Disentangling Continuous-Time Latent Dynamics: Identifiability of Latent SDEs via Diffusion Shifts

Yuanyuan Wang[1], **Wenjie Wang**[2], **Haoxuan Li**[3], **Mingming Gong**[1,2]*, **Kun Zhang**[1,4]*
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]The University of Melbourne [3]Peking University [4]Carnegie Mellon University
yuanyuan.wang@mbzuai.ac.ae, wenjiew3@student.unimelb.edu.au
hxli@stu.pku.edu.cn, mingming.gong@unimelb.edu.au, kunz1@cmu.edu

Abstract

Causal representation learning for time series has developed strong identifiability results in discrete-time latent causal models, but identifiability in continuous-time latent stochastic differential equation (SDE) models remains largely open. We address this gap using environment-induced shifts in diffusion covariance. We study additive-noise latent SDEs observed through an unknown nonlinear diffeomorphism, with shared drift but environment-specific diffusion covariance. We show that two diagonal diffusion regimes with pairwise distinct coordinate-wise variance ratios identify the latent coordinates up to permutation, coordinate-wise scaling, and a possible constant shift, without any sparsity assumption on the drift. We first prove this result for linear Ornstein–Uhlenbeck systems and then extend it to general additive-noise latent SDEs. Under mild smoothness, the instantaneous drift-Jacobian causal graph is identifiable up to the same permutation. We propose a two-stage estimator for latent disentanglement and optional graph recovery; experiments on synthetic systems confirm the predicted identifiability boundary, and an application to Hardanger Bridge monitoring data illustrates the approach on real sensor trajectories.

1 Introduction

Causal representation learning (CRL) seeks to recover latent causal variables and their structural relations from entangled observations [36]. This problem is especially compelling for time series, where causal structure underlies explanation, forecasting, and decision-making in domains such as climate science, biology, healthcare, and economics [34, 37]. In many temporal settings, however, the measured variables are not the causal variables of interest: video frames, sensor streams, and multivariate readouts are often nonlinear mixtures of latent dynamical factors. Without additional structure, those factors are not identifiable from observations alone [14, 25]. Temporal CRL therefore asks how temporal dependence, environmental variation, or interventions can break nonlinear mixing ambiguities and make latent causal variables identifiable.

Motivation. Many scientific and engineered systems evolve in continuous time, are sampled irregularly, and are intrinsically stochastic. Stochastic differential equations (SDEs) provide a standard language for such dynamics [30, 35]. Continuous-time latent and controlled dynamics are also widely used in machine learning through neural ODEs, latent ODEs, neural controlled differential equations, and neural SDEs [4, 17, 19, 33]. Continuous time also exposes objects that are blurred by a discrete-time view, including drift, diffusion, quadratic variation, and instantaneous dependence. This distinction is important for representation learning: the drift describes the deterministic law of motion, whereas the diffusion describes the scale and orientation of unresolved stochastic fluctuations.

*Co-corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In many latent dynamical systems, it is natural for the deterministic law of motion to remain stable across environments while the stochastic forcing changes. A bridge may have the same structural dynamics while wind and traffic loads change the amplitude and direction of stochastic excitation; a cardiovascular system may retain the same regulatory feedback loops while exercise or stress changes physiological variability; and a power grid may retain the same network dynamics while stochastic load fluctuations vary across demand or weather regimes. Such regime variation is often easier to observe than direct interventions on latent causal mechanisms. Such variation also provides a form of side information specific to continuous time: diffusion leaves a local quadratic-variation signature in the observed path, even though this signature is distorted by the unknown nonlinear mixing. This leads to the central identifiability question of this paper: when can diffusion shifts alone recover the latent coordinate system of a continuous-time SDE model from nonlinear observations?

Challenges. Existing identifiable nonlinear independent component analysis (ICA) and temporal CRL results show that side information can break nonlinear mixing ambiguities, but their identifiability theory is largely built for discrete-time settings, using finite-lag transitions, autoregressive mechanisms, or temporal context on a grid [15, 16, 38, 53, 54]. Conversely, work on continuous-time SDE identification and causal discovery typically assumes that the stochastic state itself is observed, rather than seen only through an unknown nonlinear observation map [9, 26, 43, 44, 56]. The latent continuous-time setting is harder: the state is observed only after an unknown nonlinear mixing, and the target is not merely a predictive state but the latent coordinate system and its instantaneous drift-Jacobian graph. What is missing is a theorem-level CRL account for continuous-time latent SDE state coordinates observed only through such a mixing. The central difficulty is that diffusion covariance is not invariant under nonlinear coordinate changes: in observed coordinates, local quadratic variation is the latent covariance transformed by the Jacobian of the unknown mixing. A single diffusion regime therefore cannot determine which latent axes are the causal coordinates. Our key observation is that two diagonal diffusion regimes with distinct coordinate-wise variance ratios impose simultaneous constraints on the same unknown Jacobian. These constraints reduce the remaining ambiguity to a monomial affine transformation (permutation, coordinate-wise scaling, and a possible constant shift), which in turn makes instantaneous drift-graph recovery well defined.

Contributions. We study additive-noise latent SDEs observed through an unknown nonlinear diffeomorphism across two environments that share the latent drift but differ in diffusion covariance. Our contributions are: (1) We formulate this continuous-time CRL problem and identify diffusion shifts as the source of side information: two diagonal regimes with coordinate-wise distinct variance ratios are sufficient to anchor the latent coordinate system. (2) We prove identifiability first for linear Ornstein–Uhlenbeck systems and then for general additive-noise latent SDEs: the latent coordinate system is recovered up to a monomial affine transformation without any sparsity assumption on the drift, and under mild smoothness the instantaneous drift-Jacobian causal graph is identifiable up to the same permutation. (3) We turn the theory into a practical two-stage estimator: diffusion-based disentanglement followed, when graph recovery is desired, by sparse drift regression. Synthetic experiments validate the predicted identifiability boundary, and a Hardanger Bridge monitoring study illustrates the estimator on real sensor trajectories.

2 Problem setup and target

We consider two environments, indexed by $e \in \{1, 2\}$, over a time horizon $T > 0$. In environment e, the latent state $z_t^{(e)} \in \mathbb{R}^d$ evolves according to the additive-noise SDE

$$dz_t^{(e)} = f(z_t^{(e)})\, dt + G_e\, dW_t^{(e)}, \qquad x_t^{(e)} = g(z_t^{(e)}), \tag{1}$$

where $x_t^{(e)} \in \mathbb{R}^d$ is the observed process, $W_t^{(e)}$ is a standard d-dimensional Brownian motion, $f : \mathbb{R}^d \to \mathbb{R}^d$ is the latent drift, $G_e \in \mathbb{R}^{d\times d}$ is the environment-specific diffusion matrix, and $g : \mathbb{R}^d \to \mathbb{R}^d$ is an unknown nonlinear observation map. The two environments share the same latent drift f and differ only through their diffusion covariances. In the linear Ornstein–Uhlenbeck special case studied first in Section 3.1, the drift takes the form $f(z) = Az$ for some matrix $A \in \mathbb{R}^{d\times d}$. Throughout, we write $\Sigma_e := G_e G_e^\top$ for the latent diffusion covariance in environment e.

The latent dynamics are autonomous: f has no explicit time dependence, so the causal mechanism is not explicitly time varying. In continuous-time causal models for SDEs, a parent relation is typically

induced when one variable enters another variable's drift or diffusion coefficient [26, 27, 43]. In our additive-noise setting, however, the diffusion matrices G_e are constant in z, so no variable enters another variable's diffusion coefficient. We therefore represent the latent causal structure by the drift alone. In the linear OU case, this reduces to the support pattern of A, which is constant over time. In the general nonlinear case, we work with the instantaneous drift-Jacobian graph at state z, namely the support pattern of $Df(z)$; this graph may depend on the current state, but not explicitly on time.

Our goal is to characterize conditions under which the pair of observed path laws from the two environments is sufficient to recover the latent coordinate system up to permutation, coordinate-wise scaling, and a possible constant shift, and to identify the corresponding latent causal structure up to the same permutation. The precise identifiability statements are given in Section 3.

3 Identifiability theory

3.1 Linear OU systems

We begin with the linear-drift case, where the identification mechanism can be seen most transparently. This case is a conceptual starting point rather than a restrictive assumption. In the OU setting, the diffusion comparison already contains the core disentanglement mechanism, and the state-space coverage step is automatic from the Gaussian transition law. This lets us isolate the role of the two regimes with minimal additional technical overhead.

Assumptions. We state the shared assumptions for two observationally equivalent latent representations: the reference parameterization and an alternative tilded parameterization. Both the OU theorem below and the general nonlinear theorem in Section 3.2 rely on these four assumptions.

Shared two-regime assumptions

(S1) (Regularity) $g, \tilde{g} : \mathbb{R}^d \to \mathbb{R}^d$ are C^2 diffeomorphisms, and $\varphi := \tilde{g}^{-1} \circ g$.

(S2) (Trajectory-law equivalence in each regime) For each $e \in \{1, 2\}$,

$$\{x_t^{(e)}\}_{t\in[0,T]} \overset{d}{=} \{\tilde{x}_t^{(e)}\}_{t\in[0,T]} \quad \text{in } C([0,T];\mathbb{R}^d).$$

(S3) (Diagonal positive-definite covariances) For each $e \in \{1, 2\}$, the diffusion covariances

$$\Sigma_e := G_e G_e^\top, \qquad \tilde{\Sigma}_e := \tilde{G}_e \tilde{G}_e^\top$$

are diagonal and positive definite.

(S4) (Two-regime separation; simple spectrum) The diagonal entries of $\Sigma_1^{-1}\Sigma_2$ are pairwise distinct; equivalently, if $\Sigma_e = \mathrm{diag}(\sigma_{e,1}^2, \ldots, \sigma_{e,d}^2)$, then

$$\frac{\sigma_{2,i}^2}{\sigma_{1,i}^2} \neq \frac{\sigma_{2,j}^2}{\sigma_{1,j}^2} \qquad \forall i \neq j.$$

Assumptions (S1) and (S2) are standard in identifiable nonlinear ICA and causal representation learning: one posits an invertible nonlinear mixing and asks whether two latent parameterizations inducing the same observed law must be equivalent, as in [15, 16, 22, 52–54]. Assumption (S3) is a nondegeneracy condition on the additive noise. It fixes coordinate-wise diffusion geometry and rules out singular directions. Related diagonal or block-diagonal noise restrictions are also common in continuous-time structure learning and path-space causal discovery [26, 27, 43].

Assumption (S4) is the separating condition: each latent coordinate carries a distinct diffusion-ratio fingerprint across environments. This is generically mild: failure occurs only when two ratios coincide exactly, a lower-dimensional subset of the diagonal covariance parameter space. It is also the continuous-time analogue of the variability conditions that drive identifiability in temporal and multi-environment CRL [22, 29, 39, 42, 53]. In the OU case, no additional support assumption is needed, because every nondegenerate Gaussian transition has full support on $\mathbb{R}^d$.

Throughout this section, $\varphi = \tilde{g}^{-1} \circ g$ denotes the change of coordinates between two observationally equivalent latent parameterizations: $\tilde{z} = \varphi(z)$ describes the same underlying state in the alternative

coordinate system. We write $D\varphi(z)$ for its Jacobian. The main identifiability question is whether φ must reduce to a monomial affine transformation.

Theorem 3.1 (OU identifiability from two diagonal diffusion regimes)**.** *For each $e \in \{1, 2\}$, consider reference and tilded latent OU parameterizations:*

$$dz_t^{(e)} = A\, z_t^{(e)}\, dt + G_e\, dW_t^{(e)}, \qquad x_t^{(e)} = g(z_t^{(e)}), \tag{2a}$$

$$d\tilde{z}_t^{(e)} = \tilde{A}\, \tilde{z}_t^{(e)}\, dt + \tilde{G}_e\, d\tilde{W}_t^{(e)}, \qquad \tilde{x}_t^{(e)} = \tilde{g}(\tilde{z}_t^{(e)}). \tag{2b}$$

Assume (S1)–(S4)*. Then there exist a constant monomial matrix $L = \Pi\Lambda$ and a constant vector $c \in \mathbb{R}^d$ such that*

$$D\varphi(z) \equiv L, \qquad \varphi(z) = Lz + c \quad (z \in \mathbb{R}^d),$$
$$\tilde{\Sigma}_e = L\Sigma_e L^\top, \qquad \tilde{A}L = LA, \qquad \tilde{A}c = 0 \quad (e \in \{1, 2\}).$$

The relation $\tilde{A}L = LA$ is equivalently $\tilde{A} = LAL^{-1}$; the additional condition $\tilde{A}c = 0$ constrains the possible translation. Hence the observed path laws identify the latent coordinate system up to permutation and scaling, together with a possible constant shift if $\varphi(0) \neq 0$. For $E_A := \{(j \to i) : A_{ij} \neq 0\}$, $E_{\tilde{A}} := \{(j \to i) : \tilde{A}_{ij} \neq 0\}$, and $\Pi e_j = e_{\pi(j)}$,

$$(j \to i) \in E_A \iff (\pi(j) \to \pi(i)) \in E_{\tilde{A}}.$$

Thus the latent causal structure in the OU case is identifiable up to the same permutation.

Proof sketch. The full proof is deferred to Appendix C. Equality of observed path laws allows a coupling with $\tilde{z}_t^{(e)} = \varphi(z_t^{(e)})$ almost surely in each regime. Itô's formula and quadratic variation give $D\varphi(z_t^{(e)})\Sigma_e D\varphi(z_t^{(e)})^\top = \tilde{\Sigma}_e$ along trajectories. For OU dynamics, the latent state at any positive time has a strictly positive Gaussian density on $\mathbb{R}^d$, so this identity extends to every state $z \in \mathbb{R}^d$. Whitening with respect to the first regime reduces the problem to an orthogonal matrix field. The second regime and the simple-spectrum condition in (S4) force that field to be a signed permutation at every z. Continuity then makes the signed permutation constant, so $D\varphi$ is a constant monomial matrix and φ is affine; substituting into the drift relation gives $\tilde{A}L = LA$, yielding the coordinate-system and graph statements.

3.2 General latent SDEs

We now turn to nonlinear drifts. As in the OU case, $\varphi := \tilde{g}^{-1} \circ g$ denotes the coordinate change between two observationally equivalent latent parameterizations, with Jacobian $D\varphi(z)$. The diffusion comparison still provides the core algebraic step. Only the support argument changes: an identity first obtained along sample paths must be extended to $\mathbb{R}^d$ without explicit Gaussian transition densities.

The shared assumptions (S1)–(S4) remain in force, and (S5) ensures well-defined continuous semimartingales and the open-set positivity step. This is a standard additive-noise SDE regularity condition: it requires continuous drifts and weak well-posedness with continuous sample paths, but not strong solutions or closed-form transition densities. Appendix C.5 gives further discussion.

Theorem 3.2 (Latent SDE identifiability from two diffusion regimes)**.** *For each $e \in \{1, 2\}$, consider reference and tilded latent SDE parameterizations:*

$$dz_t^{(e)} = f(z_t^{(e)})\, dt + G_e\, dW_t^{(e)}, \qquad x_t^{(e)} = g(z_t^{(e)}), \tag{3a}$$

$$d\tilde{z}_t^{(e)} = \tilde{f}(\tilde{z}_t^{(e)})\, dt + \tilde{G}_e\, d\tilde{W}_t^{(e)}, \qquad \tilde{x}_t^{(e)} = \tilde{g}(\tilde{z}_t^{(e)}). \tag{3b}$$

Assume the shared two-regime conditions (S1)–(S4)*, and in addition:*

Additional regularity assumption for nonlinear SDEs

(S5) (Drift regularity and well-posedness) The drifts $f, \tilde{f}$ are continuous on $\mathbb{R}^d$. For each e, both SDEs in (3a) and (3b) are weakly well-posed on $[0, T]$ with a.s. continuous sample paths.

Then there exist a constant monomial matrix $L = \Pi\Lambda$ and a constant vector $c \in \mathbb{R}^d$ such that

$$D\varphi(z) \equiv L, \qquad \varphi(z) = Lz + c \quad (z \in \mathbb{R}^d),$$
$$\tilde{\Sigma}_e = L\Sigma_e L^\top \quad (e \in \{1, 2\}), \qquad \tilde{f}(Lz + c) = Lf(z) \quad (z \in \mathbb{R}^d).$$

Consequently, the observed path laws identify the latent coordinate system up to permutation and scaling, together with a possible constant shift c.

Remark 3.3 (Support coverage is automatic). The proof only needs that, for each regime e, the law of $z_{t_\star}^{(e)}$ has full support on $\mathbb{R}^d$ at some $t_\star > 0$. Under (S3) and (S5), this follows from the open-set positivity lemma in Appendix C.3, which covers arbitrary initial laws.

Proof sketch. The proof, deferred to Appendix C, reuses the OU argument. The coupling, quadratic-variation comparison, and whitening/simple-spectrum steps are unchanged. The new ingredient is the open-set positivity lemma, which upgrades the along-trajectory covariance identity to a pointwise identity on $\mathbb{R}^d$ without Gaussian transition densities. The rest follows as in the OU case: $D\varphi$ is a constant monomial matrix, hence $\varphi(z) = Lz + c$, and the same semimartingale comparison gives $\tilde{f}(Lz + c) = Lf(z)$.

Under (S1)–(S5), Theorem 3.2 shows that any observationally equivalent latent coordinate change must be a monomial affine map. When f and $\tilde{f}$ are C^1, differentiating $\tilde{f}(Lz + c) = Lf(z)$ yields the drift-Jacobian graph statement.

Corollary 3.4 (Drift-Jacobian graph identifiability up to permutation)**.** *Under the assumptions of Theorem 3.2, assume additionally that $f, \tilde{f}$ are C^1. Let $J_f(z) := Df(z)$, $E_f(z) := \{(j \to i) : (J_f(z))_{ij} \neq 0\}$, and define $J_{\tilde{f}}, E_{\tilde{f}}$ analogously. Then*

$$D\tilde{f}(Lz + c) = L\, Df(z)\, L^{-1} \qquad \forall z \in \mathbb{R}^d,$$

so the instantaneous drift-Jacobian causal graphs are identified up to the permutation induced by Π. *Writing $\Pi e_j = e_{\pi(j)}$, equivalently,*

$$(j \to i) \in E_f(z) \quad \Longleftrightarrow \quad (\pi(j) \to \pi(i)) \in E_{\tilde{f}}(Lz + c).$$

4 Two-stage estimation

We translate the identifiability results into a two-stage estimator. Stage 1 learns the latent coordinate system from short-time transition statistics across the two regimes. Stage 2 is optional and estimates a sparse drift graph when explicit graph recovery is desired.

4.1 Model parameterization

We observe trajectories $\{x_{t_0}^{(e)}, x_{t_1}^{(e)}, \ldots, x_{t_N}^{(e)}\}$ from two regimes $e \in \{1, 2\}$ at uniform intervals Δt. The encoder h_θ maps observations into a learned latent coordinate system, playing the estimator-side role of $\tilde{g}^{-1}$ in Sections 3.1–3.2. We keep the tilde notation $(\tilde{z}, \tilde{f}, \tilde{\Sigma}_e)$ for learned coordinates, drift, and diffusion parameters, with θ, ψ denoting parametric components. Our learnable model consists of three components:

- an encoder $h_\theta : \mathbb{R}^d \to \mathbb{R}^d$, implemented as an MLP in our experiments, with $\tilde{z}_t = h_\theta(x_t)$;
- a differentiable drift model $\tilde{f}_\psi : \mathbb{R}^d \to \mathbb{R}^d$, shared across regimes and implemented as either a linear map or an MLP;
- diagonal per-regime diffusion parameters $\tilde{\Sigma}_e = \mathrm{diag}(\tilde{\sigma}_{e,1}^2, \ldots, \tilde{\sigma}_{e,d}^2)$, parameterized by $\log \tilde{\sigma}_{e,i}^2$ to ensure positivity.

We do not enforce h_θ to be globally invertible during training, so the Stage 1 likelihood below is interpreted as a pseudo-likelihood in general.

4.2 Stage 1: short-time transition fitting

Stage 1 uses the short-time behavior of additive-noise SDEs. For small Δt, the Euler–Maruyama discretization gives a Gaussian approximation to one-step transitions [30, 35]. In regime e, the encoded-space working model is

$$\tilde{z}_{t+\Delta t} \mid \tilde{z}_t, e \sim \mathcal{N}\Big(\tilde{z}_t + \tilde{f}_\psi(\tilde{z}_t)\,\Delta t, \tilde{\Sigma}_e\,\Delta t\Big). \tag{4}$$

Equivalently,

$$\tilde{z}_{t+\Delta t} = \tilde{z}_t + \tilde{f}_\psi(\tilde{z}_t)\,\Delta t + \varepsilon_t^{(e)}, \qquad \varepsilon_t^{(e)} \sim \mathcal{N}(0, \tilde{\Sigma}_e\,\Delta t).$$

The diagonal $\tilde{\Sigma}_e$ makes the one-step residual coordinates conditionally independent under this model.

If the encoded states $\tilde{z}_t$ were observed, the per-transition negative log-likelihood, up to the additive constant $\frac{d}{2}\log(2\pi\Delta t)$, is

$$\ell_e(\tilde{z}_t, \tilde{z}_{t+\Delta t}) = \frac{1}{2}\sum_{i=1}^{d} \log \tilde{\sigma}_{e,i}^2 + \frac{1}{2}\sum_{i=1}^{d} \frac{\big(\tilde{z}_{t+\Delta t,i} - \tilde{z}_{t,i} - \tilde{f}_{\psi,i}(\tilde{z}_t)\,\Delta t\big)^2}{\tilde{\sigma}_{e,i}^2\,\Delta t}. \tag{5}$$

This is the Gaussian transition score for (4).

Observation-space objective. We observe only x_t and set $\tilde{z}_t = h_\theta(x_t)$. If h_θ is a C^1 diffeomorphism, the change-of-variables formula gives

$$p_\theta(x_{t+\Delta t} \mid x_t, e) = p_{\tilde{z},e}\big(h_\theta(x_{t+\Delta t}) \mid h_\theta(x_t)\big)\,\big|\det Dh_\theta(x_{t+\Delta t})\big|,$$

where $p_{\tilde{z},e}$ is the Gaussian density in (4) and Dh_θ is the encoder Jacobian. The observed-data per-transition negative log-likelihood, up to the same additive constant, is

$$s_e(x_t, x_{t+\Delta t}) = \ell_e\big(h_\theta(x_t), h_\theta(x_{t+\Delta t})\big) - \log\big|\det Dh_\theta(x_{t+\Delta t})\big|. \tag{6}$$

In implementation, h_θ is differentiable but need not be globally invertible, so (6) is generally a *transition pseudo-likelihood*. The first term fits encoded increments to the diagonal Gaussian transition model, while the log-determinant term provides a local volume correction. If h_θ is a C^1 diffeomorphism, the same score is the exact Euler–Maruyama one-step negative log-likelihood.

Stage 1 objective. We average the per-transition score over regimes and sampled consecutive pairs:

$$\mathcal{L}_{\mathrm{S1}}(\theta, \psi, \{\tilde{\sigma}_{e,i}^2\}) = \frac{1}{2}\sum_{e=1}^{2} \mathbb{E}_{(x_t, x_{t+\Delta t})\sim\mathcal{P}_e}\Big[s_e(x_t, x_{t+\Delta t})\Big], \tag{7}$$

where $\mathcal{P}_e$ is the empirical distribution of consecutive pairs in regime e. Stage 1 jointly learns $(\theta, \psi, \{\tilde{\sigma}_{e,i}^2\})$.

Why this objective matches the theory. Over a short interval Δt, the drift contributes a mean shift of order $O(\Delta t)$, whereas the stochastic increment has standard deviation $O(\sqrt{\Delta t})$ and covariance $\tilde{\Sigma}_e\,\Delta t$. Diffusion therefore dominates the short-time second-order behavior. Our theory shows that this regime-dependent diagonal diffusion structure anchors the latent coordinate system. Thus (5)–(7) trains the encoder to find coordinates in which short-time residual increments

$$\tilde{z}_{t+\Delta t} - \tilde{z}_t - \tilde{f}_\psi(\tilde{z}_t)\,\Delta t$$

fit a shared-drift, regime-specific diagonal Gaussian model. This is the finite-step analogue of the two-regime identifiability mechanism in Section 3.

4.3 Stage 2: sparse drift regression

Stage 2 is optional and used only to estimate a sparse causal graph. For latent disentanglement alone, Stage 1 is the relevant part of the pipeline.

Velocity targets. With the encoder h_θ frozen, we form velocity targets from encoded transitions:

$$\hat{v}_t = \frac{h_\theta(x_{t+\Delta t'}) - h_\theta(x_t)}{\Delta t'}, \tag{8}$$

where $\Delta t' = s\Delta t$ for an integer stride $s \geq 1$. Larger strides reduce target variance, $\mathrm{Var}(\hat{v}_{t,i}) \approx \tilde{\sigma}_{e,i}^2/\Delta t'$, at the cost of $O(\Delta t')$ bias from time-averaging the drift.

Let $\mathcal{V}_e$ denote the empirical distribution of encoded state–velocity pairs $(\tilde{z}_t, \hat{v}_t)$ in regime e, and let $\mathcal{Z}$ denote the equally weighted empirical distribution of encoded states from the two regimes.

Sparse drift estimation. We reinitialize $\tilde{f}_\psi$ and fit it to the velocity targets with a Jacobian sparsity penalty:

$$\mathcal{L}_{\mathrm{S2}}(\psi) = \underbrace{\frac{1}{2}\sum_{e=1}^{2} \mathbb{E}_{(\tilde{z}_t,\hat{v}_t)\sim\mathcal{V}_e}\left\|\tilde{f}_\psi(\tilde{z}_t) - \hat{v}_t\right\|^2}_{\text{velocity MSE}} + \lambda_{\text{sparse}} \underbrace{\mathbb{E}_{\tilde{z}\sim\mathcal{Z}} \sum_{i=1}^{d}\sum_{j=1}^{d}\left|\frac{\partial \tilde{f}_{\psi,i}}{\partial \tilde{z}_j}(\tilde{z})\right|}_{\ell_1 \text{ drift-Jacobian penalty}}. \tag{9}$$

This stage is estimator-side only and is used when a stable sparse graph estimate is needed.

Graph extraction. By Corollary 3.4, the learned drift-Jacobian zero pattern is meaningful only up to the latent permutation induced by Π. In practice, the binary adjacency matrix is obtained by thresholding the mean absolute Jacobian at a fixed threshold $\tau > 0$:

$$\overline{|D\tilde{f}_\psi|} := \mathbb{E}_{\tilde{z}\sim\mathcal{Z}}\left[|D\tilde{f}_\psi(\tilde{z})|\right].$$

Summary. Algorithm 1 in Appendix D summarizes the training and graph-extraction procedure.

5 Experiments

5.1 Synthetic experiments

We test whether the estimator matches the identifiability boundary predicted by the theory. We use three latent drift families: dense linear OU, sparse linear OU, and nonlinear sparse drift. For each family, we compare three diffusion-regime conditions: *Distinct ratios*, where two environments have coordinate-wise distinct diffusion-variance ratios and hence satisfy the separation requirement in Assumption (S4); *One regime*, which removes environment variation; and *Proportional diffusions*, which keeps two environments but makes all variance ratios equal. The theory guarantees coordinate recovery under Distinct ratios; the controls remove the separating condition used by that guarantee.

Each drift family is tested at $d \in \{5, 7\}$ under two nonlinear mixing functions. The main-text setting uses a three-layer MLP with leaky-$\tanh$ activation $\phi(z) = \tanh(z) + 0.1z$, yielding a C^∞ diffeomorphism that satisfies the theory's smoothness assumptions. We also test LeakyReLU mixing as a robustness check beyond the C^2 setting. To keep the main text focused, we report the $d = 5$ leaky-$\tanh$ results below; LeakyReLU and $d = 7$ results are in Appendix F, with the data-generating process, training pipeline, hyperparameters, and metric definitions in Appendix E (Table 4). We report mean correlation coefficient (MCC) for one-to-one absolute-correlation alignment between learned and true coordinates; monomial score (Mon.) for how close the mean absolute Jacobian of $h \circ g$ is to a permutation-scaling matrix; and, for sparse linear and nonlinear drifts, graph match rate (GMR) for entrywise agreement between the thresholded learned drift-Jacobian graph and the ground-truth graph. Higher is better for all metrics.

Table 1 summarizes the quantitative results. Across all three drift families, Distinct ratios gives near-perfect MCC and high monomial scores, whereas both controls show substantially lower coordinate recovery and less monomial encoder Jacobians. In the sparse linear and nonlinear families, Distinct ratios also recovers the drift graph almost exactly. GMR is omitted for the dense linear family because sparse graph recovery is not the target there. The same qualitative pattern holds at $d = 7$ and under LeakyReLU mixing (Appendix F). Appendix F.4 further reports a $d = 10$ evaluation including a fully dense nonlinear drift, Appendix F.5 reports measured computational cost, and Appendix F.6 evaluates irregular sampling with a timestamp-ablated control.

Figures 1 and 2 give qualitative diagnostics for the nonlinear $d = 5$ setting, covering both latent disentanglement and graph recovery. Each column shows one regime condition, using the same randomly selected training seed across conditions. Figure 1 shows that Distinct ratios gives a near-permutation alignment between true and learned coordinates, while the controls do not. Figure 2 shows that the corresponding $\overline{|D\varphi|}$ matrix is near-monomial and the drift graph is recovered exactly under Distinct ratios; the controls produce denser encoder Jacobians and spurious or missed graph edges. Analogous figures for other drift families, dimensions, and mixing functions are in Appendix F.

Table 1: Synthetic $d = 5$ results under three-layer leaky-$\tanh$ MLP mixing. Entries are mean $\pm$ s.d. over five training seeds. **Bold** highlights the identifiable condition (Distinct ratios); the other two conditions are controls. GMR is omitted for the dense linear family because sparse graph recovery is not the target there.

Drift family	Regime condition	MCC ↑	Mon. ↑	GMR ↑
Dense linear	Distinct ratios	$\mathbf{0.9976 \pm 0.0001}$	$\mathbf{0.8697 \pm 0.0027}$	–
Dense linear	One regime	0.7819 ± 0.0160	0.4750 ± 0.0253	–
Dense linear	Proportional diffusions	0.8087 ± 0.0428	0.4940 ± 0.0344	–
Sparse linear	Distinct ratios	$\mathbf{0.9976 \pm 0.0002}$	$\mathbf{0.8920 \pm 0.0038}$	$\mathbf{1.0000 \pm 0.0000}$
Sparse linear	One regime	0.7206 ± 0.0382	0.4131 ± 0.0349	0.5760 ± 0.0784
Sparse linear	Proportional diffusions	0.7018 ± 0.0367	0.4061 ± 0.0269	0.6400 ± 0.0566
Nonlinear	Distinct ratios	$\mathbf{0.9977 \pm 0.0001}$	$\mathbf{0.8924 \pm 0.0035}$	$\mathbf{1.0000 \pm 0.0000}$
Nonlinear	One regime	0.7331 ± 0.0232	0.4234 ± 0.0280	0.6800 ± 0.0506
Nonlinear	Proportional diffusions	0.7531 ± 0.0141	0.4386 ± 0.0248	0.7840 ± 0.0742

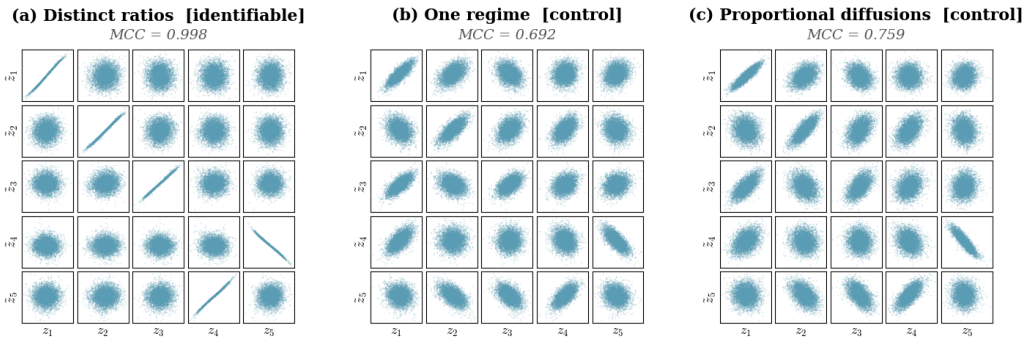

Figure 1: Latent disentanglement in the nonlinear $d = 5$ setting (three-layer leaky-$\tanh$ MLP mixing). Each column shows the 5×5 scatter matrix of true coordinates z_j (horizontal) versus learned coordinates $\tilde{z}_i$ (vertical) for one regime condition. Distinct ratios yields the expected near-permutation alignment; the controls do not recover a clean one-to-one alignment.

5.2 Real-world case study

We further test the method on operational accelerometer recordings from the Hardanger Bridge, a long-span suspension bridge instrumented for long-term wind and vibration monitoring [6, 7]. The model is trained only on bridge-deck acceleration channels; wind measurements define regimes but are not used as model inputs. This also probes robustness beyond the square diffeomorphic setting of Assumption (S1): the input has 16 sensor channels and the learned latent space has dimension $d = 5$. To obtain a real-world analogue of the diffusion-separation condition in Assumption (S4), we split segments within a matched wind-speed bin by turbulence intensity (TI). This keeps mean wind speed fixed while changing stochastic excitation, so low- and high-TI segments are intended to share structural dynamics while carrying different diffusion fingerprints.

We compare three regime constructions: *TI-split regimes (ours)*, using the low- and high-TI groups as two regimes; *One regime*, merging the two TI groups and removing regime-specific diffusion variation; and *Shuffled TI labels*, preserving the two-regime training format but destroying the turbulence-based contrast. Since ground-truth latent coordinates and graphs are unavailable, we evaluate cross-seed reproducibility rather than accuracy, using pairwise mean matched absolute correlation coefficient (MCC) and graph match rate (GMR). Table 2 reports results over $K = 10$ seeds at $d = 5$. The TI-split construction gives the highest MCC and GMR, consistent with the synthetic gap between identifiable diffusion shifts and controls. Appendix G gives the dataset construction, training protocol, aggregate diagnostics, Assumption (S4)-style diffusion-fingerprint checks, a representative drift-Jacobian interpretation, and limitations of the real-world diagnostic.

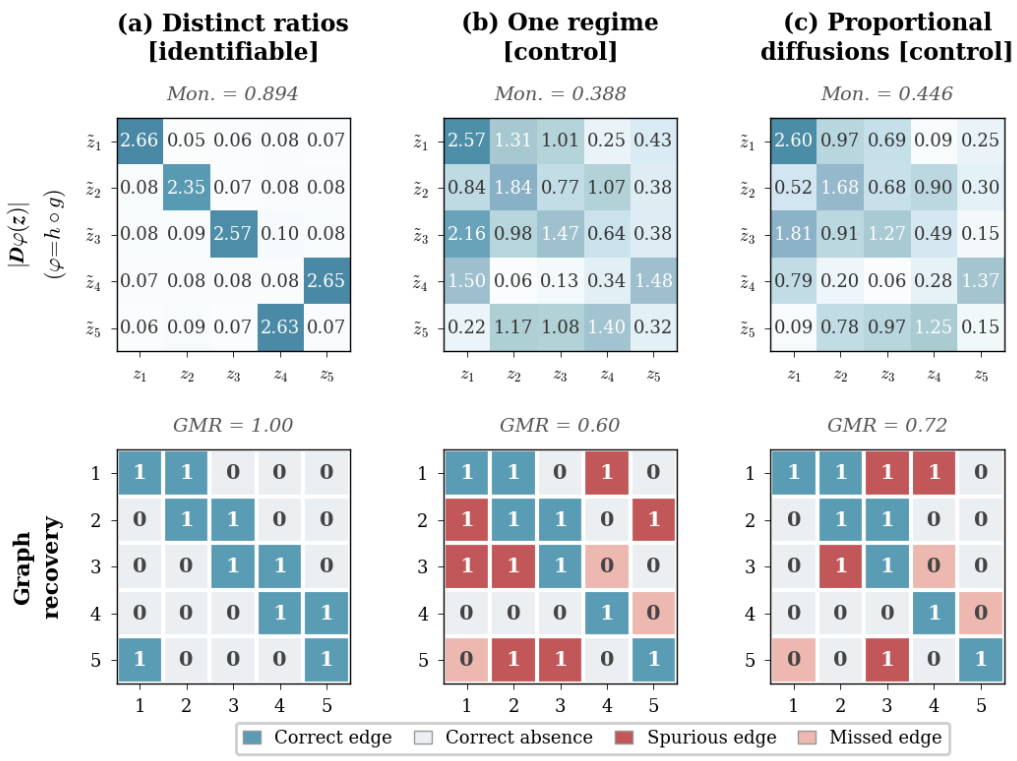

Figure 2: Structural diagnostics for the same setting and run as Figure 1. **Top row**: mean absolute encoder Jacobian $\overline{|D\varphi|}$ for $\varphi = h \circ g$; near-monomial structure indicates inversion up to permutation and scaling. **Bottom row**: learned versus ground-truth drift graph, with correct edges/absences in blue/gray and spurious/missed edges in red/pink. Distinct ratios achieves an exact graph match; both controls introduce graph errors.

Table 2: Hardanger Bridge accelerometer case study [6, 7] ($K = 10$ seeds, $d = 5$). Entries are mean pairwise cross-seed reproducibility scores (MCC, or GMR with $\tau = 0.05$) over $\binom{K}{2} = 45$ seed pairs.

Condition	MCC ↑	GMR ↑
TI-split regimes (ours)	**0.806**	**0.877**
One regime (control)	0.660	0.813
Shuffled TI labels (control)	0.601	0.830

6 Related work

Identifiable nonlinear ICA and temporal CRL show that auxiliary structure, such as nonstationarity, temporal dependence, interventions, grouping, or environment variation, can break nonlinear mixing ambiguities [3, 15, 16, 24, 38, 53, 54].

Their theorem-level temporal results, however, are largely discrete-time, based on finite-lag transitions, autoregressive mechanisms, or temporal context on a grid. Continuous-time SDE identification and causal discovery usually assume that the stochastic state is directly observed [9, 26, 43, 44, 56]; the closest latent continuous-time work treats stochastic point processes or system-level parameters rather than latent SDE state coordinates [2, 32, 50]. Recent work also studies identifiability of second-order ODE parameters from video [47], a deterministic system-parameter target rather than latent SDE state coordinates. We address this gap by identifying the latent coordinate system and drift-Jacobian graph of continuous-time SDE states observed through an unknown nonlinear map, using two environments with shared drift and different diffusion covariances. Appendix B gives a fuller comparison.

7 Discussion and limitations

This paper gives a theorem-level account of identifiable CRL for continuous-time latent SDEs. The main message is that *multi-regime diffusion structure* can anchor the latent coordinate system: with shared drift and diagonal diffusion covariances whose coordinate-wise variance ratios are distinct, nonlinear observations identify the coordinates up to a monomial affine transformation, and the instantaneous drift-Jacobian graph at corresponding states follows up to the same permutation, without drift sparsity. This is a step toward CRL for continuous latent dynamics, where most temporal identifiability results are discrete-time and most continuous-time SDE identification assumes observed states. The assumptions are intentionally simple: an invertible observation map, additive state-independent diffusion, a shared drift, and separated diffusion regimes. The diffeomorphism assumption is the strongest idealization, excluding non-invertible or nonsmooth observation maps. Our LeakyReLU experiments and Hardanger study probe robustness beyond the smooth square diffeomorphic setting. Extending the theory to weaker observation maps, state-dependent noise, and imperfect regime labels remains important future work.

Acknowledgments and Disclosure of Funding

HL was supported by the Shanghai General AI Foundation Models Program (No. 2025SHZDZX026D35), the Beijing Natural Science Foundation (No. L257007), the Beijing Major Science and Technology Project (No. Z251100008425006), and the CCF-Kuaishou Large Model Explorer Fund (No. CCF-KuaiShou 2026018). MG was supported by ARC DP240102088 and WIS–MBZUAI 142571. KZ would like to acknowledge support from NSF Award No. 2229881, the AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, Florin Court Capital, the MBZUAI–WIS Joint Program, and the Al Deira Causal Education project. The authors declare no competing interests.

References

[1] K. Ahuja, A. Mansouri, and Y. Wang. Multi-domain causal representation learning via weak distributional invariances. In *International Conference on Artificial Intelligence and Statistics*, pages 865–873. PMLR, 2024.

[2] M. W. Baumgartner, A. Lei, J. Watson, and I. Posner. Disentangling dynamical systems: Causal representation learning meets local sparse attention. *arXiv preprint arXiv:2603.14483*, 2026.

[3] G. Chen, Y. Shen, Z. Chen, X. Song, Y. Sun, W. Yao, X. Liu, and K. Zhang. Caring: learning temporal causal representation under non-invertible generation process. In *Proceedings of the 41st International Conference on Machine Learning*, pages 7236–7259, 2024.

[4] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[5] S. Fan, K. Zhang, and L. Cheng. Trace: Trajectory recovery for continuous mechanism evolution in causal representation learning. *arXiv preprint arXiv:2601.21135*, 2026.

[6] A. Fenerci, K. Andreas Kvåle, Ø. Wiig Petersen, A. Rønnquist, and O. Øiseth. Data set from long-term wind and acceleration monitoring of the hardanger bridge. *Journal of Structural Engineering*, 147(5):04721003, 2021.

[7] A. Fenerci, K. A. Kvåle, Ø. W. Petersen, A. Rønnquist, and O. Øiseth. Wind and acceleration data from the hardanger bridge. *Norges teknisk-naturvitenskapelige universitet*, 2020.

[8] L. Gresele, J. Von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34:28233–28248, 2021.

[9] V. Guan, J. Janssen, H. Rahmani, A. Warren, S. Zhang, E. Robeva, and G. Schiebinger. Identifying drift, diffusion, and causal structure from temporal snapshots. *arXiv preprint arXiv:2410.22729*, 2024.

[10] H. Hälvä and A. Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

[11] H. Hälvä, S. Le Corff, L. Lehéricy, J. So, Y. Zhu, E. Gassiat, and A. Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. *Advances in Neural Information Processing Systems*, 34:1624–1633, 2021.

[12] A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

[13] A. Hyvarinen and H. Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial intelligence and statistics*, pages 460–469. PMLR, 2017.

[14] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[15] A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pages 859–868. PMLR, 2019.

[16] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.

[17] P. Kidger, J. Morrill, J. Foster, and T. Lyons. Neural controlled differential equations for irregular time series. In *Advances in neural information processing systems*, volume 33, 2020.

[18] X. Li, J. Wang, and T. Liu. Recovery of causal graph involving latent variables via homologous surrogates. In *International Conference on Learning Representations*, volume 2025, pages 79091–79114, 2025.

[19] X. Li, T.-K. L. Wong, R. T. Chen, and D. Duvenaud. Scalable gradients for stochastic differential equations. In *International conference on artificial intelligence and statistics*, pages 3870–3882. PMLR, 2020.

[20] X. Li, K. Zhang, and T. Liu. Causal structure recovery with latent variables under milder distributional and graphical assumptions. In *International Conference on Learning Representations*, volume 2024, pages 27207–27238, 2024.

[21] Z. Li, M. Fu, J. Huang, Y. Shen, R. Cai, Y. Sun, G. Chen, and K. Zhang. Towards identifiability of hierarchical temporal causal representation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[22] Z. Li, Y. Shen, K. Zheng, R. Cai, X. Song, M. Gong, G. Chen, and K. Zhang. On the identification of temporal causal representation with instantaneous dependence. In *The Thirteenth International Conference on Learning Representations*, 2025.

[23] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2023.

[24] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and S. Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

[25] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[26] G. Manten, C. Casolo, E. Ferrucci, S. W. Mogensen, C. Salvi, and N. Kilbertus. Signature kernel conditional independence tests in causal discovery for stochastic processes. In *The Thirteenth International Conference on Learning Representations*, 2025.

[27] G. Manten, C. Casolo, S. W. Mogensen, and N. Kilbertus. An asymmetric independence model for causal discovery on path spaces. In *Causal Learning and Reasoning*, pages 64–89. PMLR, 2025.

[28] H. Morioka and A. Hyvarinen. Causal representation learning made identifiable by grouping of observational variables. In *International Conference on Machine Learning*, pages 36249–36293. PMLR, 2024.

[29] I. Ng, S. Xie, X. Dong, P. Spirtes, and K. Zhang. Causal representation learning from general environments under nonparametric mixing. In *International Conference on Artificial Intelligence and Statistics*, pages 3700–3708. PMLR, 2025.

[30] B. Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[31] P. Reizinger, S. Guo, F. Huszár, B. Schölkopf, and W. Brendel. Identifiable exchangeable mechanisms for causal structure and representation learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[32] J. Ren, Y. Wang, and B. Huang. Causal representation meets stochastic modeling under generic geometry. *arXiv preprint arXiv:2602.05033*, 2026.

[33] Y. Rubanova, R. T. Chen, and D. K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

[34] J. Runge, A. Gerhardus, G. Varando, V. Eyring, and G. Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

[35] S. Särkkä and A. Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

[36] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[37] A. Shojaie and E. B. Fox. Granger causality: A review and recent advances. *Annual review of statistics and its application*, 9:289–319, 2022.

[38] X. Song, Z. Li, G. Chen, Y. Zheng, Y. Fan, X. Dong, and K. Zhang. Causal temporal representation learning with nonstationary sparse transition. *Advances in Neural Information Processing Systems*, 37:77098–77131, 2024.

[39] X. Song, W. Yao, Y. Fan, X. Dong, G. Chen, J. C. Niebles, E. Xing, and K. Zhang. Temporally disentangled representation learning under unknown nonstationarity. *Advances in Neural Information Processing Systems*, 36:8092–8113, 2023.

[40] D. W. Stroock and S. S. Varadhan. *Multidimensional diffusion processes*. Springer, 2007.

[41] B. Varici, E. Acartürk, K. Shanmugam, and A. Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2024.

[42] J. von Kügelgen, M. Besserve, L. Wendong, L. Gresele, A. Kekić, E. Bareinboim, D. Blei, and B. Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36:48603–48638, 2023.

[43] B. Wang, J. Jennings, and W. Gong. Neural structure learning with stochastic differential equations. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] Y. Wang, X. Geng, W. Huang, B. Huang, and M. Gong. Generator identification for linear sdes with additive and multiplicative noise. *Advances in Neural Information Processing Systems*, 36:64103–64138, 2023.

[45] Y. Wang, B. Huang, W. Huang, X. Geng, and M. Gong. Identifiability analysis of linear ode systems with hidden confounders. *Advances in Neural Information Processing Systems*, 37:59054–59092, 2024.

[46] Y. Wang, W. Huang, M. Gong, X. Geng, T. Liu, K. Zhang, and D. Tao. Identifiability and asymptotics in learning homogeneous linear ode systems from discrete observations. *Journal of Machine Learning Research*, 25(154):1–50, 2024.

[47] Y. Wang, W. Wang, K. Zhang, and M. Gong. Physics from video: Identifiability of time-invariant second-order ODEs under minimal trajectory conditions. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 131032–131076. PMLR, 2026.

[48] R. Welch, J. Zhang, and C. Uhler. Identifiability guarantees for causal disentanglement from purely observational data. *Advances in Neural Information Processing Systems*, 37:102796–102821, 2024.

[49] D. Xu, D. Yao, S. Lachapelle, P. Taslakian, J. von Kügelgen, F. Locatello, and S. Magliacane. A sparsity principle for partially observable causal representation learning. In *Forty-first International Conference on Machine Learning*, 2024.

[50] D. Yao, C. Muller, and F. Locatello. Marrying causal representation learning with dynamical systems for science. *Advances in Neural Information Processing Systems*, 37:71705–71736, 2024.

[51] D. Yao, D. Rancati, R. Cadei, M. Fumero, and F. Locatello. Unifying causal representation learning with the invariance principle. In *The Thirteenth International Conference on Learning Representations*, 2025.

[52] D. Yao, D. Xu, S. Lachapelle, S. Magliacane, P. Taslakian, G. Martius, J. von Kügelgen, and F. Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024.

[53] W. Yao, G. Chen, and K. Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022.

[54] W. Yao, Y. Sun, A. Ho, C. Sun, and K. Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022.

[55] K. Zhang, S. Xie, I. Ng, and Y. Zheng. Causal representation learning from multiple distributions: a general setting. In *Proceedings of the 41st International Conference on Machine Learning*, pages 60057–60075, 2024.

[56] A. Zweig, Z. Lin, E. Azizi, and D. Knowles. Towards identifiability of interventional stochastic differential equations. *arXiv preprint arXiv:2505.15987*, 2025.

A Notation

Table 3: Summary of notation.

Symbol	Description
Spaces and indices	
d	Latent dimension (also the observed dimension in the square theory and simulations)
d_x, d_z	Observation and latent dimensions when they differ
$e \in \{1, 2\}$	Environment (regime) index
$t \in [0, T]$	Continuous time; T is the time horizon
Δt	Observation interval (constant in the regular-grid experiments)
$\Delta t' = s\Delta t$	Stage 2 velocity-target interval at stride s
Processes	
$z_t^{(e)} \in \mathbb{R}^d$	Latent state in environment e at time t
$x_t^{(e)} \in \mathbb{R}^d$	Observed state in the square theory: $x_t^{(e)} = g(z_t^{(e)})$
$\tilde{z}_t^{(e)} \in \mathbb{R}^d$	Alternative latent state under the tilded parameterization
$W_t^{(e)}$	Standard d-dimensional Brownian motion
SDE components	
$f\colon \mathbb{R}^d \to \mathbb{R}^d$	Latent drift function (shared across environments)
$A \in \mathbb{R}^{d\times d}$	Drift matrix in the linear (OU) case: $f(z) = Az$
$G_e \in \mathbb{R}^{d\times d}$	Diffusion matrix in environment e
$\Sigma_e = G_e G_e^\top$	Diffusion covariance in environment e
$\sigma_{e,i}^2$	Diagonal entry of Σ_e (variance of coordinate i in environment e)
Mixing and coordinate change	
$g\colon \mathbb{R}^d \to \mathbb{R}^d$	Unknown nonlinear observation map in the square theory (mixing diffeomorphism)
$\varphi = \tilde{g}^{-1} \circ g$	Latent coordinate change between two parameterizations
$L = \Pi\Lambda$	Monomial matrix (permutation and coordinate-wise scaling)
Π	Permutation matrix
Λ	Diagonal scaling matrix
$c \in \mathbb{R}^d$	Constant shift: $\varphi(z) = Lz + c$
Jacobians and causal graph	
$Dg(z)$	Jacobian of g at z
$D\varphi(z)$	Jacobian of the coordinate change φ at z
$Df(z)$	Jacobian of the latent drift f at z
$E_f(z)$	Instantaneous drift-Jacobian graph: $\{(j \to i) : (Df(z))_{ij} \neq 0\}$
$\hat{E}_f$	Estimated drift-Jacobian graph (from Stage 2)
$\overline{\lvert D\tilde{f}_\psi \rvert}$	Mean absolute Jacobian of the learned drift
Estimation	
$h_\theta\colon \mathbb{R}^d \to \mathbb{R}^d$	Encoder in the square theory and simulations
$\tilde{f}_\psi\colon \mathbb{R}^d \to \mathbb{R}^d$	Learned drift model
$\tilde{\sigma}_{e,i}^2$	Learned diffusion variance for coordinate i in environment e
$\hat{v}_t$	Velocity target: $(h_\theta(x_{t+\Delta t'}) - h_\theta(x_t))/\Delta t'$
$\mathcal{L}_{\mathrm{S1}}$	Stage 1 objective (transition negative log-likelihood)
$\mathcal{L}_{\mathrm{S2}}$	Stage 2 objective (velocity MSE + Jacobian L_1 penalty)
$\lambda_{\mathrm{sparse}}$	Sparsity penalty weight (Stage 2)
τ	Threshold for graph extraction from $\overline{\lvert D\tilde{f}_\psi \rvert}$
$\mathcal{P}_e$	Empirical distribution of observed transition pairs in regime e
$\mathcal{V}_e$	Empirical distribution of encoded state–velocity pairs in regime e
$\mathcal{Z}$	Equally weighted empirical distribution of encoded states across regimes
Evaluation metrics	
MCC	Mean correlation coefficient for learned/true coordinate alignment
Mon.	Monomial score for the mean absolute encoder Jacobian
GMR	Graph match rate for thresholded drift-Jacobian graphs

B Extended related work

Identifiable nonlinear ICA. General nonlinear mixing is not identifiable without extra structure [14, 25]. Identifiable nonlinear ICA uses nonstationarity, temporal dependence, or auxiliary variables [12, 13, 15, 16], with extensions to hidden Markov and structured dependent sources [10, 11]. These results show how suitable side information can break nonlinear mixing ambiguity.

Independent Mechanism Analysis (IMA) [8] imposes orthogonality on the columns of the mixing Jacobian. In contrast, Assumption (S3) constrains the latent diffusion covariance while leaving g arbitrary among C^2 diffeomorphisms; orthogonality arises only after whitening the diffusion geometry.

Temporal CRL in discrete time. A closely related CRL line studies nonlinear mixtures of latent temporal processes in discrete time. LEAP identifies latent temporal causal processes under nonstationary noise [54]; TDRL studies nonparametric time-delayed causal processes [53]; NCTRL addresses unknown nonstationarity [39]; and CtrlNS uses sparse-transition structure to recover nonstationary latent dynamics without known domain variables [38]. A parallel branch uses stronger side information: CITRIS and iCITRIS leverage interventions [23, 24], G-CaRL uses grouping information [28], and CaRiNG allows non-invertible temporal generation processes by using temporal context to recover lost information [3]. More recent extensions study instantaneous dependence via sparse latent dynamics [22], hierarchical temporal latents [21], and continuously evolving mechanisms rather than discrete switches [5]. These papers identify latent causal processes from sequential observations, but their arguments use discrete observations, transition factorizations, or auxiliary mechanism structure. Our proof instead identifies a continuous-time latent SDE path through diffusion covariance changes.

Multi-environment identification and partial observability in CRL. Environment heterogeneity and observation structure are also important sources of identifiability in CRL. Multi-environment results address unknown interventions [42], cross-domain invariance [1], general environment families [29, 41, 55], and limits of observational identification [48]; multi-view and partial-observability methods exploit additional observation structure [49, 52]. Complementary latent-variable causal-discovery work uses relaxed pure-child or homologous-surrogate conditions [18, 20]; its information comes from linear cross-sectional measurement and graph structure rather than temporal diffusion changes. Other work connects invariance-based CRL with exchangeability and causal discovery [31, 51]. Here the drift is invariant across environments, while distinct diffusion-variance ratios identify the coordinates up to a monomial affine transformation.

Continuous-time identification and causal discovery for stochastic processes. A separate literature studies what can be identified when the stochastic system itself is observed. Wang et al. analyze generator identifiability for linear SDEs [44]. SCOTCH studies observed-space structure learning with neural SDEs under irregular sampling [43]. Guan et al. identify linear drift, diffusion, and causal structure from temporal marginals [9]. Manten et al. derive path-space conditional independence constraints and corresponding tests for SDE models [26], and extend the graphical treatment to asymmetric independence on path spaces [27]. Zweig et al. analyze identifiability of interventional SDEs from stationary distributions [56]. Related deterministic-system results establish identifiability for homogeneous linear ODEs from discrete observations, linear ODEs with hidden confounders, and second-order ODE parameters from video [45–47]; these targets differ from latent SDE state coordinates and their instantaneous drift graphs. Among continuous-time latent-process CRL papers, Ren et al. [32] study identifiable CRL for continuous-time latent stochastic point processes, not latent SDE state coordinates. Closely related dynamical-systems papers connect CRL with trajectory-specific or system-parameter identification [2, 50]. Our paper differs from all of these in the recovery target: we identify continuously evolving latent *state coordinates* and their instantaneous Jacobian graph from unknown nonlinear mixtures, using two environments with shared drift and changed diffusion.

C Deferred technical material

C.1 Auxiliary lemmas

Lemma C.1 (Orthogonal conjugation of a diagonal matrix with simple spectrum)**.** *Let $K = \mathrm{diag}(k_1, \dots, k_d) \in \mathbb{R}^{d\times d}$ have real pairwise distinct diagonal entries, $k_i \neq k_j$ for all $i \neq j$. Let $Q \in \mathbb{R}^{d\times d}$ be an orthogonal matrix. If $QKQ^\top$ is diagonal, then Q is a signed permutation matrix: there exists a permutation π and signs $s_i \in \{\pm 1\}$ such that*

$$Qe_i = s_i e_{\pi(i)} \quad \textit{for all } i,$$

equivalently $Q = \Pi S$ where Π is a permutation matrix and S is diagonal with ± 1 entries. In particular, $QKQ^\top = \Pi K \Pi^\top$.

Proof. Because K is diagonal with pairwise distinct diagonal entries, its eigenvalues are exactly $\{k_1, \dots, k_d\}$ with algebraic (and geometric) multiplicity 1 each. Hence the eigenspace associated with k_i is one-dimensional and equals $\mathrm{span}\{e_i\}$.

Consider $\widehat{K} := QKQ^\top$. Since Q is orthogonal, $\widehat{K}$ is similar to K, so it has the same multiset of eigenvalues and hence also has d distinct eigenvalues. Because $\widehat{K}$ is diagonal by assumption, each eigenvalue appears as a diagonal entry of $\widehat{K}$, and its eigenspace is spanned by a standard basis vector. Concretely, there exists a permutation π such that $\widehat{K}_{\pi(i)\pi(i)} = k_i$ for all i.

Now note that

$$\widehat{K}(Qe_i) = QKe_i = Q(k_i e_i) = k_i(Qe_i),$$

so Qe_i is an eigenvector of $\widehat{K}$ with eigenvalue k_i. But the eigenspace of $\widehat{K}$ associated with k_i is $\mathrm{span}\{e_{\pi(i)}\}$, therefore $Qe_i \in \mathrm{span}\{e_{\pi(i)}\}$. Since Q is orthogonal, $\|Qe_i\| = \|e_i\| = 1$, hence $Qe_i = s_i e_{\pi(i)}$ with $s_i \in \{\pm 1\}$. This shows Q is a signed permutation matrix, and the final statement $QKQ^\top = \Pi K \Pi^\top$ follows because the sign diagonal cancels in conjugation. □

Lemma C.2 (Innovation representation in the natural filtration)**.** *Let $(z_t)_{t\in[0,T]}$ be a solution (strong or weak) on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ to*

$$dz_t = f(z_t)\,dt + G\,dW_t,$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is continuous, $G \in \mathbb{R}^{d\times d}$ is constant, and $\Sigma := GG^\top$ is positive definite. Let $(\mathcal{F}_t^z)$ denote the usual augmentation (completed and right-continuous) of the natural filtration $\sigma(z_s : 0 \le s \le t)$. Then the process

$$\hat{W}_t := \Sigma^{-1/2}\Big(z_t - z_0 - \int_0^t f(z_s)\,ds\Big), \qquad t \in [0,T],$$

is a standard d-dimensional Brownian motion with respect to $(\mathcal{F}_t^z)$, and z satisfies

$$dz_t = f(z_t)\,dt + \Sigma^{1/2}\,d\hat{W}_t$$

in the natural filtration $(\mathcal{F}_t^z)$.

Proof. Set $M_t := z_t - z_0 - \int_0^t f(z_s)\,ds = \int_0^t G\,dW_s$. Then M is a continuous $\mathcal{F}_t$-martingale with $\langle M\rangle_t = \Sigma t$, and M_t is $\mathcal{F}_t^z$-adapted because it is a measurable function of the path of z. Since $\mathcal{F}_t^z \subset \mathcal{F}_t$, the tower property gives, for $s \le t$,

$$\mathbb{E}[M_t \mid \mathcal{F}_s^z] = \mathbb{E}\big[\mathbb{E}[M_t \mid \mathcal{F}_s] \mid \mathcal{F}_s^z\big] = \mathbb{E}[M_s \mid \mathcal{F}_s^z] = M_s,$$

so M is a continuous $\mathcal{F}_t^z$-martingale. Setting $\hat{W}_t := \Sigma^{-1/2} M_t$, we have $\langle \hat{W}\rangle_t = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} t = I_d t$. By Lévy's characterization theorem, $\hat{W}$ is a standard $\mathcal{F}_t^z$-Brownian motion. □

Remark C.3 (Common natural filtration under diffeomorphic reparameterization). If $\tilde{z}_t = \varphi(z_t)$ where $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ is a homeomorphism (in particular, a diffeomorphism), then $\mathcal{F}_t^z = \mathcal{F}_t^{\tilde{z}}$: indeed $\tilde{z}_s = \varphi(z_s)$ gives $\mathcal{F}_t^{\tilde{z}} \subset \mathcal{F}_t^z$, and $z_s = \varphi^{-1}(\tilde{z}_s)$ gives $\mathcal{F}_t^z \subset \mathcal{F}_t^{\tilde{z}}$. In particular, a process that is a (local) martingale with respect to one filtration is automatically a (local) martingale with respect to the other.

C.2 Proof of Theorem 3.1

Proof of Theorem 3.1. We prove that equality in law of observed OU trajectories across two diffusion regimes forces the latent reparameterization $\varphi = \tilde{g}^{-1} \circ g$ to be affine with a constant monomial Jacobian.

Step 0: Coupling and a pathwise state relation.

Fix a regime $e \in \{1, 2\}$. By (S2), the observed trajectories $\{x_t^{(e)}\}_{t\in[0,T]}$ and $\{\tilde{x}_t^{(e)}\}_{t\in[0,T]}$ have the same law on the Polish path space $C([0, T]; \mathbb{R}^d)$. We may therefore realize both as the canonical coordinate process on $C([0, T]; \mathbb{R}^d)$ equipped with their common law, yielding a single process $\{X_t^{(e)}\}_{t\in[0,T]}$ satisfying

$$X^{(e)} = x^{(e)} = \tilde{x}^{(e)} \quad \text{a.s. as elements of } C([0, T]; \mathbb{R}^d).$$

Since g and $\tilde{g}$ are diffeomorphisms $\mathbb{R}^d \to \mathbb{R}^d$ by (S1), we can invert $x_t^{(e)} = g(z_t^{(e)})$ and $\tilde{x}_t^{(e)} = \tilde{g}(\tilde{z}_t^{(e)})$ pointwise to obtain

$$z_t^{(e)} = g^{-1}(X_t^{(e)}), \qquad \tilde{z}_t^{(e)} = \tilde{g}^{-1}(X_t^{(e)}), \qquad t \in [0, T].$$

Therefore, with $\varphi := \tilde{g}^{-1} \circ g$,

$$\tilde{z}_t^{(e)} = \varphi(z_t^{(e)}) \qquad \text{a.s. for all } t \in [0, T]. \tag{10}$$

Step 1: Quadratic variation identifies an equation for $D\varphi$.

Fix a regime e. Recall that under regime e the latent OU satisfies

$$dz_t^{(e)} = Az_t^{(e)}\, dt + G_e\, dW_t^{(e)}, \qquad \Sigma_e := G_e G_e^\top.$$

Since $\varphi \in C^2$ by (S1), Itô's formula applies to $\tilde{z}_t^{(e)} = \varphi(z_t^{(e)})$ and gives

$$\begin{aligned} d\tilde{z}_t^{(e)} &= D\varphi(z_t^{(e)})\, dz_t^{(e)} + \frac{1}{2} \sum_{k,\ell=1}^{d} (\Sigma_e)_{k\ell}\, \partial_{k\ell}\varphi(z_t^{(e)})\, dt \\ &= \Big(D\varphi(z_t^{(e)}) Az_t^{(e)} + \frac{1}{2} \sum_{k,\ell=1}^{d} (\Sigma_e)_{k\ell}\, \partial_{k\ell}\varphi(z_t^{(e)}) \Big) dt + D\varphi(z_t^{(e)}) G_e\, dW_t^{(e)}. \end{aligned}$$

If a d-dimensional process Y_t has the form $dY_t = b_t\, dt + H_t\, dW_t$, then only the noise term contributes to quadratic variation, and

$$d\langle Y \rangle_t = H_t H_t^\top\, dt.$$

Applying this with $Y_t = \tilde{z}_t^{(e)}$ and $H_t = D\varphi(z_t^{(e)})G_e$, we obtain

$$d\langle \tilde{z}^{(e)} \rangle_t = D\varphi(z_t^{(e)})\, \Sigma_e\, D\varphi(z_t^{(e)})^\top\, dt. \tag{11}$$

On the other hand, under the alternative OU model (2b), the diffusion covariance is constant and equals $\tilde{\Sigma}_e = \tilde{G}_e \tilde{G}_e^\top$, hence

$$d\langle \tilde{z}^{(e)} \rangle_t = \tilde{\Sigma}_e\, dt. \tag{12}$$

Equating (11) and (12) yields

$$D\varphi(z_t^{(e)})\, \Sigma_e\, D\varphi(z_t^{(e)})^\top = \tilde{\Sigma}_e, \tag{13}$$

holding for Lebesgue-a.e. $t \in [0, T]$ and almost surely.

Define the matrix-valued function along each sample path

$$F(t, \omega) := D\varphi(z_t^{(e)}(\omega))\, \Sigma_e\, D\varphi(z_t^{(e)}(\omega))^\top.$$

Because $t \mapsto z_t^{(e)}$ is continuous a.s. and $D\varphi$ is continuous, $t \mapsto F(t, \omega)$ is continuous on $[0, T]$ for almost every ω. Moreover, (13) holds for a.e. t and a.s. in ω; by a standard Fubini-type argument,

there exists an event Ω_0 with $\mathbb{P}(\Omega_0) = 1$ such that for each $\omega \in \Omega_0$, (13) holds for Lebesgue-a.e. $t \in [0, T]$. Fix such an ω. Since $F(\cdot, \omega)$ is continuous and equals the constant matrix $\tilde{\Sigma}_e$ for a.e. t, it must equal $\tilde{\Sigma}_e$ for all $t \in [0, T]$. Therefore (13) holds for all $t \in [0, T]$ a.s.

Step 2: Strictly positive density upgrades the pathwise identity to a pointwise identity on $\mathbb{R}^d$.

Fix $e \in \{1, 2\}$ and any $t_\star \in (0, T]$. Define the continuous matrix-valued function on $\mathbb{R}^d$ by

$$B_e(z) := D\varphi(z)\, \Sigma_e\, D\varphi(z)^\top - \tilde{\Sigma}_e.$$

Evaluating (13) at $t = t_\star$ gives $B_e(z_{t_\star}^{(e)}) = 0$ a.s.

We claim this implies $B_e(z) \equiv 0$ on $\mathbb{R}^d$. Suppose not. Then there exists $z_0 \in \mathbb{R}^d$ with $B_e(z_0) \neq 0$. By continuity of B_e, there exist $\varepsilon > 0$ and an open neighborhood V of z_0 such that $\|B_e(z)\|_F > \varepsilon$ for all $z \in V$. Since V is open, it has positive Lebesgue measure. Now, the OU solution satisfies $z_{t_\star}^{(e)} = e^{At_\star} z_0^{(e)} + \xi_{t_\star}^{(e)}$, where $\xi_{t_\star}^{(e)} := \int_0^{t_\star} e^{A(t_\star - s)} G_e\, dW_s^{(e)}$ is a zero-mean Gaussian with covariance

$$P_{t_\star} = \int_0^{t_\star} e^{A(t_\star - s)} \Sigma_e e^{A^\top (t_\star - s)}\, ds \succ 0.$$

Indeed, for any $v \neq 0$,

$$v^\top P_{t_\star} v = \int_0^{t_\star} \|\Sigma_e^{1/2} e^{A^\top (t_\star - s)} v\|^2\, ds > 0,$$

since $\Sigma_e^{1/2}$ is invertible and $s \mapsto e^{A^\top (t_\star - s)} v$ is continuous and nonvanishing. Hence the conditional density $p(z_{t_\star}^{(e)} \mid z_0^{(e)})$ is strictly positive on all of $\mathbb{R}^d$ for every value of $z_0^{(e)}$, and therefore the marginal density of $z_{t_\star}^{(e)}$ is also strictly positive on $\mathbb{R}^d$. Thus

$$\mathbb{P}(z_{t_\star}^{(e)} \in V) > 0,$$

contradicting $B_e(z_{t_\star}^{(e)}) = 0$ a.s. Hence $B_e \equiv 0$ on $\mathbb{R}^d$, i.e.

$$D\varphi(z)\, \Sigma_e\, D\varphi(z)^\top = \tilde{\Sigma}_e, \qquad \forall z \in \mathbb{R}^d,\ \ e \in \{1, 2\}. \tag{14}$$

Step 3: Reduce to an orthogonal conjugation constraint.

Use (14) with $e = 1$. Since Σ_1 and $\tilde{\Sigma}_1$ are diagonal positive definite by (S3), their square roots $\Sigma_1^{1/2}$ and $\tilde{\Sigma}_1^{1/2}$ are well-defined diagonal positive definite. Define

$$M(z) := \tilde{\Sigma}_1^{-1/2}\, D\varphi(z)\, \Sigma_1^{1/2}.$$

Then $M(z) M(z)^\top = I$, so $M(z) \in \mathbb{R}^{d \times d}$ is an orthogonal matrix for all $z \in \mathbb{R}^d$. Using (14) with $e = 2$ yields

$$M(z)\, K\, M(z)^\top = \tilde{K}, \tag{15}$$

where

$$K := \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} = \operatorname{diag}\Big(\frac{\sigma_{2,1}^2}{\sigma_{1,1}^2}, \ldots, \frac{\sigma_{2,d}^2}{\sigma_{1,d}^2}\Big), \qquad \tilde{K} := \tilde{\Sigma}_1^{-1/2} \tilde{\Sigma}_2 \tilde{\Sigma}_1^{-1/2},$$

and $\tilde{K}$ is diagonal positive definite. By (S4), K has simple spectrum.

Step 4: Conclude $M(z)$ is a signed permutation, hence $D\varphi(z)$ is monomial.

Because $M(z) \in \mathbb{R}^{d \times d}$ is an orthogonal matrix and $M(z) K M(z)^\top$ is diagonal, Lemma C.1 implies that $M(z)$ is a signed permutation matrix for each $z \in \mathbb{R}^d$. Hence

$$D\varphi(z) = \tilde{\Sigma}_1^{1/2} M(z) \Sigma_1^{-1/2}$$

is monomial for each $z \in \mathbb{R}^d$.

Step 5: Connectedness forces $D\varphi$ to be constant on $\mathbb{R}^d$.

The map $z \mapsto M(z)$ is continuous on $\mathbb{R}^d$ and takes values in a finite (discrete) set (signed permutations), so it must be constant on connected $\mathbb{R}^d$. Thus $M(z) \equiv M_0$ for a fixed signed permutation M_0, and therefore

$$D\varphi(z) \equiv L := \tilde{\Sigma}_1^{1/2} M_0 \Sigma_1^{-1/2} \qquad \forall z \in \mathbb{R}^d.$$

Since $\Sigma_1^{1/2}$ and $\tilde{\Sigma}_1^{1/2}$ are diagonal and M_0 is a signed permutation, L is a constant monomial matrix.

Step 6: Constant Jacobian implies φ is affine; identify $\tilde{A}$.

Since $D\varphi \equiv L$ on $\mathbb{R}^d$, φ is affine:

$$\varphi(z) = Lz + c, \qquad z \in \mathbb{R}^d,$$

for some constant $c \in \mathbb{R}^d$.

Fix a regime e. Since φ is a global C^2 diffeomorphism by (S1), Itô's formula applied to $\tilde{z}_t^{(e)} = \varphi(z_t^{(e)})$ gives

$$\begin{aligned} d\tilde{z}_t^{(e)} &= D\varphi(z_t^{(e)})\, dz_t^{(e)} + \frac{1}{2} \sum_{k,\ell=1}^{d} (\Sigma_e)_{k\ell} \partial_{k\ell} \varphi(z_t^{(e)})\, dt \\ &= \Big(D\varphi(z_t^{(e)}) A z_t^{(e)} + \tfrac{1}{2}\, \Sigma_e : D^2 \varphi(z_t^{(e)}) \Big)\, dt + D\varphi(z_t^{(e)}) G_e\, dW_t^{(e)}, \end{aligned}$$

where $\Sigma_e : D^2\varphi$ denotes $\sum_{k,\ell} (\Sigma_e)_{k\ell} \partial_{k\ell} \varphi$.

On the other hand, the alternative OU model gives

$$d\tilde{z}_t^{(e)} = \tilde{A} \tilde{z}_t^{(e)}\, dt + \tilde{G}_e\, d\tilde{W}_t^{(e)}.$$

By Lemma C.2, we may replace $G_e\, dW_t^{(e)}$ in the Itô expansion with $\Sigma_e^{1/2}\, d\hat{W}_t^{(e)}$, where $\hat{W}^{(e)}$ is a standard $\mathcal{F}_t^{z^{(e)}}$-Brownian motion. The noise term in the expansion thus becomes $D\varphi(z_t^{(e)}) \Sigma_e^{1/2}\, d\hat{W}_t^{(e)}$, a continuous $\mathcal{F}_t^{z^{(e)}}$-local martingale. Likewise, Lemma C.2 applied to the alternative model gives $d\tilde{z}_t^{(e)} = \tilde{A} \tilde{z}_t^{(e)}\, dt + \tilde{\Sigma}_e^{1/2}\, d\hat{\tilde{W}}_t^{(e)}$ with $\hat{\tilde{W}}^{(e)}$ an $\mathcal{F}_t^{\tilde{z}^{(e)}}$-Brownian motion. By Remark C.3, $\mathcal{F}_t^{z^{(e)}} = \mathcal{F}_t^{\tilde{z}^{(e)}}$ since $\tilde{z}^{(e)} = \varphi(z^{(e)})$ with φ a diffeomorphism, so both noise integrals are continuous local martingales in the common natural filtration. Subtracting the two representations of $d\tilde{z}^{(e)}$, the difference has the form $\int_0^t H(s)\, ds + M_t$, where $\int_0^t H(s)\, ds$ is of finite variation and M_t is a continuous local martingale in the common filtration. A continuous process that is both of finite variation and a local martingale must be constant; since it starts at 0, it is identically 0. Hence $H(t) = 0$ for Lebesgue-a.e. $t \in [0, T]$, almost surely. Comparing drift terms yields

$$\tilde{A}\, \tilde{z}_t^{(e)} = D\varphi(z_t^{(e)}) A z_t^{(e)} + \tfrac{1}{2}\, \Sigma_e : D^2 \varphi(z_t^{(e)}) \quad \text{for a.e. } t \text{ and a.s.}$$

Define

$$H(t) := \tilde{A}\, \tilde{z}_t^{(e)} - \Big(D\varphi(z_t^{(e)}) A z_t^{(e)} + \tfrac{1}{2}\, \Sigma_e : D^2 \varphi(z_t^{(e)}) \Big).$$

We have shown that $H(t) = 0$ for a.e. $t \in [0, T]$, a.s. Since $t \mapsto z_t^{(e)}$ and $t \mapsto \tilde{z}_t^{(e)}$ are a.s. continuous and $\varphi \in C^2$, the map $t \mapsto H(t)$ is a.s. continuous. Hence $H(t) = 0$ for all $t \in [0, T]$, a.s. In particular, fix any $t_\star \in (0, T]$; the identity holds at $t = t_\star$.

Since $D\varphi \equiv L$ and $D^2\varphi \equiv 0$ on $\mathbb{R}^d$ (from Steps 4–5), evaluating at $t = t_\star$ gives

$$\tilde{A}\, \tilde{z}_{t_\star}^{(e)} = L A z_{t_\star}^{(e)} \quad \text{a.s.}$$

Since $\tilde{z}_{t_\star}^{(e)} = \varphi(z_{t_\star}^{(e)}) = L z_{t_\star}^{(e)} + c$, this becomes

$$\tilde{A}(L z_{t_\star}^{(e)} + c) = L A z_{t_\star}^{(e)} \quad \text{a.s.}$$

We now upgrade this to a pointwise identity on $\mathbb{R}^d$. Define the continuous function $F(z) := \tilde{A}(Lz + c) - LAz$ on $\mathbb{R}^d$. We have shown $F(z_{t_\star}^{(e)}) = 0$ a.s. If $F(z_0) \neq 0$ for some $z_0 \in \mathbb{R}^d$, then by

continuity there exists an open set $V \ni z_0$ with $\|F(z)\| > 0$ for all $z \in V$, contradicting $F(z_{t_\star}^{(e)}) = 0$ a.s. since $z_{t_\star}^{(e)}$ has a strictly positive density on $\mathbb{R}^d$ (as shown in Step 2). Therefore,

$$\tilde{A}(Lz + c) = LAz \qquad \forall z \in \mathbb{R}^d.$$

Expanding the left-hand side gives $\tilde{A}Lz + \tilde{A}c = LAz$ for all $z \in \mathbb{R}^d$. Matching coefficients of z and constant terms yields

$$\tilde{A}L = LA, \qquad \tilde{A}c = 0.$$

Finally, $\tilde{\Sigma}_e = L\Sigma_e L^\top$ follows from (14) with $D\varphi \equiv L$. Since $L = \Pi\Lambda$ with Λ invertible diagonal,

$$\tilde{A} = LAL^{-1} = \Pi\big(\Lambda A \Lambda^{-1}\big)\Pi^{-1}.$$

Left and right multiplication by invertible diagonal matrices preserves the zero pattern of A, while conjugation by Π relabels coordinates by the permutation π. Therefore

$$(j \to i) \in E_A \quad \Longleftrightarrow \quad (\pi(j) \to \pi(i)) \in E_{\tilde{A}},$$

which proves the graph-support statement. □

C.3 Open-set positivity for additive-noise SDEs

Lemma C.4 (Open-set positivity under nondegenerate additive noise)**.** *Let $(z_t)_{t\in[0,T]}$ be a weak solution with continuous sample paths to the SDE*

$$dz_t = f(z_t)\,dt + \Sigma^{1/2}\,dW_t,$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is continuous, $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite, and the initial condition z_0 has an arbitrary law μ on $\mathbb{R}^d$. Then for every $t \in (0, T]$ and every nonempty open set $V \subset \mathbb{R}^d$,

$$\mathbb{P}(z_t \in V) > 0.$$

Equivalently, the law of z_t charges every nonempty open subset of $\mathbb{R}^d$.

Proof. Fix $t \in (0, T]$ and a nonempty open set $V \subset \mathbb{R}^d$. We first prove the claim for a deterministic initial condition, and then pass to an arbitrary initial law by conditioning on z_0.

Step 1: deterministic initial condition. Fix $x \in \mathbb{R}^d$ and suppose first that $z_0 = x$ almost surely. Pick $y \in V$ and choose $r > 0$ such that $B(y, r) \subset V$.

Let $(\mathcal{F}_s)_{0\le s\le T}$ denote the filtration on the underlying weak-solution probability space. Choose $R > 0$ so large that $x \in B(0, R)$ and $B(y, r) \subset B(0, R)$. Define the exit time

$$\tau_R := \inf\big\{s \in [0, T] : \|z_s\| \ge R\big\},$$

with the convention $\inf \varnothing = \infty$. Since z has continuous adapted paths and starts from $x \in B(0, R)$, we have $\tau_R > 0$ almost surely.

Because f is continuous, it is bounded on the compact set $\overline{B(0, R)}$, hence

$$M_R := \sup_{\|u\|\le R} \big\|\Sigma^{-1/2} f(u)\big\| < \infty.$$

Define

$$\theta_s := \Sigma^{-1/2} f(z_s)\,\mathbf{1}_{\{s\le\tau_R\}}, \qquad 0 \le s \le t.$$

Then θ is progressively measurable and

$$\int_0^t \|\theta_s\|^2\,ds \le M_R^2 t \qquad \text{a.s.}$$

Therefore Novikov's condition holds:

$$\mathbb{E}\Big[\exp\Big(\frac{1}{2}\int_0^t \|\theta_s\|^2\,ds\Big)\Big] \le e^{M_R^2 t/2} < \infty.$$

Define

$$\Lambda_t := \exp\Big(- \int_0^t \theta_s^\top \, dW_s - \frac{1}{2} \int_0^t \|\theta_s\|^2 \, ds \Big), \qquad \left. \frac{d\mathbb{Q}}{d\mathbb{P}} \right|_{\mathcal{F}_t} := \Lambda_t.$$

Then $\Lambda_t > 0$ almost surely and $\mathbb{E}_{\mathbb{P}}[\Lambda_t] = 1$, so $\mathbb{P}$ and $\mathbb{Q}$ are equivalent on $\mathcal{F}_t$.

By Girsanov's theorem, the process

$$\widetilde{W}_s := W_s + \int_0^s \theta_u \, du, \qquad 0 \le s \le t,$$

is a standard d-dimensional Brownian motion under $\mathbb{Q}$. Moreover, for every $s \in [0,t]$,

$$\begin{aligned} z_{s\wedge\tau_R} &= x + \int_0^{s\wedge\tau_R} f(z_u) \, du + \Sigma^{1/2} W_{s\wedge\tau_R} \\ &= x + \int_0^s f(z_u) \, \mathbf{1}_{\{u \le \tau_R\}} \, du + \Sigma^{1/2} \Big(\widetilde{W}_{s\wedge\tau_R} - \int_0^{s\wedge\tau_R} \theta_u \, du \Big) \\ &= x + \Sigma^{1/2} \widetilde{W}_{s\wedge\tau_R}, \end{aligned}$$

since $\Sigma^{1/2}\theta_u = f(z_u)\,\mathbf{1}_{\{u\le\tau_R\}}$ for Lebesgue-a.e. u.

Now define

$$X_s := x + \Sigma^{1/2} \widetilde{W}_s, \qquad 0 \le s \le t.$$

Under $\mathbb{Q}$, the law of $\widetilde{W}$ on $C_0([0,t];\mathbb{R}^d) := \{\eta \in C([0,t];\mathbb{R}^d) : \eta(0) = 0\}$ is Wiener measure, which has full support on $C_0([0,t];\mathbb{R}^d)$. Since

$$X = \Psi(\widetilde{W}), \qquad \Psi(\eta)(s) := x + \Sigma^{1/2}\eta(s),$$

and Ψ is a homeomorphism from $C_0([0,t];\mathbb{R}^d)$ onto

$$C_x([0,t];\mathbb{R}^d) := \{\eta \in C([0,t];\mathbb{R}^d) : \eta(0) = x\},$$

the law of X under $\mathbb{Q}$ has full support on $C_x([0,t];\mathbb{R}^d)$.

Consider the straight-line path

$$\gamma(s) := x + \frac{s}{t}(y - x), \qquad s \in [0,t].$$

Because $B(0,R)$ is convex and $x, y \in B(0,R)$, we have $\gamma([0,t]) \subset B(0,R)$. Since $\gamma([0,t])$ is compact and $B(0,R)$ is open, there exists $\delta > 0$ such that

$$\delta < r \qquad \text{and} \qquad \sup_{0\le s\le t} \|\gamma(s)\| + \delta < R.$$

Define the event

$$E := \Big\{ \sup_{0\le s\le t} \|X_s - \gamma(s)\| < \delta \Big\}.$$

Then $E \in \mathcal{F}_t$, and by the support statement above,

$$\mathbb{Q}(E) > 0.$$

We claim that

$$E \subset \{z_t \in B(y,r)\} \subset \{z_t \in V\}.$$

Indeed, fix $\omega \in E$. Then for every $s \in [0,t]$,

$$\|X_s(\omega)\| \le \|\gamma(s)\| + \delta < R.$$

If $\tau_R(\omega) \le t$, then by continuity of z and the definition of τ_R,

$$\|z_{\tau_R(\omega)}(\omega)\| = R.$$

But from the stopped identity above,

$$z_{\tau_R(\omega)}(\omega) = z_{\tau_R(\omega)\wedge\tau_R(\omega)}(\omega) = X_{\tau_R(\omega)}(\omega),$$

which contradicts $\|X_{\tau_R(\omega)}(\omega)\| < R$. Hence $\tau_R(\omega) > t$. Therefore

$$z_t(\omega) = z_{t\wedge\tau_R}(\omega) = X_t(\omega),$$

and so

$$\|z_t(\omega) - y\| = \|X_t(\omega) - \gamma(t)\| < \delta < r.$$

Thus $z_t(\omega) \in B(y, r) \subset V$, proving the claim.

Since $\mathbb{P}$ and $\mathbb{Q}$ are equivalent on $\mathcal{F}_t$,

$$\mathbb{P}(z_t \in V) > 0$$

whenever $z_0 = x$ almost surely. This proves the deterministic-start claim.

Step 2: arbitrary initial law. Now return to the original weak solution with arbitrary initial law $\mu = \mathcal{L}(z_0)$. Realize the weak solution on the canonical product path space

$$\Omega := C([0,T];\mathbb{R}^d) \times C([0,T];\mathbb{R}^d),$$

with canonical coordinates (Z, W) and raw canonical filtration

$$\mathcal{G}_t := \sigma(Z_s, W_s : 0 \le s \le t), \qquad 0 \le t \le T.$$

Let $\mathbf{P}$ denote the joint law of (z, W) on Ω. Since Ω and $\mathbb{R}^d$ are Polish, there exists a regular conditional law $(\mathbf{P}_x)_{x\in\mathbb{R}^d}$ of (Z, W) given $Z_0 = x$.

We first identify a μ-full set of initial states on which the canonical pair solves the same SDE with deterministic start. By definition of the conditional law, there exists a Borel set $E_0 \subset \mathbb{R}^d$ with $\mu(E_0) = 1$ such that for every $x \in E_0$,

$$Z_0 = x \qquad \mathbf{P}_x\text{-a.s.}$$

Since $W_0 = 0$ $\mathbf{P}$-a.s., after shrinking E_0 if needed we may also assume

$$W_0 = 0 \qquad \mathbf{P}_x\text{-a.s.}$$

for every $x \in E_0$.

Let A_{SDE} be the measurable set of path pairs on which

$$Z_t = Z_0 + \int_0^t f(Z_s)\,ds + \Sigma^{1/2} W_t$$

holds for all rational $t \in [0, T]$. Since this identity holds $\mathbf{P}$-a.s., there exists a Borel set $E_1 \subset \mathbb{R}^d$ with $\mu(E_1) = 1$ such that $\mathbf{P}_x(A_{\mathrm{SDE}}) = 1$ for all $x \in E_1$. By continuity of Z and W, the integral identity then holds for all $t \in [0, T]$, $\mathbf{P}_x$-a.s., for every $x \in E_1$.

It remains to verify the Brownian property. For each rational $0 \le s < t \le T$, let $\mathcal{C}_s$ be a countable π-system generating $\mathcal{G}_s$ (for example, finite cylinder sets at rational times with rational rectangles), and let $\mathcal{I}$ be the countable π-system of rectangles with rational endpoints generating $\mathcal{B}(\mathbb{R}^d)$. Under $\mathbf{P}$, the canonical second coordinate W is a Brownian motion with respect to $(\mathcal{G}_t)$: this follows because on the original filtered space the Brownian increment is independent of the underlying filtration, hence in particular independent of the smaller sigma-field generated by the past of (z, W). Therefore, for every rational $s < t$, every $A \in \mathcal{C}_s$, every $B \in \mathcal{I}$, and every bounded Borel function $\phi : \mathbb{R}^d \to \mathbb{R}$,

$$\begin{aligned}
\int_{\mathbb{R}^d} \phi(x)\,\mathbf{P}_x\big(A \cap \{W_t - W_s \in B\}\big)\,\mu(dx) &= \mathbb{E}_{\mathbf{P}}\big[\phi(Z_0)\,\mathbf{1}_A\,\mathbf{1}_{\{W_t - W_s \in B\}}\big] \\
&= \mathbb{E}_{\mathbf{P}}\big[\phi(Z_0)\,\mathbf{1}_A\big]\,\gamma_{t-s}(B) \\
&= \int_{\mathbb{R}^d} \phi(x)\,\mathbf{P}_x(A)\,\gamma_{t-s}(B)\,\mu(dx),
\end{aligned}$$

where $\gamma_{t-s} = N(0, (t-s)I_d)$. Hence, for each fixed rational $s < t$, $A \in \mathcal{C}_s$, and $B \in \mathcal{I}$, there exists a Borel set $E_{s,t,A,B} \subset \mathbb{R}^d$ with $\mu(E_{s,t,A,B}) = 1$ such that

$$\mathbf{P}_x\big(A \cap \{W_t - W_s \in B\}\big) = \mathbf{P}_x(A)\,\gamma_{t-s}(B) \qquad \forall x \in E_{s,t,A,B}.$$

Intersecting over the countable family of such tuples gives a Borel set $E_2 \subset \mathbb{R}^d$ with $\mu(E_2) = 1$ such that for every $x \in E_2$ and every rational $0 \le s < t \le T$, the increment $W_t - W_s$ under $\mathbf{P}_x$ is independent of $\mathcal{G}_s$ and has law γ_{t-s}. By a monotone-class argument, this extends from $\mathcal{C}_s$ and $\mathcal{I}$ to all $A \in \mathcal{G}_s$ and all Borel $B \subset \mathbb{R}^d$. Since W has continuous paths, the Brownian property extends from rational times to all real times. Hence, for every $x \in E_2$, W is a Brownian motion with respect to the raw canonical filtration under $\mathbf{P}_x$, and therefore also with respect to its $\mathbf{P}_x$-usual augmentation.

Set $E := E_0 \cap E_1 \cap E_2$. Then $\mu(E) = 1$, and for every $x \in E$ the pair (Z, W) is a weak solution of the same SDE with deterministic initial condition x. Applying Step 1 under $\mathbf{P}_x$ gives

$$\mathbf{P}_x(Z_t \in V) > 0 \qquad \text{for every } x \in E.$$

Integrating over the initial law,

$$\mathbb{P}(z_t \in V) = \mathbf{P}(Z_t \in V) = \int_{\mathbb{R}^d} \mathbf{P}_x(Z_t \in V)\, \mu(dx) \ge \int_E \mathbf{P}_x(Z_t \in V)\, \mu(dx) > 0.$$

The last inequality holds because the integrand is measurable and strictly positive on the full-measure set E. This proves the lemma. □

C.4 Proof of Theorem 3.2 and Corollary 3.4

Proof of Theorem 3.2. The proof is arranged to reuse the OU argument as much as possible. More precisely, Steps 0, 1, 3, 4, and 5 of the proof of Theorem 3.1, together with the implication "constant Jacobian ⇒ affine map" at the start of Step 6 there, depend only on (S1)–(S4) and on having a pointwise version of the quadratic-variation constraint. They do not use the linear form of the drift. We therefore record only the places where the nonlinear setting requires a modified argument.

Step 0: Coupling and the along-path covariance constraint.

By exactly the same coupling construction as in Step 0 of the proof of Theorem 3.1, now applied to (3a) and (3b), we may realize the two observed path laws on a common canonical path space and obtain

$$\tilde{z}_t^{(e)} = \varphi(z_t^{(e)}) \qquad \text{a.s. for all } t \in [0,T],\ e \in \{1,2\}.$$

By (S5), the latent paths are weak solutions with a.s. continuous sample paths; as SDE solutions, they are continuous semimartingales. Hence Itô's formula and quadratic variation apply exactly as in the OU proof. Repeating Step 1 of Theorem 3.1 word for word, we obtain for each regime e:

$$d\langle \tilde{z}^{(e)} \rangle_t = D\varphi(z_t^{(e)}) \Sigma_e D\varphi(z_t^{(e)})^\top\, dt, \qquad \Sigma_e := G_e G_e^\top, \tag{16}$$

while under the alternative model,

$$d\langle \tilde{z}^{(e)} \rangle_t = \tilde{\Sigma}_e\, dt, \qquad \tilde{\Sigma}_e := \tilde{G}_e \tilde{G}_e^\top. \tag{17}$$

Therefore, with the same continuity-in-time upgrade used in the OU proof,

$$D\varphi(z_t^{(e)}) \Sigma_e D\varphi(z_t^{(e)})^\top = \tilde{\Sigma}_e \qquad \text{a.s. for all } t \in [0,T]. \tag{18}$$

Step 1: Full support upgrades the along-path identity to a pointwise identity on $\mathbb{R}^d$.

Fix $e \in \{1,2\}$ and any $t_\star \in (0,T]$. Define the continuous matrix-valued function on $\mathbb{R}^d$ by

$$B_e(z) := D\varphi(z) \Sigma_e D\varphi(z)^\top - \tilde{\Sigma}_e.$$

Evaluating (18) at $t = t_\star$ gives $B_e(z_{t_\star}^{(e)}) = 0$ a.s.

We claim $B_e \equiv 0$ on $\mathbb{R}^d$. Suppose not. Then there exists $z_0 \in \mathbb{R}^d$ with $B_e(z_0) \neq 0$. By continuity, there exist $\varepsilon > 0$ and an open neighborhood V of z_0 such that $\|B_e(z)\|_F > \varepsilon$ for all $z \in V$. By Lemma C.4—proved by first establishing positivity conditional on a deterministic initial state and then integrating over the law of $z_0^{(e)}$—we have $\mathbb{P}(z_{t_\star}^{(e)} \in V) > 0$, contradicting $B_e(z_{t_\star}^{(e)}) = 0$ a.s. Hence

$$D\varphi(z)\, \Sigma_e\, D\varphi(z)^\top = \tilde{\Sigma}_e \qquad \forall z \in \mathbb{R}^d,\ \forall e \in \{1,2\}. \tag{19}$$

Step 2: Reuse the OU monomial-identifiability core.

Once (19) is available, Steps 3, 4, and 5 of the proof of Theorem 3.1 apply verbatim. Define

$$M(z) := \tilde{\Sigma}_1^{-1/2} D\varphi(z) \Sigma_1^{1/2}. \tag{20}$$

Then $M(z)$ is orthogonal for every $z \in \mathbb{R}^d$, and using regime $e = 2$ gives

$$M(z)\, K\, M(z)^\top = \tilde{K}, \qquad K := \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}, \quad \tilde{K} := \tilde{\Sigma}_1^{-1/2}\tilde{\Sigma}_2\tilde{\Sigma}_1^{-1/2}. \tag{21}$$

By (S3), $\tilde{K}$ is diagonal, and by (S4), K has simple spectrum. The same whitening/simple-spectrum argument as in the OU proof therefore shows that $M(z)$ is a constant signed permutation matrix. Consequently,

$$D\varphi(z) \equiv L := \tilde{\Sigma}_1^{1/2} M_0 \Sigma_1^{-1/2} \qquad \forall z \in \mathbb{R}^d, \tag{22}$$

for a constant monomial matrix $L = \Pi\Lambda$. Exactly as at the start of Step 6 in the OU proof, $D\varphi \equiv L$ implies that $\varphi(z) = Lz + c$ for some constant $c \in \mathbb{R}^d$. Substituting $D\varphi \equiv L$ into (19) gives

$$\tilde{\Sigma}_e = L\Sigma_e L^\top \qquad \text{for each } e \in \{1, 2\}.$$

Step 3: Identify the drift transformation.

This step parallels the latter half of Step 6 in the OU proof, with Az and $\tilde{A}\tilde{z}$ replaced by the general drifts $f(z)$ and $\tilde{f}(\tilde{z})$. Fix a regime e. Since φ is a global C^2 diffeomorphism by (S1), applying Itô's formula to $\tilde{z}_t^{(e)} = \varphi(z_t^{(e)})$ yields

$$\begin{aligned} d\tilde{z}_t^{(e)} &= D\varphi(z_t^{(e)})\, dz_t^{(e)} + \frac{1}{2}\sum_{k,\ell=1}^{d} (\Sigma_e)_{k\ell}\partial_{k\ell}\varphi(z_t^{(e)})\, dt \\ &= \Big(D\varphi(z_t^{(e)}) f(z_t^{(e)}) + \tfrac{1}{2}\,\Sigma_e : D^2\varphi(z_t^{(e)})\Big)\, dt + D\varphi(z_t^{(e)}) G_e\, dW_t^{(e)}. \end{aligned}$$

Under the alternative model (3b),

$$d\tilde{z}_t^{(e)} = \tilde{f}(\tilde{z}_t^{(e)})\, dt + \tilde{G}_e\, d\tilde{W}_t^{(e)}.$$

To justify comparing drift terms, we use Lemma C.2: replacing $G_e\, dW_t^{(e)}$ with $\Sigma_e^{1/2}\, d\hat{W}_t^{(e)}$ (where $\hat{W}^{(e)}$ is a standard $\mathcal{F}_t^{z^{(e)}}$-Brownian motion), the noise term in the Itô expansion becomes $D\varphi(z_t^{(e)})\Sigma_e^{1/2}\, d\hat{W}_t^{(e)}$, a continuous $\mathcal{F}_t^{z^{(e)}}$-local martingale. Likewise, the alternative model's innovation representation yields a noise integral that is an $\mathcal{F}_t^{\tilde{z}^{(e)}}$-local martingale. By Remark C.3, these filtrations coincide since $\tilde{z}^{(e)} = \varphi(z^{(e)})$ with φ a diffeomorphism. Subtracting the two representations of $d\tilde{z}^{(e)}$, the difference has the form $\int_0^t H(s)\, ds + M_t$, where $\int_0^t H(s)\, ds$ is of finite variation and M_t is a continuous local martingale in the common filtration. A continuous process that is both of finite variation and a local martingale must be constant; since it starts at 0, it is identically 0. Hence $H(t) = 0$ for Lebesgue-a.e. $t \in [0, T]$, almost surely. Therefore,

$$\tilde{f}(\tilde{z}_t^{(e)}) = D\varphi(z_t^{(e)}) f(z_t^{(e)}) + \tfrac{1}{2}\,\Sigma_e : D^2\varphi(z_t^{(e)}) \quad \text{for a.e. } t \text{ and a.s.}$$

Define

$$H(t) := \tilde{f}(\tilde{z}_t^{(e)}) - \Big(D\varphi(z_t^{(e)}) f(z_t^{(e)}) + \tfrac{1}{2}\,\Sigma_e : D^2\varphi(z_t^{(e)})\Big).$$

As above, $H(t) = 0$ for a.e. t a.s., and H is a.s. continuous in t. Therefore $H(t) = 0$ for all $t \in [0, T]$ a.s. In particular, fix any $t_\star \in (0, T]$; the identity holds at $t = t_\star$.

Since $D\varphi \equiv L$ and $D^2\varphi \equiv 0$ on $\mathbb{R}^d$ (from Step 2), evaluating at $t = t_\star$ gives

$$\tilde{f}(\varphi(z_{t_\star}^{(e)})) = L f(z_{t_\star}^{(e)}) \quad \text{a.s.}$$

Since $\varphi(z) = Lz + c$ for all $z \in \mathbb{R}^d$, this becomes

$$\tilde{f}(Lz_{t_\star}^{(e)} + c) = L f(z_{t_\star}^{(e)}) \quad \text{a.s.}$$

We now upgrade this almost-sure equality to a pointwise identity on $\mathbb{R}^d$. Define the continuous function $F(z) := \tilde{f}(Lz + c) - Lf(z)$ on $\mathbb{R}^d$ (continuous by (S5)). We have shown $F(z_{t_\star}^{(e)}) = 0$ a.s. Suppose $F(z_0) \neq 0$ for some $z_0 \in \mathbb{R}^d$. By continuity of F, there exist $\varepsilon > 0$ and an open neighborhood V of z_0 such that $\|F(z)\| > \varepsilon$ for all $z \in V$. Applying Lemma C.4 once more (again via conditioning on the initial state $z_0^{(e)}$ and integrating over its law) gives $\mathbb{P}(z_{t_\star}^{(e)} \in V) > 0$, contradicting $F(z_{t_\star}^{(e)}) = 0$ a.s. Therefore $F \equiv 0$ on $\mathbb{R}^d$, i.e.

$$\tilde{f}(Lz + c) = Lf(z) \qquad \forall z \in \mathbb{R}^d.$$

□

Proof of Corollary 3.4. Assume f and $\tilde{f}$ are C^1 on $\mathbb{R}^d$. From Theorem 3.2 we have, for all $z \in \mathbb{R}^d$,

$$\tilde{f}(\tilde{z}) = Lf(z), \qquad \text{where } \tilde{z} := \varphi(z) = Lz + c.$$

Equivalently, using $z = L^{-1}(\tilde{z} - c)$, we may write

$$\tilde{f}(\tilde{z}) = L\, f\big(L^{-1}(\tilde{z} - c)\big), \qquad \forall\, \tilde{z} \in \mathbb{R}^d.$$

Differentiating with respect to $\tilde{z}$ and applying the chain rule yields, for all $\tilde{z} \in \mathbb{R}^d$,

$$D\tilde{f}(\tilde{z}) = L\, Df\big(L^{-1}(\tilde{z} - c)\big)\, L^{-1}.$$

Equivalently, writing $\tilde{z} = Lz + c$, we have

$$D\tilde{f}(Lz + c) = L\, Df(z)\, L^{-1}, \qquad \forall z \in \mathbb{R}^d.$$

Write $L = \Pi\Lambda$ with Π a permutation matrix and Λ an invertible diagonal matrix. Then

$$L\, Df(z)\, L^{-1} = \Pi\big(\Lambda\, Df(z)\, \Lambda^{-1}\big)\Pi^{-1}.$$

Left/right multiplication by an invertible diagonal matrix preserves the zero pattern, hence $\Lambda Df(z)\Lambda^{-1}$ has the same support as $Df(z)$. Therefore the Jacobian edge set is permuted by Π:

$$(\partial_j f_i(z) \neq 0) \quad \Longleftrightarrow \quad \big(\partial_{\pi(j)} \tilde{f}_{\pi(i)}(\tilde{z}) \neq 0\big), \qquad \tilde{z} = Lz + c.$$

Equivalently,

$$(j \to i) \in E_f(z) \quad \Longleftrightarrow \quad (\pi(j) \to \pi(i)) \in E_{\tilde{f}}(Lz + c).$$

This proves the claimed Jacobian-graph identifiability. □

C.5 Why weak well-posedness suffices

Remark C.5 (Why weak well-posedness suffices). (S5) requires only *weak* well-posedness of the SDE (existence of a weak solution that is unique in law), rather than the stronger notion of a *strong* (pathwise) solution. This distinction is worth clarifying.

A **strong solution** is constructed on a given probability space with a given Brownian motion W_t: one seeks a process z_t that is adapted to the filtration of W and satisfies the integral equation

$$z_t = z_0 + \int_0^t f(z_s)\, ds + \int_0^t G_e\, dW_s.$$

Uniqueness is *pathwise*: for each sample path of W, the solution $z_t(\omega)$ is uniquely determined. A standard sufficient condition for strong existence and pathwise uniqueness is global Lipschitz continuity, together with a linear-growth condition.

A **weak solution** allows freedom to choose the probability space and the driving Brownian motion: one seeks some probability space supporting both a Brownian motion W_t and a process z_t that together satisfy the SDE. Uniqueness is *in law*: any two weak solutions induce the same law on path space $C([0, T]; \mathbb{R}^d)$. Weak existence can hold under much milder conditions than strong existence; see, for example, [40]. In this paper we directly assume weak well-posedness in (S5) and do not rely on a specific existence theorem.

The identifiability argument in Theorem 3.2 depends only on the *distributional* properties of the observed trajectories:

- (S2) compares the laws of $\{x_t^{(e)}\}$ and $\{\tilde{x}_t^{(e)}\}$;
- the coupling in Step 0 is a distributional construction;
- Itô's formula and quadratic variation apply to any continuous semimartingale, including weak solutions;
- the full-support argument in Steps 1 and 3 uses the law of $z_{t_\star}^{(e)}$, not its pathwise construction.

Requiring strong solutions would unnecessarily restrict the class of admissible drifts without strengthening the identifiability conclusion.

D Estimation algorithm

Algorithm 1 summarizes the two-stage estimator used in the experiments. Stage 1 fits the shared-drift, regime-specific diagonal-diffusion transition model in the learned coordinate system. Stage 2 is optional and is used only when a sparse drift-Jacobian graph estimate is needed.

Algorithm 1 Two-stage latent SDE estimation with optional sparse drift-Jacobian graph recovery

Require: Multi-regime trajectories $\{x_{t_0}^{(e)}, \dots, x_{t_N}^{(e)}\}_{e=1}^{2}$; time step Δt; stride s; hyperparameter λ_{sparse}; threshold τ

Ensure: Encoder h_θ; learned drift model $\tilde{f}_\psi$; optional drift-Jacobian graph estimate $\hat{E}_f$

Stage 1: short-time transition fitting in learned coordinates
Initialize encoder h_θ, drift model $\tilde{f}_\psi$, and diffusion parameters $\{\log \tilde{\sigma}_{e,i}^2\}_{e,i}$
for optimization step $k = 1, \dots, T_1$ **do**
  for $e = 1, 2$ **do**
    Sample mini-batch $(x_t, x_{t+\Delta t}) \sim \mathcal{P}_e$
    Encode $\tilde{z}_t \leftarrow h_\theta(x_t)$ and $\tilde{z}_{t+\Delta t} \leftarrow h_\theta(x_{t+\Delta t})$
    Compute the per-regime transition score from (6)
  end for
  Update $(\theta, \psi, \{\log \tilde{\sigma}_{e,i}^2\})$ using $\nabla \mathcal{L}_{\text{S1}}$
end for

Optional Stage 2: sparse drift-Jacobian graph recovery
Freeze h_θ and $\{\tilde{\sigma}_{e,i}^2\}$; reinitialize $\tilde{f}_\psi$
Compute velocity targets $\hat{v}_t \leftarrow \big(h_\theta(x_{t+s\Delta t}) - h_\theta(x_t)\big)/(s\Delta t)$
for optimization step $k = 1, \dots, T_2$ **do**
  Sample mini-batches $(\tilde{z}_t, \hat{v}_t)$ from both regimes
  Update ψ using $\nabla \mathcal{L}_{\text{S2}}$
end for
Estimate $\overline{|D\tilde{f}_\psi|}_{ij} \leftarrow \mathbb{E}_{\tilde{z} \sim \mathcal{Z}}\big[|\partial \tilde{f}_{\psi,i}/\partial \tilde{z}_j(\tilde{z})|\big]$
Return the thresholded drift-Jacobian edge set $\hat{E}_f = \{(j \to i) : \overline{|D\tilde{f}_\psi|}_{ij} > \tau\}$ if graph recovery is desired

E Simulation setup

This section specifies the data-generating process, observation model, training procedure, evaluation metrics, and hyperparameters for all simulation experiments.

E.1 Latent process

Each experiment generates N_{traj} independent trajectories from a d-dimensional latent stochastic differential equation

$$\mathrm{d}z_t^{(e)} = f(z_t^{(e)})\,\mathrm{d}t + \Sigma_e^{1/2}\,\mathrm{d}W_t^{(e)},$$

where $f\colon \mathbb{R}^d \to \mathbb{R}^d$ is the latent drift, $\Sigma_e = \mathrm{diag}(\sigma_{e,1}^2, \ldots, \sigma_{e,d}^2)$ is the environment-specific diagonal diffusion covariance, and $W_t^{(e)}$ is a standard d-dimensional Brownian motion. Trajectories are discretized via the Euler–Maruyama scheme with step size Δt. In the implementation, T is a nominal horizon parameter and $n_{\text{steps}} = \lfloor T/\Delta t \rfloor = 20$ points are stored including the initial condition. Thus the last stored time is $(n_{\text{steps}} - 1)\Delta t$ and each trajectory yields 19 adjacent one-step transition pairs. Initial conditions are drawn i.i.d. from $\mathcal{N}(0,\, 0.25\, I_d)$.

We consider three drift families: dense linear OU, sparse linear OU, and nonlinear sparse drift, whose constructions are detailed in Sections F.1 to F.3. Within each drift family and dimension, the drift f is fixed and shared across all regime conditions and all training seeds.

E.2 Regime conditions and diffusion covariances

Each drift family is evaluated under three regime conditions. These conditions vary only the number and choice of diagonal diffusion covariances; all other components are held fixed. In all conditions, the first environment uses

$$\sigma_{1,j}^2 = 0.5\, j, \qquad j = 1, \ldots, d,$$

i.e. $\Sigma_1 = \mathrm{diag}(0.5,\, 1.0,\, 1.5, \ldots, 0.5d)$.

Distinct ratios (identifiable). Two environments ($e \in \{1, 2\}$) with Σ_2 chosen so that the variance ratios $\sigma_{2,j}^2/\sigma_{1,j}^2$ are pairwise distinct across all coordinates $j = 1, \ldots, d$. The diagonal entries of Σ_2 are

$$\begin{aligned} d = 5\colon &\quad \sigma_{2,j}^2 = (2.0,\ 1.5,\ 0.8,\ 0.5,\ 0.4), \\ d = 7\colon &\quad \sigma_{2,j}^2 = (2.5,\ 0.3,\ 2.1,\ 0.35,\ 1.8,\ 0.43,\ 1.6). \end{aligned}$$

One regime (control). A single environment with diffusion covariance Σ_1 only. Since there is only one environment, the distinct-ratios condition is not applicable.

Proportional diffusions (control). Two environments with $\sigma_{2,j}^2 = 2\,\sigma_{1,j}^2$ for all j, so that all variance ratios $\sigma_{2,j}^2/\sigma_{1,j}^2$ equal 2. This violates the two-regime separation requirement in Assumption (S4), since all ratios are identical.

For $d \in \{5, 7\}$, the same diffusion-covariance choices are used across all three drift families. The separate $d = 10$ setup is specified in Section F.4.

E.3 Observation model

The learner observes $x_t^{(e)} = g(z_t^{(e)})$. We implement $g\colon \mathbb{R}^d \to \mathbb{R}^d$ as an invertible MLP mixing map with orthogonal linear layers and componentwise activations. We test two choices of activation within the same m-layer orthogonal MLP architecture:

$$y_1 = \phi(R_1 z + b_1), \qquad y_k = \phi(R_k y_{k-1} + b_k)\ \ (2 \le k \le m-1), \qquad g(z) = R_m y_{m-1},$$

where each $R_k \in \mathrm{O}(d)$ is a random orthogonal matrix (drawn via the Haar measure), $b_k \sim \mathcal{N}(0, I_d)$ for $k = 1, \ldots, m-1$ are bias vectors, and ϕ is a componentwise nonlinearity. The final layer is a pure rotation with no bias or nonlinearity. In all experiments $m = 3$.

Leaky-$\tanh$ (main text). $\phi(z) = \tanh(z) + 0.1\, z$. Since $\phi'(z) = \operatorname{sech}^2(z) + 0.1 > 0$ everywhere and $\phi \in C^\infty$, the composition g is a C^∞ diffeomorphism, consistent with the smoothness assumptions of our identifiability results.

LeakyReLU (appendix). $\phi(z) = \max(z,\, 0.2\, z)$. This map is invertible and piecewise linear, but it is not C^1 at zero and hence not C^2, providing a robustness check beyond the smoothness assumptions.

The orthogonal matrices and biases are drawn with a fixed seed, so the same g is used across all regime conditions and training seeds within each fixed drift family, dimension, and activation choice.

E.4 Training procedure

We follow the two-stage procedure described in Section 4 and summarized in Algorithm 1.

Stage 1: Short-time transition fitting. The encoder $h_\theta \colon \mathbb{R}^d \to \mathbb{R}^d$ is a fully connected MLP with hidden dimension 256, four hidden layers, and ELU activations. We minimize the Stage 1 objective $\mathcal{L}_{\mathrm{S1}}$ (Eq. 7), including the change-of-variables log-determinant term from Eq. 6:

$$-\log \left| \det Dh_\theta(x_{t+\Delta t}) \right|.$$

The log-determinant term is the local volume correction in the observation-space pseudo-likelihood and penalizes collapse of the encoder Jacobian. All parameters $(\theta, \psi, \{\tilde{\sigma}^2_{e,i}\})$ are optimized jointly with Adam (learning rate 10^{-3}, batch size $4{,}096$) for $20{,}000$ stochastic optimization steps, each using a fresh minibatch.

Stage 2: Sparse drift-Jacobian graph recovery (sparse linear and nonlinear only). After Stage 1, the encoder is frozen and we minimize $\mathcal{L}_{\mathrm{S2}}$ (Eq. 9) to fit a drift model to the encoded velocity targets. For the sparse linear family, the drift model is a linear map $\tilde{f}_\psi(\tilde{z}) = \hat{A}\tilde{z}$, with sparsity penalty $\sum_{i,j} |\hat{A}_{ij}|$. For the nonlinear family, $\tilde{f}_\psi$ is a three-layer MLP with ELU activations, with the Jacobian L_1 penalty in Eq. 9 evaluated on encoded samples. We optimize with Adam (learning rate 5×10^{-4}). Dense linear experiments skip Stage 2 because the drift is dense by design, making sparse drift-Jacobian graph recovery inapplicable.

Computational resources. The reported experiments were run as standard Slurm jobs requesting one NVIDIA L40S GPU and 32 GB memory per job. For a fixed setting, the five training seeds used in the synthetic tables typically completed in a few minutes. The implementation does not rely on specialized hardware and can also run on CPU or on other CUDA GPUs, with CPU runs expected to be slower.

Graph thresholding. Following Algorithm 1, the learned drift-Jacobian graph is obtained by thresholding the mean absolute Jacobian $\overline{|D\tilde{f}_\psi|}$ at a fixed threshold $\tau = 0.05$ in the encoded space, before any scaling correction. The resulting binary adjacency is then permuted back to the original coordinate ordering for comparison with the ground truth.

Evaluation. Each experiment is repeated over five independent training seeds. We report mean $\pm$ standard deviation of three metrics defined below. For diagnostic plots (scatter matrices, Jacobian heatmaps, graph comparisons), we use one representative seed, held fixed across the regime conditions shown in each figure.

Mean correlation coefficient (MCC). Given n samples of the true latent coordinates $z \in \mathbb{R}^{n \times d}$ and the learned coordinates $\tilde{z} \in \mathbb{R}^{n \times d}$, we compute the $d \times d$ absolute Pearson correlation matrix C with entries $C_{ij} = \left|\operatorname{corr}(z_{\cdot,i},\, \tilde{z}_{\cdot,j})\right|$. The optimal one-to-one assignment π^* is found by the Hungarian algorithm:

$$\pi^* = \arg\max_{\pi \in S_d} \sum_{i=1}^{d} C_{i,\pi(i)}, \qquad \mathrm{MCC} = \frac{1}{d} \sum_{i=1}^{d} C_{i,\pi^*(i)}.$$

A value near 1 indicates that each learned coordinate aligns with exactly one true latent.

Monomial score (Mon.). Let $\bar{J} = \overline{|D(h_\theta \circ g)(z)|}$ denote the mean absolute Jacobian of the composition of the encoder and observation map, a $d \times d$ non-negative matrix. We normalize $\bar{J}$ along rows and columns separately:

$$R_{ij} = \frac{\bar{J}_{ij}}{\sum_k \bar{J}_{ik}}, \qquad Q_{ij} = \frac{\bar{J}_{ij}}{\sum_k \bar{J}_{kj}},$$

and define the row and column concentration as the mean of the row-wise and column-wise maxima:

$$\rho_{\text{row}} = \frac{1}{d}\sum_{i=1}^{d} \max_j R_{ij}, \qquad \rho_{\text{col}} = \frac{1}{d}\sum_{j=1}^{d} \max_i Q_{ij}.$$

The monomial score is $\text{Mon.} = \frac{1}{2}(\rho_{\text{row}} + \rho_{\text{col}})$. For a perfect monomial matrix every row and column has exactly one nonzero entry, giving $\text{Mon.} = 1$; a uniform matrix yields $\text{Mon.} = 1/d$.

Graph match rate (GMR). The mean absolute Jacobian of the learned drift, $\overline{|D\tilde{f}_\psi|}$, is binarized at threshold τ: $\hat{G}^{\text{enc}}_{ij} = \mathbf{1}\big[\overline{|D\tilde{f}_\psi|}_{ij} > \tau\big]$. Let p denote the encoder permutation extracted from the encoder Jacobian, where $p(k)$ is the true coordinate matched to encoded coordinate k. This learned drift-Jacobian adjacency is then permuted back to the original coordinate ordering: $\hat{G}_{ij} = \hat{G}^{\text{enc}}_{p^{-1}(i),\,p^{-1}(j)}$. In practice, p is obtained by applying the Hungarian algorithm to the mean absolute encoder Jacobian $\bar{J}$. The ground-truth adjacency G^* includes self-loops ($G^*_{ii} = 1$ for all i) because every coordinate has a self-regulating drift term. The graph match rate is the fraction of all d^2 entries that agree:

$$\text{GMR} = \frac{1}{d^2}\sum_{i,j=1}^{d} \mathbf{1}\big[\hat{G}_{ij} = G^*_{ij}\big].$$

Table 4 summarizes the key simulation and training hyperparameters. Values that differ across drift families are noted explicitly.

Table 4: Simulation and training hyperparameters. All values are shared across the three drift families unless noted otherwise.

	$d = 5$	$d = 7$
Data generation		
Step size Δt	0.005	0.0025
Nominal horizon parameter T	0.1	0.05
Number of trajectories per regime N_{traj}	10,000	20,000
Stage 1 (encoder)		
Hidden dim / layers	256 / 4	256 / 4
Learning rate	10^{-3}	10^{-3}
Optimization steps	20,000	20,000
Batch size	4,096	4,096
Stage 2 (drift, sparse linear & nonlinear only)		
Drift MLP dim / layers[†]	128 / 3	128 / 3
Learning rate	5×10^{-4}	5×10^{-4}
Optimization steps (sparse linear)	10,000	10,000
Optimization steps (nonlinear)	20,000	20,000
Velocity stride s	1	1
Jacobian batch size (nonlinear)	256	512
λ_{sparse} (sparse linear)	0.3	0.25
λ_{sparse} (nonlinear)[‡]	0.5	0.4

[†]Nonlinear experiments only; sparse linear uses a linear drift model $\tilde{f}_\psi(\tilde{z}) = \hat{A}\tilde{z}$.

[‡]Leaky-tanh mixing; the LeakyReLU experiments use $\lambda_{\text{sparse}} = 0.9$ for both $d = 5$ and $d = 7$.

F Additional simulation results

Table 5 reports results under the same three drift families and regime conditions as in the main text (Table 1), but with a three-layer MLP using LeakyReLU activations as the mixing function. The same identifiability gap persists: Distinct ratios yields near-perfect disentanglement across all drift families, whereas both controls substantially reduce MCC and monomial score.

Table 5: Simulation results on $d = 5$ under a three-layer MLP with LeakyReLU activations. All entries are mean $\pm$ standard deviation over five seeds. **Bold** highlights the Distinct ratios condition; the other two conditions are controls. GMR is omitted for the dense linear family because graph sparsity is not the target there.

Drift family	Regime condition	MCC ↑	Mon. ↑	GMR ↑
Dense linear	Distinct ratios	$\mathbf{0.9977 \pm 0.0002}$	$\mathbf{0.8429 \pm 0.0068}$	–
Dense linear	One regime	0.7363 ± 0.0516	0.4178 ± 0.0304	–
Dense linear	Proportional diffusions	0.6978 ± 0.0234	0.3877 ± 0.0113	–
Sparse linear	Distinct ratios	$\mathbf{0.9979 \pm 0.0003}$	$\mathbf{0.8667 \pm 0.0088}$	$\mathbf{1.0000 \pm 0.0000}$
Sparse linear	One regime	0.7364 ± 0.0491	0.4129 ± 0.0273	0.5840 ± 0.0543
Sparse linear	Proportional diffusions	0.7153 ± 0.0600	0.4081 ± 0.0375	0.5680 ± 0.0588
Nonlinear	Distinct ratios	$\mathbf{0.9979 \pm 0.0002}$	$\mathbf{0.8660 \pm 0.0068}$	$\mathbf{1.0000 \pm 0.0000}$
Nonlinear	One regime	0.7387 ± 0.0647	0.4269 ± 0.0332	0.6160 ± 0.0650
Nonlinear	Proportional diffusions	0.7336 ± 0.0495	0.4142 ± 0.0328	0.7200 ± 0.1043

Tables 6 and 7 extend the analysis to $d = 7$ under leaky-$\tanh$ and LeakyReLU mixing, respectively. The same qualitative pattern holds: Distinct ratios yields near-perfect disentanglement across all three drift families, whereas both controls substantially reduce coordinate recovery.

Table 6: Simulation results on $d = 7$ under a three-layer MLP with leaky-$\tanh$ activations. All entries are mean $\pm$ standard deviation over five seeds. **Bold** highlights the Distinct ratios condition; the other two conditions are controls. GMR is omitted for the dense linear family because graph sparsity is not the target there.

Drift family	Regime condition	MCC ↑	Mon. ↑	GMR ↑
Dense linear	Distinct ratios	$\mathbf{0.9949 \pm 0.0041}$	$\mathbf{0.8283 \pm 0.0196}$	–
Dense linear	One regime	0.6745 ± 0.0555	0.3520 ± 0.0406	–
Dense linear	Proportional diffusions	0.6319 ± 0.0276	0.3252 ± 0.0225	–
Sparse linear	Distinct ratios	$\mathbf{0.9974 \pm 0.0005}$	$\mathbf{0.8616 \pm 0.0090}$	$\mathbf{0.9959 \pm 0.0082}$
Sparse linear	One regime	0.6441 ± 0.0535	0.3424 ± 0.0279	0.6327 ± 0.0428
Sparse linear	Proportional diffusions	0.6350 ± 0.0575	0.3270 ± 0.0351	0.6776 ± 0.0523
Nonlinear	Distinct ratios	$\mathbf{0.9972 \pm 0.0002}$	$\mathbf{0.8564 \pm 0.0053}$	$\mathbf{0.9755 \pm 0.0153}$
Nonlinear	One regime	0.6256 ± 0.0468	0.3264 ± 0.0239	0.7061 ± 0.0277
Nonlinear	Proportional diffusions	0.6387 ± 0.0484	0.3230 ± 0.0311	0.8204 ± 0.0327

F.1 Dense linear drift

The dense linear experiments use an Ornstein–Uhlenbeck process $\mathrm{d}z = Az\,\mathrm{d}t + G\,\mathrm{d}W$ with a symmetric negative-definite drift matrix A. This family evaluates coordinate recovery without using graph sparsity as an evaluation target. We construct A as

$$A = -(MM^\top + \lambda I), \qquad M_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1),$$

with $\lambda = 0.5$ to ensure a stability margin. The random matrix M is drawn with a fixed seed, so the same A is shared across all training seeds and all three regime conditions. For $d = 5$ the resulting drift matrix, rounded to two decimals, is

$$A = \begin{pmatrix} -4.16 & 0.00 & -2.79 & 1.30 & -4.78 \\ 0.00 & -1.74 & -1.56 & 1.26 & 1.57 \\ -2.79 & -1.56 & -5.71 & 2.56 & -0.36 \\ 1.30 & 1.26 & 2.56 & -3.51 & 0.15 \\ -4.78 & 1.57 & -0.36 & 0.15 & -10.27 \end{pmatrix}.$$

Table 7: Simulation results on $d = 7$ under a three-layer MLP with LeakyReLU activations. All entries are mean $\pm$ standard deviation over five seeds. **Bold** highlights the Distinct ratios condition; the other two conditions are controls. GMR is omitted for the dense linear family because graph sparsity is not the target there.

Drift family	Regime condition	MCC ↑	Mon. ↑	GMR ↑
Dense linear	Distinct ratios	$\mathbf{0.9948 \pm 0.0006}$	$\mathbf{0.7815 \pm 0.0039}$	–
Dense linear	One regime	0.6750 ± 0.0519	0.3364 ± 0.0259	–
Dense linear	Proportional diffusions	0.6452 ± 0.0550	0.3123 ± 0.0324	–
Sparse linear	Distinct ratios	$\mathbf{0.9960 \pm 0.0005}$	$\mathbf{0.8023 \pm 0.0114}$	$\mathbf{1.0000 \pm 0.0000}$
Sparse linear	One regime	0.6542 ± 0.0410	0.3327 ± 0.0238	0.5714 ± 0.0671
Sparse linear	Proportional diffusions	0.6319 ± 0.0251	0.3213 ± 0.0180	0.5837 ± 0.0909
Nonlinear	Distinct ratios	$\mathbf{0.9961 \pm 0.0007}$	$\mathbf{0.8095 \pm 0.0078}$	$\mathbf{0.9388 \pm 0.0316}$
Nonlinear	One regime	0.6343 ± 0.0367	0.3276 ± 0.0168	0.5429 ± 0.0988
Nonlinear	Proportional diffusions	0.5969 ± 0.0278	0.3044 ± 0.0136	0.6612 ± 0.1625

Figure 3 shows the pairwise scatter plots of true latent coordinates z_j versus learned representations $\tilde{z}_i$ for each regime condition. Under Distinct ratios, each learned coordinate aligns tightly with exactly one true coordinate, producing a near-permutation pattern. Both controls lose this structure.

Figure 4 displays the mean absolute Jacobian $\overline{|D\varphi|}$ of the composition $\varphi = h \circ g$. Under Distinct ratios the matrix is near-monomial, confirming that the encoder inverts the mixing up to coordinate permutation and scaling. Both controls yield less monomial, more mixed encoder Jacobians. Graph recovery is not shown because the drift is dense, so sparse graph recovery is not the evaluation target.

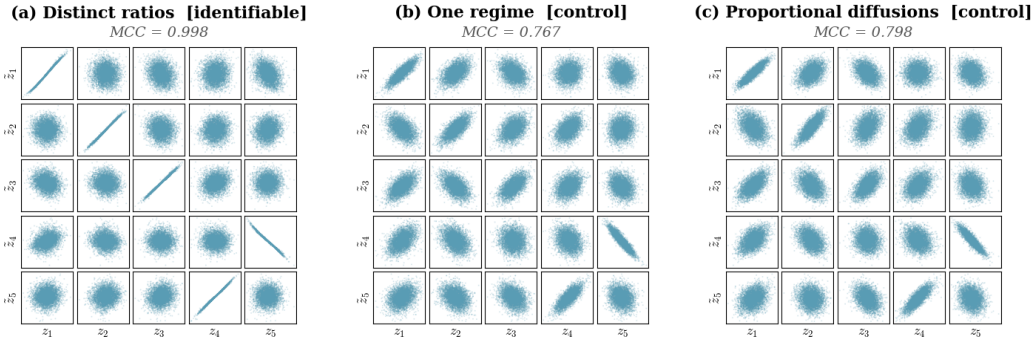

Figure 3: Latent recovery in the dense linear $d = 5$ setting (three-layer MLP, leaky-$\tanh$ mixing). Each column shows the 5×5 scatter matrix of true latent coordinates z_j (horizontal) versus learned representations $\tilde{z}_i$ (vertical) for one regime condition. Under Distinct ratios, each $\tilde{z}_i$ aligns with exactly one z_j. Both controls fail to disentangle.

The $d = 7$ case uses the same construction with identical $\lambda = 0.5$. The resulting drift matrix, rounded to two decimals, is

$$A = \begin{pmatrix} -4.89 & 1.11 & -2.12 & -0.13 & 0.98 & -2.94 & 4.09 \\ 1.11 & -5.21 & -2.30 & -0.70 & 4.00 & -1.50 & 0.93 \\ -2.12 & -2.30 & -9.39 & -0.37 & -0.35 & -5.45 & 4.37 \\ -0.13 & -0.70 & -0.37 & -8.22 & 4.36 & -3.82 & -1.67 \\ 0.98 & 4.00 & -0.35 & 4.36 & -8.66 & 3.26 & -1.52 \\ -2.94 & -1.50 & -5.45 & -3.82 & 3.26 & -7.56 & 4.88 \\ 4.09 & 0.93 & 4.37 & -1.67 & -1.52 & 4.88 & -8.83 \end{pmatrix}.$$

Figures 5 and 6 show the $d = 7$ results. The same qualitative pattern holds: Distinct ratios achieves near-perfect disentanglement, while both controls show substantially less coordinate recovery.

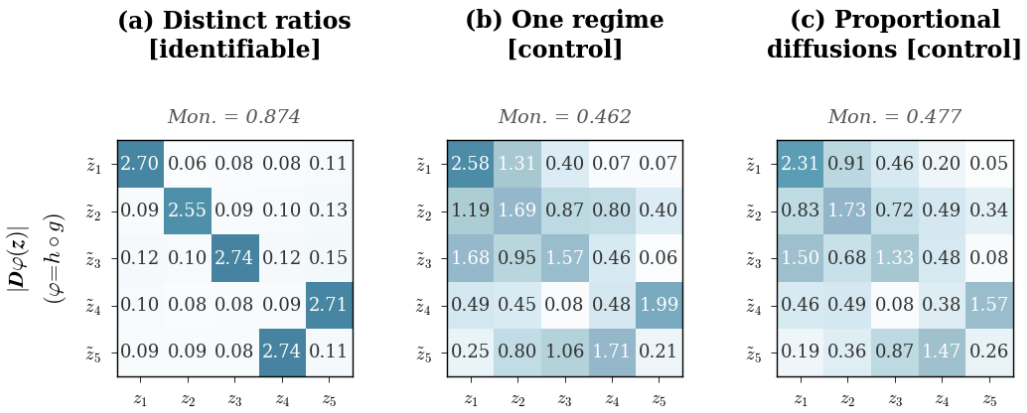

Figure 4: Mean absolute encoder Jacobian $\overline{|D\varphi|}$ ($\varphi = h \circ g$) for the dense linear $d = 5$ setting. All three regime conditions use the same representative seed. A near-monomial matrix confirms successful inversion of the mixing under Distinct ratios. Both controls yield less monomial, more mixed encoder Jacobians.

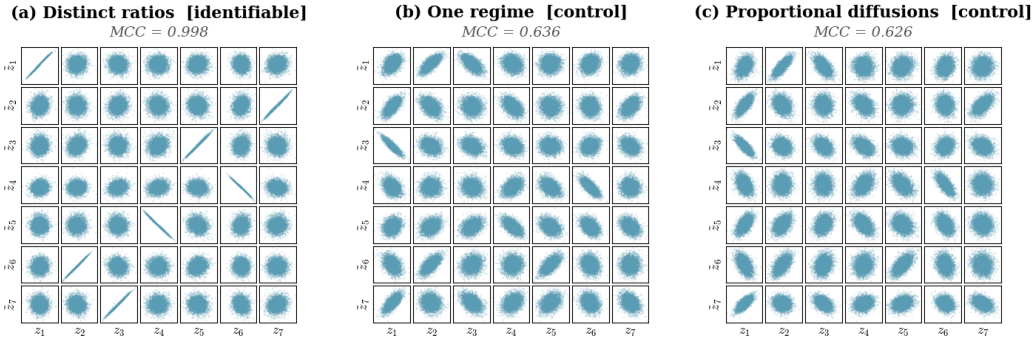

Figure 5: Latent recovery in the dense linear $d = 7$ setting (three-layer MLP, leaky-$\tanh$ mixing). Layout follows Figure 3. Under Distinct ratios, each $\tilde{z}_i$ aligns with exactly one z_j. Both controls fail to disentangle.

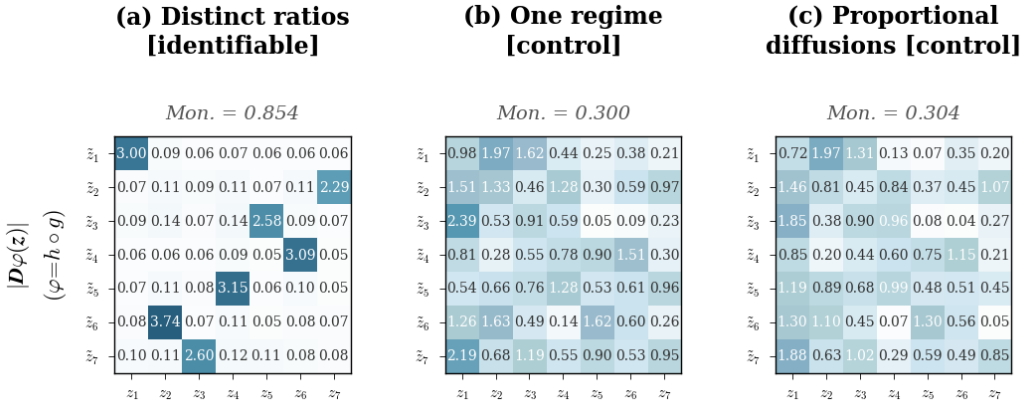

Figure 6: Mean absolute encoder Jacobian $\overline{|D\varphi|}$ ($\varphi = h \circ g$) for the dense linear $d = 7$ setting. All three regime conditions use the same representative seed. A near-monomial matrix confirms successful inversion under Distinct ratios.

F.2 Sparse linear drift

The sparse linear experiments use an Ornstein–Uhlenbeck process with a sparse drift matrix A built from a directed cycle support. The nonzero off-diagonal support entries are placed at $(i, i+1)$ for $i = 1, \ldots, d-1$ and at $(d, 1)$, and the drift matrix is constructed as

$$A_{ii} = -\alpha, \qquad A_{ij} = \beta\, c_{ij} \text{ if } \mathrm{adj}(i,j) = 1, \qquad A_{ij} = 0 \text{ otherwise},$$

where $\alpha = 1.0$, $\beta = 1.5$, and $c_{ij} \overset{\text{iid}}{\sim} \mathrm{Uniform}(0.8, 1.2)$ with a fixed seed. The same A is shared across all training seeds and all three regime conditions.

For $d = 5$ the drift matrix, rounded to two decimals, is

$$A = \begin{pmatrix} -1.00 & 1.77 & 0 & 0 & 0 \\ 0 & -1.00 & 1.72 & 0 & 0 \\ 0 & 0 & -1.00 & 1.33 & 0 \\ 0 & 0 & 0 & -1.00 & 1.37 \\ 1.57 & 0 & 0 & 0 & -1.00 \end{pmatrix}.$$

Figure 7 shows the pairwise scatter plots for the sparse linear $d = 5$ setting. Figure 8 displays the encoder Jacobian and graph recovery. Under Distinct ratios, the encoder inverts the mixing up to permutation and scaling, and the thresholded drift-Jacobian graph is recovered exactly (GMR $= 1$ for the representative seed).

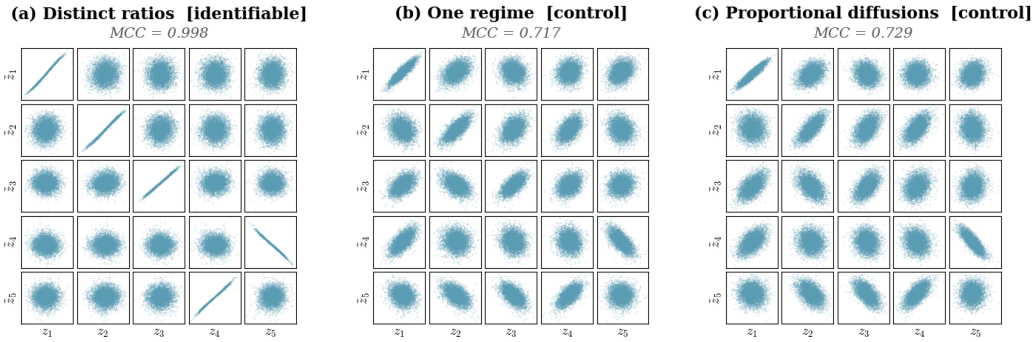

Figure 7: Latent recovery in the sparse linear $d = 5$ setting (three-layer MLP, leaky-$\tanh$ mixing). Each column shows the 5×5 scatter matrix of true latent coordinates z_j (horizontal) versus learned representations $\tilde{z}_i$ (vertical) for one regime condition.

The $d = 7$ case uses the same construction. The drift matrix, rounded to two decimals, is

$$A = \begin{pmatrix} -1.00 & 1.77 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1.00 & 1.62 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1.00 & 1.51 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1.00 & 1.67 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1.00 & 1.77 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1.00 & 1.50 \\ 1.22 & 0 & 0 & 0 & 0 & 0 & -1.00 \end{pmatrix}.$$

Figures 9 and 10 show the corresponding results for $d = 7$. The same qualitative pattern holds: Distinct ratios gives near-monomial encoder Jacobians and near-perfect graph recovery, whereas the controls show mixed encoder Jacobians and lower GMR.

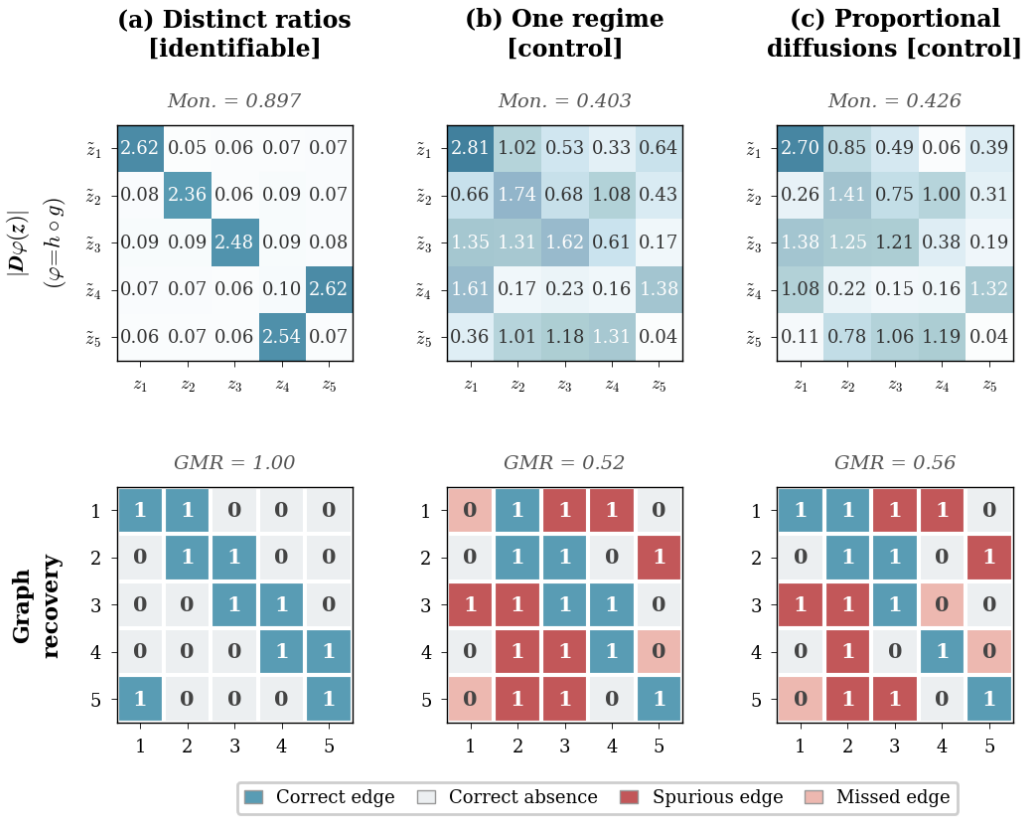

Figure 8: Structural diagnostics for the sparse linear $d = 5$ setting. All three regime conditions use the same representative seed. **Top row**: mean absolute Jacobian $\overline{|D\varphi|}$. **Bottom row**: learned causal graph versus the ground-truth graph.

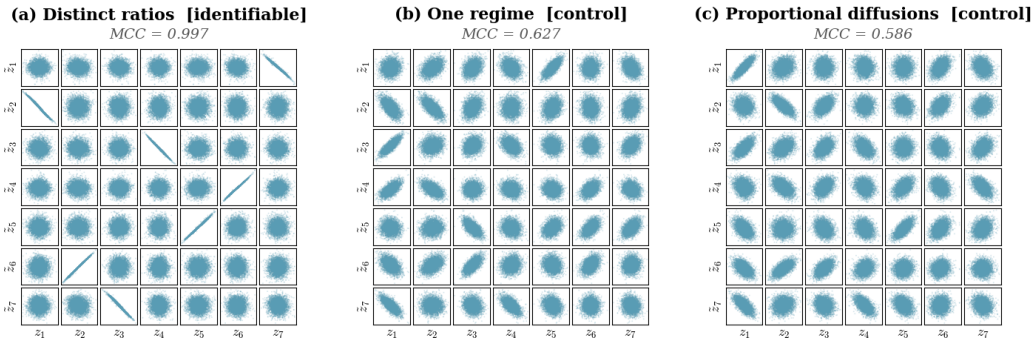

Figure 9: Latent recovery in the sparse linear $d = 7$ setting (three-layer MLP, leaky-$\tanh$ mixing). Layout follows Figure 7.

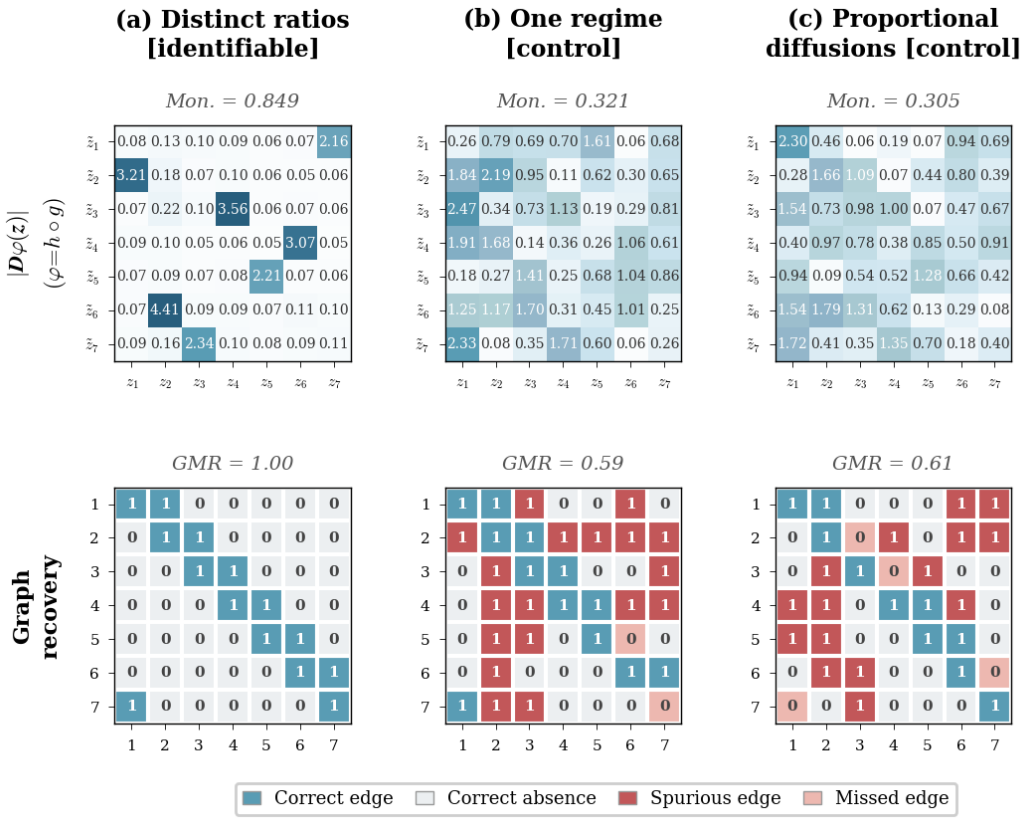

Figure 10: Structural diagnostics for the sparse linear $d = 7$ setting. All three regime conditions use the same representative seed. **Top row**: mean absolute encoder Jacobian $\overline{|D\varphi|}$. **Bottom row**: learned causal graph versus the ground-truth graph.

F.3 Nonlinear sparse drift

The nonlinear experiments use the same directed cycle support as the sparse linear case (Section F.2), but replace the linear coupling with a $\tanh$ nonlinearity. Concretely, the drift function is

$$f_i(z) \;=\; -\alpha\, z_i \;+\; \beta \sum_{j:\,\mathrm{adj}(i,j)=1} c_{ij}\, \tanh(z_j),$$

where $\alpha = 1.0$ enforces mean-reversion along the diagonal and $\beta = 1.5$ controls the coupling strength. The coefficients c_{ij} are drawn independently from $\mathrm{Uniform}(0.8, 1.2)$ with a fixed seed and are nonzero only where the adjacency is 1. For $d = 5$ the nonzero coupling coefficients, rounded to two decimals, are

$$c_{12} = 1.18, \quad c_{23} = 1.15, \quad c_{34} = 0.88, \quad c_{45} = 0.92, \quad c_{51} = 1.04.$$

Here $\mathrm{adj}(i, j) = 1$ means that coordinate z_j enters the ith drift component f_i.

For $d = 7$ the nonzero off-diagonal support entries are again placed at $(i, i + 1)$ for $i = 1, \ldots, 6$ and at $(7, 1)$, and the nonzero coupling coefficients, rounded to two decimals, are

$$c_{12} = 1.09, \quad c_{23} = 0.95, \quad c_{34} = 1.13, \quad c_{45} = 1.06, \quad c_{56} = 1.03, \quad c_{67} = 0.97, \quad c_{71} = 1.08.$$

The Jacobian of f evaluated at the origin recovers the same sparse support pattern as in Section F.2, but away from the origin the $\tanh$ saturates, making the dynamics genuinely nonlinear.

For each fixed dimension, the same drift function is shared across the leaky-$\tanh$ and LeakyReLU mixing experiments. The main text (Figures 1 and 2) reports the $d = 5$ results under leaky-$\tanh$ mixing. Here we show the corresponding results under LeakyReLU mixing for both $d = 5$ and $d = 7$. All three regime conditions in each figure use the same representative training seed for a fair comparison. The identifiability gap is consistent with the leaky-$\tanh$ results in the main text. Under Distinct ratios, the displayed run recovers clear coordinate alignment and a substantially more accurate thresholded graph; the controls show mixed coordinates and graph errors.

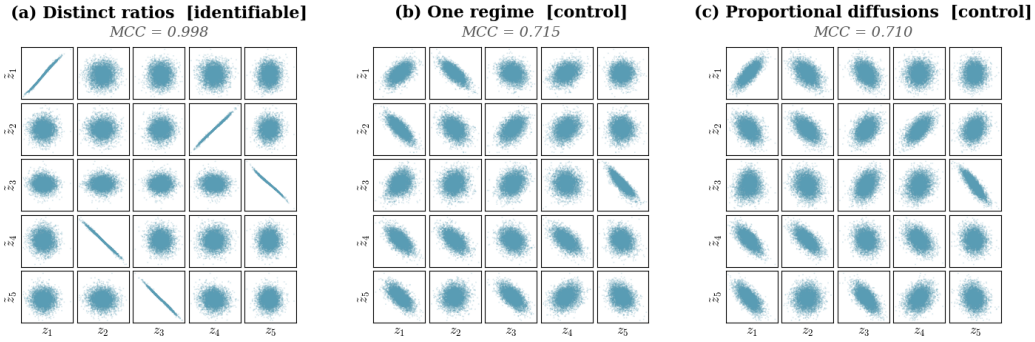

Figure 11: Latent recovery in the nonlinear $d = 5$ setting (three-layer MLP, LeakyReLU mixing). Each column shows the 5×5 scatter matrix of true latent coordinates z_j (horizontal) versus learned representations $\tilde{z}_i$ (vertical) for one regime condition.

Figures 13 and 14 show the $d = 7$ results under LeakyReLU mixing. The identifiability gap remains clear: Distinct ratios achieves near-perfect disentanglement and high graph match rate, whereas both controls show weaker coordinate recovery and less reliable graph recovery.

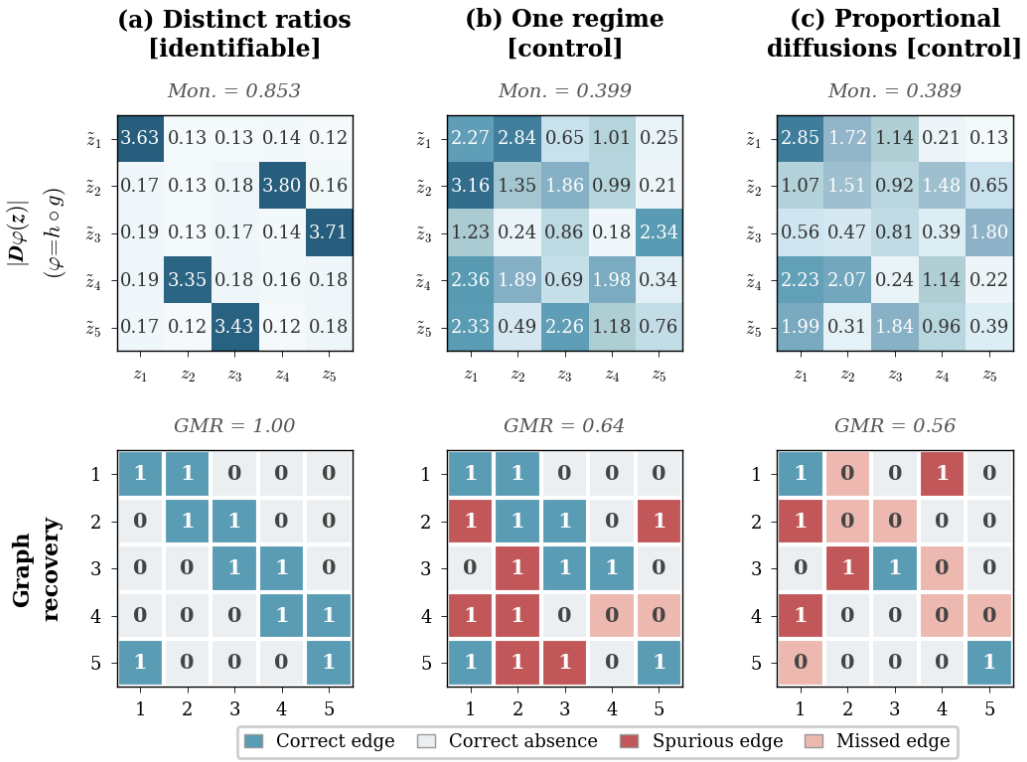

Figure 12: Structural diagnostics for the nonlinear $d = 5$ setting with LeakyReLU mixing. All three regime conditions use the same representative seed. **Top row**: mean absolute encoder Jacobian $\overline{|D\varphi|}$, where $\varphi = h \circ g$. **Bottom row**: learned causal graph versus the ground-truth graph.

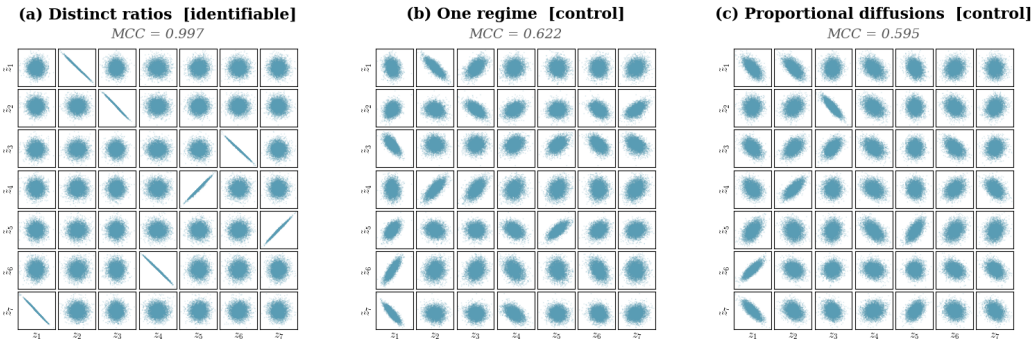

Figure 13: Latent recovery in the nonlinear $d = 7$ setting (three-layer MLP, LeakyReLU mixing). Each column shows the 7×7 scatter matrix of true latent coordinates z_j (horizontal) versus learned representations $\tilde{z}_i$ (vertical) for one regime condition. Under Distinct ratios, each $\tilde{z}_i$ aligns with exactly one z_j. Both controls fail to disentangle.

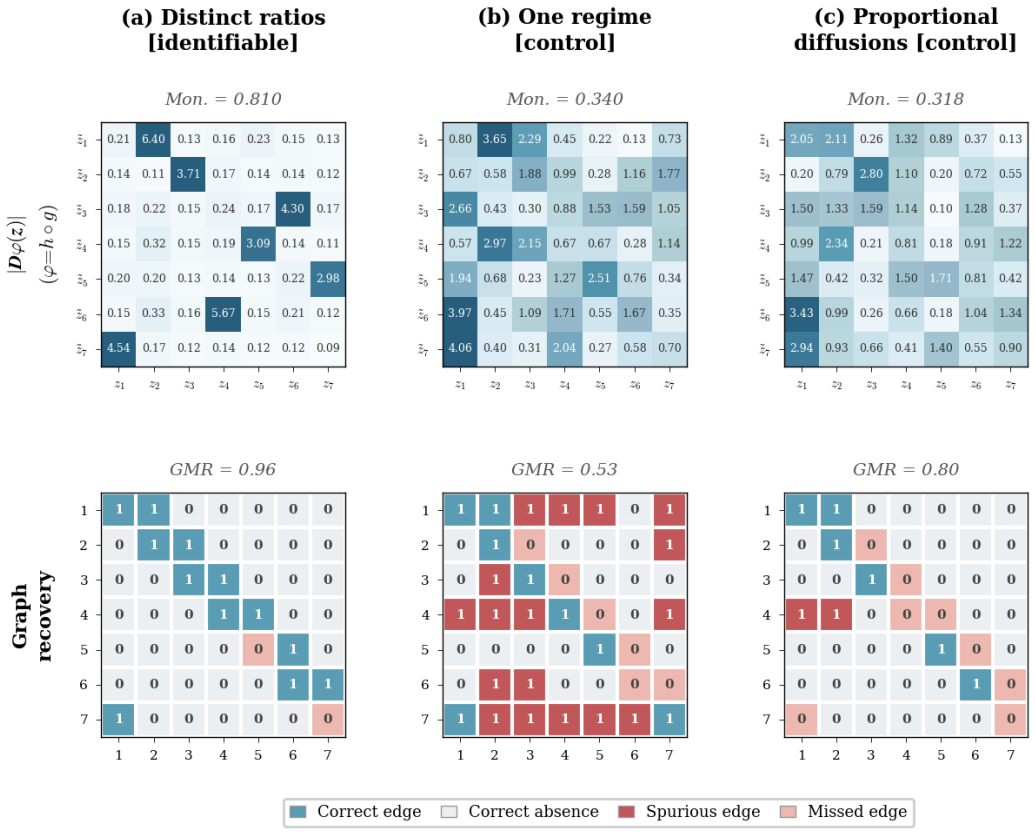

Figure 14: Structural diagnostics for the nonlinear $d = 7$ setting with LeakyReLU mixing. All three regime conditions use the same representative seed. **Top row**: mean absolute encoder Jacobian $\overline{|D\varphi|}$, where $\varphi = h \circ g$. **Bottom row**: learned causal graph versus the ground-truth graph.

F.4 $d = 10$ evaluation and dense nonlinear drift

We additionally evaluate $d = 10$ under the same smooth three-layer leaky-$\tanh$ MLP mixing used in the main experiments. The evaluation covers dense linear, sparse linear, sparse nonlinear, and fully dense nonlinear drifts under all three regime conditions. Table 8 reports means $\pm$ s.d. over five training seeds.

For reproducibility, these runs use $\Delta t = 0.001$, nominal horizon $T = 0.02$, $N_{\text{traj}} = 50{,}000$ per regime, a width-512 encoder with four hidden layers, and 30,000 Stage 1 optimization steps. In the Distinct-ratios condition, all families use $\text{diag}(\Sigma_2) = (2.5, 0.3, 2.1, 0.35, 1.8, 0.43, 1.6, 0.5, 1.1, 0.38)$. For dense linear, $\text{diag}(\Sigma_1) = (0.5, 1.0, \ldots, 5.0)$; the other three families use $\text{diag}(\Sigma_1) = (0.5, 0.8, 1.1, 1.4, 1.7, 2.0, 2.3, 2.5, 2.7, 3.0)$. The One-regime control retains only the corresponding Σ_1, and the Proportional-diffusions control uses $\Sigma_2 = 2\Sigma_1$. Sparse-linear and sparse-nonlinear Stage 2 use $\lambda_{\text{sparse}} = 0.2$ and 0.8, respectively; the dense families omit Stage 2.

Table 8: $d = 10$ results under three-layer leaky-$\tanh$ MLP mixing. Bold entries indicate the identifiable Distinct-ratios condition. GMR is omitted for dense drift families, for which sparse graph recovery is not the target.

Drift family	Regime condition	MCC ↑	Mon. ↑	GMR ↑
Dense linear	Distinct ratios	$\mathbf{0.9810 \pm 0.0067}$	$\mathbf{0.7709 \pm 0.0086}$	–
Dense linear	One regime	0.5548 ± 0.0178	0.2566 ± 0.0087	–
Dense linear	Proportional diffusions	0.5701 ± 0.0221	0.2622 ± 0.0097	–
Sparse linear	Distinct ratios	$\mathbf{0.9962 \pm 0.0016}$	$\mathbf{0.8207 \pm 0.0153}$	$\mathbf{0.9880 \pm 0.0147}$
Sparse linear	One regime	0.5696 ± 0.0263	0.2483 ± 0.0176	0.5820 ± 0.0449
Sparse linear	Proportional diffusions	0.5510 ± 0.0260	0.2401 ± 0.0180	0.6540 ± 0.0206
Sparse nonlinear	Distinct ratios	$\mathbf{0.9845 \pm 0.0227}$	$\mathbf{0.8085 \pm 0.0188}$	$\mathbf{0.8940 \pm 0.0258}$
Sparse nonlinear	One regime	0.5966 ± 0.0127	0.2604 ± 0.0155	0.7880 ± 0.0791
Sparse nonlinear	Proportional diffusions	0.5654 ± 0.0108	0.2522 ± 0.0100	0.8200 ± 0.0126
Dense nonlinear	Distinct ratios	$\mathbf{0.9947 \pm 0.0037}$	$\mathbf{0.8113 \pm 0.0130}$	–
Dense nonlinear	One regime	0.5777 ± 0.0237	0.2525 ± 0.0158	–
Dense nonlinear	Proportional diffusions	0.5763 ± 0.0067	0.2534 ± 0.0021	–

For the dense nonlinear family, we use

$$f_i(z) = -1.5 z_i + \sum_{j \neq i} C_{ij} \tanh(z_j), \qquad \|C\|_2 = 1,$$

where all off-diagonal coefficients C_{ij} are nonzero. Its drift Jacobian therefore has all 90 directed off-diagonal couplings and all 10 diagonal self-effects nonzero, for $100/100$ nonzero entries. Under Distinct ratios it achieves MCC 0.9947 ± 0.0037 and Monomial score 0.8113 ± 0.0130, compared with approximately $0.577/0.253$ for both controls. Across all four drift families, the Distinct-ratios condition retains high coordinate recovery, whereas both controls remain substantially lower.

For the sparse nonlinear family, the $d = 10$ graph contains only 20 edges among 100 entries, so full-matrix GMR is influenced by correctly matched absences. Pooling the off-diagonal decisions across the five seeds gives F1 0.716 under Distinct ratios, compared with 0.200 and 0.039 for One regime and Proportional diffusions, respectively. These results are a $d = 10$ stress test rather than a claim of asymptotic scalability.

F.5 Computational cost and scaling

In the synthetic experiments, the observation and latent dimensions are both d, so the encoder Jacobian is a $d \times d$ matrix. Our finite-difference implementation evaluates the encoder once at the sampled state and once after perturbing each input coordinate, giving $d + 1$ forward evaluations. It explicitly stores the resulting Jacobian, requiring $O(d^2)$ storage, and computes its log-determinant in $O(d^3)$ arithmetic per sampled state, excluding the cost of the encoder forward evaluations themselves. Sparse-nonlinear Stage 2 differentiates the drift Jacobian with higher-order automatic differentiation and is the main measured bottleneck; sparse-linear Stage 2 avoids this cost.

Table 9 reports training-only wall time and peak GPU memory on one NVIDIA L40S. Entries are means over five seeds; timing s.d. is below 2 s throughout, and memory is the peak allocation across stages.

Table 9: Measured computational cost for the identifiable-condition configurations.

Drift family	d	Stage 1 (s)	Stage 2 (s)	Peak MiB
Sparse linear	5	113.7	17.8	162
Sparse linear	7	131.4	18.9	166
Sparse linear	10	267.1	43.7	334
Sparse nonlinear	5	122.9	111.6	187
Sparse nonlinear	7	141.1	151.4	191
Sparse nonlinear	10	289.5	350.2	385

Dense families have comparable Stage 1 cost at the same d and omit Stage 2. Because the $d = 10$ configurations use wider networks and more optimization steps, the measured increase combines dimension, model capacity, and optimization effort rather than estimating an asymptotic scaling exponent. Together with the $d = 10$ recovery results in Table 8, these measurements establish practical feasibility at this scale and identify nonlinear Stage 2 as the dominant measured cost.

F.6 Irregular sampling

This experiment asks whether one learned continuous-time drift, together with the regime-specific learned diffusions, can describe transitions at observation gaps not seen during training. We use the dense-linear system at $d = 5$ over five seeds. Data are generated with exact Ornstein–Uhlenbeck transitions, while the learner retains the local Euler Stage 1 objective and the same smooth three-layer MLP mixing as in the main experiments. Training uses balanced gaps $\{0.002, 0.008\}$. At test time, 0.005 is an unseen interpolation gap, whereas 0.001 and 0.010 lie outside the training range.

We compare a gap-aware estimator, which uses the observed elapsed time in each transition mean and covariance, with an otherwise identical mean-gap ablation that substitutes 0.005 for every elapsed time. The latter is a controlled analogue of a fixed-lag discrete-time model that ignores timestamps. Before scoring, we freeze the encoder and drift and refit each variant's diagonal diffusion on independent validation data. Lower held-out observation-space transition pseudo-NLL is better. We report the reduction $\mathrm{NLL}_{\text{mean-gap}} - \mathrm{NLL}_{\text{gap-aware}}$ in nats (natural-log units) per transition, so a positive value favors the gap-aware model. Calibration is the mean squared normalized residual: one is ideal, a value below one means that the predictive variance is too large, and a value above one means that it is too small.

Table 10: Held-out transition quality at gaps not used for training. Pseudo-NLL reduction is mean-gap minus gap-aware, in natural-log units per transition; positive values favor the gap-aware model. Calibration s.d. is at most 0.031 across the five seeds.

Test gap	Gap-aware calibration	Mean-gap calibration	Pseudo-NLL reduction
0.001	1.058	0.263	2.074 ± 0.028
0.005	1.032	1.022	0.013 ± 0.011
0.010	1.001	2.068	0.969 ± 0.051

The gap-aware model remains close to ideal calibration at the interpolation point and at both extrapolation points. The mean-gap ablation predicts too much variance at the shorter gap and too little at the longer gap. At 0.005, where the ablation happens to use the correct elapsed time, the two models have nearly identical calibration and pseudo-NLL. This matched point is a sanity check: it shows that the gains at 0.001 and 0.010 come from using the correct elapsed time rather than from a systematic difference in training quality. Coordinate recovery remains high for both variants (Pearson MCC 0.9972 ± 0.0003 versus 0.9967 ± 0.0003), as expected because this experiment targets transition calibration rather than representation accuracy. The result therefore illustrates the practical role of the continuous-time parameterization: the same shared drift and regime-specific diffusions can be evaluated across observation gaps, including gaps outside the training range.

G Real-world case study: Hardanger Bridge

This appendix gives the setup and diagnostics behind the Hardanger Bridge experiment reported in Section 5.2. The goal of this experiment is not to evaluate recovery against ground-truth latent coordinates, which are unavailable in the real bridge data. Instead, we test whether the TI-split two-regime construction produces a more reproducible latent representation and drift structure than controls that remove the turbulence-intensity (TI) contrast. The appendix first describes the data and training protocol, then reports cross-seed reproducibility, diffusion-fingerprint diagnostics, and a representative learned drift structure.

G.1 Dataset and regime construction

We use operational acceleration recordings from the Hardanger Bridge, a long-span suspension bridge monitored with wind and acceleration sensors [6, 7]. The dataset DOI/repository record lists a Creative Commons Attribution 4.0 International (CC BY 4.0) license [7]. The model receives only bridge-deck accelerometer channels as observations. Wind measurements are used only to construct regimes, not as model inputs.

To keep the shared-drift assumption physically meaningful, we restrict the data to a single mean-wind-speed bin, 4–$6\,\mathrm{m/s}$. Within this bin, each raw recording is divided into non-overlapping $120\,\mathrm{s}$ sub-windows. For each sub-window we compute the turbulence intensity

$$\mathrm{TI} = \frac{\mathrm{std}(\text{wind speed})}{\mathrm{mean}(\text{wind speed})}.$$

The two regimes are then defined by a quantile split of TI: the bottom 20% of sub-windows form the low-TI regime and the top 20% form the high-TI regime; the middle 60% is discarded. This construction is the real-world analogue of the distinct-diffusion-regime condition in the synthetic experiments: the bridge dynamics are approximately shared within a matched wind speed bin, while turbulence intensity changes the stochastic excitation.

Table 11 summarizes the resulting dataset. The accelerometer signals are detrended, low-pass filtered at $2.5\,\mathrm{Hz}$, resampled to $10\,\mathrm{Hz}$, and standardized with one scaler fitted jointly across both regimes. The selected $120\,\mathrm{s}$ sub-windows are then converted into overlapping $12.8\,\mathrm{s}$ trajectory windows with $6.4\,\mathrm{s}$ stride for training.

Table 11: Hardanger data construction for the reported 4–$6\,\mathrm{m/s}$, 16-channel experiment. The same training tensors are used for the TI-split-regime experiment and for both controls, except that the controls alter or remove the regime labels as described in Section G.2.

Quantity	Value
Mean-wind-speed bin	4–$6\,\mathrm{m/s}$
Accelerometer channels	16 bridge-deck channels
Regime statistic window	$120\,\mathrm{s}$, non-overlapping
Low-TI threshold	0.0336
High-TI threshold	0.0794
Selected sub-windows	987 low-TI and 987 high-TI
Trajectory window	$12.8\,\mathrm{s}$ (128 samples at $10\,\mathrm{Hz}$)
Trajectory stride	$6.4\,\mathrm{s}$
Training trajectories	16,779 per regime
Input dimension / latent dimension	$d_x = 16$, $d_z = 5$

G.2 Training protocol and controls

We use the same two-stage estimator as in Section 4. The encoder is a three-layer MLP with hidden dimension 64, and the drift model is linear in the learned $d_z = 5$ coordinates. Both stages are trained for up to 10,000 optimization steps with early stopping patience 100. Stage 2 uses stride $s = 6$, sparse-Jacobian penalty $\lambda_{\mathrm{sparse}} = 1.0$, and graph threshold $\tau = 0.05$. All three regime constructions use the same list of ten random training seeds. The Hardanger runs use the same single-GPU compute setup as the simulation experiments in Appendix E.4.

Because the Hardanger encoder maps $\mathbb{R}^{16}$ to $\mathbb{R}^5$, its Jacobian is rectangular. In Stage 1 we replace the square-map volume term by the row-volume regularizer $-\frac{1}{2}\log\det\big(Dh_\theta(x)Dh_\theta(x)^\top\big)$. This reduces to $-\log|\det Dh_\theta(x)|$ in the square case, but for the rectangular encoder it is a regularizer within the transition pseudo-likelihood rather than an exact change-of-variables likelihood.

We compare three regime constructions:

- *TI-split regimes (ours)*: the bottom and top TI quantiles are used as two regimes.
- *One regime (control)*: the two TI groups are merged into a single environment. This preserves the total amount of data but removes regime-specific diffusion parameters.
- *Shuffled TI labels (control)*: the two TI groups are pooled and randomly re-split into two equally sized groups. This preserves the two-regime training format but destroys the turbulence-based contrast.

G.3 Evaluation metrics

Because ground-truth latent states are not available, we evaluate reproducibility across random seeds. For each seed, we encode the same set of $5{,}000$ evaluation points. For every pair of seeds, we align learned coordinates by Hungarian matching on absolute correlation and compute the mean matched absolute correlation coefficient (MCC). We then use the same alignment to compare the thresholded drift-Jacobian graphs and report the graph match rate (GMR). With ten seeds, this gives $\binom{10}{2} = 45$ seed pairs. These are reproducibility diagnostics rather than accuracy metrics: high values mean that independent trainings recover compatible coordinate systems and drift graphs. As an auxiliary within-seed diagnostic, we also report the mean absolute off-diagonal correlation among learned latent coordinates; lower values indicate less redundant latent coordinates.

We also report diffusion-fingerprint diagnostics for the two-regime conditions. For parametric fingerprints, we use the learned diagonal variances $\tilde{\sigma}^2_{e,k}$ from Stage 1 and compute $r_k^{\mathrm{par}} = \tilde{\sigma}^2_{2,k}/\tilde{\sigma}^2_{1,k}$. For empirical fingerprints, we compute residual covariance matrices C_e after Stage 2 and use $r_k^{\mathrm{emp}} = C_{2,kk}/C_{1,kk}$. Since learned coordinates may differ across runs by permutation, scaling, and a possible shift, and since the sign of the log-ratio follows the internal regime ordering, we summarize contrast magnitude by $|\log r_k|$ pooled over seeds and latent dimensions.

G.4 Cross-seed consistency

Table 12 expands the main-text result with descriptive standard deviations across seed pairs. TI-split regimes yield the highest latent reproducibility and the highest drift-graph reproducibility. The MCC gain is substantial relative to both controls: $+0.146$ over One regime and $+0.205$ over Shuffled TI labels. The same construction also gives the lowest mean latent off-diagonal correlation, suggesting that the learned coordinates are less redundant within each learned representation. Because the thresholded drift graphs are mostly diagonal with sparse off-diagonal edges, absolute GMR values can be helped by matching many absent edges; we therefore interpret the relative TI-split gain together with MCC and the diffusion fingerprints below.

Table 12: Cross-seed consistency for the Hardanger experiment. Entries are descriptive means $\pm$ s.d. For MCC and GMR, the s.d. is across the 45 pairwise seed comparisons, not a standard error and not a standard deviation over ten independent seeds. For mean latent off-diagonal correlation, the s.d. is across the ten seeds. GMR thresholds the mean absolute drift Jacobian at $\tau = 0.05$.

Regime construction	MCC ↑	GMR ↑	Mean latent offdiag. corr. ↓
TI-split regimes (ours)	$\mathbf{0.806 \pm 0.061}$	$\mathbf{0.877 \pm 0.081}$	$\mathbf{0.182 \pm 0.022}$
One regime (control)	0.660 ± 0.051	0.813 ± 0.077	0.250 ± 0.040
Shuffled TI labels (control)	0.601 ± 0.067	0.830 ± 0.070	0.239 ± 0.052

G.5 Diffusion-fingerprint diagnostics

Table 13 shows that the TI-split regimes produce a strong learned diffusion contrast, whereas Shuffled TI labels collapse the fingerprint ratios to nearly one. This is the intended behavior of the control:

it preserves the existence of two batches in the objective but removes the physical TI-based regime information. Thus the reproducibility gap in Table 12 is accompanied by the expected difference in learned diffusion fingerprints. The large TI-split contrasts show that the construction supplies nontrivial learned diffusion variation, which is necessary but not sufficient for Assumption (S4); the shuffled-label control does not. In particular, this pooled contrast does not establish the pairwise distinctness of all coordinate-wise ratios. It is therefore a diagnostic rather than a formal verification of Assumption (S4) in the absence of ground-truth latent coordinates.

The One-regime control is omitted from the fingerprint table because it has only one diffusion covariance and therefore no variance ratio.

Table 13: Diffusion-fingerprint diagnostics. Larger contrast magnitudes are more favorable for the separation required by Assumption (S4). The parametric column uses $|\log(\tilde{\sigma}^2_{2,k}/\tilde{\sigma}^2_{1,k})|$, while the empirical column uses $|\log(C_{2,kk}/C_{1,kk})|$. Entries are mean $\pm$ s.d. over all seeds and latent dimensions (10×5 values); these are descriptive contrast magnitudes, not confidence intervals.

Regime construction	Parametric contrast ↑	Empirical contrast ↑
TI-split regimes (ours)	$\mathbf{0.786 \pm 0.144}$	$\mathbf{0.797 \pm 0.150}$
Shuffled TI labels (control)	0.004 ± 0.002	0.003 ± 0.002

G.6 Representative learned structure

Figure 15 shows the learned drift Jacobians for a randomly selected seed, shared across the three regime constructions and ordered as in the main text. In each heatmap, row i corresponds to drift component $\tilde{f}_{\psi,i}$ and column j corresponds to source coordinate $\tilde{z}_j$; the plotted value is the evaluation mean of $|\partial \tilde{f}_{\psi,i}/\partial \tilde{z}_j|$. Each panel is diagonal-dominant, but the off-diagonal structure differs across regime constructions. This is the qualitative pattern one would expect if the learned coordinates approximate a modal-like coordinate system for bridge vibrations: self-dynamics should dominate, while cross-coordinate effects should appear as a small number of structured couplings rather than as dense or scattered dependence. In this representative TI-split-regime run, the largest off-diagonal entries form a reciprocal pair between two learned coordinates: one drift component depends on the other coordinate with magnitude about 0.48, and the reverse dependence has magnitude about 0.40. A reciprocal pair of this kind is consistent with a learned two-dimensional dynamical subspace, as can arise from nearby vibration modes, non-proportional damping, aeroelastic or environmental coupling, or residual mixing in modal-like coordinates. By contrast, the controls under the same representative seed also contain off-diagonal entries, but these appear as more scattered one-way couplings across different coordinate pairs rather than as a single coherent coupled block. The TI-split structure is therefore more naturally aligned with this modal-like physical interpretation, while still serving only as an illustrative structural diagnostic that complements the aggregate reproducibility and fingerprint results above.

G.7 Limitations of the real-world diagnostic

This experiment should be read as a real-world diagnostic rather than a ground-truth recovery benchmark. The Hardanger data do not provide true latent coordinates or a true causal graph, so MCC and GMR measure cross-seed reproducibility after alignment, not absolute accuracy; absolute GMR values can also be influenced by matching many absent edges in sparse thresholded graphs. The diffusion-fingerprint ratios support the presence of TI-dependent diffusion variation, but they do not formally verify Assumption (S4): the true latent coordinates and diffusion covariances are unobserved, and the pooled contrast does not itself test pairwise distinctness of all ratios. Likewise, the modal-like interpretation of the representative drift Jacobian is not a substitute for external validation against finite-element models or modal analysis labels. Finally, the TI quantile split is a proxy for diffusion-regime variation. Its usefulness depends on maintaining enough stochastic-forcing contrast while keeping the deterministic structural dynamics approximately shared, exactly the tradeoff made explicit by the theory.

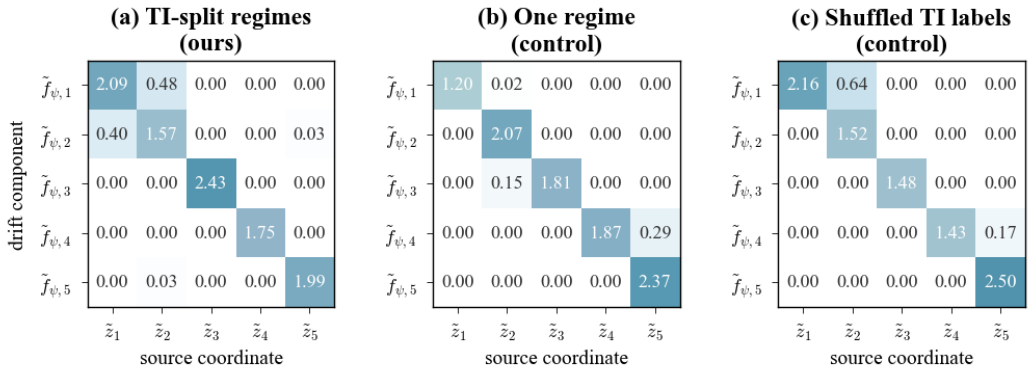

Figure 15: Drift-Jacobian diagnostics for a randomly selected seed, shared across the Hardanger regime constructions. Panels follow the main-text order: TI-split regimes (ours), One regime (control), and Shuffled TI labels (control). Rows index drift components $\tilde{f}_{\psi,i}$ and columns index source coordinates $\tilde{z}_j$; cell (i, j) is the mean absolute derivative $\overline{|\partial \tilde{f}_{\psi,i}/\partial \tilde{z}_j|}$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state the main scope, assumptions, theoretical claims, estimator, and experiments. The assumptions and limitations are further qualified in Sections 3, 4, 5, and 7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses the main limitations of the work. Appendix G.7 separately discusses the limitations of the real-world Hardanger diagnostic.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions for the OU and nonlinear SDE results are stated in Sections 3.1 and 3.2. The corresponding proofs are provided in Appendix C, with proof sketches in the main text.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The estimator is specified in Section 4 and Algorithm 1. Appendix E gives the synthetic data-generating process, observation model, training details, hyperparameters, metrics, and training-seed protocol, while Appendix G gives the Hardanger data construction and evaluation protocol.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code for reproducing the experiments is included in the supplementary material. The synthetic data are procedurally generated from Appendix E, and the real-world Hardanger recordings are available from the cited dataset DOI/repository described in Appendix G.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix E specifies the synthetic data generation, regimes, mixing functions, optimizer, training schedule, hyperparameters, and metrics. Appendix G specifies the Hardanger preprocessing, regime construction, controls, model settings, and evaluation diagnostics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The synthetic tables report mean $\pm$ standard deviation over five training seeds, as stated in Section 5 and Appendix E. The main Hardanger table reports mean cross-seed scores, and Appendix G reports descriptive standard deviations across seed pairs or seeds, with the relevant variability source specified there.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E.4 reports the compute setup used for the experiments: standard Slurm jobs requesting one NVIDIA L40S GPU and 32 GB memory per job, with five-seed synthetic batches typically completing in a few minutes. Appendix G.2 states that the Hardanger runs use the same single-GPU setup. The appendix also notes that the implementation can run on CPU or other CUDA GPUs, with CPU execution expected to be slower.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is theoretical and experimental, using synthetic simulations and non-human bridge sensor recordings. It does not involve human subjects or private personal data, and the Hardanger study is a diagnostic case study rather than a deployed decision system.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational work on identifiability for latent stochastic dynamics and is not tied to a deployed decision system or a high-risk application. We do not identify a direct societal harm beyond generic concerns that may apply to causal modeling tools.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, generative systems, scraped datasets, or other assets that require controlled access safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The Hardanger Bridge data source is credited through the original references in Section 5.2 and Appendix G. Appendix G states that the dataset DOI/repository record lists a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce a new dataset, benchmark, or pretrained model asset. The synthetic data used in the experiments are generated procedurally from the specifications in Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology, theory, and experiments do not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.