densities. The sets $\mathrm{Pa}_z(i) \subseteq \{1, \dots, d\}$ and $\mathrm{Pa}_a(i) \subseteq \{1, \dots, q\}$ contain the latent and auxiliary parents of coordinate i; subscripts indexed by these sets select the corresponding subvectors. Equation (2) describes time-delayed dependencies, including self-lags and reciprocal influences across successive times. Throughout the paper, G_z represents $Z^{t-1} \to Z^t$ and G_a represents $A^{t-1} \to Z^t$; both belong to the temporal mechanism structure.

Let $G_z^\star \in \{0,1\}^{d\times d}$ and $G_a^\star \in \{0,1\}^{d\times q}$ encode the minimal conditional dependencies in (2). Throughout the paper, *rows index targets and columns index sources*:

$$(G_z^\star)_{ij} = 1 \iff j \in \mathrm{Pa}_z(i), \qquad (G_a^\star)_{ik} = 1 \iff k \in \mathrm{Pa}_a(i). \tag{3}$$

This convention follows the mechanism-sparsity framework. Because every edge points forward in time, the unrolled graph is acyclic even with a general lag adjacency matrix $G_z^\star$.

The variational posterior has parameters ϕ and takes the form $q_\phi(z^t \mid x^t) = \mathcal{N}(\mu_\phi(x^t), \operatorname{diag} v_\phi(x^t))$. Here $\mu_\phi(x) \in \mathbb{R}^d$ and $v_\phi(x) \in (0,\infty)^d$ are the posterior mean and variance vectors, and $\operatorname{diag} v$ denotes the diagonal matrix with diagonal v. We use the posterior mean as the encoder representation, $U^t = \mu_\phi(X^t)$. The decoder density p_ϑ and transition density p_ψ have parameters ϑ and ψ. The mask matrices $M_z \in [0,1]^{d\times d}$ and $M_a \in [0,1]^{d\times q}$ are binary or continuous relaxations of binary matrices. The ith conditional mechanism receives

$$M_{z,i:} \odot z^{t-1} + (1 - M_{z,i:}) \odot d_i^z, \qquad M_{a,i:} \odot a^{t-1} + (1 - M_{a,i:}) \odot d_i^a, \tag{4}$$

Here $M_{z,i:}$ and $M_{a,i:}$ denote the corresponding row vectors, $\odot$ is componentwise multiplication, and $d_i^z \in \mathbb{R}^d, d_i^a \in \mathbb{R}^q$ are fixed reference vectors, set to zero in the experiments. Subtraction from 1 is componentwise. The network parameterizes a Gaussian conditional mechanism. Categorical contexts, when present, can be represented by a masked embedding.

3.2 Latent Distributional Refinement

We summarize our latent-space denoising module here and give a self-contained derivation in Appendix D.1. A local linearization of the encoder motivates an additive approximation to latent perturbations. We use Gaussian convolution as a local approximation to potentially state-dependent

perturbations. For a fixed encoder, let ν be the encoded-observation density and ρ a candidate clean latent density. With a centered Gaussian kernel κ_H of covariance $H \succ 0$, define

$$q_\rho = \rho * \kappa_H, \qquad \mathcal{F}_\nu(\rho) = \mathrm{KL}(q_\rho \| \nu) + \lambda_H \,\mathrm{Ent}(\rho), \qquad \mathrm{Ent}(\rho) = \int_{\mathbb{R}^d} \rho(v) \log \rho(v)\, dv. \quad (5)$$

Here $*$ denotes convolution, $\lambda_H \geq 0$ is the entropy weight, and $\mathrm{KL}(r\|s) = \int r \log(r/s)$, with value $+\infty$ when absolute continuity fails. We consider densities for which (5) is finite. The KL term fits the noise-convolved candidate to the encoded observations, while the negative-entropy term discourages concentration. The objective estimates a latent distribution without paired clean targets. Its weighted Wasserstein descent motivates the particle update below.

3.3 Latent-Space Denoising and Joint Estimation

With the encoder fixed, N persistent particles $P = \{V_m\}_{m=1}^N$ represent the candidate distribution. Their updates use a smoothed score difference and an entropy score computed from kernel density estimates. We then hold the particles, noisy query points, and bandwidths fixed and update the encoder using $\widehat{\mathcal{R}}_{\mathrm{LDMS}}$. Only the negative log-density of the encoded observations at these queries contributes to the encoder gradient. The particles provide a distributional training signal, while encoder outputs remain paired across time for mechanism learning. Equation (9) couples this regularizer to variational and sparse learning. Appendix D.2 defines the update and surrogate, and Algorithm 1 gives the full procedure.

3.4 Variational Learning and Sparsity

We use the variational formulation of latent-variable learning (Kingma & Welling, 2014; 2019), with the sparse temporal parameterization of Lachapelle et al. (2026). For a two-step window $x^- = x^{t-1}, x^+ = x^t$ and its preceding auxiliary input $a = a^{t-1}$, let $q_- = q_\phi(\cdot \mid x^-)$ and $q_+ = q_\phi(\cdot \mid x^+)$. Here $Z^- = Z^{t-1}$ and $Z^+ = Z^t$; $\mathbb{E}_{q_s}$ denotes expectation under q_s. With a nonnegative transition weight β, define

$$\mathcal{L}_{\mathrm{rec}} = -\frac{1}{2p} \sum_{s\in\{-,+\}} \mathbb{E}_{q_s} \log p_\vartheta(x^s \mid Z^s), \quad (6)$$

$$\mathcal{K}_{\mathrm{tr}} = \mathbb{E}_{q_-} \,\mathrm{KL}\big(q_+ \| p_\psi(\cdot \mid Z^-, a; M_z, M_a)\big)\,, \quad (7)$$

$$\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{rec}} + \frac{\beta}{p}\mathcal{K}_{\mathrm{tr}}. \quad (8)$$

Equation (8) defines a transition-regularized window objective. For the temporal experiments, we include an initial-prior term, replacing $\mathcal{K}_{\mathrm{tr}}$ by $(\mathcal{K}_{\mathrm{tr}} + \mathrm{KL}(q_-\|p_0))/2$. Here p_0 is the initial latent-state prior. At $\beta = 1$, this yields the negative two-step evidence lower bound (ELBO) divided by $2p$, as derived in Appendix C. For longer trajectories, the full-sequence ELBO includes all initial and transition terms.

Let $\Pi_z \in [0,1]^{d\times d}$ and $\Pi_a \in [0,1]^{d\times q}$ be the matrices of edge-inclusion probabilities, and set $\mathcal{S}_z = \sum_{ij}(\Pi_z)_{ij}$ and $\mathcal{S}_a = \sum_{ik}(\Pi_a)_{ik}$. The penalty formulation minimizes the stochastic objective

$$\mathcal{J} = \mathcal{L}_{\mathrm{VAE}} + \lambda_D \widehat{\mathcal{R}}_{\mathrm{LDMS}} + \frac{\lambda_z}{p}\mathcal{S}_z + \frac{\lambda_a}{p}\mathcal{S}_a. \quad (9)$$

The nonnegative coefficients $\lambda_D, \lambda_z, \lambda_a$ weight denoising and the two sparsity penalties. The quantities $\mathcal{S}_z, \mathcal{S}_a$ are expected edge counts. Mask samples or differentiable relaxations enter the VAE term. In the alternative constrained formulation, the explicit sparsity penalties are replaced by $\mathcal{S}_z \leq s_z$ and/or $\mathcal{S}_a \leq s_a$, where $s_z, s_a \geq 0$ are prescribed upper bounds on expected edge counts; these constraints are enforced by nonnegative Lagrange multipliers. When categorical inputs are present, their masks also contribute to the auxiliary sparsity penalty or constraint.

Training alternates particle updates, performed with the encoder fixed, and gradient updates of the encoder, decoder, transition model, and masks. The denoising term refines the encoder distribution, and the updated representations guide transition and mask learning through the variational objective. Evaluation uses the trained encoder means and learned masks. Algorithm 1 summarizes the training procedure.

4 THEORETICAL ANALYSIS

We study when latent denoising is compatible with the sparse temporal mechanisms targeted by the original method. Our results relate conditional representation quality to mechanism strengths and graph recovery in fixed, aligned latent coordinates. Appendix B contains auxiliary results, full consistency conditions, and proofs. We use Law for distribution, W_2 for quadratic Wasserstein distance, and $\|\cdot\|_{L^2(\mu)}$ for the L^2 norm under μ.

Mechanisms and representation convention. Consider conditional mean mechanisms with exogenous transition noise,

$$Z_i^t = f_i(Z^{t-1}_{\mathrm{Pa}_z(i)}, A^{t-1}_{\mathrm{Pa}_a(i)}) + \eta_i^t, \qquad i = 1, \dots, d. \tag{10}$$

Here $f = (f_1, \dots, f_d)^\top$ is the conditional mean and $\eta^t = (\eta_1^t, \dots, \eta_d^t)^\top$ is the exogenous transition-noise vector. Each η_i^t is the disturbance affecting the transition of latent coordinate i. We also write $f_i(z, a)$ as a function of the full input vector. All comparisons use fixed latent coordinates and scales, within the graph-preserving identifiability framework of Lachapelle et al. (2026). For $s \in \{t-1, t\}$, let $\widetilde{Z}^s$ denote the encoder representation after LDMS regularization, aligned in these reference coordinates. We compare its distribution with that of Z^s directly, without imposing an additive representation-noise model. The analysis applies to either posterior means or posterior samples, provided the same convention is used throughout each application. Define

$$W = (Z^{t-1}, A^{t-1}), \qquad \widetilde{W} = (\widetilde{Z}^{t-1}, A^{t-1}), \qquad \mathsf{P} = \mathrm{Law}(W), \quad \mathsf{Q} = \mathrm{Law}(\widetilde{W}). \tag{11}$$

Assumption 1 (Exogenous transition noise and moment conditions)**.** *The clean and denoised variables $Z^{t-1}, Z^t, \widetilde{Z}^{t-1}, \widetilde{Z}^t, A^{t-1}$ have finite second moments in the fixed reference coordinates. The exogenous noise η^t has finite second moment and mean zero, is independent of (Z^{t-1}, A^{t-1}), and has mutually independent coordinates. Thus $f_i(W) = \mathbb{E}[Z_i^t \mid W]$, and the transition factorizes as in* (2)*. The mechanism derivatives appearing below are square integrable under* P *and* Q*.*

Representation errors may depend on the state. We characterize them through the joint law of the refined representation and the auxiliary variable. The derivative integrability condition holds, for example, for derivatives of at most linear growth under the stated moment assumptions.

4.1 MECHANISM STRENGTH AND STABILITY

For continuous auxiliary inputs, define the latent and auxiliary mechanism strengths by

$$M^z_{ij} = \left\|\partial_{z_j} f_i\right\|_{L^2(\mathsf{P})}, \qquad M^a_{ik} = \left\|\partial_{a_k} f_i\right\|_{L^2(\mathsf{P})}. \tag{12}$$

The strength matrices are $M^z = (M^z_{ij}) \in [0, \infty)^{d \times d}$ and $M^a = (M^a_{ik}) \in [0, \infty)^{d \times q}$. For the input-type label $b \in \{z, a\}$, M^b denotes M^z when $b = z$ and M^a when $b = a$. Define the sets of candidate edges $\mathcal{I}_z = \{(z, i, j) : 1 \le i, j \le d\}$, $\mathcal{I}_a = \{(a, i, k) : 1 \le i \le d,\ 1 \le k \le q\}$, and $\mathcal{I} = \mathcal{I}_z \cup \mathcal{I}_a$. For $e = (z, i, j)$, write $i(e) = i$, $\partial_e = \partial_{z_j}$, and $M_e = M^z_{ij}$; for $e = (a, i, k)$, use $\partial_e = \partial_{a_k}$ and $M_e = M^a_{ik}$. The same indexing applies to refined and estimated strengths.

Set $(G^\star_z)_{ij} = \mathbf{1}\{M^z_{ij} > 0\}$ and $(G^\star_a)_{ik} = \mathbf{1}\{M^a_{ik} > 0\}$. These supports encode minimal functional parent sets for C^1 mechanisms on an open rectangular domain with a density positive almost everywhere (Appendix B.7). Superscripted M^z, M^a denote strengths; subscripted M_z, M_a denote architectural masks. Latent-input derivatives also permit discrete auxiliary inputs; Appendix B.14 uses finite contrasts for categorical-input strengths.

With $P_A = \mathrm{Law}(A^{t-1})$, measure representation quality by

$$\Delta_z^2 = \int W_2^2\Big(\mathrm{Law}(\widetilde{Z}^{t-1} \mid A^{t-1} = a), \mathrm{Law}(Z^{t-1} \mid A^{t-1} = a)\Big)\, dP_A(a). \tag{13}$$

This discrepancy uses the coupling formulation of Wasserstein distance (Santambrogio, 2015, Ch. 5), with the auxiliary input held fixed. It is finite under Assumption 1 and reduces to the marginal W_2 distance when auxiliary inputs are absent (Lemma 2).

Assumption 2 (Sensitivity regularity). *Each f_i is C^1. For all evaluated z, z' and a common a, there are finite $L_z, L_a \geq 0$ such that*

$$\|\nabla_z f_i(z,a) - \nabla_z f_i(z',a)\| \leq L_z \|z - z'\|, \tag{14}$$

$$\|\nabla_a f_i(z,a) - \nabla_a f_i(z',a)\| \leq L_a \|z - z'\|. \tag{15}$$

These conditions control sensitivities in the latent argument affected by denoising. Appendix B.5 gives a local version with tail control.

Lemma 1 (Mechanism-strength stability). *Let $\widetilde{M}^z_{ij} = \left\|\partial_{z_j} f_i\right\|_{L^2(\mathsf{Q})}$ and $\widetilde{M}^a_{ik} = \|\partial_{a_k} f_i\|_{L^2(\mathsf{Q})}$. Under Assumptions 1–2,*

$$|\widetilde{M}^z_{ij} - M^z_{ij}| \leq L_z \Delta_z, \qquad |\widetilde{M}^a_{ik} - M^a_{ik}| \leq L_a \Delta_z. \tag{16}$$

Thus the same conditional representation discrepancy controls changes in both types of mechanism strength.

Joint transition risk. To analyze estimation of the transition mechanisms, we also compare the joint distributions of consecutive latent states conditional on the auxiliary variable:

$$\mathfrak{D}^2 = \int W_2^2\Big(\mathrm{Law}((\widetilde{Z}^{t-1}, \widetilde{Z}^t) \mid A^{t-1} = a), \mathrm{Law}((Z^{t-1}, Z^t) \mid A^{t-1} = a)\Big)\, dP_A(a). \tag{17}$$

The cost in this Wasserstein distance is the squared Euclidean distance on $\mathbb{R}^{2d}$, so $\Delta_z \leq \mathfrak{D}$. For a candidate conditional mean g on the joint input domain, define the population risks $\mathcal{R}_i(g) = \mathbb{E}[(Z^t_i - g(W))^2]$ and $\widetilde{\mathcal{R}}_i(g) = \mathbb{E}[(\widetilde{Z}^t_i - g(\widetilde{W}))^2]$.

Proposition 1 (Stability of the transition risk under denoising). *If g is Lipschitz in z with constant $K \geq 0$, uniformly in a, and both risks are finite, then*

$$\left|\sqrt{\widetilde{\mathcal{R}}_i(g)} - \sqrt{\mathcal{R}_i(g)}\right| \leq \sqrt{1 + K^2}\, \mathfrak{D}. \tag{18}$$

Under the joint law of the clean and encoded states, if f_i is K_i-Lipschitz in z,

$$\left\|\widetilde{Z}^t_i - f_i(\widetilde{W}) - \eta^t_i\right\|_{L^2} \leq \left\|\widetilde{Z}^t_i - Z^t_i\right\|_{L^2} + K_i \left\|\widetilde{Z}^{t-1} - Z^{t-1}\right\|_{L^2}. \tag{19}$$

The proposition quantifies the contribution of denoising at both consecutive times. A smaller discrepancy $\mathfrak{D}$ gives a tighter bound on the change in population transition risk. Equation (19) separates the contributions of the current latent response and the preceding latent state.

4.2 Preservation of Sparse Mechanisms

We apply entrywise thresholding, as used in sparse matrix estimation (Bickel & Levina, 2008), to mechanism strengths. For a fixed finite nonempty graph, let the weakest nonzero strength be

$$\gamma = \min\left(\{M^z_{ij} : (G^\star_z)_{ij} = 1\} \cup \{M^a_{ik} : (G^\star_a)_{ik} = 1\}\right) > 0. \tag{20}$$

Theorem 1 (Sparse-support preservation). *Under Assumptions 1–2, choose $0 < \tau < \gamma$ and threshold the refined strengths: $\widetilde{G}_z = \mathbf{1}\{\widetilde{M}^z > \tau\}$ and $\widetilde{G}_a = \mathbf{1}\{\widetilde{M}^a > \tau\}$, entrywise. If*

$$\max\{L_z, L_a\}\Delta_z < \min\{\tau, \gamma - \tau\}, \tag{21}$$

then $(\widetilde{G}_z, \widetilde{G}_a) = (G^\star_z, G^\star_a)$.

The choice $\tau = \gamma/2$ maximizes the admissible perturbation. The theorem shows that sufficiently accurate refined representations preserve the original sparse parent sets when the true mechanisms are evaluated under the refined input law.

For fitted mechanisms $\widehat{f}_i$, let $\widehat{M}^z, \widehat{M}^a$ estimate their derivative norms under Q. Suppose every norm estimate has absolute evaluation error at most b_{ev} on an event of conditional probability at least $1 - \delta$. Define

$$e_z = \max_{ij} \left\|\partial_{z_j} \widehat{f}_i - \partial_{z_j} f_i\right\|_{L^2(\mathsf{Q})}, \qquad e_a = \max_{ik} \left\|\partial_{a_k} \widehat{f}_i - \partial_{a_k} f_i\right\|_{L^2(\mathsf{Q})}. \tag{22}$$

Proposition 2 (Recovery with estimation error)**.** *Under Assumptions 1–2, on the stated evaluation event,*

$$|\widehat{M}^z_{ij} - M^z_{ij}| \le B_z := L_z\Delta_z + e_z + b_{\mathrm{ev}}, \qquad |\widehat{M}^a_{ik} - M^a_{ik}| \le B_a := L_a\Delta_z + e_a + b_{\mathrm{ev}}. \quad (23)$$

If $\max\{B_z, B_a\} < \min\{\tau, \gamma - \tau\}$*, then* $\widehat{G}^\partial_z = \mathbf{1}\{\widehat{M}^z > \tau\}$ *and* $\widehat{G}^\partial_a = \mathbf{1}\{\widehat{M}^a > \tau\}$ *recover both true supports.*

The bound separates representation, mechanism-fitting, and evaluation errors. With the latter two fixed, reducing Δ_z tightens the support-recovery bound. This provides a conditional guarantee for adding latent denoising: once representations and mechanism estimates are sufficiently accurate, the added module preserves the mechanism sparsity targeted by the original method. Appendix B.9 gives a finite-sample evaluation bound, and Appendix B.13 states the additional conditions linking sensitivity graphs to learned hard masks.

4.3 Consistency of Joint Estimation

Let n be the training sample size, $\widehat{f}_{i,n}$ the fitted mechanism, and Q_n the refined input law conditional on the training data. Define $\Delta_{z,n}$ and $\mathfrak{D}_n$ by (13) and (17).

Assumption 3 imposes three conditions: (i) a continuously differentiable mechanism class containing f_i, compact in a metric that controls functions and gradients, on an open rectangular domain where the input density is positive almost everywhere; (ii) uniform Lipschitz bounds in the latent input for mechanisms and their gradients, with bounded population risk; and (iii) uniform convergence of empirical risks and approximate minimization of the penalized empirical criterion. The empirical-risk conditions follow the standard M-estimation framework (van der Vaart, 1998, Ch. 5); the gradient conditions link function and derivative estimation. The full statement is in Appendix B.2.

Write $u_{i,n}$ for the uniform empirical-risk error, $r_{i,n}$ for the range of penalty values over the class, and $\varepsilon_{i,n}$ for optimization error. Let K_i, H_b be the Lipschitz constants for mechanisms and type-b gradients in the latent input, and C_i the upper bound on the square root of clean population risk, as specified in Assumption 3.

Define $c_{i,n} = \sqrt{1 + K_i^2}\,\mathfrak{D}_n$ and

$$E_{i,n} = 2u_{i,n} + 4C_i c_{i,n} + 2c_{i,n}^2 + r_{i,n} + \varepsilon_{i,n}. \quad (24)$$

Theorem 2 (Consistency of joint denoising and mechanism estimation)**.** *Under Assumptions 1 and 3,*

$$\left\|\widehat{f}_{i,n} - f_i\right\|^2_{L^2(\mathsf{P})} \le E_{i,n}. \quad (25)$$

There is a deterministic nondecreasing modulus of continuity $\Omega_i : [0,\infty) \to [0,\infty)$*, with* $\Omega_i(s) \to 0$ *as* $s \downarrow 0$*, such that*

$$\left\|\nabla\widehat{f}_{i,n} - \nabla f_i\right\|_{L^2(\mathsf{P})} \le \Omega_i(\sqrt{E_{i,n}}). \quad (26)$$

Consequently, if $\mathfrak{D}_n \to_p 0$ *and* $r_{i,n}, \varepsilon_{i,n} \to_p 0$*, the joint transition distribution and the estimated mechanisms converge to their population counterparts in the stated metrics. With the edge notation defined above, the errors satisfy*

$$\max_{e\in\mathcal{I}_b}\left|\left\|\partial_e\widehat{f}_{i(e),n}\right\|_{L^2(\mathsf{Q}_n)} - M_e\right| \le H_b\Delta_{z,n} + \max_i \Omega_i(\sqrt{E_{i,n}}) \xrightarrow{p} 0. \quad (27)$$

The error bound separates errors due to denoising, sampling, regularization, and optimization. Reducing $\mathfrak{D}_n$ decreases both denoising terms in $E_{i,n}$. If strength-evaluation errors also vanish, Corollary 1 gives recovery of both temporal graphs with probability tending to one. These results formalize how sufficiently accurate latent denoising and mechanism estimation preserve the original sparse structure.

Corollary 1 (Asymptotic recovery of both temporal graphs)**.** *Under the convergence conditions of Theorem 2, suppose the strength evaluation bound holds with error* $b_{\mathrm{ev},n} \to_p 0$ *and conditional failure probability* $\delta_n \to 0$*. For a fixed nonempty graph and any fixed* $0 < \tau < \gamma$*,*

$$\mathbb{P}\Big((\widehat{G}^\partial_{z,n}, \widehat{G}^\partial_{a,n}) = (G^\star_z, G^\star_a)\Big) \longrightarrow 1. \quad (28)$$

For population evaluation of the true mechanisms, the same conclusion follows directly from $\Delta_{z,n} \to_p 0$ *and Theorem 1.*

5 EXPERIMENTS

We evaluate recovery of latent causal variables and their dependency graphs from noisy observations. Synthetic temporal datasets assess dependencies between successive latent states, while action-driven datasets assess dependencies from auxiliary inputs to latent variables. Known ground truth allows separate evaluation of representation and graph recovery. InTaSet additionally assesses reconstruction and probabilistic modeling on real hydraulic observations.

5.1 EXPERIMENTAL SETUP

We compare LDMS with a mechanism-sparsity VAE baseline without the denoising regularizer. For each synthetic dataset and method, we run five trials with resampled data and new initializations, and report the mean and sample standard deviation across trials. Both methods use the same model architecture and training budget. In the three temporal comparisons marked LDMS$^{\triangle}$, LDMS also uses a lower-triangular support prior that is absent from the baseline. Appendices E.1 and E.4 give the comparison protocol and hyperparameters.

For InTaSet, both methods share a chronological data split, training-only standardization, model architecture, and final expected-edge budget of eight. LDMS$^{\triangle}$ combines latent denoising with a lower-triangular candidate support, and the baseline uses an unrestricted candidate support. We report means and sample standard deviations over five training seeds on this fixed split. For each method and seed, validation negative ELBO selects a final model from ten training budgets between 10k and 100k updates; all selected models use 30k updates. Appendix F.1 gives the detailed protocol.

5.2 DATASETS

Synthetic datasets. The nine datasets follow the temporal and action families of Lachapelle et al. (2026), with an additional temporal chain. Diagonal, non-diagonal, and block structures provide different patterns of dependence and different degrees of latent factor separation. Appendix E.2 describes the data-generating processes and ground-truth graphs.

InTaSet. InTaSet records a physical interconnected four-tank system (Putri et al., 2026). We use the closed-loop normal-operation sequence from InTaSet-FD, containing 11,971 records with four tank-level and two pump-related signals. The sequence is divided chronologically into training, validation, and test blocks, and adjacent observations form two-step windows. The dataset is publicly available. Appendix F describes preprocessing and evaluation.

5.3 EVALUATION METRICS

For synthetic data, we report the mean correlation coefficient (MCC), graph-constrained recovery score (R_{con}), and structural Hamming distance (SHD). MCC measures alignment with individual latent variables; R_{con} evaluates representation recovery using graph-preserving predictor sets; and SHD counts errors in the aligned graph support. Higher MCC and R_{con} indicate better representation recovery, while lower SHD indicates better graph recovery. Appendix E.3 defines the metrics and their aggregation across trials.

For InTaSet, MSE, RMSE, and NRMSE assess reconstruction from the encoder posterior mean, averaged over the four tank channels and summarized across seeds. Test negative ELBO evaluates reconstruction and posterior–prior matching over all six observation channels. Lower values are better. Appendix F.2 defines the metrics, normalization, and aggregation rules.

5.4 EXPERIMENTAL RESULTS

Table 1 presents the four datasets on which LDMS achieves better mean values for all three metrics. Appendix E.5 reports the remaining five datasets, completing the comparison across all nine datasets.

Temporal mechanisms. On TimeDiag, LDMS recovers the ground-truth graph in every trial, giving zero SHD. On TimeNonDiag–chain, mean MCC and R_{con} increase from 0.619 to 0.721, while SHD decreases from 12.2 to 9.2. The reduction in SHD reflects improved self-lag recovery. On

Table 1: Results for the four datasets on which LDMS achieves better mean values for all three reported metrics. Entries show the mean ± sample standard deviation over five trials per method. Boldface indicates the better mean. LDMS$^{\triangle}$ additionally uses a lower-triangular support prior absent from the baseline. Results for the other datasets appear in Table 6.

Dataset	Method	MCC ↑	R_{con} ↑	SHD ↓
TimeDiag	Baseline	0.6780 ± 0.0740	0.6780 ± 0.0740	5.6000 ± 0.9000
	LDMS$^{\triangle}$	$\mathbf{0.6850 \pm 0.0660}$	$\mathbf{0.6850 \pm 0.0660}$	$\mathbf{0.0000 \pm 0.0000}$
TimeNonDiag–chain	Baseline	0.6190 ± 0.0400	0.6190 ± 0.0400	12.2000 ± 1.5000
	LDMS$^{\triangle}$	$\mathbf{0.7210 \pm 0.0690}$	$\mathbf{0.7210 \pm 0.0690}$	$\mathbf{9.2000 \pm 0.4000}$
TimeNonDiag–triangular	Baseline	0.6310 ± 0.0370	0.6310 ± 0.0370	40.6000 ± 4.9000
	LDMS$^{\triangle}$	$\mathbf{0.6440 \pm 0.0860}$	$\mathbf{0.6440 \pm 0.0860}$	$\mathbf{40.0000 \pm 12.1000}$
ActionDiag	Baseline	0.7460 ± 0.0450	0.7460 ± 0.0450	9.4000 ± 0.5000
	LDMS	$\mathbf{0.7680 \pm 0.0900}$	$\mathbf{0.7680 \pm 0.0900}$	$\mathbf{9.2000 \pm 1.1000}$

TimeNonDiag–triangular, LDMS also improves all three mean values. These comparisons use latent denoising together with the lower-triangular support prior specified in the setup.

Action-driven mechanisms. On ActionDiag, both methods use the same candidate mask support. Mean MCC and R_{con} increase from 0.746 to 0.768, and SHD decreases from 9.4 to 9.2. These results show higher mean factor alignment and lower mean graph error under the same candidate support. Appendix E.5 provides the results for the other action and temporal datasets.

Real hydraulic observations. Table 2 reports the validation-selected InTaSet models. LDMS$^{\triangle}$ lowers mean MSE from 0.1584 to 0.0676, RMSE from 0.3598 to 0.2319, and NRMSE from 0.6656 to 0.4391, corresponding to reductions of 57.3%, 35.5%, and 34.0%. Mean test negative ELBO decreases from 58.95 to 21.10, a 64.2% reduction. These results show improved tank-level reconstruction and a lower variational objective across all six channels. Across all ten examined training budgets, LDMS achieves lower means for all four metrics while remaining close to the common final sparsity target after the constraint schedule ends. Appendix F.3 provides the remaining budget results and curves.

Table 2: InTaSet results for validation-selected models (30k updates). Entries are means ± sample standard deviations over five training seeds. Lower values are better; boldface marks the lower mean. LDMS$^{\triangle}$ includes the lower-triangular support prior.

Metric	Baseline	LDMS$^{\triangle}$
MSE ↓	0.1584 ± 0.0424	$\mathbf{0.0676 \pm 0.0456}$
RMSE ↓	0.3598 ± 0.0584	$\mathbf{0.2319 \pm 0.0721}$
NRMSE ↓	0.6656 ± 0.1271	$\mathbf{0.4391 \pm 0.1453}$
Negative ELBO ↓	58.9500 ± 17.8700	$\mathbf{21.1000 \pm 13.7600}$

6 Limitations and Future Work

The denoising module estimates latent densities and scores with Gaussian kernels. Its accuracy depends on latent dimension, sample size, bandwidth selection, and the particle budget, which limit estimation quality and computational efficiency. The Gaussian corruption model is also a local approximation and may not fully describe state-dependent latent errors. Future work will study alternative density and score estimators, adaptive particle budgets, and integration with other causal representation learning models. Broader evaluation should include physical temporal benchmarks, different noise levels, and matched structural priors to isolate the contribution of denoising. Our theoretical guarantees require sufficiently accurate representations and mechanism estimates; they do not ensure that finite-step training reaches these conditions. Connecting the particle approximation and joint optimization to these error bounds remains an open direction.

AI USE STATEMENT

In preparing this paper, we used OpenAI's ChatGPT for language polishing, translation, rephrasing, and simplifying some technical descriptions. It was also used to assist with parts of the proof derivation and the generation and refinement of some figures. The LLM was used only as an auxiliary tool. All research ideas, methods, results, and final verification were completed and reviewed by the authors.

ETHICS STATEMENT

This study uses synthetic data and publicly available measurements from an interconnected four-tank system. The reported experiments concern simulated variables and physical sensor signals, rather than human participants or personal records. We describe the experimental settings and provide code to support transparent and reproducible research. Applications that use learned dependencies to guide decisions should include domain-specific validation and appropriate human oversight.

REPRODUCIBILITY STATEMENT

Notation and probability conventions appear in Appendix A; proofs and supplementary analyses are in Appendix B. Appendix C derives the variational objective, and Appendix D details the denoising method and training algorithm. Appendices E.1–E.4 provide the comparison protocol, data generation, evaluation metrics, and hyperparameters, respectively; Appendix E.5 reports additional results. Appendix F gives the InTaSet protocol, metric definitions, and additional training-budget results. Our code is included in the supplementary materials to improve the transparency and reproducibility of this study.

REFERENCES

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, 2023.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2008.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2013.

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 2022.

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*, 2024.

Juan L. Gamella, Simon Bing, and Jakob Runge. Sanity checking causal representation learning on a simple real-world system. In *International Conference on Machine Learning*, 2025a.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7:107–118, 2025b.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, 2nd revised edition, 1993.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvärinen. Disentangling identifiable features from noisy data with structured nonlinear ICA. In *Advances in Neural Information Processing Systems*, 2021.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, 2016.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. Denoising criterion for variational auto-encoding framework. In *AAAI Conference on Artificial Intelligence*, 2017.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29:1–17, 1998.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12:307–392, 2019.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan M. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E. 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*, 2022.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *Journal of Machine Learning Research*, 27:1–90, 2026.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

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*, 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*, 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*, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal representation learning from binary interactions. In *Conference on Uncertainty in Artificial Intelligence*, 2023a.

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*, 2023b.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

Hiroshi Morioka, Hermanni Hälvä, and Aapo Hyvärinen. Independent innovation analysis for nonlinear vector autoregressive process. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden (eds.), *Handbook of Econometrics*, volume 4, pp. 2111–2245. Elsevier, 1994.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, 2017.

Saskia A. Putri, Daniela Vilacres, Nazia Raza, Oluwabunmi Iwakin, and Faegheh Moazeni. InTaSet: A benchmark dataset for data-driven system identification and fault detection in an interconnected water system. In *World Environmental and Water Resources Congress*, 2026.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 2015.

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

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*, 2024.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, 2008.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, 2024.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In *Advances in Neural Information Processing Systems*, 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*, 2022b.

A NOTATION

Table 3: Principal notation. All adjacency matrices have response coordinates in rows and input coordinates in columns.

Symbol	Meaning
X^t, Z^t, A^{t-1}	Observation, latent state, and preceding auxiliary variable
p, d, q	Observation, latent, and auxiliary-input dimensions
$G_z^\star, G_a^\star$	True adjacency matrices of dimensions $d \times d$ and $d \times q$
$M_z, M_a; \Pi_z, \Pi_a$	Sampled/relaxed masks; edge-inclusion probabilities
ϕ, ϑ, ψ	Encoder, decoder, and transition parameters
U^t	Encoder posterior mean $\mu_\phi(X^t)$
$P = \{V_m\}_{m=1}^N$	Persistent particles representing the candidate distribution
ν, ρ, q_ρ	Noisy, candidate clean, and convolved candidate densities
$H; b_q, b_\rho, b_\nu$	Covariance in the corruption model; kernel variances
$\mathcal{F}_\nu, \mathrm{Ent}$	Population free energy and negative differential entropy
s_r, Σ_m, α	Score function, diagonal mobility, and particle step size
N, M, K_{iter}	Particle count, embedding-group size, and inner update count
Y_m	Detached noisy query generated from particle V_m
$\widehat{\mathcal{R}}_{\mathrm{LDMS}}$	Empirical denoising regularizer for the encoder
$\beta, \lambda_D, \lambda_H$	KL weight, LDMS weight, and entropy-score weight
$f_i, \widehat{f}_i, v_i$	True conditional mean, estimated mean, and transition variance
ϵ^t, η^t	Observation noise and exogenous transition noise
$\widetilde{Z}^t$	Aligned encoder representation after LDMS regularization
$W, \widetilde{W}; \mathsf{P}, \mathsf{Q}$	Clean and refined preceding inputs and their distributions
M^z, M^a	Latent and continuous-auxiliary strengths ($d \times d$, $d \times q$)
$b, M^b; e, M_e$	Input-type label and matrix; edge index and scalar strength
$\mathcal{I}_z, \mathcal{I}_a; i(e), \partial_e$	Candidate-edge sets; target coordinate and input derivative
$\widetilde{M}^z, \widetilde{M}^a$	Population mechanism strengths under the denoised input law
$\widehat{M}^z, \widehat{M}^a$	Evaluated strengths of the learned mechanisms
$\Delta_z, \mathfrak{D}$	Conditional input and joint-transition Wasserstein discrepancies
$L_z, L_a; \gamma, \tau$	Sensitivity constants; minimum true strength and threshold
$e_z, e_a, b_{\mathrm{ev}}; B_z, B_a$	Derivative fitting/evaluation errors; support-error bounds
$\mathcal{H}_i, d_{1,i}, \Omega_i$	Mechanism class, function/gradient metric, and continuity modulus
$\mathcal{R}_i, \widetilde{\mathcal{R}}_i$	Clean and denoised transition risks
$u_{i,n}, r_{i,n}, \varepsilon_{i,n}$	Sampling, penalty-oscillation, and optimization errors
$E_{i,n}; H_z, H_a$	Mechanism estimation error bound; fitted sensitivity regularity constants
B, m, J, δ	Clipping level, evaluation size, entry count, and failure probability
$G^{\mathrm{mask}}, \widehat{G}^\partial$	Thresholded mask graph and sensitivity graph
$\mathcal{C}_k, M_{ik}^{a,\Delta}$	Finite-contrast operator and categorical auxiliary strength

Notation and probabilistic conventions. For a random vector V, $\mathrm{Law}(V)$ denotes its probability law. For a probability measure μ and a measurable vector-valued function h, write $\|h\|_{L^2(\mu)} = (\int \|h(v)\|^2 \, d\mu(v))^{1/2}$; $\|\cdot\|$ is the Euclidean norm. An L^2 norm without a measure subscript is evaluated under the joint law of the random variables in its argument. Expectations and probabilities are denoted by $\mathbb{E}$ and $\mathbb{P}$. The notation $\mathcal{N}(m, C)$ denotes a Gaussian law with mean m and covariance C, I_r is the $r \times r$ identity matrix, and $\top$ denotes transpose. A subscript on $\mathbb{E}$ specifies the law used for integration. Let $\mathcal{P}_2(\mathbb{R}^r)$ be the probability measures with finite second moment on $\mathbb{R}^r$. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^r)$, the quadratic Wasserstein distance is

$$W_2(\mu, \nu) = \inf_{\pi \in \Gamma(\mu,\nu)} \left(\int \|x - y\|^2 \, d\pi(x, y) \right)^{1/2},$$

where $\Gamma(\mu, \nu)$ is the set of probability measures with marginals μ, ν, called couplings. The dimension r is generic in this definition.

The generative model is defined in Section 3.1; Table 3 summarizes the notation. Random vectors and their realizations use uppercase and lowercase letters, respectively. Deterministic matrices, probability laws, and sample sizes are defined separately. In particular, P denotes the particle system, P the clean input law, and P_O the observation law in the empirical-process argument. The empirical observation law is $\mathsf{P}_{O,n}$; $\mathbb{P}$ is reserved for probability.

B Proofs and Supplementary Analysis

The proofs below keep the auxiliary input A^{t-1} in every coupling. For the nonasymptotic statements, the estimated representation, mechanism, and coordinate alignment are fixed. For a training sample size n, let $\mathcal{T}_n$ be the sigma-field generated by the training observations and training randomness. The population quantities in the asymptotic analysis are conditional on $\mathcal{T}_n$ and evaluated on an independent transition observation.

Representation convention. Define $f = (f_1, \dots, f_d)^\top$ and $\eta^t = (\eta_1^t, \dots, \eta_d^t)^\top$, where η^t is the exogenous transition-noise vector. We also write $f_i(z, a)$ for the same function on the full input vector, with dependence restricted to its parent coordinates. Both parent sets refer to the preceding time. In the experimental Gaussian family, f_i is the transition mean and the exogenous noise variance is input independent. All results use a fixed latent parameterization satisfying (10), with coordinate alignment and scales held fixed. This convention specifies the representative in which derivatives and functional dependencies are compared; the graph-preserving equivalence relations of Lachapelle et al. (2026) provide the relevant identifiability framework. For $s \in \{t-1, t\}$, let $\widetilde{Z}^s$ denote the encoder representation after LDMS regularization, aligned in the fixed reference coordinates. No additive representation-noise model or observation-wise particle-refinement map is required: the proofs use the conditional input distributions and joint distributions of consecutive representations. The distributional statements apply to either posterior means or posterior samples, with one convention fixed throughout each application. The transition model and regularity assumptions hold in these reference coordinates. The discrepancies Δ_z and $\mathfrak{D}$ quantify achieved representation quality, and the recovery results specify sufficient bounds on these discrepancies and on mechanism-estimation errors.

B.1 Conditional Discrepancy and Entrywise Recovery

Lemma 2 (Well-defined distortion and paired control)**.** *Under Assumption 1, Δ_z is finite and*

$$W_2(\mathrm{Law}(\widetilde{Z}^{t-1}), \mathrm{Law}(Z^{t-1})) \le \Delta_z \le \left\| \widetilde{Z}^{t-1} - Z^{t-1} \right\|_{L^2}. \tag{29}$$

Entrywise support conditions. The proof of Proposition 2 also gives individual-edge guarantees. Write $L_b = L_z$ for $b = z$ and $L_b = L_a$ for $b = a$, and likewise for B_b. For an entry $e \in \mathcal{I}_b$, an absent edge is correctly excluded when $B_b < \tau$, and a true edge is correctly included when $B_b < M_e - \tau$. If fitting and evaluation errors are fixed, replacing $\Delta_z^{\mathrm{noisy}}$ by a smaller Δ_z^{LDMS} reduces B_b by $L_b(\Delta_z^{\mathrm{noisy}} - \Delta_z^{\mathrm{LDMS}})$. When fitting errors differ, the bound still improves if their increase is smaller than this reduction.

B.2 Full Consistency Conditions and Results

Let n be the training sample size and consider a sequence of joint LDMS estimators. Conditional on the training data, let Q_n be the law of the denoised preceding input for an independent observation, and let $\Delta_{z,n}, \mathfrak{D}_n$ be the corresponding conditional discrepancies. Population risks below are also conditional on the training data. For response coordinate i, let $\mathcal{H}_i$ be a class of candidate conditional means and $\widehat{f}_{i,n} \in \mathcal{H}_i$ the estimated mechanism. Write $\widehat{\mathcal{R}}_{i,n}$ for the empirical mean squared transition error, taking the posterior expectation when latent variables are sampled. The associated population risk is $\widetilde{\mathcal{R}}_{i,n}$. Convergence in probability, denoted by $\to_p$, is with respect to the training data and any estimation randomness.

Assumption 3 (Regularity and accuracy of mechanism estimation)**.** *For each i, the class $\mathcal{H}_i$ contains f_i and satisfies the following conditions.*

(i) The functions are continuously differentiable on an open rectangular domain $\mathcal{D} \subseteq \mathbb{R}^{d+q}$ in which the clean and denoised inputs take values almost surely. The measure P has a density positive almost everywhere on $\mathcal{D}$. The class is compact under the metric $d_{1,i}(g,h) = \|g-h\|_{L^2(\mathsf{P})} + \|\nabla g - \nabla h\|_{L^2(\mathsf{P})}$.

(ii) For a finite constant $K_i \geq 0$, every $g \in \mathcal{H}_i$ is K_i-Lipschitz in z uniformly in a, and its z- and a-gradients are Lipschitz in z with finite common constants $H_z, H_a \geq 0$. Set $C_i = \sup_{g\in\mathcal{H}_i} \sqrt{\mathcal{R}_i(g)} < \infty$.

(iii) The empirical criterion satisfies $u_{i,n} = \sup_{g\in\mathcal{H}_i} |\widehat{\mathcal{R}}_{i,n}(g) - \widetilde{\mathcal{R}}_{i,n}(g)| \to_p 0$. For a possibly data-dependent penalty $\mathcal{P}_{i,n} : \mathcal{H}_i \to \mathbb{R}$ of finite oscillation $r_{i,n} = \sup_{g\in\mathcal{H}_i} \mathcal{P}_{i,n}(g) - \inf_{g\in\mathcal{H}_i} \mathcal{P}_{i,n}(g)$, the estimator satisfies the approximate minimization inequality, with optimization error $\varepsilon_{i,n} \geq 0$,

$$\widehat{\mathcal{R}}_{i,n}(\widehat{f}_{i,n}) + \mathcal{P}_{i,n}(\widehat{f}_{i,n}) \leq \inf_{g\in\mathcal{H}_i} \{\widehat{\mathcal{R}}_{i,n}(g) + \mathcal{P}_{i,n}(g)\} + \varepsilon_{i,n}. \tag{30}$$

The full gradient $\nabla = (\nabla_z, \nabla_a)$ is used in the metric $d_{1,i}$; H_b denotes H_z for $b = z$ and H_a for $b = a$. The uniform-risk and approximate-minimization requirements are standard in M-estimation (van der Vaart, 1998, Ch. 5). The quantity $u_{i,n}$ measures the uniform deviation of empirical and population risks, and $r_{i,n}$ is the oscillation of the penalty over $\mathcal{H}_i$. A density that is positive almost everywhere ensures that continuous functions equal in $L^2(\mathsf{P})$ are equal everywhere on $\mathcal{D}$. Compactness in $d_{1,i}$ provides the regularity needed to infer convergence of derivatives from convergence of functions. This condition holds for a class with a compact parameter space when the functions and gradients are continuous in the parameters and have square-integrable envelopes. Appendix B.10 relates (30) to the Gaussian transition term in joint LDMS training and gives uniform-law conditions for adaptive encoders. Local regularity and uniform tail control provide an alternative to global derivative bounds, as detailed in Appendix B.5.

The finite-sample bounds are stated in Theorem 2 in the main text. The bound separates denoising discrepancy, empirical-process deviation, regularization, and optimization error. Reducing $\mathfrak{D}_n$ decreases both the linear and quadratic denoising terms in $E_{i,n}$. Statistical consistency requires an accurate joint temporal distribution, including both the preceding inputs and the current latent response. The nonasymptotic bound also applies when the denoising discrepancy is nonzero. The regularization condition permits hard sparsity constraints containing the true mechanism, or penalties whose oscillation vanishes with sample size.

The graph-recovery result is stated in Corollary 1 in the main text.

B.3 Proof of Lemma 2

Proof. We use the coupling characterization of W_2 (Santambrogio, 2015, Ch. 5). Finite second moments imply that the conditional latent laws have finite second moments for P_A-almost every a. The conditional law of the actual pair $(\widetilde{Z}^{t-1}, Z^{t-1})$, given $A^{t-1} = a$, is an admissible coupling, so

$$W_2^2(\mathrm{Law}(\widetilde{Z}^{t-1} \mid a), \mathrm{Law}(Z^{t-1} \mid a)) \leq \mathbb{E}[\left\|\widetilde{Z}^{t-1} - Z^{t-1}\right\|^2 \mid A^{t-1} = a].$$

Integration proves the upper bound and finiteness. For the lower bound, choose conditional optimal couplings measurably and integrate against P_A. Such selections exist for Euclidean probability laws; measurable selections of arbitrarily near-optimal couplings give the same conclusion by taking a limit. The resulting joint law has cost Δ_z^2 and the two specified latent marginals, so it is an admissible coupling of those marginals. The marginal optimal cost is at most its cost. Taking square roots proves (29).

If the auxiliary variable is absent, the conditional laws reduce to their marginals. If each latent law is independent of A^{t-1}, its conditional law is constant in a, so the integral is again the marginal optimal cost. Applying the same construction to complete windows also shows that (17) is finite. Projecting each window coupling onto its previous-state coordinates proves $\Delta_z \leq \mathfrak{D}$. □

B.4 Proof of Lemma 1

Proof. Let π be the joint law of coordinate variables $(z, \widetilde{z}, a)$ under a coupling that holds a fixed, with input marginals P, Q, and latent cost Δ_z^2; alternatively take a sequence with costs decreasing to this value. Both sensitivities can be viewed as random variables on this common probability space. The reverse triangle inequality in $L^2(\pi)$ and (14) give

$$\begin{aligned}|\widetilde{M}_{ij}^z - M_{ij}^z| &\le \left\|\partial_{z_j} f_i(\widetilde{z}, a) - \partial_{z_j} f_i(z, a)\right\|_{L^2(\pi)} \\ &\le \left\|\nabla_z f_i(\widetilde{z}, a) - \nabla_z f_i(z, a)\right\|_{L^2(\pi)} \le L_z \Delta_z.\end{aligned}$$

For the auxiliary edge, use (15) on the same coupling:

$$\begin{aligned}|\widetilde{M}_{ik}^a - M_{ik}^a| &\le \left\|\partial_{a_k} f_i(\widetilde{z}, a) - \partial_{a_k} f_i(z, a)\right\|_{L^2(\pi)} \\ &\le \left\|\nabla_a f_i(\widetilde{z}, a) - \nabla_a f_i(z, a)\right\|_{L^2(\pi)} \le L_a \Delta_z.\end{aligned}$$

Thus both bounds follow with the same auxiliary-preserving distortion. Square integrability from Assumption 1 justifies all norms and applications of the reverse triangle inequality. □

B.5 Local Regularity and Tail Control

The global regularity conditions in the main results can be replaced by local conditions with tail control. For $e \in \mathcal{I}$, put $h_e = \partial_e f_{i(e)}$. For a radius $R > 0$, suppose there is a finite constant $L_{e,R}$ such that

$$|h_e(z, a) - h_e(z', a)| \le L_{e,R} \left\|z - z'\right\|^{\alpha_e}, \qquad 0 < \alpha_e \le 1,$$

on $\|z\|, \|z'\|, \|a\| \le R$. For any auxiliary-preserving coupling π, let E_R^c denote the complement of E_R, where $E_R = \{\|z\|, \|\widetilde{z}\|, \|a\| \le R\}$, $d_\pi^2 = \mathbb{E}_\pi \left\|\widetilde{z} - z\right\|^2$, and $T_{e,R}(\pi) = \left\|[h_e(\widetilde{z}, a) - h_e(z, a)]\mathbf{1}_{E_R^c}\right\|_{L^2(\pi)}$. Then

$$|\widetilde{M}_e - M_e| \le L_{e,R} d_\pi^{\alpha_e} + T_{e,R}(\pi). \tag{31}$$

Indeed, split the derivative difference over E_R and its complement, apply the L^2 triangle inequality, and use $\mathbb{E}\left\|\widetilde{z} - z\right\|^{2\alpha_e} \le (\mathbb{E}\left\|\widetilde{z} - z\right\|^2)^{\alpha_e}$. For a fixed coupling, square integrability gives $T_{e,R}(\pi) \to 0$ by dominated convergence. Taking an optimal coupling sets $d_\pi = \Delta_z$; taking the actual sample coupling gives the paired version directly.

For a sequence of fitted models, uniform tail control and common local moduli give the corresponding convergence statement: first choose R so the tails are small, and then let the distortion tend to zero. For the learning class in Assumption 3, impose these local moduli and vanishing tails uniformly over $g \in \mathcal{H}_i$. They yield

$$\sup_{g \in \mathcal{H}_i} \left| \|\partial_e g\|_{L^2(\mathsf{Q}_n)} - \|\partial_e g\|_{L^2(\mathsf{P})} \right| \to_p 0,$$

which replaces $H_b \Delta_{z,n}$ in (27). This allows regularity to vary across the latent state space while retaining the same proof structure.

B.6 Proof of Proposition 1

Proof. Consider an auxiliary-preserving window coupling with clean coordinate variables (z^-, z^+, a) and denoised coordinate variables $(\widetilde{z}^-, \widetilde{z}^+, a)$, where the signs indicate the preceding and current times. The difference between the residuals satisfies

$$\begin{aligned}|[\widetilde{z}_i^+ - g(\widetilde{z}^-, a)] - [z_i^+ - g(z^-, a)]| &\le |\widetilde{z}_i^+ - z_i^+| + K \left\|\widetilde{z}^- - z^-\right\| \\ &\le \sqrt{1 + K^2} \sqrt{\left\|\widetilde{z}^+ - z^+\right\|^2 + \left\|\widetilde{z}^- - z^-\right\|^2}.\end{aligned}$$

The last step uses the Cauchy–Schwarz inequality in two dimensions. Taking L^2 norms, applying the reverse triangle inequality to the residuals, and minimizing the transport cost proves (18). For (19), use the actual data coupling and the identity

$$\widetilde{Z}_i^t - f_i(\widetilde{W}) - \eta_i^t = (\widetilde{Z}_i^t - Z_i^t) + f_i(W) - f_i(\widetilde{W}).$$

The L^2 triangle inequality and the Lipschitz bound on f_i complete the proof. □

B.7 Proof of Theorem 1

Proof. The argument separates zero entries from nonzero entries using a uniform perturbation bound, as in entrywise thresholding for sparse matrices (Bickel & Levina, 2008). Here Lemma 1 supplies that bound. For a true latent edge, $M^z_{ij} \geq \gamma$, and Lemma 1 gives

$$\widetilde{M}^z_{ij} \geq M^z_{ij} - L_z\Delta_z \geq \gamma - L_z\Delta_z > \tau.$$

Thus it is retained. For an absent latent edge, $M^z_{ij} = 0$, so $\widetilde{M}^z_{ij} \leq L_z\Delta_z < \tau$ and it is excluded. For a true auxiliary edge, applying the bound for its partial derivative gives

$$\widetilde{M}^a_{ik} \geq M^a_{ik} - L_a\Delta_z \geq \gamma - L_a\Delta_z > \tau.$$

For an absent auxiliary edge, $\widetilde{M}^a_{ik} \leq L_a\Delta_z < \tau$. Both adjacency matrices therefore equal the true adjacency matrices. If the true graph is empty, the sufficient conditions reduce to $L_z\Delta_z < \tau$ and $L_a\Delta_z < \tau$ for an arbitrary $\tau > 0$. □

Observable support and signal size. The graph defined by positive mechanism strengths describes dependencies on the support of the input distribution. To relate it to functional parents, suppose $f_i \in C^1$ on an open rectangular domain $\mathcal{D}$ and P has a density positive almost everywhere there. If f_i is constant in an input coordinate, the corresponding derivative vanishes. Conversely, if its squared derivative has zero expectation, that derivative is zero almost everywhere on $\mathcal{D}$. By continuity it is zero everywhere: a nonzero value would give a nonempty open neighborhood of nonzero values and hence positive probability. Along every coordinate segment in the rectangular domain, the fundamental theorem of calculus then shows that f_i is constant in that coordinate. Thus minimal functional parents coincide with positive strengths. The minimum over finitely many such edges is positive for each fixed problem; weak edges have a correspondingly smaller admissible perturbation.

B.8 Proof of Proposition 2

Proof. For a latent edge, add and subtract the fitted and true derivative norms under Q:

$$\begin{aligned}
|\widehat{M}^z_{ij} - M^z_{ij}| &\leq \left| \widehat{M}^z_{ij} - \left\| \partial_{z_j} \widehat{f}_i \right\|_{L^2(\mathsf{Q})} \right| \\
&\quad + \left| \left\| \partial_{z_j} \widehat{f}_i \right\|_{L^2(\mathsf{Q})} - \left\| \partial_{z_j} f_i \right\|_{L^2(\mathsf{Q})} \right| + |\widetilde{M}^z_{ij} - M^z_{ij}| \\
&\leq b_{\mathrm{ev}} + e_z + L_z\Delta_z.
\end{aligned}$$

The difference between the fitted and true derivative norms is bounded by the reverse triangle inequality. The auxiliary calculation uses ∂_{a_k}, e_a, and L_a and gives the second bound in (23). Applying the arguments for present and absent edges from Theorem 1 with B_z, B_a proves both the common-margin and edgewise conclusions. □

Reduction of the support-estimation error bound. For type $b \in \{z, a\}$, write L_b, e_b, B_b for the corresponding latent or auxiliary quantities. Subtracting the two estimation bounds gives

$$B_b^{\mathrm{LDMS}} - B_b^{\mathrm{noisy}} = -L_b(\Delta_z^{\mathrm{noisy}} - \Delta_z^{\mathrm{LDMS}}) + (e_b^{\mathrm{LDMS}} - e_b^{\mathrm{noisy}}) + (b_{\mathrm{ev}}^{\mathrm{LDMS}} - b_{\mathrm{ev}}^{\mathrm{noisy}}).$$

For equal evaluation error bounds, a reduction in $L_b\Delta_z$ that exceeds the increase in derivative estimation error gives a smaller support-estimation error bound. The sufficient recovery condition then holds for any dependency whose separation from the threshold exceeds the smaller LDMS bound. With a common fitting error this reduction is exactly $L_b(\Delta_z^{\mathrm{noisy}} - \Delta_z^{\mathrm{LDMS}})$.

B.9 Finite-Sample Evaluation of the Learned Strengths

Let $J = |\mathcal{I}| = d^2 + dq$; for a model with a single type of input, restrict $\mathcal{I}$ accordingly. Define $h_e = \partial_e f_{i(e)}$ and $\widehat{h}_e = \partial_e \widehat{f}_{i(e)}$. Let $\widetilde{W}_1, \ldots, \widetilde{W}_m$ be m evaluation inputs, independent of training and conditionally independent with common law Q. For a truncation level $B > 0$, define

$$C_B(x) = \max\{-B, \min\{x, B\}\}, \qquad \widehat{M}_{e,B} = \left(\frac{1}{m} \sum_{\ell=1}^{m} C_B(\widehat{h}_e(\widetilde{W}_\ell))^2 \right)^{1/2}. \tag{32}$$

The conditional population version is $\left\|C_B(\widehat{h}_e)\right\|_{L^2(\mathbb{Q})}$.

Lemma 3 (Uniform evaluation accuracy)**.** *For $B > 0$ and $0 < \delta < 1$, with conditional probability at least $1 - \delta$, all J entries satisfy*

$$\left|\widehat{M}_{e,B} - \left\|C_B(\widehat{h}_e)\right\|_{L^2(\mathbb{Q})}\right| \le b_{m,B,\delta} := 2B\sqrt{\frac{\log(2J/\delta)}{m}}. \tag{33}$$

Proof. For a fixed edge, let $X_\ell = C_B(\widehat{h}_e(\widetilde{W}_\ell))^2$, $\mu = \mathbb{E}X_\ell$, and $\bar{X} = m^{-1}\sum X_\ell$. Then $0 \le X_\ell \le B^2$ and $\mathrm{Var}(X_\ell) \le B^2\mu$. We use the exponential-moment proof of Bernstein's inequality (Boucheron et al., 2013, Ch. 2). For a centered variable Y with $|Y| \le c$ and variance at most v, the exponential series and $k! \ge 2 \cdot 3^{k-2}$ imply

$$\log \mathbb{E}e^{sY} \le \frac{vs^2}{2(1 - |s|c/3)}, \qquad |s| < 3/c.$$

Indeed, $|\mathbb{E}Y^k| \le c^{k-2}v$ bounds each term of the series, and $\log(1+x) \le x$ completes the bound. Independence and the exponential Markov inequality, with $s = u/(B^2\mu + B^2u/3)$, give

$$\mathbb{P}(|\bar{X} - \mu| \ge u) \le 2\exp\left(-\frac{mu^2}{2(B^2\mu + B^2u/3)}\right).$$

If $\mu = 0$, all X_ℓ vanish almost surely. Otherwise, for $v_0 > 0$, choosing $u = \sqrt{2B^2\mu v_0/m} + 2B^2v_0/(3m)$ gives an event of probability at least $1 - 2e^{-v_0}$ on which

$$|\bar{X} - \mu| \le \sqrt{2}ac + 2c^2/3, \quad a = \sqrt{\mu}, \quad c = B\sqrt{v_0/m}.$$

This implies $\sqrt{\bar{X}} \le a + 2c$. If $a \le 2c$, the lower bound $\sqrt{\bar{X}} \ge a - 2c$ follows from nonnegativity. If $a > 2c$, it follows from $a^2 - \sqrt{2}ac - 2c^2/3 \ge (a - 2c)^2$. Hence $|\sqrt{\bar{X}} - \sqrt{\mu}| \le 2c$. Use $v_0 = \log(2J/\delta)$ and a union bound. Independence between edges is unnecessary because the union bound uses only their individual tail bounds. □

For Proposition 2, use $b_{\mathrm{ev}} = b_{m,B,\delta}$ and replace each derivative fitting error by $\left\|C_B(\widehat{h}_e) - h_e\right\|_{L^2(\mathbb{Q})}$. Writing $(x)_+ = \max\{x, 0\}$, the nonexpansiveness of the truncation map C_B gives

$$\left\|C_B(\widehat{h}_e) - h_e\right\|_{L^2(\mathbb{Q})} \le \left\|\widehat{h}_e - h_e\right\|_{L^2(\mathbb{Q})} + \left\|(|h_e| - B)_+\right\|_{L^2(\mathbb{Q})}. \tag{34}$$

Alternatively, relative to the untruncated fitted strength, the evaluation error is at most

$$b_{\mathrm{ev}} = b_{m,B,\delta} + \max_e \left\|(|\widehat{h}_e| - B)_+\right\|_{L^2(\mathbb{Q})}. \tag{35}$$

These two decompositions make the clipping error explicit and allow unbounded, square-integrable sensitivities.

B.10 Relation to the Joint LDMS Objective

The LDMS transition objective uses both the encoded input and the encoded response. For a fixed encoder and evaluated deterministic mask, let S^-, S^+ be its reparameterized posterior draws at the two times, and let target i have Gaussian transition variance $v_i > 0$, independent of the inputs. Write $\mu^+_{\phi,i} = \mu_{\phi,i}(x^+)$ and $v^+_{\phi,i} = v_{\phi,i}(x^+)$ for the posterior mean and variance of the response coordinate. For a candidate mean $g \in \mathcal{H}_i$, the coordinatewise KL term, conditional on the observed pair and S^-, is

$$\frac{1}{2}\left[\log\frac{v_i}{v^+_{\phi,i}} + \frac{v^+_{\phi,i} + (\mu^+_{\phi,i} - g(S^-, a))^2}{v_i} - 1\right].$$

Because $\mathbb{E}[(S^+_i - g(S^-, a))^2 \mid x^-, x^+, a, S^-] = v^+_{\phi,i} + (\mu^+_{\phi,i} - g(S^-, a))^2$, the part depending on g is the posterior expected squared transition loss multiplied by $1/(2v_i)$. Reconstruction and the detached particle regularizer remain fixed during this conditional comparison. Therefore, comparing

a jointly fitted model with the same encoder, decoder, particle system, and variances but a different target mechanism gives (30), with the sparsity penalty and optimization error scaled by $2v_i$. Fixed positive normalization factors in (9) are handled by the same rescaling. Correctly specified hard support constraints define the feasible class $\mathcal{H}_i$. Equation (30) records the achieved comparison accuracy for the evaluated mechanism; its tolerance includes any additional error introduced when a trained stochastic mask is fixed for this evaluation.

This argument applies with $\widetilde{Z}$ chosen as the posterior draws used by the transition objective. The distributional perturbation bounds also apply to the encoder means used for reported evaluation, using the corresponding mean-based laws and fitting criterion. Each application retains a fixed representation convention. A posterior-mean fitting diagnostic or a conditional refit supplies the mean-based comparison in (30). Both conventions use the encoder learned with denoising regularization.

Adaptive encoders and the uniform law. The condition $u_{i,n} \to_p 0$ accounts for the data dependence of the fitted representation. One way to satisfy this condition is to fix the encoder and use independent windows to fit the mechanism. Another uses a uniform law over the joint encoder–mechanism family. To see the latter, let $O = (X^{t-1}, A^{t-1}, X^t)$ be an observed transition and $\ell_{\theta_D,\phi,g}(O)$ its posterior expected squared transition loss. Here θ_D denotes any additional denoising parameters used to define the representation, P_O is the law of O and $\mathsf{P}_{O,n} = n^{-1}\sum_{\ell=1}^{n}\delta_{O_\ell}$ is its empirical measure for independent observations $O_1, \ldots, O_n$; δ_o denotes unit mass at o. For a measure μ, $\mu\ell$ denotes $\int \ell\, d\mu$. If

$$\sup_{\theta_D,\phi,g} |(\mathsf{P}_{O,n} - \mathsf{P}_O)\ell_{\theta_D,\phi,g}| \to_p 0,$$

the same bound holds after substituting the data-dependent encoder parameters and taking the supremum over g. Any numerical integration or Monte Carlo error is included in $u_{i,n}$.

The following compactness and integrable-envelope argument is a standard uniform law of large numbers (Newey & McFadden, 1994, Lemma 2.4). For completeness, let Θ be a compact metric parameter space, with generic parameter θ and loss $\ell_\theta(O)$ continuous in θ almost surely. Suppose $|\ell_\theta(O)| \le F(O)$ for an envelope F with $\mathsf{P}_O F < \infty$. For $r > 0$, let $w_r(O)$ be the supremum of $|\ell_\theta(O) - \ell_{\theta'}(O)|$ over parameter pairs at distance at most r. Compactness gives $w_r \to 0$ almost surely, and the integrable envelope gives $\mathbb{E}w_r \to 0$. An r-net $\{\theta_1, \ldots, \theta_{N_r}\}$ of finite size N_r yields

$$\sup_{\theta\in\Theta} |(\mathsf{P}_{O,n} - \mathsf{P}_O)\ell_\theta| \le \max_{j\le N_r} |(\mathsf{P}_{O,n} - \mathsf{P}_O)\ell_{\theta_j}| + \mathsf{P}_{O,n}w_r + \mathsf{P}_O w_r.$$

By the strong law of large numbers, the finite maximum tends to zero and the remaining terms converge to $2\mathbb{E}w_r$. Letting $r \downarrow 0$ proves the uniform law. Growing model classes require a corresponding complexity bound. Similarly, continuity of a smooth parameterized mechanism and its gradient into $L^2(\mathsf{P})$ maps a compact parameter set to a compact class under $d_{1,i}$. Dominated convergence with square-integrable function and derivative envelopes is sufficient for that continuity.

B.11 Proof of Theorem 2

Proof. We first establish all inequalities conditional on the training data and randomness, so the refined laws and $\widehat{f}_{i,n}$ are fixed. Proposition 1 gives, uniformly for $g \in \mathcal{H}_i$,

$$|\sqrt{\widetilde{\mathcal{R}}_{i,n}(g)} - \sqrt{\mathcal{R}_i(g)}| \le c_{i,n}.$$

Since $\sqrt{\mathcal{R}_i(g)} \le C_i$, multiplication by the sum of the two square roots gives

$$|\widetilde{\mathcal{R}}_{i,n}(g) - \mathcal{R}_i(g)| \le 2C_i c_{i,n} + c_{i,n}^2 =: \chi_{i,n}. \tag{36}$$

Following the risk-comparison argument for approximate M-estimators (van der Vaart, 1998, Ch. 5), use $f_i \in \mathcal{H}_i$ as the reference function in (30). Inserting the uniform empirical and distortion bounds twice yields

$$\begin{aligned}\mathcal{R}_i(\widehat{f}_{i,n}) &\le \widehat{\mathcal{R}}_{i,n}(\widehat{f}_{i,n}) + u_{i,n} + \chi_{i,n}\\ &\le \widehat{\mathcal{R}}_{i,n}(f_i) + r_{i,n} + \varepsilon_{i,n} + u_{i,n} + \chi_{i,n}\\ &\le \mathcal{R}_i(f_i) + 2u_{i,n} + 2\chi_{i,n} + r_{i,n} + \varepsilon_{i,n}.\end{aligned}$$

Using $Z_i^t = f_i(W) + \eta_i^t$ and $\mathbb{E}[\eta_i^t \mid W] = 0$,

$$\mathcal{R}_i(g) - \mathcal{R}_i(f_i) = \mathbb{E}[(g(W) - f_i(W))^2].$$

The cross term vanishes by conditional expectation; its integrability follows from Cauchy–Schwarz. Substitution proves (25) with $E_{i,n}$ from (24).

For $s \geq 0$, define the deterministic modulus

$$\Omega_i(s) = \sup\{\|\nabla g - \nabla f_i\|_{L^2(\mathsf{P})} : g \in \mathcal{H}_i,\ \|g - f_i\|_{L^2(\mathsf{P})} \leq s\}.$$

It is finite by compactness and nondecreasing by set inclusion. We show $\Omega_i(s) \to 0$. Otherwise there are $s_k \downarrow 0$ and $g_k \in \mathcal{H}_i$ with vanishing function-estimation error and derivative error bounded below by a positive constant. A subsequence converges to $g_\infty \in \mathcal{H}_i$ in $d_{1,i}$. Hence $g_\infty = f_i$ holds P-almost surely. By continuity and positive density on the open domain, they are equal everywhere on that domain, so their gradients are equal there. Convergence in $d_{1,i}$ then forces the derivative error of this subsequence to zero, a contradiction. This establishes the stated modulus of continuity and (26).

For $e \in \mathcal{I}_b$, put $i = i(e)$. The uniform derivative Lipschitz constant H_b permits the coupling argument of Lemma 1 with the fitted function $\widehat{f}_{i,n}$ itself. Thus

$$\begin{aligned}\left|\left\|\partial_e \widehat{f}_{i,n}\right\|_{L^2(\mathsf{Q}_n)} - M_e\right| &\leq \left|\left\|\partial_e \widehat{f}_{i,n}\right\|_{L^2(\mathsf{Q}_n)} - \left\|\partial_e \widehat{f}_{i,n}\right\|_{L^2(\mathsf{P})}\right| \\ &\quad + \left\|\partial_e \widehat{f}_{i,n} - \partial_e f_i\right\|_{L^2(\mathsf{P})} \\ &\leq H_b \Delta_{z,n} + \Omega_i(\sqrt{E_{i,n}}).\end{aligned}$$

Taking the maximum over finitely many entries proves (27). The convergence assumptions imply $E_{i,n} \to_p 0$ and $\Delta_{z,n} \leq \mathfrak{D}_n \to_p 0$, so the bound converges to zero in probability. A finite union bound establishes joint convergence of the distributional and mechanism-estimation errors. □

Approximation error and regularization. The nonasymptotic bound can also use an approximating function $g_{i,n}^\star \in \mathcal{H}_i$ by adding $\left\|g_{i,n}^\star - f_i\right\|_{L^2(\mathsf{P})}^2$ to its right-hand side. Derivative consistency in an expanding class additionally requires a common compactness or equicontinuity argument. For a penalty $\mathcal{P}_{i,n} = \lambda_n \mathcal{S}_i$ where $\mathcal{S}_i$ is a bounded sparsity functional, $\lambda_n \geq 0$, and $K_{\mathcal{S}_i} = \sup_g \mathcal{S}_i(g) - \inf_g \mathcal{S}_i(g)$, the penalty oscillation satisfies $r_{i,n} \leq \lambda_n K_{\mathcal{S}_i}$. A vanishing penalty scale gives the stated consistency regime, while a fixed penalty contributes its finite term to (24). Hard sparsity constraints are handled directly by the feasible class. This decomposition quantifies the roles of denoising accuracy, mechanism regularity, and sparsity regularization in LDMS.

B.12 Proof of Corollary 1

Proof. By Theorem 2 and the evaluation assumption, the maximum absolute error of all estimated strengths tends to zero in probability, since both the evaluation error bound and its failure probability tend to zero. Let $m_\star = \min\{\tau, \gamma - \tau\} > 0$. If every error is smaller than $m_\star$, absent strengths remain below τ and true strengths remain above τ. Therefore

$$\mathbb{P}\big((\widehat{G}_{z,n}^\partial, \widehat{G}_{a,n}^\partial) \neq (G_z^\star, G_a^\star)\big) \leq \mathbb{P}\left(\max_e |\widehat{M}_{e,n} - M_e| \geq m_\star\right) \longrightarrow 0.$$

For population evaluation of the true functions, Lemma 1 makes this maximum at most $\max\{L_z, L_a\}\Delta_{z,n}$, proving the final statement. An empty true graph uses any positive threshold and the same absent-edge argument. □

A concrete consistent evaluation scheme. Choose independent evaluation sample sizes $m_n \to \infty$, failure probabilities $\delta_n \to 0$, and clipping levels $B_n \to \infty$ with $B_n\sqrt{\log(2J/\delta_n)/m_n} \to 0$. For example, $B_n = m_n^{1/4}$ and $\delta_n = m_n^{-1}$ satisfy this condition. Compactness under $d_{1,i}$ makes the derivative class compact in $L^2(\mathsf{P})$, hence its clipping tails vanish uniformly. To verify this last statement, cover the derivative class by finitely many L^2 balls of radius ϵ. The map $h \mapsto (|h| - B)_+$

is 1-Lipschitz in L^2, so its norm is at most ϵ plus the largest tail norm among the finitely many centers. Each center has vanishing tail by square integrability. Sending $B \to \infty$ and then $\epsilon \downarrow 0$ proves uniformity.

For the fitted derivative on Q_n, an auxiliary-preserving optimal coupling and the same 1-Lipschitz property give

$$\left\|(|\widehat{h}_{e,n}| - B_n)_+\right\|_{L^2(\mathsf{Q}_n)} \le H_b \Delta_{z,n} + \sup_{g \in \mathcal{H}_{i(e)}} \|(|\partial_e g| - B_n)_+\|_{L^2(\mathsf{P})} \to_p 0.$$

Thus (35) and Lemma 3 give $b_{\mathrm{ev},n} \to_p 0$, including the vanishing failure probability. This establishes the evaluation condition in the corollary under the stated regularity assumptions.

B.13 Connection to the Learned Hard Masks

Let $G^{\mathrm{mask}} = (G_z^{\mathrm{mask}}, G_a^{\mathrm{mask}})$ denote the graphs obtained by thresholding mask probabilities at $1/2$, after alignment that preserves the masked input structure. Re-evaluate the fitted means with these hard masks, and form the sensitivity graph $\widehat{G}^{\partial,\mathrm{hard}}$ at a positive threshold. Every input path passes through (4). Hence a zero mask replaces the input by a constant and makes its derivative zero, giving $\widehat{G}^{\partial,\mathrm{hard}} \subseteq G^{\mathrm{mask}}$. Identify each graph with its set of directed edges. For two such graphs G, H, define $\mathrm{SHD}(G,H) = |G \setminus H| + |H \setminus G|$, where $|\cdot|$ is cardinality and $\setminus$ is set difference. Then

$$\mathrm{SHD}(G^{\mathrm{mask}}, G^\star) \le \mathrm{SHD}(\widehat{G}^{\partial,\mathrm{hard}}, G^\star) + |G^{\mathrm{mask}} \setminus \widehat{G}^{\partial,\mathrm{hard}}|. \tag{37}$$

Indeed, $|G^\star \setminus G^{\mathrm{mask}}| \le |G^\star \setminus \widehat{G}^{\partial,\mathrm{hard}}|$ and $|G^{\mathrm{mask}} \setminus G^\star| \le |\widehat{G}^{\partial,\mathrm{hard}} \setminus G^\star| + |G^{\mathrm{mask}} \setminus \widehat{G}^{\partial,\mathrm{hard}}|$. Adding these inequalities proves the bound. Evaluating the estimation-error terms on the hard-masked means makes Proposition 2 applicable. Exact mask recovery follows when the hard-masked sensitivity graph is correct and every selected entry passes its threshold. This is the additional support-selection condition needed to connect mechanism recovery to the mask graphs evaluated in the experiments.

B.14 Auxiliary-Only Tasks and Discrete Inputs

An auxiliary-only example is a transition from an observed input to a latent response, $Z_i^t = f_i(A^{t-1}) + \eta_i^t$. Latent denoising leaves the input distribution unchanged, so $\Delta_z = 0$ and the input derivative norms are unchanged. The response component of $\mathfrak{D}$ and the learned-mechanism error remain in the fitting bound. Thus denoising the latent response can improve estimation of auxiliary mechanisms through the response component of the transition risk.

For a categorical auxiliary coordinate k, let $\mathcal{A}$ be the auxiliary state space and let $\mathcal{K}_k(da' \mid a)$ be a fixed probability kernel on admissible alternative inputs $a' \in \mathcal{A}$ satisfying $a'_\ell = a_\ell$ for $\ell \neq k$. The same kernel is used for clean and denoised inputs. Define the augmented laws

$$\mathsf{P}_k^\Delta(dz, da, da') = \mathsf{P}(dz, da)\mathcal{K}_k(da' \mid a), \qquad \mathsf{Q}_k^\Delta(dz, da, da') = \mathsf{Q}(dz, da)\mathcal{K}_k(da' \mid a).$$

Thus the alternative input a' is conditionally independent of the latent input given a. The finite-contrast operator and its strengths are

$$\mathcal{C}_k g(z, a, a') = g(z, a') - g(z, a), \qquad M_{ik}^{a,\Delta} = \|\mathcal{C}_k f_i\|_{L^2(\mathsf{P}_k^\Delta)}, \quad \widetilde{M}_{ik}^{a,\Delta} = \|\mathcal{C}_k f_i\|_{L^2(\mathsf{Q}_k^\Delta)}.$$

The superscript Δ labels finite contrasts. Assume these contrasts are square integrable. If the two evaluations of f_i are uniformly Lipschitz in z with constants $K_{ik}^{(1)}$ and $K_{ik}^{(0)}$, respectively, set $L_{ik}^\Delta = K_{ik}^{(1)} + K_{ik}^{(0)}$. Then

$$|\widetilde{M}_{ik}^{a,\Delta} - M_{ik}^{a,\Delta}| \le L_{ik}^\Delta \Delta_z.$$

Indeed, extend an auxiliary-preserving coupling of $(z, \widetilde{z}, a)$ by drawing the same a' from $\mathcal{K}_k(\cdot \mid a)$. This leaves its latent transport cost unchanged. The contrast difference is at most $L_{ik}^\Delta \|\widetilde{z} - z\|$, so the reverse L^2 triangle inequality and minimization over couplings prove the bound.

For a fitted mechanism $\widehat{f}_i$, define the contrast estimation error $e_{ik}^\Delta = \left\|\mathcal{C}_k \widehat{f}_i - \mathcal{C}_k f_i\right\|_{L^2(\mathsf{Q}_k^\Delta)}$. If $\widehat{M}_{ik}^{a,\Delta}$ estimates $\left\|\mathcal{C}_k \widehat{f}_i\right\|_{L^2(\mathsf{Q}_k^\Delta)}$ with absolute error at most b_{ev}^Δ, the same triangle inequality gives

$$|\widehat{M}_{ik}^{a,\Delta} - M_{ik}^{a,\Delta}| \le L_{ik}^\Delta \Delta_z + e_{ik}^\Delta + b_{\mathrm{ev}}^\Delta.$$

Thresholding therefore recovers the support defined by positive contrast strengths whenever this bound is smaller than both the threshold and the separation of every positive strength from that threshold. Vanishing denoising, contrast estimation, and evaluation errors give the corresponding fixed-graph recovery limit. This statement uses estimation accuracy under the specified augmented laws. Identification of a categorical parent on the full support additionally requires the contrast kernel to assign positive probability to the relevant category comparisons, with the mechanism defined at both evaluated inputs.

The latent-input derivative $\partial_{z_j} f_i(z, a)$ remains well defined for categorical a whenever $f_i(\cdot, a)$ is differentiable. Its stability bound follows by the same coupling argument under the latent-gradient regularity condition, uniformly in a. For mixed auxiliary inputs, continuous coordinates use derivatives and categorical coordinates use the finite contrasts above.

C Variational Objective and Normalization

For the two-step generative density $p_0(z^-)p_\psi(z^+ \mid z^-, a)p_\vartheta(x^- \mid z^-)p_\vartheta(x^+ \mid z^+)$ and variational law q_-q_+, let $p(x^-, x^+ \mid a)$ be the corresponding marginal observation density. Jensen's inequality yields

$$\log p(x^-, x^+ \mid a) \geq \mathcal{B} = \sum_{s\in\{-,+\}} \mathbb{E}_{q_s} \log p_\vartheta(x^s \mid Z^s) - \mathrm{KL}(q_-\|p_0) \\ - \mathbb{E}_{q_-} \mathrm{KL}(q_+\|p_\psi(\cdot \mid Z^-, a)). \tag{38}$$

Here $\mathcal{B}$ denotes the ELBO, and the factorization assumes the initial prior p_0 is independent of a. Dividing $-\mathcal{B}$ by $2p$ gives the temporal experimental objective at $\beta = 1$. Equation (8) specifies the reconstruction-plus-transition formulation. Reconstruction expectations are estimated with reparameterized samples, while the diagonal-Gaussian KL is evaluated analytically conditional on the sampled previous state. For scalar Gaussian distributions with means $\mu_q, \mu_p \in \mathbb{R}$ and variances $v_q, v_p > 0$,

$$\mathrm{KL}(\mathcal{N}(\mu_q, v_q)\|\mathcal{N}(\mu_p, v_p)) = \frac{1}{2}\left[\log\frac{v_p}{v_q} + \frac{v_q + (\mu_q - \mu_p)^2}{v_p} - 1\right]. \tag{39}$$

The transition term sums this expression over coordinates and averages over samples of the previous state. The joint objective augments this variational criterion with LDMS regularization and sparsity. Generative evaluation reports the normalized negative ELBO, computed from reconstruction and KL terms.

D LDMS: Denoising Principle and Training Algorithm

D.1 From Latent Corruption to Distributional Descent

This subsection derives the denoising principle introduced in Section 3.2. The following subsections describe the particle approximation and its use in encoder training. Write $h_\phi = \mu_\phi$ and $X_0^t = g^\star(Z^t)$, so that $X^t = X_0^t + \epsilon^t$. If the encoder Jacobian J_{h_ϕ} is locally Lipschitz, Taylor expansion gives

$$U^t = h_\phi(X_0^t) + J_{h_\phi}(X_0^t)\epsilon^t + r^t, \qquad \|r^t\| \leq \tfrac{1}{2}L_h\|\epsilon^t\|^2, \tag{40}$$

where L_h bounds the Jacobian's Lipschitz constant along the perturbation segment. Thus even independent observation noise generally induces state-dependent latent noise. Treating the Jacobian as locally constant motivates an additive-noise approximation; using κ_H further approximates that noise by a Gaussian kernel. This is a working model, not an exact consequence of nonlinear encoding. The candidate ρ models clean embeddings in the learned coordinates, without assuming they already equal the true factors. The denoising loss does not require paired clean observations.

First variation and transport direction. Hold ν, H, and λ_H fixed. Assume smooth positive densities, finite energy, and sufficient integrability and decay to justify differentiation, interchanging integrals, and integration by parts. For a mass-preserving perturbation η with $\int \eta = 0$, differentiating (5) and using the symmetry of κ_H gives $D\mathcal{F}_\nu(\rho)[\eta] = \int \chi_\rho \eta$, where, up to an additive constant,

$$\chi_\rho = \kappa_H * \log(q_\rho/\nu) + \lambda_H \log\rho, \qquad \nabla\chi_\rho = \kappa_H * (s_{q_\rho} - s_\nu) + \lambda_H s_\rho, \tag{41}$$

and $s_r = \nabla \log r$ denotes a density score. The first term accounts for the smoothed score difference, and the second arises from negative entropy. For a smooth positive-definite mobility $\Sigma(v)$, choose the transport velocity $w_s = -\Sigma \nabla \chi_{\rho_s}$. The continuity equation and vanishing boundary flux yield

$$\partial_s \rho_s = \nabla \cdot (\rho_s \Sigma \nabla \chi_{\rho_s}), \qquad \frac{d}{ds} \mathcal{F}_\nu(\rho_s) = -\int \rho_s (\nabla \chi_{\rho_s})^\top \Sigma \nabla \chi_{\rho_s} \, dv \le 0. \tag{42}$$

Here s is optimization time, distinct from the observation index t. This is the mobility-weighted Wasserstein descent of the fixed population objective, using the standard continuity-equation framework (Jordan et al., 1998; Ambrosio et al., 2008). Transporting particles with w_s and taking an Euler step motivates (44).

From the population flow to LDMS. The implementation replaces the densities with kernel density estimates (KDEs), approximates convolution with a finite kernel operator, and estimates a diagonal mobility matrix from the particles. The population identity (42) motivates these particle updates. The encoder surrogate below uses the refined particles during alternating training, while the conditional discrepancies in Section 4 quantify the representation quality sufficient for mechanism recovery.

D.2 Particle Approximation and Encoder Regularization

LDMS approximates the candidate latent distribution by a particle system $P = \{V_m\}_{m=1}^N$, where $V_m \in \mathbb{R}^d$ and N is the number of particles. The particles are initialized by perturbing encoder means with Gaussian noise and are retained across minibatches. During particle optimization, the encoder means remain fixed. For M current encoder means $\mathcal{U} = \{U_r\}_{r=1}^M$, define the kernel density estimates (KDEs)

$$\begin{aligned} \widehat{\rho}_P(v) &= \frac{1}{N} \sum_m \varphi_{b_\rho}(v - V_m), \qquad \widehat{\nu}_\phi(v) = \frac{1}{M} \sum_r \varphi_{b_\nu}(v - U_r), \\ \widehat{q}_P(v) &= \frac{1}{N} \sum_m \varphi_{b_q + b_\rho}(v - V_m). \end{aligned} \tag{43}$$

where φ_b is the centered Gaussian density with covariance bI_d, I_d is the $d \times d$ identity matrix, and $v \in \mathbb{R}^d$ is the density argument. The positive parameters b_ρ, b_ν, b_q are, respectively, the particle-density, encoder-density, and convolution variances. Here the working corruption covariance is $H = b_q I_d$. Sums over m and r range over $1, \dots, N$ and $1, \dots, M$. Adaptive bandwidth selection is specified in Appendix D.4.

For a differentiable positive density r, let $s_r = \nabla \log r$ denote its score function. With step size $\alpha > 0$ and positive diagonal mobility matrix $\Sigma_m \in \mathbb{R}^{d \times d}$, the particle update is

$$V_m \leftarrow V_m - \alpha \Sigma_m \left[\mathsf{T}_{b_q, P} s_{\widehat{q}_P}(V_m) - \mathsf{T}_{b_q, \mathcal{U}} s_{\widehat{\nu}_\phi}(V_m) + \lambda_H s_{\widehat{\rho}_P}(V_m) \right]. \tag{44}$$

Here $\mathsf{T}_{b, \mathcal{C}}$ is the empirical kernel operator with variance b and centers $\mathcal{C}$, defined in (47); Σ_m is specified in (48). Particle updates and bandwidth selection treat the encoder states as constants, so no gradients pass through them. The particles approximate a latent distribution. The encoder outputs remain associated with their observations and preserve the temporal pairing needed to estimate mechanisms.

After the update, LDMS draws $\zeta_m \sim \mathcal{N}(0, I_d)$ and sets $Y_m = V_m + \sqrt{b_q} \zeta_m$. The loss for each group is

$$\widehat{\mathcal{R}}_{\mathrm{LDMS}} = \frac{1}{N} \sum_m \left[\log \widehat{q}_P(Y_m) - \log \widehat{\nu}_\phi(Y_m) \right] - \frac{\lambda_H}{N} \sum_m \log \widehat{\rho}_{P, -m}(V_m), \tag{45}$$

Here Y_m is a noisy particle evaluation point, the ζ_m are independent standard Gaussian vectors, and $\widehat{\rho}_{P,-m}(v) = (N-1)^{-1} \sum_{j \neq m} \varphi_{b_\rho}(v - V_j)$ is the leave-one-out KDE, with $N \ge 2$. During the encoder update, P, Y_m, and all bandwidths are held fixed. Thus only the $-\log \widehat{\nu}_\phi$ term contributes a direct encoder gradient. This density-estimation step regularizes the same encoder that supplies the transition model. Section 4.1 quantifies how changes in the encoded inputs and responses affect mechanism estimation.

The query mixture in (45) has component variance b_q, whereas $\widehat{q}_P$ has variance $b_q + b_\rho$. Equation (45) defines the training surrogate. Its first and last terms are constant with respect to encoder parameters, so the cross-entropy term supplies the encoder gradient. Particle refinement uses the positive entropy-score coefficient in (44), consistent with (41). The particle trajectory is detached during encoder updates.

The experiments pool means from both times into a global particle cloud. For discrete contexts, separate particle clouds may be used, with losses averaged equally across groups. Continuous auxiliary inputs condition the transition model, not the global KDE. The encoder retains paired states and auxiliary inputs for conditional mechanism learning.

D.3 Kernel Operator and Mobility

The Gaussian KDE score for centers $\mathcal{C} = \{v_j\}_{j=1}^{J_c}$, with J_c centers $v_j \in \mathbb{R}^d$, evaluation point $u \in \mathbb{R}^d$, and variance $b > 0$, is

$$s_{b,\mathcal{C}}(u) = \frac{\sum_j w_j(u) v_j - u}{b}, \qquad w_j(u) = \frac{e^{-\|u-v_j\|^2/(2b)}}{\sum_\ell e^{-\|u-v_\ell\|^2/(2b)}}. \tag{46}$$

For a vector-valued function $h : \mathbb{R}^d \to \mathbb{R}^d$, the finite smoothing operator is

$$\mathsf{T}_{b,\mathcal{C}} h(u) = \frac{1}{J_c} \sum_{j=1}^{J_c} \exp\left[-\frac{\|u - v_j\|^2 - \min_\ell \|u - v_\ell\|^2}{2b} \right] h(v_j). \tag{47}$$

This operator evaluates the function at the kernel centers, rescales the weights using the nearest center, and divides the weighted sum by the number of centers. These choices specify the finite kernel approximation used in (44); its accuracy contributes to the quality of the refined latent distribution analyzed in Section 4.

For query V_m, let w_{mj} be the normalized Gaussian weights over the particle centers with variance b_q. Set

$$\bar{v}_m = \sum_j w_{mj} V_j, \qquad (\Sigma_m)_{kk} = \max\left\{ \sum_j w_{mj} (V_j)_k^2 - (\bar{v}_m)_k^2, \epsilon_{\mathrm{mob}} \right\}. \tag{48}$$

Here $\bar{v}_m$ is the weighted local particle mean, $(V_j)_k$ is the kth coordinate of particle V_j, and the off-diagonal entries of Σ_m are zero. Numerical denominators are stabilized with $\epsilon_{\mathrm{num}} = 10^{-12}$ and $\epsilon_{\mathrm{mob}} = 10^{-12}$. The positive floor stabilizes each coordinate of the local mobility.

D.4 Adaptive Bandwidths and Particle Initialization

The bandwidths b_q, b_ρ, b_ν are adapted to the latent scale. At each selection step, the particle-density and encoder-density kernel variances are chosen by maximizing the leave-one-out log-likelihood; b_q is selected by the negative log-density of observed embeddings under particle centers with effective variance $b_q + b_\rho$. The reference grid uses

$$b_0 = \left(\frac{4}{d+2} \right)^{2/(d+4)} s^2 N_{\mathrm{bw}}^{-2/(d+4)}, \tag{49}$$

where N_{bw} is the number of points used in the bandwidth estimate and s^2 is their mean coordinate variance. The search evaluates 25 logarithmically spaced candidates from $b_0/4$ to $4b_0$. The convolution search uses its own pooled-scale reference minus b_ρ, clipped to a positive floor. For each group, the updated value is a weighted average of the previous value and the new estimate, with weights 0.8 and 0.2. Additional scale floors prevent extremely small variances. These adaptive smoothing parameters track changes in the latent scale during training.

A cloud is initialized from current encoder means with isotropic Gaussian perturbations of standard deviation 0.01 and persists across minibatches. The experiments use a global cloud with $N = 512$ particles; the complete numerical settings appear in Table 5. For discrete window-level contexts, one may instead maintain a cloud for each group containing at least eight embeddings and average the eligible group losses uniformly. Continuous auxiliary inputs enter the transition model, and the reported experiments pool their encoder means in the global particle system.

D.5 Algorithm

Let K_{iter} be the number of particle updates per minibatch. Algorithm 1 summarizes joint LDMS estimation. Stage A refines the persistent particle distribution and forms the denoising signal; stages B and C integrate that signal with sparse mechanism learning and update the shared model.

Algorithm 1: LDMS: joint latent denoising and sparse mechanism learning

Input: Noisy observation windows, transition inputs, and training settings
Output: Encoder, decoder, transition model, and sparse masks

```
Initialize model parameters; leave the persistent particle system empty
for each training minibatch do
    Encode observations to obtain posterior parameters and means U^t
    A. Latent-space denoising
    Pool means globally, or within eligible discrete context groups
    for each group do
        Initialize its particle cloud from the encoder means if needed
        Detach the means and update the kernel bandwidths
        for k = 1, ..., K_iter do
            Compute density scores and mobility; update particles by (44)
        Retain the cloud and sample detached noisy queries Y_m
        Form the denoising loss R̂_LDMS using (45)
    Average the denoising losses over eligible groups
    B. Sparse conditional-mechanism learning
    Sample latent states and masks; compute L_VAE and sparsity terms
    C. Joint loss and optimization
    Combine the VAE, denoising, and sparsity terms using (9)
    Update model parameters; update multipliers if sparsity is constrained
```

For evaluation, use the encoder means and learned masks. The denoising loss propagates gradients through the encoder density while holding the particle system fixed. Factor-based coordinate alignment is used only when reference factors are available for evaluation.

D.6 Numerical Approximation and Computational Cost

LDMS uses a population density objective, an empirical particle update, and a surrogate loss for encoder training with fixed particles. The kernel operator in (47) and query law in (45) determine the numerical approximation. The denoising and estimation error bounds in Section 4.3 summarize how representation quality and achieved optimization accuracy affect mechanism learning. Finite-step accuracy also depends on the integration step size (Hairer et al., 1993). With M embeddings and N particles, a dense particle update has a computational cost of order $d(M^2 + N^2 + MN)$, with bandwidth searches adding repeated kernel evaluations. Holding the particles fixed during encoder updates avoids differentiating through their trajectories. The particle system is retained across minibatches. Runtime comparisons should accompany the stated update and particle budgets.

E Additional Experimental Results

This appendix supplements Section 5 with the comparison protocol, data-generating mechanisms, metric definitions, and hyperparameters for all nine synthetic datasets. It then reports and discusses results for the five datasets not shown in the main text.

E.1 Implementation and Comparison Protocol

The baseline implements mechanism-sparsity learning without the denoising regularizer, whereas LDMS adds the particle-based encoder regularizer described in Section 3.3. This comparison tests whether latent denoising improves representation and graph recovery under a common model and

training budget. The additional support restriction in the three LDMS$^{\triangle}$ comparisons is discussed in Appendix E.5.

The baseline and LDMS use the same encoder, decoder, and conditional-mechanism architectures (Table 5). The posterior is a diagonal Gaussian, the decoder has learned isotropic variance, and the conditional mechanisms have learned input-independent variances. The latent dimension is set to its true value. LDMS additionally uses the particle regularizer and, in the three marked temporal datasets, a lower-triangular support prior.

Temporal runs use the negative two-step ELBO divided by $2p$, including the initial-prior term, with $\beta = 1$. Action runs use the single-observation conditional negative ELBO divided by p. LDMS pools all available encoder means in a minibatch into one global particle system. Continuous actions still enter the conditional mechanism network, but the particle regularizer is not conditioned on them. Its parameters and adaptive bandwidths are specified in Table 5 and Appendix D.4. Evaluation uses the trained encoder means directly, without a separate particle-refinement pass for each test example.

Sampling and repeated trials. For each dataset, the graph and generating mechanism are held fixed across five trials. Each trial independently generates 800,000 training, 100,000 validation, and 100,000 test examples and trains both methods from new initializations. Actions and innovations are sampled independently across the three splits. The resulting variability reflects both dataset resampling and training randomness under a fixed mechanism.

E.2 Data-Generating Mechanisms and Graphs

We adopt the action and temporal families of Lachapelle et al. (2026) and additionally consider a sparse temporal chain. All datasets have $d = 10$ latent factors and $p = 20$ observed coordinates. The observation model uses an injective neural mixing map with three hidden layers of width 20, following the nonlinear construction in their Appendix C.1. The coordinates produced by the mixing map are standardized separately within each split before independent observation noise $\epsilon^t \sim \mathcal{N}(0, 10^{-4} I_p)$ is added. Hence 0.01 is the observation-noise standard deviation after normalization. Innovation noise is $\eta^t \sim \mathcal{N}(0, 10^{-4} I_d)$, independent of the preceding inputs, with independent coordinates.

For temporal data, draw $Z^- \sim \mathcal{N}(0, I_d)$ and set $Z^+ = f(Z^-) + \eta^t$, where $Z^- = Z^{t-1}$ and $Z^+ = Z^t$. Each example is an independently sampled two-step window, not a segment of one long trajectory. TimeDiag uses $f_i(z) = z_i + \frac{1}{2}\sin z_i$. The other temporal datasets use

$$f_i(z) = z_i + \frac{1}{2}\sum_{j=1}^{d} (G_z^\star)_{ij} \sin(\omega_i z_j + \delta_i), \qquad \omega_i = \frac{i+2}{\pi}, \quad \delta_i = i - 1. \tag{50}$$

Indices are one-based. Every temporal graph contains its diagonal, so the residual term is consistent with the declared support. These self-lags connect successive times, rather than a variable to itself within one time slice. The conditional-mean network is learned from the masked preceding inputs.

For action data, draw $A^{t-1} \sim \mathrm{Unif}([-2, 2]^q)$ and set $Z^t = f(A^{t-1}) + \eta^t$. ActionDiag uses $q = d$ and $f_i(a) = \sin a_i$. The remaining action datasets use

$$f_i(a) = \sum_{k=1}^{q} (G_a^\star)_{ik} \sin(\omega_i a_k + \delta_i). \tag{51}$$

These auxiliary-only examples test the input–response structure $A^{t-1} \to Z^t$ without a preceding latent input. The temporal and action mean functions follow the corresponding non-diagonal constructions in Lachapelle et al. (2026); graph changes select the terms present in each mechanism.

Rows of every adjacency matrix index targets and columns index sources, as in (3). TimeDiag has $G_z^\star = I_d$. TimeNonDiag–chain has $G_z^\star = I_d + \mathrm{diag}_{-1}(\mathbf{1}_{d-1})$, giving 10 self-lags and nine edges $j \to j + 1$. Here diag_{-1} places a vector on the first subdiagonal. TimeNonDiag–triangular is the complete lower triangle, with 55 edges; it corresponds to TimeNonDiag in the reference paper, while the chain is an additional dataset. TimeBlockDiag comprises five 2×2 all-ones diagonal blocks. TimeBlockNonDiag adds sources $9, 10$ to targets $1, 2$, and sources $1, 2, 5, 6$ to targets $9, 10$, for 32 edges in total.

Table 4: Reference graphs and sparsity budgets, shared across all five trials per method. The non-chain graphs follow Lachapelle et al. (2026). A temporal matrix is $G_z^\star$; an auxiliary matrix is $G_a^\star$. Both methods use the same numerical budget for each dataset. The last column indicates whether a lower-triangular support restriction is applied to LDMS only.

Dataset	Shape	True edges	Budget	LDMS triangular
TimeDiag	10×10	10	10	Yes
TimeNonDiag–chain	10×10	19	20	Yes
TimeNonDiag–triangular	10×10	55	55	Yes
TimeBlockDiag	10×10	20	20	No
TimeBlockNonDiag	10×10	32	30	No
ActionDiag	10×10	10	20	No
ActionNonDiag	10×10	20	20	No
ActionBlockDiag	10×5	10	10	No
ActionBlockNonDiag	10×5	12	10	No

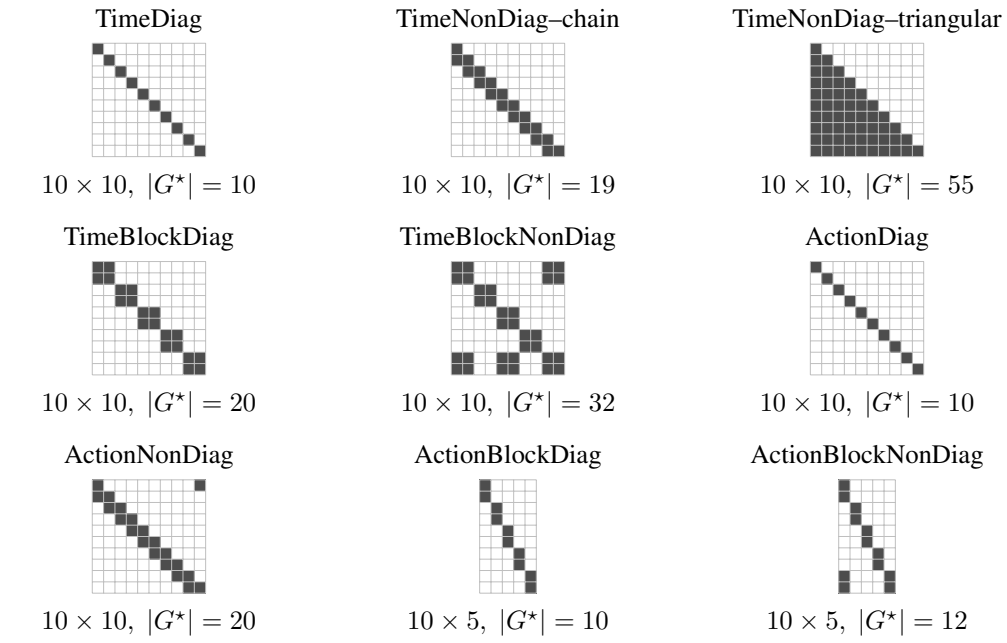


Figure 2: True adjacency matrices for the synthetic datasets. Rows run from target 1 at the top to target 10 at the bottom; columns run from source 1 on the left to source d (temporal) or q (action) on the right. Dark cells indicate edges. The non-chain graphs reproduce the synthetic constructions of Lachapelle et al. (2026); the chain defines an additional sparse temporal dataset.

ActionDiag has $G_a^\star = I_d$. ActionNonDiag connects each target to the action with the same index and the preceding action in cyclic order (target 1 also receives action 10), giving 20 edges. In ActionBlockDiag, action k drives targets $2k-1, 2k$. ActionBlockNonDiag additionally connects action 1 to targets 9 and 10. The two block-action graphs have $q = 5$ and contain 10 and 12 edges, respectively. These block constructions are those used to study partial disentanglement in Lachapelle et al. (2026).

E.3 Evaluation Metrics and Aggregation

MCC assesses coordinate alignment, R_{con} assesses recovery within graph-preserving predictor sets, and SHD assesses the learned graph support. We additionally report unrestricted linear recovery R in the supplementary comparisons. The definitions below specify how each score is computed before averaging across trials.

Let $\{(z_n, u_n)\}_{n=1}^{n_{\mathrm{ev}}}$ be true factors and encoder means from the evaluation sample. All factor coordinates are scalar. Define a matching σ and its permutation matrix P_σ by

$$\sigma \in \arg\max_{\sigma' \in \mathfrak{S}_d} \sum_{i=1}^{d} \left|\mathrm{Corr}(Z_i, U_{\sigma'(i)})\right|, \qquad (P_\sigma)_{i,\sigma(i)} = 1. \tag{52}$$

Here Z_i and U_j denote coordinates in the evaluation sample, with the time index omitted; z_{ni} is coordinate i of sample z_n. The number of pairs is n_{ev}, $\mathfrak{S}_d$ is the set of permutations of $\{1, \ldots, d\}$, and the remaining entries of P_σ are zero. The operator Corr denotes sample Pearson correlation, and the matching is found using the Hungarian algorithm. The mean correlation coefficient is

$$\mathrm{MCC} = \frac{1}{d} \sum_{i=1}^{d} \left|\mathrm{Corr}(Z_i, U_{\sigma(i)})\right|. \tag{53}$$

The absolute correlations account for sign changes, and the matching accounts for coordinate permutations. MCC measures recovery in these matched scalar coordinates.

For a predictor index set $S \subseteq \{1, \ldots, d\}$, let

$$\mathcal{A}_i(S) = \left[1 - \frac{\min_{b,w} \sum_{n=1}^{n_{\mathrm{ev}}} (z_{ni} - b - w^\top u_{n,S})^2}{\sum_{n=1}^{n_{\mathrm{ev}}} (z_{ni} - \bar{z}_i)^2}\right]^{1/2}, \qquad R = \frac{1}{d} \sum_{i=1}^{d} \mathcal{A}_i(\{1, \ldots, d\}). \tag{54}$$

Here $b \in \mathbb{R}$ and $w \in \mathbb{R}^{|S|}$ are a regression intercept and coefficient vector, $u_{n,S}$ is the subvector indexed by S, and $\bar{z}_i = n_{\mathrm{ev}}^{-1} \sum_n z_{ni}$ is the sample mean. The symbol $\top$ denotes matrix or vector transpose. Each regression is fitted and scored on the same evaluation sample, and the unrestricted regression first standardizes each coordinate. Thus R is the mean square root of the in-sample, per-factor coefficient of determination; all evaluated factors have positive sample variance.

Following the scalar-factor evaluation protocol of Lachapelle et al. (2026), form the graph-preserving predictor pattern

$$C_z = \left[\mathbf{1} - G_z^\star(\mathbf{1} - (G_z^\star)^\top)\right]_+ \odot \left[\mathbf{1} - (G_z^\star)^\top(\mathbf{1} - G_z^\star)\right]_+, \tag{55}$$

$$C_a = \left[\mathbf{1} - G_a^\star(\mathbf{1} - (G_a^\star)^\top)\right]_+. \tag{56}$$

Here $[\cdot]_+$ is elementwise truncation at zero, $\mathbf{1}$ denotes an all-ones matrix of the required shape, and matrix products are ordinary products. Use $C = C_z$ for temporal-only experiments and $C = C_a$ for auxiliary-only experiments; when both types of dependencies are present, use the elementwise product of the two patterns. Set $S_i = \{k : (CP_\sigma)_{ik} = 1\}$. The graph-constrained recovery score is

$$R_{\mathrm{con}} = \frac{1}{d} \sum_i \mathcal{A}_i(S_i). \tag{57}$$

When $C = I_d$, this score equals MCC up to numerical precision. For block datasets, C permits additional predictors, so MCC and R_{con} measure different aspects of recovery.

For graph evaluation, let $M_z^{\mathrm{hard}} = \mathbf{1}\{\Pi_z > 0.5\}$ and $M_a^{\mathrm{hard}} = \mathbf{1}\{\Pi_a > 0.5\}$. Align them as $\widehat{G}_z = P_\sigma M_z^{\mathrm{hard}} P_\sigma^\top$ and $\widehat{G}_a = P_\sigma M_a^{\mathrm{hard}}$. These are the aligned mask graphs $G_z^{\mathrm{mask}}, G_a^{\mathrm{mask}}$ in

Appendix B.13. For either aligned binary graph $\widehat{G}$ and its true graph $G^\star$,

$$\mathrm{FN} = \sum_{ij} G^\star_{ij}(1 - \widehat{G}_{ij}), \qquad \mathrm{FP} = \sum_{ij}(1 - G^\star_{ij})\widehat{G}_{ij}, \qquad \mathrm{SHD} = \mathrm{FN} + \mathrm{FP}. \tag{58}$$

Here FN and FP are the numbers of false negative and false positive edges; SHD is their sum. Temporal self-lags are included. Reversing one directed relationship incurs two matrix-entry errors. Graph matching uses the coordinate permutation selected by MCC. SHD therefore evaluates support in those matched scalar coordinates.

Representation evaluation uses 100,000 factor vectors: all action test examples, or the first 100,000 vectors after flattening both times of temporal test windows. The matching and regressions are fitted separately for each trial. True factors are used only for evaluation. The graph metric counts errors in the MCC-aligned mask graph; it is not a direct evaluation of the derivative-strength graph in Section 4.2. Unlike the normalized SHD in Lachapelle et al. (2026), equation (58) is not divided by the number of candidate edges.

For each metric v, compute its final value v_s for trial $s \in \{0, 1, 2, 3, 4\}$ and report

$$\bar{v} = \frac{1}{5}\sum_{s=0}^{4} v_s, \qquad \mathrm{SD}(v) = \left[\frac{1}{4}\sum_{s=0}^{4}(v_s - \bar{v})^2\right]^{1/2}. \tag{59}$$

The standard deviation describes variation across resampled datasets and training initializations for one fixed graph and mechanism. It is not a standard error or confidence interval. We average trial-level metrics, rather than pooling latent arrays or thresholding an averaged graph. No seed or intermediate checkpoint is selected by its test score.

E.4 Training Settings and Hyperparameters

Table 5 summarizes the common architecture and optimization settings, together with the additional particle parameters used by LDMS. Dataset-specific sparsity budgets are listed in Table 4. These settings are fixed across the five trials of each comparison.

Table 5: Common model and optimization settings. Dataset-specific sparsity budgets are listed in Table 4; the scope of the comparisons is discussed in Appendix E.5.

Setting	**Value**
Encoder / decoder	6 hidden layers, width 512; Leaky-ReLU
Transition network	5 hidden layers, width 512 per factor
Posterior / transition variance	Diagonal Gaussian / input-independent
Decoder variance	Learned isotropic Gaussian
Batch size / evaluation batch size	1024 / 1024
Adam learning rate / dual SGD rate	5×10^{-4} / 5×10^{-4}
KL coefficient	$\beta = 1$
Total updates / end of budget schedule	20,000 / 10,000 updates
Trials per dataset and method	Five, with independent data sampling and initialization
Particle count / inner updates	$N = 512$ / $K_{\mathrm{iter}} = 1$
Particle step / entropy-score weight	$\alpha = 0.0100$ / $\lambda_H = 0.1000$
LDMS weight	$\lambda_D = 0.0500$ for LDMS; 0 for baseline
Initial kernel variances	$b_q = b_\rho = b_\nu = 0.5000$
Bandwidths / grouping	Adaptive / global, all available times
Graph threshold	$(\Pi_z)_{ij} > 0.5000$ or $(\Pi_a)_{ik} > 0.5000$

The temporal or auxiliary mask is learned using the constrained formulation in Section 3.4. The sparsity schedule reaches its final budget after 10,000 updates, and every trial completes 20,000 updates. The constraints bound sums of edge probabilities rather than the number of thresholded edges. Model parameters are optimized with Adam at a constant learning rate, while nonnegative dual variables enforce the sparsity constraints.

E.5 Additional Results

Table 6 reports the five datasets not included in the main text. Together with Table 1, it covers all nine datasets without repeating any comparison. Each entry reports the mean and sample standard deviation over five trials, evaluated at the end of the common training budget. The additional score R measures factor recovery by unrestricted linear regression.

Table 6: Results for the five datasets not included in Table 1. Entries show the mean $\pm$ sample standard deviation over five trials per method. Boldface indicates the better mean.

Dataset	Method	MCC ↑	R_{con} ↑	R ↑	SHD ↓
TimeBlockDiag	Baseline	$\mathbf{0.6460 \pm 0.0650}$	$\mathbf{0.7270 \pm 0.0450}$	$\mathbf{0.9480 \pm 0.0004}$	13.4000 ± 2.1000
	LDMS	0.6190 ± 0.0280	0.6980 ± 0.0460	0.9449 ± 0.0005	$\mathbf{11.8000 \pm 1.5000}$
TimeBlock NonDiag	Baseline	$\mathbf{0.6460 \pm 0.0270}$	$\mathbf{0.7180 \pm 0.0350}$	$\mathbf{0.9490 \pm 0.0010}$	22.6000 ± 0.5000
	LDMS	0.6190 ± 0.0360	0.6980 ± 0.0210	0.9456 ± 0.0006	$\mathbf{22.0000 \pm 0.0000}$
ActionNonDiag	Baseline	$\mathbf{0.7490 \pm 0.0290}$	$\mathbf{0.7490 \pm 0.0290}$	$\mathbf{0.9481 \pm 0.0005}$	$\mathbf{13.0000 \pm 0.7000}$
	LDMS	0.7320 ± 0.0670	0.7320 ± 0.0670	0.9418 ± 0.0005	13.4000 ± 2.3000
ActionBlockDiag	Baseline	$\mathbf{0.7440 \pm 0.0470}$	$\mathbf{0.8890 \pm 0.0300}$	$\mathbf{0.9468 \pm 0.0020}$	7.0000 ± 0.7000
	LDMS	0.7240 ± 0.0360	0.8760 ± 0.0260	0.9417 ± 0.0020	$\mathbf{6.4000 \pm 0.5000}$
ActionBlock NonDiag	Baseline	$\mathbf{0.7000 \pm 0.0330}$	$\mathbf{0.8710 \pm 0.0380}$	$\mathbf{0.9427 \pm 0.0028}$	$\mathbf{6.4000 \pm 1.1000}$
	LDMS	0.6900 ± 0.0540	0.8530 ± 0.0230	0.9345 ± 0.0027	7.6000 ± 0.9000

E.5.1 Temporal Datasets

On TimeBlockDiag, LDMS reduces mean SHD from 13.4 to 11.8; on TimeBlockNonDiag, it reduces SHD from 22.6 to 22.0. Table 6 reports the accompanying MCC, R_{con}, and unrestricted recovery score R, providing a complete comparison of factor alignment and graph recovery for these block structures.

E.5.2 Action Datasets

On ActionBlockDiag, LDMS reduces mean SHD from 7.0 to 6.4. The table also includes ActionNonDiag and ActionBlockNonDiag, completing the evaluation across the four action-graph families. Reporting all metrics for each dataset makes the representation and structural comparisons explicit.

Structural settings and aggregation. The three temporal comparisons marked $\triangle$ use a lower-triangular candidate support for LDMS and an unrestricted candidate support for the baseline. Both methods share the numerical sparsity budgets in Table 4. These budgets constrain sums of edge probabilities; the graph metric uses thresholded masks. All means summarize five trials with fixed generating mechanisms and resampled observations. Sample standard deviations describe variation across trials under the stated training budget and observation-noise level.

F Additional Results on InTaSet

This appendix describes the real-data protocol in Section 5 and reports results across all examined training budgets.

F.1 Dataset and Training Protocol

InTaSet comprises measurements from a physical interconnected four-tank platform operating in open-loop and closed-loop configurations (Putri et al., 2026). We use the closed-loop normal-operation sequence in InTaSet-FD, with 11,971 records. Each observation contains four tank-level signals and two pump-related signals. Time stamps determine the observation order. The model uses all six signal channels.

The sequence is divided chronologically into approximately 70%, 15%, and 15% training, validation, and test blocks. Channel means and standard deviations are estimated from the training records

and applied to all three splits. Adjacent-record windows of length $T = 2$ and stride one are then formed within each split, yielding 8,378 training, 1,794 validation, and 1,796 test windows. This construction preserves temporal order and keeps every window within its assigned split.

Both methods use six latent coordinates, multilayer perceptron encoders and decoders with hidden width 64, and first-order latent transitions. Training and evaluation batch sizes are 128 and 512, respectively, and the KL weight is $\beta = 1$. The graph constraint bounds the sum of edge-inclusion probabilities. Its upper bound is tightened linearly to eight over the first 15,000 updates and then held fixed. The baseline uses an unrestricted candidate graph. LDMS$^{\triangle}$ adds the particle-based encoder regularizer and a lower-triangular candidate support.

For each method, five seeds $s \in \{0, 1, 2, 3, 4\}$ control initialization and stochastic training at ten budgets, from 10k to 100k updates in steps of 10k. All runs share the preprocessed data and data-loader seed. Evaluation covers every validation and test window. For each method and seed, validation negative ELBO selects the final model from one budget; all selections use 30k updates. All four main-table metrics use these selected models. The supplementary comparisons evaluate the final model at each training budget.

F.2 Metrics and Statistical Aggregation

Let $N_{\text{test}} = 1796$, $T = 2$, and $p = 6$. For tank channel $j \in \{1, 2, 3, 4\}$, let $x_{n\ell j}$ be the standardized observation at position ℓ in test window n. Under seed s, deterministic reconstruction is

$$\widehat{x}^{(s)}_{n\ell} = g_{\vartheta_s}(\mu_{\phi_s}(x_{n\ell})), \qquad \text{MSE}_{s,j} = \frac{1}{N_{\text{test}} T} \sum_{n,\ell} (\widehat{x}^{(s)}_{n\ell j} - x_{n\ell j})^2. \tag{60}$$

Set $\text{RMSE}_{s,j} = \sqrt{\text{MSE}_{s,j}}$. With σ_j^{test} denoting the population standard deviation of channel j across the test windows, define

$$\text{NRMSE}_{s,j} = \frac{\text{RMSE}_{s,j}}{a_j}, \qquad a_j = \begin{cases} \sigma_j^{\text{test}}, & \sigma_j^{\text{test}} > 10^{-12}, \\ 1, & \text{otherwise.} \end{cases} \tag{61}$$

MSE and RMSE use the standardized observation scale. NRMSE measures error relative to the variation of each tank signal. Each reconstruction score is first averaged across the four tank channels. In particular, the reported RMSE averages the four channel-level square roots.

The test negative ELBO evaluates the six-channel variational model. With the seed index suppressed, its normalized estimate is

$$\widehat{\mathcal{L}}_{\text{test}} = \frac{1}{N_{\text{test}}} \sum_n (\widehat{\mathcal{R}}_n + \widehat{\mathcal{K}}_n), \qquad \widehat{\mathcal{R}}_n = -\frac{1}{Tp} \sum_{\ell=1}^{T} \log p_\vartheta(x_{n\ell} \mid z_{n\ell}), \tag{62}$$

where $z_{n\ell} \sim q_\phi(\cdot \mid x_{n\ell})$. For $T = 2$, the prior contribution is

$$\widehat{\mathcal{K}}_n = \frac{1}{2p} \left[\text{KL}(q_\phi(\cdot \mid x_{n1}) \| p_0) + \text{KL}(q_\phi(\cdot \mid x_{n2}) \| p_\psi(\cdot \mid z_{n1}, M_n)) \right]. \tag{63}$$

Here p_0 is the initial latent prior and M_n is the sampled graph mask. The Gaussian reconstruction density includes all six observed channels, and each KL term sums over the six latent coordinates. The decoder has a learned shared standard deviation $\exp(\ell_{\text{dec}}/2) + 10^{-4}$. Evaluation samples latent states and graph masks, and computes the diagonal-Gaussian KL terms analytically. The criterion combines observation fit with posterior–prior matching through the sparse transition model.

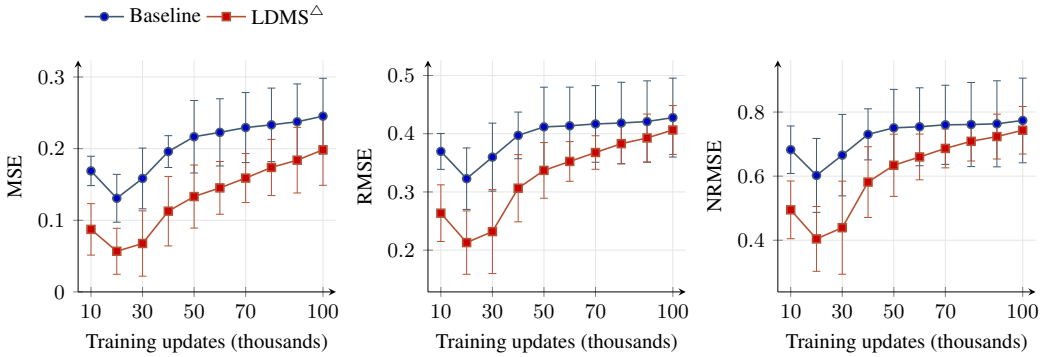
Metrics are summarized by their mean and sample standard deviation over five seeds (variance denominator $5 - 1$). The standard deviation measures training variability on the fixed data split. Figure points summarize separate budget-specific runs; error bars show one sample standard deviation.

F.3 Results across Training Budgets

The comparisons across training budgets complement the 30k results in the main text. LDMS has lower mean MSE, RMSE, NRMSE, and negative ELBO at every examined budget. At 20k updates, mean MSE decreases from 0.1307 to 0.0567 and mean RMSE from 0.3228 to 0.2130. At 100k,

LDMS retains reductions of 19.1% in MSE, 5.0% in RMSE, and approximately 4.0% in NRMSE. Its RMSE standard deviation at that endpoint is also lower, 0.0420 compared with 0.0678.

Figure 3 shows the reconstruction comparisons, and Figure 4 shows the test variational objective. The lowest mean reconstruction errors occur at 20k, while both methods attain their lowest mean test negative ELBO at 30k. At 100k, negative ELBO remains lower for LDMS (119.71 versus 230.69). LDMS therefore has lower mean errors across all examined training budgets.



(a) MSE (b) RMSE (c) NRMSE

Figure 3: InTaSet reconstruction across training budgets: (a) MSE, (b) RMSE, and (c) NRMSE. Points show means and error bars show sample standard deviations over five training seeds. Metrics average the four tank channels. LDMS$^{\triangle}$ includes the lower-triangular support prior.

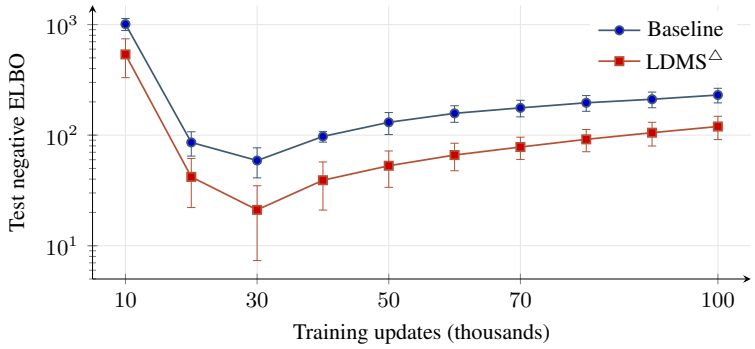


Figure 4: Test negative ELBO on all six InTaSet channels. Points show means, and error bars span one sample standard deviation on either side of each mean; the vertical axis is logarithmic. Each point summarizes five budget-specific training runs.

Variational-objective decomposition. At 30k updates, the mean reconstruction negative log-likelihood decreases from 57.2871 for the baseline to 19.3522 for LDMS; the corresponding KL terms are 1.6665 and 1.7495. The reconstruction term accounts for the reduction in the total objective. At 100k, the reconstruction terms are 228.9974 and 118.0681, with KL terms 1.6954 and 1.6448. These values connect the lower variational objective with improved observation fit.

F.4 Sparsity-Constraint Behavior

Figure 5 reports the expected-edge constraint and its Lagrange multiplier. At 10k, while the sparsity schedule is still tightening, the mean constraint values are 26.6548 for the baseline and 16.0489 for LDMS. The schedule reaches its final target of eight at 15k updates. Across the 20k–100k endpoints,

baseline means range from 7.9997 to 8.0148, and LDMS means range from 7.9987 to 8.0104. The largest absolute deviation from eight is 0.0148, or approximately 0.19% of the target.

The mean Lagrange multipliers decrease from 37.9682 and 17.1368 at 10k to small values at subsequent endpoints. From 20k onward, the baseline values are approximately 8.5×10^{-5}–1.7×10^{-3}, and the LDMS values are 3.3×10^{-4}–2.2×10^{-3}. Together, the constraint values and test metrics show that LDMS improves observation modeling while keeping the expected-edge count close to the prescribed target.

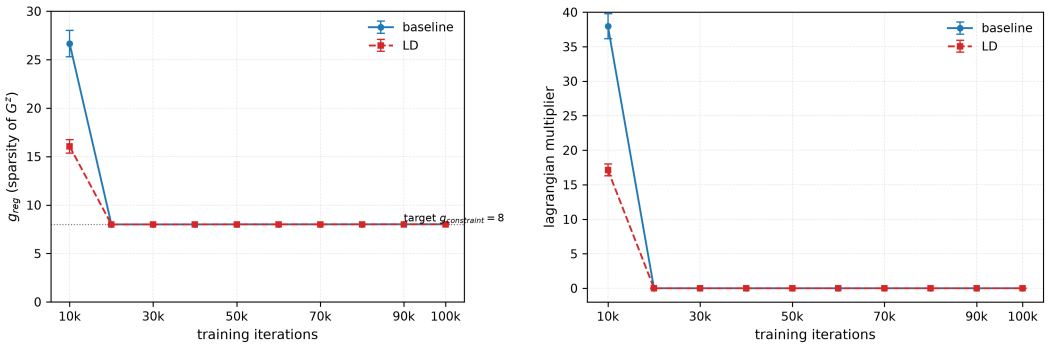


Figure 5: Expected-edge constraint (left) and Lagrange multiplier (right) across InTaSet training budgets. The original plot label LD denotes LDMS$^{\triangle}$. Points and error bars summarize five seeds. The dashed line at eight marks the final sparsity target, reached by the schedule at 15k updates.