When Does LeJEPA Learn a World Model?

David Klindt
Cold Spring Harbor Laboratory
klindt@cshl.edu

Yann LeCun
New York University
yann.lecun@nyu.edu

Randall Balestriero
Brown University
randall_balestriero@brown.edu

Abstract

A representation that scrambles the true degrees of freedom of the world cannot support reliable planning or compositional generalization. We prove that LeJEPA (alignment plus Gaussian regularization) linearly recovers the world's latent variables from nonlinear observations, a property known as *linear identifiability*, in a class of worlds where latents evolve under stationary, additive-noise transitions. Our main result is that among all such worlds, the Gaussian is the *unique* latent distribution for which this guarantee holds. The forward direction rests on a spectral decomposition in which each degree of nonlinearity is strictly penalized by alignment, making the linear map the optimum; the converse rules out every non-Gaussian alternative. We further prove an *approximate identifiability* result where the guarantee degrades gracefully, and show that linear, orthogonal identifiability enables *optimal latent-space planning*. We validate the theory with experiments ranging from 2D examples to 4096-dimensional latents, including ablations, pixel-based robotic control, and natural images from ImageNet-X. Our theory turns an empirically successful recipe into a mathematical guarantee, providing the foundation for building World Models that provably recover the structure of the world.

1 Introduction

The promise of self-supervised learning (SSL) is that we can learn useful representations of the world without labeled data, just by observing and predicting. Joint-Embedding Predictive Architectures [JEPAs, 1] pursue this vision by training a representation to produce similar embeddings for related views of the same input, while a regularizer prevents the representation from collapsing to a trivial constant [2, 3]. The resulting representations have proven remarkably effective across image [4], video [5, 6], and latent-space planning [7, 8, 9]. However, a deeper question remains: *When is a learned representation a World Model, i.e., a faithful map of the world's latent structure?*

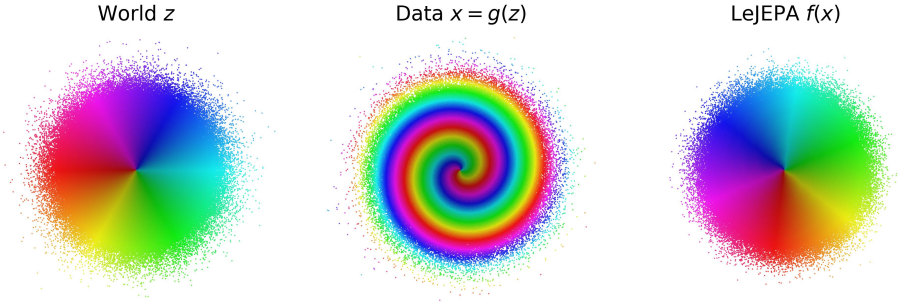


Figure 1: **LeJEPA learns the World Model. (left)** A world with independent Gaussian *latent variables*. **(center)** An unknown nonlinear process scrambles them into the *data* we observe. **(right)** LeJEPA [2] recovers the latent variables up to rotation. We prove this is the *unique* optimum.

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

Our answer: *When it linearly recovers the world's latent variables* (Fig. 1). A representation that entangles unrelated latent variables, scrambling an object's position with its color, or mixing velocity with texture, may score well on narrow tasks but will fail when the world changes [10]. What we need is *linear identifiability*, i.e., a mathematical guarantee that the learned representation recovers the underlying latent variables, up to simple symmetries [11]. Practitioners already test for this routinely. Every time a representation is evaluated via *linear probing* [12], the implicit question is whether the model has learned a *linear representation* of the latent variables [13, 14, 15, 16, 17, 18]. Without this, linear probes cannot recover latent variables exactly. Linear identifiability is thus a *necessary*, albeit not *sufficient* [19], condition for faithful linear probing.

Yet no identifiability results exist for JEPAs. This has remained out of reach because prior methods rely on implicit collapse prevention [20, 21, 22] with unspecified embedding distributions. A recent proposal changes the picture: LeJEPA [2] prevents collapse by explicitly regularizing the embedding distribution toward an isotropic Gaussian via Sketched Isotropic Gaussian Regularization (SIGReg). Along with an alignment loss this enables stable end-to-end training from raw pixels [9], but whether the resulting representation actually recovers the world's latent variables still remains open.

Contributions. We close this gap with the first identifiability result for JEPAs (Fig. 2). We consider a class of worlds with Gaussian latent variables and positive pairs from a stationary process with independent, additive-noise transitions, which are maximum-entropy choices [23]. We prove that LeJEPA learns a linearly identifiable representation *if and only if* the latent variables are *Gaussian*. The forward direction rests on a spectral decomposition that shows how every degree of nonlinearity is penalized; the converse rules out every non-Gaussian alternative. We further prove that identifiability degrades gracefully when the objectives are only *approximately* satisfied, and that linear identifiability suffices for *optimal latent-space planning*. We validate empirically across 2D mixings, up to 4096-dim latents, distributional ablations, pixel-based robotic control, and the approximate bound, and probe beyond the assumptions with a misspecified latent dimension, learned predictors [9], and natural images from ImageNet-X [24] (Sec. 6).[1]

2 Related Work

Representation Learning. JEPAs [1, 4, 5, 6] predict in representation space rather than pixel space, avoiding capacity wasted on irrelevant detail. More generally, SSL pulls positive views of the same content together while preventing collapse. Contrastive methods [25, 26, 27], such as InfoNCE [28], use negatives explicitly; non-contrastive methods substitute stop-gradient teachers [21, 20], covariance regularization [VICReg, 3, 29], or self-distillation with feature clustering [22, 30]. LeJEPA [2] adds an explicit Gaussianity regularizer (SIGReg); LeWorldModel [9] scales the recipe to action-conditioned control. InfoNCE, VICReg, and LeJEPA span a hierarchy of 'Gaussianity' constraints, from implicit [31, 32, 33, 34] through second-moment to full; we test all three (Tab. 1, App. H.7).

World Models. The internal-model concept spans cognitive science, control theory, and modern ML. Cognitive scientists framed mental simulation as the substrate of reasoning, perception, and motor control [35, 36, 37, 38, 39], with free-energy [40] and probabilistic-program [41] formalizations of the brain as a generative model. Cybernetics extended this to engineered systems [42, 43], while classical control developed dynamics on a given state space [44, 45]. Neural world models progressed from recurrent controller-model pairs [46, 47, 48] through latent dynamics from pixels [49, 50, 51, 52, 53], value-equivalent prediction [54], generative [55] and joint-embedding predictive [1, 9] architectures, to large generative video simulators [56, 57]. The JEPA argument is that pixel-perfect prediction wastes capacity [1]. We address the encoder side of this picture; classical control applies in the learned coordinates (Thm. 4).

Identifiability. Nonlinear ICA is unidentifiable without additional structure [58, 59]. Identifiability becomes possible when the world supplies it: non-stationarity [60], temporal dependence [61, 62], auxiliary variables [63, 64], contrastive learning [65], augmentation [66], mechanism sparsity [67], interventions [68, 69], or supervision [70, 71]. These results typically constrain the representation to a smooth diffeomorphism; our Hermite approach works for arbitrary measurable maps. Most closely related is slow feature analysis (SFA) [72], which JEPA objectives empirically recover

[1] Code, Lean proofs, and demo: `https://github.com/klindtlab/lejepa-identifiability`.

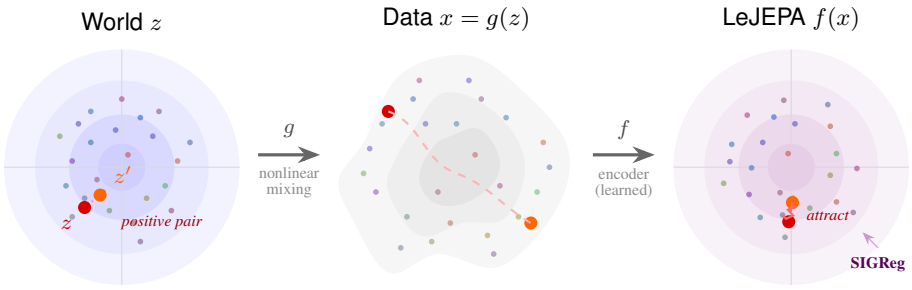


Figure 2: **LeJEPA Theory Illustration. (left)** The world has clean latent structure (Gaussian, disentangled) with correlated positive pairs. **(center)** An unknown nonlinear process produces the observations we actually see, scrambling the latent structure. **(right)** LeJEPA trains a representation with two objectives: pull positive pairs together (*attract*) and keep the embedding distribution Gaussian (SIGReg) to prevent collapse. We prove that the learned representation must be a rotation of the true latents, i.e., the representation is forced to learn the correct World Model (Thm. 1).

[73]; App. F contrasts in depth with SFA theory [74]. Similar spectral analysis in representation space has previously characterized other SSL objectives [75], and our Thm. 3 connects to recent quantitative stability work [76, 77]. The common lesson is that identifiability is always a joint statement about the *World* ('data generating process') and the *Learner* ('learning objective', LeJEPA).

3 The World and the Learner

What does it mean to *Learn the World Model?* Suppose the world has latent variables $z \in \mathbb{R}^N$,[2] for instance, position, velocity, color or lighting. These degrees of freedom are called *latent variables / sources* in the ICA literature, or *factors of variation* [78] in representation learning. We never observe z directly. Instead, an unknown process g generates the data / observations $x = g(z)$.[3] Think of g as rendering a 3D scene into an image, or mapping physical states to sensor readings. The process g can be highly nonlinear, it scrambles the clean latent structure into complicated, entangled observations. We train a representation f that maps observations back to representations. The ideal outcome is that f undoes g: the composed map $h = f \circ g$ should recover the original latent variables z. Of course, perfect recovery is too much to hope for, there are symmetries that cannot be resolved like the rotation invariance of a Gaussian. We will show that h must be a *linear function* of the true latents: $h(z) = Qz$. This is a necessary condition for linear probes to work, and it is what lets planning in the learned latent space cost the same as planning in the true latent space (Thm. 4).

3.1 The World

The world specifies a joint distribution $p(z, z')$ over positive pairs $(g(z), g(z'))$. In SSL, these are views of the same underlying content, e.g., two frames of a video, two augmentations of an image, two nearby time steps of a trajectory. Throughout, z_i denotes the i-th coordinate of $z \in \mathbb{R}^N$, i.e., one scalar latent variable, and z' the second view. A class of worlds is defined by three assumptions:

Assumptions (World)

(i) **Independence.** $p(z_i) \perp p(z_j)$ and transitions $p(z'_i \mid z_i) \perp p(z'_j \mid z_j)$ for all $i \neq j$.

(ii) **Stationarity.** Both views share the same marginal: $p(z) = p(z')$.

(iii) **Additive noise.** $z'_i = m_i(z_i) + \eta_i$ with η_i independent of z_i.

Independent latent variables are the standard assumption shared by ICA and disentangled representation learning [11, 78]. Stationarity means the generative process does not change between views. Additive noise is the simplest perturbation model: random fluctuations are added on top of a deterministic signal, as in sensor noise, Brownian motion, or data augmentation by jitter. It is also what continuous-time dynamics produce over short intervals: an Itô diffusion $dz = b(z)\,dt + \sigma\,dW$ with

[2] To help readers, all notation is summarized in the Glossary (App. 8).

[3] More philosophically, we only ever observe the *shadows* (x) of *reality* (z), *projected* (g) onto the walls of *Plato*'s cave.

constant volatility has, to first order in Δt, the transition $z' = z + b(z)\Delta t + \sigma\sqrt{\Delta t}\,\eta$, which is (iii) with $m(z) = z + b(z)\Delta t$, whatever the drift. Together, these assumptions define a general class of World Models. Our *forward* result (Sec. 5.1) will specialize this class by choosing a specific latent variable distribution (Gaussian). Our *converse* result (Sec. 5.2) shows that this is the *unique* choice yielding linear identifiability in this class of worlds.

3.1.1 The Gaussian World

We now make a specific distributional choice within the framework above. We assume *Gaussian latents*, i.e., $z \sim \mathcal{N}(0, I_N)$. This is the maximum-entropy distribution for a given mean and covariance [23]; it assumes as little structure as possible. Moreover, task-relevant latents are often aggregates of many micro-variables, which, by the central limit theorem, tend toward Gaussianity. Within this class we take the transition to be linear in the latent, $m_i(z_i) = \rho\, z_i$. Stationarity then requires $z' \sim \mathcal{N}(0, I_N)$, and by Cramér's decomposition theorem the independent noise must itself be Gaussian, which yields the Ornstein–Uhlenbeck (OU) transition [79, 80]:

$$z' = \rho\, z + \sqrt{1-\rho^2}\,\eta, \qquad \eta \sim \mathcal{N}(0, I_N), \quad \eta \perp z, \tag{1}$$

where $\rho \in (0, 1)$ controls the correlation between views. One can verify: $\mathbb{E}[z'] = 0$, $\mathrm{Var}(z') = \rho^2 I_N + (1-\rho^2) I_N = I_N$, and $\mathrm{Cov}(z, z') = \rho I_N$. The Gaussian is the *unique* distribution for which a linear channel with independent Gaussian noise preserves the marginal, a consequence of the Gaussian being a fixed point of convolution up to rescaling. The independence assumption is satisfied because the components of z are independent and the noise η has diagonal covariance, so each (z_i, z_i') evolves independently.

3.2 The Learner: LeJEPA

The representation is characterized by the composed map $h = f \circ g : \mathbb{R}^N \to \mathbb{R}^N$ with g the unknown generative function of the world and f our learned representation. The two components of LeJEPA [2] training are an invariance loss pulling positive pairs together and a regularizer that shapes the embedding distribution to prevent collapse, i.e., h (specifically, f in $h = f \circ g$) is trained to

$$\min_h \mathcal{L}(h) = \underbrace{\mathbb{E}\big[\|h(z') - h(z)\|^2\big]}_{\textbf{Alignment}} \qquad \text{s.t.} \qquad \underbrace{h(z) \sim \mathcal{N}(0, I_N)}_{\textbf{Gaussianity}} \tag{2}$$

We model the situation where SIGReg has succeeded, i.e., $h(z)$ matches the target Gaussian; in practice this holds approximately (Sec. 5.3). LeJEPA has no predictor, since SIGReg alone prevents collapse, so this is the objective as used in practice [2]; adding a linear or MLP predictor [9] leaves identifiability unchanged empirically (Sec. 6.5). We require nothing else about h except measurability so expectations are defined. The result thus applies to any neural network, with output dimension taken to equal the latent dimension N throughout (a mismatched output dimension is studied in Sec. 6.5). In addition, *whitening* (i.e., $\mathrm{Cov}(h(z)) = I_N$) fixes $\mathbb{E}[\|h(z)\|^2] = \mathbb{E}[\|h(z')\|^2] = N$, so

$$\mathcal{L}(h) = 2N - 2\sum_{i=1}^{N} \mathbb{E}[h_i(z')\, h_i(z)]\,. \tag{3}$$

Thus, *minimizing distance is equivalent to maximizing correlation* between the two views. The question becomes: among all measure-preserving maps h with $h(z) \sim \mathcal{N}(0, I_N)$, which achieves the highest correlation between positive pairs $h(z), h(z')$?

4 Spectral Analysis of the World

Before proving anything, let us build intuition for the mathematical tools behind our main results.

Transition Operator. The world's transition from z to z' (3.1) acts on any scalar function h_i through the *transition operator* $(Th_i)(z) = \mathbb{E}[h_i(z') \mid z]$, the expected value of h_i at the next view. This is the (stochastic) *Koopman operator* of the dynamics [81, 82]: linear even when the dynamics are not, with eigenfunctions $T\varphi_k = \lambda_k \varphi_k$ and, for the diffusion-type worlds we consider, eigenvalues $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Eigenfunctions with larger eigenvalues are more correlated across positive pairs, i.e., the most predictable features of the latent variables; this spectral view underlies both our forward and converse results.

Gaussian World: Hermite Polynomials. For Gaussian worlds (Sec. 3.1.1), the eigenfunctions are known in closed form: they are the Hermite polynomials $\{\mathrm{He}_k\}_{k\geq 0}$, the natural orthogonal basis for functions of Gaussian variables, analogous to Fourier modes for periodic functions. The eigenvalue of a degree-d Hermite polynomial is exactly ρ^d, a consequence of Mehler's formula [83]. This means any function $h_i(z)$ with zero mean and unit variance can be decomposed into a linear part (degree 1), a quadratic part (degree 2), a cubic part (degree 3), and so on, with variance fractions $w_1, w_2, w_3, \ldots$ summing to 1. The correlation across positive pairs decomposes:

$$\mathbb{E}[h_i(z')\, h_i(z)] = w_1 \cdot \rho + w_2 \cdot \rho^2 + w_3 \cdot \rho^3 + \cdots \leq \rho, \tag{4}$$

with equality if and only if $w_1 = 1$, i.e., h_i is linear. In words: *any nonlinear distortion of the representation strictly reduces the correlation between positive pairs*. This is the key intuition.

General Case: Sturm-Liouville Theory. For a general latent variable distribution (not necessarily Gaussian) evolving under constant diffusion, the eigenfunctions of the transition operator are characterized by a classical Sturm–Liouville (SL) equation [74]. The first non-constant eigenfunction φ_1 is always monotonic, providing identifiability up to a monotonic transformation. But *linear* identifiability requires φ_1 to be affine, which places a strong constraint on the latent variable distribution. This is the engine behind our converse result (Sec. 5.2): we show that only the Gaussian satisfies this constraint. Full details in App. A.2 (Gauß/Hermite) and App. F (SL connection).

5 Theory

We state four results that together characterize when LeJEPA learns the World Model. Thm. 1 shows that a Gaussian world is *linearly identifiable* with LeJEPA. Thm. 2 establishes that, within the class of worlds defined in Sec. 3.1, the Gaussian is the *unique* distribution with this property. Thm. 3 bounds the recovery error when objectives are not fully satisfied. Thm. 4 shows that linear identifiability enables *optimal planning* in latent space. All proofs have been verified in Lean 4 theorem prover modulo standard background lemmas axiomatized from the literature (App. G).

5.1 Forward Direction: LeJEPA Learns the World Model

Theorem 1 (LeJEPA Linear Identifiability)

Consider the Gaussian world (Sec. 3.1.1). Let $h : \mathbb{R}^N \to \mathbb{R}^N$ be any measurable map with $h(z) \sim \mathcal{N}(0, I_N)$. Then $\mathcal{L}(h) \geq 2(1-\rho)\, N$, with equality if and only if $h(z) = Qz$ for some orthogonal $Q \in O(N)$. At any such optimum, $h(z') \mid h(z) \sim \mathcal{N}(\rho\, h(z), (1-\rho^2)\, I_N)$.

Proof Sketch. Minimizing $\mathcal{L}$ is equivalent to maximizing $\sum_i \mathbb{E}[h_i(z')h_i(z)]$ (3). The spectral bound (4) gives $\mathbb{E}[h_i(z')h_i(z)] \leq \rho$ for each component, with equality if and only if h_i is linear. At equality, $h(z) = Qz$ for a matrix Q with unit-norm rows; Gaussianity forces $QQ^\top = I_N$, so Q is orthogonal. The transition then follows by direct substitution: $h(z') = \rho\, h(z) + \sqrt{1-\rho^2}\, Q\eta$, and since Q is orthogonal, $Q\eta \sim \mathcal{N}(0, I_N)$ independently of $h(z)$. (Proof in App. A.)

Interpretation. The representation has no choice but to learn the full World Model. Any representation satisfying the two LeJEPA objectives must recover a rotation/reflection of the true latent variables *and* the true transition dynamics. The only remaining ambiguity is a global rotation, which is inherent to the isotropic Gaussian. We provide an alternative proof via Dirichlet energy and the Mazur–Ulam theorem in App. E, operating in a more theoretical ($\rho \to 1$) regime.

5.2 Converse Direction: The Gaussian is Unique

Thm. 1 uses only that $h(z)$ has whitened covariance ($\mathrm{Cov}(h(z)) = I_N$, entailed by SIGReg, Fig. 11). Does the specific choice of Gaussian matter, or would any white distribution work? The optimal representation extracts the *slowest* features of the latent process [73]. Under additive noise, Sturm–Liouville theory orders these as increasingly oscillatory eigenfunctions [74]; the first is always monotonic, giving identifiability up to a monotonic transformation for *any* latent distribution. For *linear* identifiability this eigenfunction must be affine, which only the Gaussian satisfies.

Theorem 2 (Gaussian Uniqueness)

Consider any world satisfying Assumptions 3.1. Suppose every minimizer of (3) with $\mathrm{Cov}(h(z)) = I_N$ is linear, $h(z) = Qz$. Then z is Gaussian.

Proof Sketch. Demanding an affine eigenfunction forces the *score function* $(\log p)'$ of the latent distribution to be linear. If $\varphi(z_i) = az_i + b$ is an eigenfunction, then φ' is constant, and the eigenvalue equation collapses to a linear ODE for $(\log p)'$ in z_i. Solving it yields $\log p(z_i) \propto -(z_i - \mu)^2$ (the sign fixed by normalizability), which is Gaussian. The argument is per-component; independence of the latent variables lifts the conclusion to the joint distribution. (Proof in App. B.)

5.3 Approximate Identifiability

The preceding results assume exact optimality. In practice, both objectives are only approximately satisfied: alignment reaches a value near but not equal to its minimum, and the regularizer enforces approximate whitening. Identifiability degrades gracefully in both quantities:

Theorem 3 (Approximate Identifiability)

Consider the Gaussian world (3.1.1). Let $h : \mathbb{R}^N \to \mathbb{R}^N$ be measurable with $\mathbb{E}[h(z)] = 0$, satisfying: *Approximate alignment:* $\mathcal{L}(h) \le 2(1-\rho)\,\mathrm{tr}(\mathrm{Cov}(h(z))) + \delta$; *Approximate whitening:* $\|\mathrm{Cov}(h(z)) - I_N\|_F \le \varepsilon$. Define $D = \delta/(2\rho(1-\rho))$. Then there exists $Q \in O(N)$ such that

$$\mathbb{E}\big[\|h(z) - Qz\|^2\big] \;\le\; D \;+\; (\varepsilon + D)^2 \tag{5}$$

Interpretation. The bound has two pieces: a nonlinearity term D measuring how far h is from linear, and a distortion term $(\varepsilon + D)^2$ measuring how far the linear part is from orthogonal. Both vanish in the exact case ($\delta = \varepsilon = 0$), recovering Thm. 1. In practice the first dominates, so recovery error scales as $\delta/(2\rho(1-\rho))$: alignment is hard, whitening essentially free. (Proof in App. C.)

5.4 Optimal Latent Planning

Theorems 1–3 characterize what the encoder recovers. We now ask what orthogonal identifiability buys for planning, a key motivation of World Models. In control, an action a_t moves the latent from z_{t-1} to $z_t \sim p(\cdot \mid z_{t-1}, a_t)$. A plan $a_{1:T}$ from z_0 costs $\mathbb{E}\big[\ell_T(z_T) + \sum_{t=1}^{T} \ell(z_{t-1}, a_t)\big]$, with stage cost ℓ and terminal cost ℓ_T; let $V^*(z_0)$ denote the minimal cost and $a^*_{1:T}(z_0)$ the optimal plan.

Theorem 4 (Optimal Latent Planning)

Let $h(z) = Qz$ with $Q \in O(N)$, and let the costs be rotation-invariant:

$$\ell(Rz, a) = \ell(z, a), \qquad \ell_T(Rz) = \ell_T(z), \qquad \forall\, R \in O(N). \tag{6}$$

Then planning in the learned latent $\hat{z} = h(z)$, with transitions p mapped through h and the same costs, yields the same value and plan: $\hat{V}^*\big(h(z_0)\big) = V^*(z_0)$ and $\hat{a}^*_{1:T}\big(h(z_0)\big) = a^*_{1:T}(z_0)$.

Interpretation. Linear identifiability turns the learned representation into a useful World Model for planning: trajectories planned in the learned latent are mathematically identical to trajectories planned in the true world, with the same actions and the same value. (Proof in App. D.)

6 Experiments

We validate each result: linear identifiability on Gaussian latent variables (Sec. 6.1, Thm. 1); converse on a latent-distribution sweep and RL-policy [9] latents (Sec. 6.2, Thm. 2); approximate bound across all runs (Sec. 6.3, Thm. 3); and near-optimal latent planning [84] (Sec. 6.4, Thm. 4). Sec. 6.5 then tests what happens beyond the assumptions: a misspecified latent dimension, a learned predictor, and natural images from ImageNet-X [24].

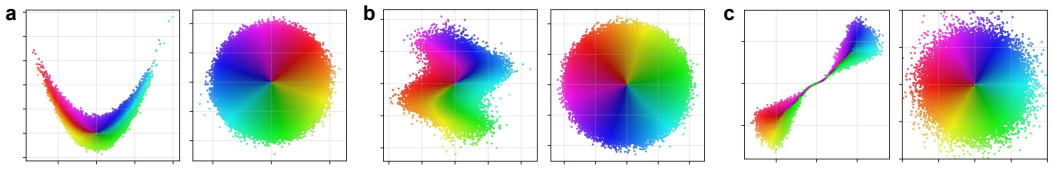


Figure 3: **2D Simulations.** Points colored by the polar angle and radius of the ground-truth latent variables $z \sim \mathcal{N}(0, I_2)$ (like Fig. 1). Observations (**a-c** left) $x = g(z)$ after nonlinear mixing: parabolic shear, sinusoidal shear, RealNVP coupling layer [85]. Learned embeddings (**a-c** right): LeJEPA recovers the isotropic Gaussian structure up to rotation, consistent with Thm. 1.

6.1 Forward: Linear Identifiability from Gaussian Latent Variables

We first verify Thm. 1 in a controlled 2D setting. We sample latents $z \sim \mathcal{N}(0, I_2)$ and apply four nonlinear mixing functions (Figures 1, 3): a norm-dependent rotation $g(z) = R(\pi\|z\|_2)\, z$ producing a spiral [c.f. 76], a sinusoidal shear $g(z_1, z_2) = (z_1 + \sin(1.5\, z_2),\ z_2)$, a parabolic shear $g(z_1, z_2) = (z_1,\ z_2 + z_1^2)$, and a RealNVP-style coupling layer [85]. All four are diffeomorphisms. In addition, the spiral is measure-preserving, demonstrating that Gaussianity of the observations alone does not suffice for identifiability; alignment is essential. We train a 4-layer MLP with the LeJEPA loss $\mathcal{L} = \lambda\, \mathcal{L}_{\text{SIG}} + (1 - \lambda)\, \mathcal{L}_{\text{inv}}$, where positive pairs are generated via the OU transition in (1). Fig. 3 shows that the learned representation inverts each nonlinear mixing up to rotation, consistent with Thm. 1. A grid search over λ and ρ (App. H.6) shows: too much Gaussianity ($\lambda = 0.5$) collapses the representation, while the best recovery occurs at low λ and high ρ.

Scaling to High Dimensions. We next sweep the latent dimension $N \in \{2^1, \ldots, 2^{12}\}$ (beyond, e.g., DINOv3's 768 embedding dimensions), using a RealNVP mixing and matched encoder [85] so that any failure is attributable to optimization rather than encoder expressivity. We test all three Gaussianity-enforcing classes on the same setup: SIGReg [2], VICReg [3], and InfoNCE [28]. Table 1: batch-statistic estimators (SIGReg, VICReg) maintain $R^2 > 0.999$ up to $N = 2048$; InfoNCE degrades at scale under fixed kernel width. At $N = 4096$ the default batch of 256 becomes the bottleneck for both batch-statistic estimators, and a batch of 1024 restores SIGReg to $R^2 = 0.9994$ (App. H.10). Thm. 1 is thus supported across mixing functions, dimensions, and hyperparameters; where methods differ, the cause is estimation (kernel width, batch size), not identifiability.

Table 1: **Scaling Comparison Across Regularizers** (mean ± std, 5 seeds). Three Gaussianity-enforcing objectives on a shared RealNVP mixing and matched encoder. SIGReg and VICReg maintain $R^2 > 0.999$ up to N=2048; InfoNCE matches at low N but degrades at scale under fixed kernel width σ=1. At N=4096 all three fail at batch size 256; †raising the batch size to 1024 restores SIGReg (3 seeds). Per-method details in Tabs. 6–8.

	Mixing	**Linear identifiability** $R^2(h \to z)$		
N	$R^2(x \to z)$ $\pm$std $\times 10^{-3}$	SIGReg $\pm$std $\times 10^{-7}$	VICReg $\pm$std $\times 10^{-7}$	InfoNCE $\pm$std $\times 10^{-3}$
2	$0.781_{\pm 2.1}$	$0.999998_{\pm 3.4}$	$0.999996_{\pm 8.4}$	$0.950961_{\pm 1.6}$
4	$0.727_{\pm 24}$	$0.999996_{\pm 12}$	$0.999987_{\pm 54}$	$0.910871_{\pm 8.2}$
8	$0.728_{\pm 10}$	$0.999993_{\pm 9.0}$	$0.999988_{\pm 4.8}$	$0.886818_{\pm 42}$
16	$0.734_{\pm 6.3}$	$0.999988_{\pm 4.9}$	$0.999987_{\pm 4.6}$	$0.999880_{\pm 0.01}$
32	$0.737_{\pm 2.3}$	$0.999981_{\pm 7.2}$	$0.999981_{\pm 9.4}$	$0.907809_{\pm 26}$
64	$0.737_{\pm 1.5}$	$0.999966_{\pm 7.4}$	$0.999968_{\pm 8.1}$	$0.648496_{\pm 3.1}$
128	$0.739_{\pm 0.61}$	$0.999938_{\pm 3.2}$	$0.999942_{\pm 7.2}$	$0.566955_{\pm 6.6}$
256	$0.742_{\pm 0.49}$	$0.999884_{\pm 7.9}$	$0.999889_{\pm 7.2}$	$0.696587_{\pm 0.49}$
512	$0.749_{\pm 0.30}$	$0.999775_{\pm 6.7}$	$0.999785_{\pm 6.9}$	$0.704393_{\pm 0.26}$
1024	$0.763_{\pm 0.17}$	$0.999561_{\pm 12}$	$0.999582_{\pm 11}$	$0.720241_{\pm 0.20}$
2048	$0.768_{\pm 0.05}$	$0.999043_{\pm 7.9}$	$0.999094_{\pm 7.5}$	$0.727140_{\pm 0.10}$
4096	$0.768_{\pm 0.03}$	$0.663 \to 0.999449^{\dagger}$	0.652	$0.727111_{\pm 0.02}$

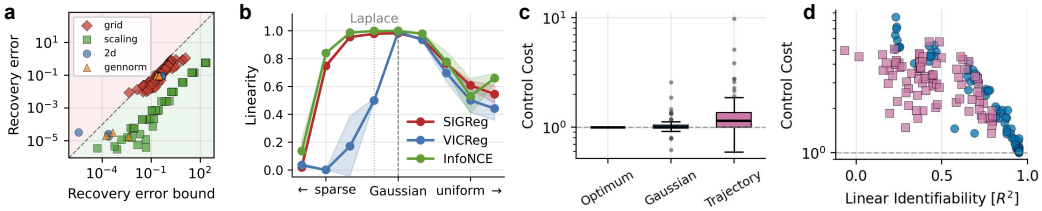


Figure 4: **Experimental Results. (a)** *Bound Verification.* SIGReg runs across grid, 2D, scaling, and gennorm α=2 lie below the diagonal, confirming Thm. 3. Two near-zero outliers reflect finite-sample noise. **(b)** *Gaussian Optimality.* Linear recovery, $R^2(h \to z)$, peaks at Gaussian, illustrating Thm. 2. SIGReg's Gaussianization of h is more robust to non-Gaussian latent variable distributions than whitening. **(c)** Control cost over $K = 30$ random start-goal pairs (path length ≥ 1, ideal 1). The Gaussian encoder is statistically indistinguishable from the oracle; the Trajectory encoder is biased upward. **(d)** Control cost decreases with linear identifiability R^2, supporting Thm. 4.

6.2 Converse: Non-Gaussian Latent Variables Break Linear Identifiability

Next, we validate Thm. 2's predictions that non-Gaussian latents break linear identifiability.

Latent-Distribution Sweep. We sweep the latent variable through the generalized normal family with shape parameter α ($\alpha \to 0$ heavy-tailed, $\alpha = 1$ Laplace, $\alpha = 2$ Gaussian, $\alpha \to \infty$ uniform). Linear recovery peaks at $\alpha = 2$ across all three objectives (Fig. 4b, App. H.7), illustrating Thm. 2; SIGReg and InfoNCE retain a wider plateau than VICReg for heavy-tailed latents.

Pixel-Based RL Trajectories. The DMC Reacher [84] has two joints, giving a 2D latent state $z = (\theta_0, \theta_1)$ (Fig. 13). We train a CNN encoder with LeJEPA (App. H.11) under two conditions sharing the same rendering pipeline but different distributions: *(i) OU:* Gaussian samples $z \sim \mathcal{N}(0, I_2)$ as before (1); *(ii) Trajectory:* joint-angle pairs $(z_t, z_{t+\delta})$ with δ frames separation from 10k RL episodes [9]. Table 2 (left) shows that OU pairs attain $R^2 = 0.95$ at $\rho = 0.99$, with the two joint dimensions linearly recovered. In contrast, real trajectories break the Gaussian assumption (App. H.11, Figs. 14,15). Table 2 (right) shows per-dimension R^2 is anisotropic, and total R^2 never exceeds 0.5, consistent with Thm. 2. The trajectory condition violates several theory assumptions at once (non-Gaussian marginals, anisotropic $\rho_0 \neq \rho_1$, joint-limit wrapping).

6.3 Approximation Bound and Loss Predict Identifiability

Thm. 3 bounds the recovery error by the whitening error $\varepsilon = \|\mathrm{Cov}(h(z)) - I\|_F$ and the alignment gap $\delta = \mathcal{L}(h) - 2(1-\rho)\,\mathrm{tr}(\mathrm{Cov}(h(z)))$. For each run we compute ε, δ, the bound $D + (\varepsilon + D)^2$, and the actual recovery error $\min_{Q \in O(N)} \mathbb{E}[\|h(z) - Qz\|^2]$ (Fig. 4a). The bound holds across grid search, 2D mixings, scaling, and the latent-distribution sweep, supporting Thm. 3 empirically. As a practical corollary, training loss is a reliable proxy for identifiability (App. H.9).

Table 2: **Linear Identifiability on RL Trajectories.** (mean ± std, 3 seeds). **(left)** Gaussian OU pairs recover true latents with rising R^2. **(right)** RL-policy trajectories [9] (δ: frame separation) show reduced identifiability because of anisotropy ($\rho_0 \neq \rho_1$) and non-Gaussian transitions (Figs. 14,15).

OU (Gaussian)			**Trajectory (non-Gaussian)**					
ρ	$R^2(z \to h)$	$R^2(h \to z)$	δ	ρ_0	ρ_1	$R^2(z \to h)$	$R^2(h \to z_0)$	$R^2(h \to z_1)$
0.30	0.67±2e-02	0.67±2e-02	1	1.000	0.999	-0.39±1e-01	0.71±3e-02	-0.03±4e-02
0.50	0.86±1e-02	0.86±1e-02	2	0.999	0.996	-0.47±5e-02	0.73±4e-03	0.01±2e-03
0.70	0.93±7e-03	0.93±7e-03	4	0.997	0.991	-0.05±3e-01	0.51±1e-01	0.43±9e-02
0.80	0.94±3e-04	0.94±3e-04	8	0.992	0.982	0.50±2e-02	0.80±2e-03	0.78±6e-03
0.90	0.95±7e-04	0.95±7e-04	16	0.981	0.963	0.44±4e-02	0.63±2e-02	0.87±2e-02
0.95	0.95±2e-04	0.95±2e-04	32	0.959	0.928	0.45±6e-02	0.62±2e-02	0.81±1e-02
0.99	0.95±4e-04	0.95±4e-04	64	0.915	0.863	0.44±4e-02	0.55±3e-02	0.77±2e-02

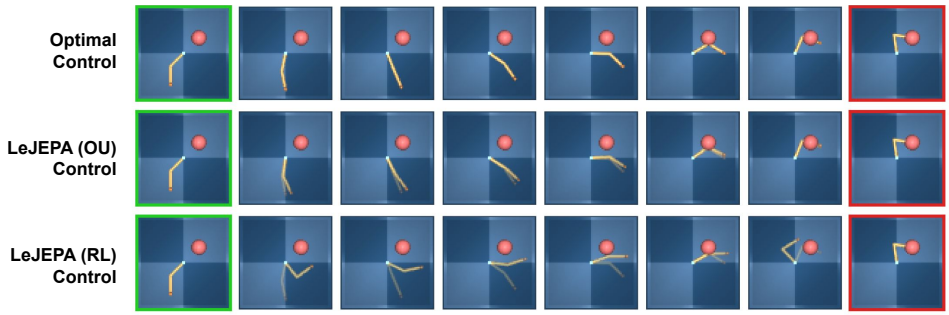


Figure 5: **Linear Identifiability Enables Latent-Space Planning.** Interpolation in each encoder's latent space between fixed start and goal frames, decoded by nearest-neighbor retrieval (Fig. 19). **top)** Oracle (joint-space straight line) [84]. **middle)** Gaussian encoder (Sec. 3.1.1: OU, $\rho = 0.99$) tracks the oracle (overlaid) closely. **bottom)** RL trajectory [9] encoder (stride $\delta = 8$) deviates.

6.4 Linear Identifiability Enables Latent-Space Planning

Thm. 4 predicts that any planner using a rotation-invariant cost attains identical performance in the learned and true latent. We test a natural instance: goal-reaching, where the straight line $\hat{z}_0 \to \hat{z}^*$ is the cost-minimizing trajectory. For an encoder satisfying $h(z) \approx Qz$, this latent straight line decodes to a near-straight path in the true latent; without, the same plan induces a curved path (Fig. 19). Figs. 4c,d; 5 confirm both. The Gaussian encoder's straight-latent plans decode to oracle-quality joint-space trajectories (Figs. 5, 4c), while the Trajectory encoder inflates control cost. Across all models, control cost tracks linear identifiability R^2 monotonically (Fig. 4d). *Thus, linear identifiability is the structural property that turns a faithful World Model into a useful planner.*

6.5 Beyond the Assumptions: Latent Dimension, Natural Images, Predictors

Unknown latent dimension. Our theorems assume the encoder output dimension M equals the latent dimension N. What if it does not (Fig. 6)? Below N, the encoder keeps an M-dimensional subspace of the latents intact rather than compressing all of them, so recovery tracks M/N. Above N, recovery saturates. The true N can be read off without ground truth from the participation ratio of the embedding covariance, $\mathrm{PR} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$: in both our settings, $\mathrm{PR} = M$ exactly up to $M{=}N$ and falls below M beyond (App. H.13). Which subspace is kept, and when the system resorts to *superposition* [15, 86, 18], remains open.

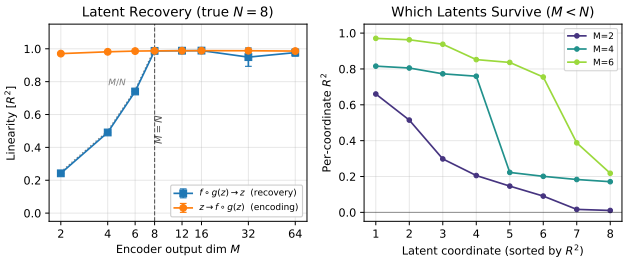


Figure 6: **Misspecified latent dimension** (true $N{=}8$, 5 seeds). **Left:** recovery tracks M/N below N and saturates above. **Right:** for $M < N$, M latents are recovered well and the rest poorly.

Natural images. Without ground-truth latents, we test the observable consequence: annotated factors should be linearly decodable. Following Reizinger et al. [70, Fig. 3], we probe the ImageNet-X factors [24] in ViT-B/16 encoders trained on ImageNet-1k with LeJEPA, DINO [22], and iBOT [87] under a shared codebase and schedule. LeJEPA leads (Tab. 3), also against DINOv2 [88] and DINOv3 [30] trained on far more data, while meeting SIGReg only approximately, the regime of Thm. 3 (App. H.16).

Table 3: **ImageNet-X Linear Probes.** Mean balanced accuracy over the nine factors with $>$ 1000 balanced samples (5 repetitions; full table in App. H.16).

	LeJEPA	DINO	iBOT	v2	v3	init	pixels
data		IN-1k		142M	1.7B	–	–
acc.	**.677**	.589	.628	.656	.645	.545	.533

Predictors. LeJEPA has no predictor, but action-conditioned World Models do [9]. Aligning $h(z')$ with a learned prediction $P(h(z))$ leaves recovery unchanged (0.989 without, 0.987 linear, 0.988 MLP predictor; N=8, 5 seeds), and the linear P converges to the true transition operator ($\|P - \rho I\|_F = 0.011$). For the MLP this is empirical only, as an expressive predictor could absorb a nonlinear distortion of h (App. H.14).

7 Limitations

Are the latents Gaussian? The Gaussian is the maximum-entropy distribution for a given mean and covariance [23], making it the least-assumption prior. Whether real-world latents are Gaussian is unknowable from observations alone, but the same applies to the non-Gaussianity assumptions of classical ICA [89] (maybe a structuralist perspective [90] is more agnostic). There is a scale-of-description argument in favor: individual micro-variables may be non-Gaussian, but task-relevant latent variables are often aggregates that tend toward Gaussianity by the central limit theorem. These are arguments rather than evidence, but Thm. 3 does not need them, and we test non-Gaussian latents directly (Fig. 4b). Dependent, non-Euclidean and other latents remain open questions.

Finite samples and optimization. Our result is a population-level statement about the global optimum. Thm. 3 shows the guarantee degrades continuously with respect to alignment gap and covariance deviation, but does not address how these scale with sample size [91] or training dynamics [92]. Empirically, recovery emerges abruptly after an initial plateau of a few hundred to a few thousand steps (App. H.15). The few bound violations we observe (Fig. 4a) are consistent with finite-sample estimation noise in ε and δ.

8 Discussion

Outlook. Linear identifiability addresses the *state* side of a World Model; the action-conditioned transition $\hat{p}(\hat{z}' \mid \hat{z}, a)$ must still be learned. Lifting identifiability to that setting connects naturally to interventional causal representation learning, where actions act as interventions on the latent variables [68, 69], and onward to causal graphs [93]. Actions may also relax the Gaussian assumption: action-conditioned World Models can prevent collapse through inverse-dynamics regularization instead of an isotropic Gaussian target [94], a promising route to identifiability for non-Gaussian latents that are subject to control. Our assumptions may not hold exactly in practice, but Thm. 3 shows the guarantee degrades gracefully.

Practical Implications. Two implications, one for data and one for objective. *Data:* the Reacher result shows that the same physical system supports identifiability when sampled isotropically (OU) but not under a goal-directed policy, whose marginals collapse onto a low-entropy region of latent space (Tab. 2); for self-supervised pretraining, exploration approximating an isotropic random walk keeps the data in the regime our theory covers. *Objective:* SIGReg, VICReg, and InfoNCE all yield linear identifiability when their assumptions hold, and our experiments suggest they fail in different ways: pair-based estimators are sensitive to kernel choice at scale (Tab. 1), moment-based estimators to non-Gaussian latents (App. H.7). Which estimator is preferable in practice, and how the choice interacts with batch size, architecture, and data regime, is left to future work.

A Theoretical Foundation for World Models. Our theory inverts the classical ICA narrative: in *linear* ICA the Gaussian is the one distribution where source separation fails [89]; in our nonlinear setting it is exactly what enables it. The payoff is in control: linear identifiability turns a learned representation into a usable *state* for a control system, so any orthogonally-invariant cost transfers from the true world to the learned latent without modification, and simple planning (Thm. 4) and linear probing come for free. *This is what it means to provably learn a World Model.*

Acknowledgments

We thank Aapo Hyvärinen, Patrik Reizinger, Attila Juhos, Heejeong (Hazel) Nam, Christian Internò and Habon Issa for their feedback and inputs. DK acknowledges the CSHL GPU cluster with assistance from the US National Institutes of Health Grant S10OD028632-01.

References

[1] Yann LeCun. A path towards autonomous machine intelligence, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. (Cited on pages 1 and 2.)

[2] Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. (Cited on pages 1, 2, 4, 7, 34, 37, 46, and 57.)

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *CoRR*, abs/2105.04906, 2021. URL https://arxiv.org/abs/2105.04906. (Cited on pages 1, 2, and 7.)

[4] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. (Cited on pages 1 and 2.)

[5] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. (Cited on pages 1 and 2.)

[6] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. (Cited on pages 1 and 2.)

[7] Vlad Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Stress-testing offline reward-free reinforcement learning: A case for planning with latent dynamics models. In *7th Robot Learning Workshop: Towards Robots with Human-Level Abilities*, 2025. URL https://openreview.net/forum?id=jON7H6A9UU. (Cited on page 1.)

[8] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, 2025. (Cited on page 1.)

[9] Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026. (Cited on pages 1, 2, 4, 6, 8, 9, 10, 27, 40, 44, 55, and 57.)

[10] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. (Cited on page 2.)

[11] Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1):1–33, 2024. (Cited on pages 2 and 3.)

[12] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. (Cited on page 2.)

[13] David E Rumelhart and Adele A Abrahamson. A model for analogical reasoning. *Cognitive Psychology*, 5(1):1–28, July 1973. ISSN 0010-0285. doi: 10.1016/0010-0285(73)90023-6. URL https://www.sciencedirect.com/science/article/pii/0010028573900236. (Cited on page 2.)

[14] Geoffrey E Hinton. Learning distributed representations of concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 8, 1986. (Cited on page 2.)

[15] Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216, 1990. URL https://www.sciencedirect.com/science/article/pii/000437029090007M. Publisher: Elsevier. (Cited on pages 2 and 9.)

[16] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear Algebraic Structure of Word Senses, with Applications to Polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, December 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00034. URL https://direct.mit.edu/tacl/article/43451. (Cited on page 2.)

[17] Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models, July 2024. URL http://arxiv.org/abs/2311.03658. arXiv:2311.03658 [cs, stat]. (Cited on page 2.)

[18] David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, and Nina Miolane. From superposition to sparse codes: interpretable representations in neural networks. *arXiv preprint arXiv:2503.01824*, 2025. (Cited on pages 2 and 9.)

[19] Vitória Barin Pacela, Shruti Joshi, Isabela Camacho, Simon Lacoste-Julien, and David Klindt. Stop probing, start coding: Why linear probes and sparse autoencoders fail at compositional generalisation, 2026. URL https://arxiv.org/abs/2603.28744. (Cited on page 2.)

[20] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. (Cited on page 2.)

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. (Cited on page 2.)

[22] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. (Cited on pages 2, 9, 45, and 57.)

[23] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957. (Cited on pages 2, 4, and 10.)

[24] Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. ImageNet-X: Understanding model mistakes with factor of variation annotations. In *International Conference on Learning Representations*, 2023. (Cited on pages 2, 6, 9, 45, 55, and 57.)

[25] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. (Cited on page 2.)

[26] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. (Cited on page 2.)

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. (Cited on page 2.)

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. (Cited on pages 2 and 7.)

[29] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *International Conference on Machine Learning*, 2021. (Cited on page 2.)

[30] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni,

Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. (Cited on pages 2, 9, 45, and 57.)

[31] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. (Cited on page 2.)

[32] Haochen Li, Xin Zhou, Luu Anh Tuan, and Chunyan Miao. Rethinking negative pairs in code search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12760–12774, 2023. (Cited on page 2.)

[33] Daniel Eftekhari and Vardan Papyan. On the importance of gaussianizing representations. *arXiv preprint arXiv:2505.00685*, 2025. (Cited on page 2.)

[34] Roy Betser, Eyal Gofer, Meir Yossef Levi, and Guy Gilboa. InfoNCE induces Gaussian distribution. In *International Conference on Learning Representations*, 2026. (Cited on page 2.)

[35] Kenneth J W Craik. *The Nature of Explanation*. Cambridge University Press, 1943. (Cited on page 2.)

[36] Edward C Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. (Cited on page 2.)

[37] Philip N Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, 1983. (Cited on page 2.)

[38] Daniel M Wolpert, Zoubin Ghahramani, and Michael I Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995. (Cited on page 2.)

[39] Richard L Gregory. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038):181–197, 1980. (Cited on page 2.)

[40] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. (Cited on page 2.)

[41] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. (Cited on page 2.)

[42] Roger C Conant and W Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970. (Cited on page 2.)

[43] B A Francis and W M Wonham. The internal model principle of control theory. *Automatica*, 12(5):457–465, 1976. (Cited on page 2.)

[44] Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957. (Cited on page 2.)

[45] Rudolph E Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering*, 82(Series D):35–45, 1960. (Cited on page 2.)

[46] Derrick H Nguyen and Bernard Widrow. Neural networks for self-learning control systems. *IEEE Control systems magazine*, 10(3):18–23, 1990. (Cited on pages 2 and 28.)

[47] Jürgen Schmidhuber. Making the world differentiable: On using fully recurrent self-supervised neural networks for dynamic reinforcement learning and planning in non-stationary environments. Technical Report FKI-126-90, Institut für Informatik, Technische Universität München, 1990. (Cited on page 2.)

[48] Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 216–224, 1990. (Cited on page 2.)

[49] Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in Neural Information Processing Systems*, 2015. (Cited on page 2.)

[50] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in Atari games. In *Advances in Neural Information Processing Systems*, 2015. (Cited on page 2.)

[51] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, 2016. (Cited on page 2.)

[52] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3): 440, 2018. (Cited on pages 2 and 28.)

[53] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR, 2019. (Cited on page 2.)

[54] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020. (Cited on page 2.)

[55] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 2025. (Cited on page 2.)

[56] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Technical Report, 2024. URL `https://openai.com/index/video-generation-models-as-world-simulators/`. (Cited on page 2.)

[57] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C Y Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning*, pages 4603–4623. PMLR, 2024. (Cited on page 2.)

[58] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. (Cited on page 2.)

[59] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019. (Cited on page 2.)

[60] Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, 2016. (Cited on page 2.)

[61] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*, 2017. (Cited on page 2.)

[62] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021. (Cited on page 2.)

[63] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2020. (Cited on page 2.)

[64] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, April 2019. URL `https://proceedings.mlr.press/v89/hyvarinen19a.html`. (Cited on page 2.)

[65] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, 2021. (Cited on page 2.)

[66] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021. (Cited on page 2.)

[67] Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, 2022. (Cited on page 2.)

[68] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, 2023. (Cited on pages 2, 10, and 27.)

[69] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36:45419–45462, 2023. (Cited on pages 2, 10, and 27.)

[70] Patrik Reizinger, Alice Bizeul, Attila Juhos, Julia E Vogt, Randall Balestriero, Wieland Brendel, and David Klindt. Cross-entropy is all you need to invert the data generating process. *arXiv preprint arXiv:2410.21869*, 2024. (Cited on pages 2, 9, and 45.)

[71] Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021. (Cited on page 2.)

[72] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002. (Cited on pages 2 and 30.)

[73] Vlad Sobal, S V Jyothir, Siddhartha Jalagam, Nicolas Carion, Kyunghyun Cho, and Yann LeCun. Joint embedding predictive architectures focus on slow features. *arXiv preprint arXiv:2211.10831*, 2022. (Cited on pages 3, 5, and 30.)

[74] Henning Sprekeler, Tiziano Zito, and Laurenz Wiskott. An extension of slow feature analysis for nonlinear blind source separation. *Journal of Machine Learning Research*, 15:921–947, 2014. (Cited on pages 3, 5, 23, 30, 31, and 32.)

[75] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods, 2022. URL `https://arxiv.org/abs/2205.11508`. (Cited on pages 3 and 32.)

[76] Simon Buchholz and Bernhard Schölkopf. Robustness of nonlinear representation learning. *arXiv preprint arXiv:2503.15355*, 2025. (Cited on pages 3, 7, and 28.)

[77] Beatrix MG Nielsen, Emanuele Marconato, Andrea Dittadi, and Luigi Gresele. When does closeness in distribution imply representational similarity? an identifiability perspective. *arXiv preprint arXiv:2506.03784*, 2025. (Cited on page 3.)

[78] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013. (Cited on page 3.)

[79] George E Uhlenbeck and Leonard S Ornstein. On the theory of the Brownian motion. *Physical Review*, 36(5):823–841, 1930. (Cited on page 4.)

[80] Joseph L Doob. The Brownian movement and stochastic equations. *Annals of Mathematics*, 43(2):351–369, 1942. (Cited on page 4.)

[81] Bernard O. Koopman. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931. (Cited on page 4.)

[82] Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1–3):309–325, 2005. (Cited on page 4.)

[83] Ferdinand Gustav Mehler. Über die entwicklung einer function von beliebig vielen variablen nach Laplaceschen functionen höherer ordnung. *Journal für die reine und angewandte Mathematik*, 66:161–176, 1866. (Cited on pages 5 and 33.)

[84] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. (Cited on pages 6, 8, 9, and 39.)

[85] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. (Cited on pages 7, 34, 35, and 38.)

[86] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy Models of Superposition, September 2022. URL `http://arxiv.org/abs/2209.10652`. arXiv:2209.10652 [cs]. (Cited on page 9.)

[87] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. (Cited on pages 9, 45, and 57.)

[88] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. (Cited on pages 9, 45, and 57.)

[89] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3): 287–314, 1994. (Cited on page 10.)

[90] Shruti Joshi, Aaron Mueller, David Klindt, Wieland Brendel, Patrik Reizinger, and Dhanya Sridhar. Causality is key for interpretability claims to generalise. *arXiv preprint arXiv:2602.16698*, 2026. (Cited on page 10.)

[91] Qi Lyu and Xiao Fu. On finite-sample identifiability of contrastive learning-based nonlinear independent component analysis. In *International Conference on Machine Learning*, pages 14582–14600. PMLR, 2022. (Cited on page 10.)

[92] James B Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J Fetterman, and Joshua Albrecht. On the stepwise nature of self-supervised learning. In *International Conference on Machine Learning*, pages 31852–31876. PMLR, 2023. (Cited on page 10.)

[93] Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-jepa: Learning world models through object-level latent interventions. *arXiv preprint arXiv:2602.11389*, 2026. (Cited on page 10.)

[94] Petr Ivashkov, Randall Balestriero, and Bernhard Schölkopf. Sensorimotor world models: Perception for action via inverse dynamics. *arXiv preprint arXiv:2606.20104*, 2026. (Cited on page 10.)

[95] Arthur E. Bryson and Yu-Chi Ho. *Applied Optimal Control: Optimization, Estimation, and Control*. Hemisphere Publishing / Taylor & Francis, revised edition, 1975. (Cited on page 28.)

[96] Stanisław Mazur and Stanisław Ulam. Sur les transformations isométriques d'espaces vectoriels normés. *Comptes Rendus de l'Académie des Sciences de Paris*, 194:946–948, 1932. (Cited on page 29.)

[97] Mathias Franzius, Henning Sprekeler, and Laurenz Wiskott. Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computational Biology*, 3(8):e166, 2007. (Cited on page 30.)

[98] Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S Zimmermann, and Wieland Brendel. Infonce: Identifying the gap between theory and practice. *arXiv preprint arXiv:2407.00143*, 2024. (Cited on page 31.)

[99] Hans Gebelein. Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *ZAMM – Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941. doi: 10.1002/zamm.19410210604. (Cited on page 31.)

[100] Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10(3–4):441–451, 1959. doi: 10.1007/BF02024507. (Cited on page 31.)

[101] H. O. Lancaster. The structure of bivariate distributions. *The Annals of Mathematical Statistics*, 29(3):719–736, 1958. doi: 10.1214/aoms/1177706532. (Cited on pages 31 and 33.)

[102] Henning Sprekeler. On the relation of slow feature analysis and Laplacian eigenmaps. *Neural Computation*, 23(12):3287–3302, 2011. (Cited on page 32.)

[103] Amit Singer and Ronald R Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008. (Cited on page 32.)

Appendix Overview

	Contents	Page
Appendix A	Proof of Thm. 1 (linear identifiability), incl. Hermite background	19
Appendix B	Proof of Thm. 2 (Gaussian uniqueness)	23
Appendix C	Proof of Thm. 3 (approximate identifiability)	24
Appendix D	Proof and discussion of Thm. 4 (optimal latent planning)	26
Appendix E	Alternative proof via Dirichlet energy	28
Appendix F	Prior work: connection to Slow Feature Analysis	30
Appendix G	Lean 4 formal verification	32
Appendix H	Experimental details and additional results	34

Symbol	Description
World Model	
$z, z' \in \mathbb{R}^N$	True latent variables and positive pair (second view)
z_i	i-th coordinate of z, one scalar latent variable
N	Latent dimension
M	Encoder output dimension ($M = N$ in theorems; App. H.13)
m_i	Deterministic part of the additive-noise transition, $z'_i = m_i(z_i) + \eta_i$
$\rho \in (0, 1)$	Autocorrelation of the Ornstein–Uhlenbeck transition, $m_i(z_i) = \rho\, z_i$
$\eta \sim \mathcal{N}(0, I_N)$	Independent transition noise
$K > 0$	Diffusion coefficient (Sturm–Liouville setting)
Maps	
g	Unknown nonlinear mixing (generative) function; $x = g(z)$
f	Learned encoder; $y = f(x)$
$h = f \circ g$	Composed map from latent space to representation space
$J_h(z)$	Jacobian matrix of h at z
Linear algebra	
$O(N)$	Orthogonal group (matrices with $QQ^\top = I_N$)
$Q \in O(N)$	Orthogonal recovery matrix (exact identifiability)
Losses and objectives	
$\mathcal{L}(h)$	Alignment loss: $\mathbb{E}[\|h(z') - h(z)\|^2]$
$\mathcal{L}_{\text{inv}}$	Empirical invariance (alignment) loss
$\mathcal{L}_{\text{SIG}}$	SIGReg Gaussianity regularizer
λ	Regularization weight balancing $\mathcal{L}_{\text{SIG}}$ and $\mathcal{L}_{\text{inv}}$
Hermite polynomials and spectral decomposition	
$\mathrm{He}_k(x)$	Probabilist's Hermite polynomial of degree k
$\hat{h}_k(x)$	Normalized Hermite polynomial: $\mathrm{He}_k(x)/\sqrt{k!}$
$\alpha \in \mathbb{N}^N$	Multi-index $(\alpha_1, \ldots, \alpha_N)$; $\|\alpha\| = \alpha_1 + \cdots + \alpha_N$ is the total degree
$H_\alpha(z)$	Multivariate Hermite basis function: $\prod_j \hat{h}_{\alpha_j}(z_j)$
$c_\alpha^{(i)}$	Hermite coefficient of h_i at multi-index α
$w_d^{(i)}$	Variance fraction of h_i at degree d: $\sum_{\|\alpha\|=d} (c_\alpha^{(i)})^2$
T	Transition operator: $(T\varphi)(z) = \mathbb{E}[\varphi(z') \mid z]$
Approximate identifiability	
$\delta \geq 0$	Alignment gap (excess loss above the linear optimum)
$\varepsilon \geq 0$	Covariance deviation: $\|\mathrm{Cov}(h(z)) - I_N\|_F$
D	Normalized alignment gap: $\delta/(2\rho(1-\rho))$

A Proof of Theorem 1 (Forward Direction)

Theorem 1 (LeJEPA Linear Identifiability)

Consider the Gaussian world (Sec. 3.1.1). Let $h : \mathbb{R}^N \to \mathbb{R}^N$ be any measurable map with $h(z) \sim \mathcal{N}(0, I_N)$. Then $\mathcal{L}(h) \geq 2(1-\rho)\, N$, with equality if and only if $h(z) = Qz$ for some orthogonal $Q \in O(N)$. At any such optimum, $h(z') \mid h(z) \sim \mathcal{N}(\rho\, h(z), (1-\rho^2)\, I_N)$.

A.1 Intuition: Why Should This Force Linearity?

Before the formal proof, let us build intuition for why linearity is forced. A linear orthogonal map $h(z) = Qz$ clearly preserves the Gaussian and achieves correlation ρ per component, giving $\mathcal{L} = 2(1-\rho)N$. *Could a nonlinear map do better?*

Here is a suggestive information-theoretic argument. The variables z and z' are jointly Gaussian with correlation ρ, so $I(z; z') = -\frac{N}{2}\ln(1-\rho^2)$. For Gaussians, mutual information depends only on the correlation. If a nonlinear h achieved higher correlation while keeping Gaussian marginals, and if the joint $(h(z), h(z'))$ were also Gaussian, then $I(h(z); h(z')) > I(z; z')$, violating the data processing inequality. So nonlinear maps cannot beat linear ones.

This argument has the right conclusion but two gaps: (1) a nonlinear h can produce a non-Gaussian *joint* even when both marginals are Gaussian, so the mutual information formula does not apply; (2) even if nonlinear maps cannot beat linear ones, perhaps they can *tie*. To close both gaps, we need to understand precisely how nonlinearity interacts with the transition between positive pairs. The right tool turns out to be the Hermite polynomials, the natural orthogonal basis for functions of Gaussian variables. Just as Fourier analysis decomposes a signal into frequencies, the Hermite expansion decomposes any function into a linear part, a quadratic part, a cubic part, and so on. The key payoff is that the transition between positive pairs acts on each degree separately, and attenuates higher degrees more. This lets us compute the correlation of any function in closed form, closing both gaps simultaneously.

A.2 Hermite Polynomial Background

The proof rests on the Hermite polynomials, the natural orthogonal basis for functions of Gaussian random variables. This subsection collects the definitions and properties used throughout the paper.

Definitions and orthogonality. The probabilist's Hermite polynomials $\mathrm{He}_k(x)$ are defined by the Rodrigues formula:

$$\mathrm{He}_k(x) = (-1)^k\, e^{x^2/2}\, \frac{d^k}{dx^k}\, e^{-x^2/2}, \qquad k = 0, 1, 2, \ldots \tag{7}$$

The first several are:

$$\mathrm{He}_0(x) = 1, \quad \mathrm{He}_1(x) = x, \quad \mathrm{He}_2(x) = x^2 - 1, \quad \mathrm{He}_3(x) = x^3 - 3x \tag{8}$$

Each He_k is a degree-k polynomial. Note that $\mathrm{He}_2(x) = x^2 - 1$ includes a correction term that makes it orthogonal to the constant and linear terms under the Gaussian measure. More generally, these polynomials satisfy the orthogonality relation under the standard Gaussian measure γ with density $p(x) = (2\pi)^{-1/2} e^{-x^2/2}$:

$$\mathbb{E}_{x\sim\gamma}[\mathrm{He}_j(x)\, \mathrm{He}_k(x)] = k!\, \delta_{jk}. \tag{9}$$

Normalizing by $\hat{h}_k(x) := \mathrm{He}_k(x)/\sqrt{k!}$ gives an orthonormal basis of $L^2(\mathbb{R}, \gamma)$: $\mathbb{E}[\hat{h}_j(x)\, \hat{h}_k(x)] = \delta_{jk}$. The completeness of this basis means that any $f \in L^2(\mathbb{R}, \gamma)$ can be uniquely expanded as $f(x) = \sum_{k=0}^{\infty} a_k \hat{h}_k(x)$ with $\sum_k a_k^2 = \mathbb{E}[f(x)^2] < \infty$ (Parseval's identity).

Multivariate extension. For $z = (z_1, \ldots, z_N) \in \mathbb{R}^N$ with independent standard Gaussian components, the multivariate Hermite basis is constructed as tensor products. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathbb{N}^N$ with total degree $|\alpha| = \alpha_1 + \cdots + \alpha_N$, define:

$$H_\alpha(z) = \prod_{j=1}^{N} \hat{h}_{\alpha_j}(z_j). \tag{10}$$

The family $\{H_\alpha\}_{\alpha\in\mathbb{N}^N}$ is an orthonormal basis of $L^2(\mathbb{R}^N, \gamma_N)$, where γ_N is the standard Gaussian measure on $\mathbb{R}^N$. The degree $|\alpha|$ captures the "order of nonlinearity":

- Degree 0: $H_0(z) = 1$. The constant function.
- Degree 1: $H_{e_j}(z) = z_j$ for $j = 1, \ldots, N$. The N linear coordinate functions.
- Degree 2: terms like $z_i z_j$ (for $i \neq j$) and $(z_j^2 - 1)/\sqrt{2}$. Quadratic nonlinearities.
- Degree d: captures d-th order nonlinear structure.

Hermite expansion of measure-preserving maps. Any component $h_i : \mathbb{R}^N \to \mathbb{R}$ of a map satisfying $h(z) \sim \mathcal{N}(0, I_N)$ has $\mathbb{E}[h_i(z)] = 0$ and $\mathbb{E}[h_i(z)^2] = 1$. By completeness, it expands as:

$$h_i(z) = \sum_{|\alpha|\geq 1} c_\alpha^{(i)} H_\alpha(z), \qquad \sum_{|\alpha|\geq 1} (c_\alpha^{(i)})^2 = 1. \tag{11}$$

The constant term ($|\alpha| = 0$) vanishes because $\mathbb{E}[h_i] = 0$. The normalization $\sum (c_\alpha^{(i)})^2 = 1$ follows from Parseval's identity and $\mathbb{E}[h_i^2] = 1$. Grouping by total degree defines the *variance at degree* d:

$$w_d^{(i)} = \sum_{|\alpha|=d} (c_\alpha^{(i)})^2, \qquad \sum_{d\geq 1} w_d^{(i)} = 1. \tag{12}$$

If h_i is a linear function of z, then $w_1^{(i)} = 1$ and $w_d^{(i)} = 0$ for all $d \geq 2$. If h_i has any nonlinear component, then $w_d^{(i)} > 0$ for some $d \geq 2$.

A.3 Proof

Setup and notation. Since $z \sim \mathcal{N}(0, I_N)$, the multivariate Hermite polynomials $\{H_\alpha\}_{\alpha\in\mathbb{N}^N}$ form an orthonormal basis of $L^2(\mathbb{R}^N, \gamma_N)$ (Appendix A.2). Each component h_i has $\mathbb{E}[h_i(z)] = 0$ and $\mathbb{E}[h_i(z)^2] = 1$ by the Gaussianity constraint, and admits the expansion

$$h_i(z) = \sum_{|\alpha|\geq 1} c_\alpha^{(i)} H_\alpha(z), \qquad \sum_{|\alpha|\geq 1} (c_\alpha^{(i)})^2 = 1. \tag{13}$$

Grouping by total degree $d = |\alpha|$, let $w_d^{(i)} = \sum_{|\alpha|=d} (c_\alpha^{(i)})^2$ be the fraction of variance at degree d. Then $w_1^{(i)} + w_2^{(i)} + w_3^{(i)} + \cdots = 1$. If h_i is linear, then $w_1^{(i)} = 1$; any nonlinearity spills variance into $d \geq 2$. By equation (3), minimizing $\mathcal{L}$ is equivalent to maximizing $\sum_i \mathbb{E}[h_i(z')h_i(z)]$. The proof establishes that each term satisfies $\mathbb{E}[h_i(z')h_i(z)] \leq \rho$, with equality if and only if h_i is linear. The argument has three main steps. First, we need to know what the transition $z \to z'$ does to each Hermite component individually. Second, we need the cross-correlation structure between all components across the two views: does the degree-2 part of $h_i(z')$ correlate with the degree-1 part of $h_i(z)$? Third, we combine these to bound the total correlation any function can achieve.

Step 1: The transition attenuates each Hermite degree by ρ^d. We show that $\mathbb{E}_\eta[\mathrm{He}_k(z')] = \rho^k \mathrm{He}_k(z)$ for $z' = \rho z + \sqrt{1-\rho^2}\,\eta$ with $\eta \sim \mathcal{N}(0, 1)$ independent of $z \sim \mathcal{N}(0, 1)$.

We use the exponential generating function for the Hermite polynomials:

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} \mathrm{He}_k(x) = \exp\left(tx - \frac{t^2}{2}\right). \tag{14}$$

This identity encodes all Hermite polynomials into a single exponential, allowing us to compute the effect of noise on all degrees simultaneously.

Evaluate (14) at $x = z' = \rho z + \sqrt{1-\rho^2}\,\eta$:

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} \mathrm{He}_k(z') = \exp\left(t(\rho z + \sqrt{1-\rho^2}\,\eta) - \frac{t^2}{2}\right). \tag{15}$$

Since η is independent of z, the right side factors:

$$\exp\left(t\rho z - \frac{t^2}{2}\right) \cdot \exp\left(t\sqrt{1-\rho^2}\,\eta\right). \tag{16}$$

The first factor depends only on z and comes out of $\mathbb{E}_\eta$. For the second factor, we use the Gaussian moment generating function: if $\eta \sim \mathcal{N}(0,1)$, then $\mathbb{E}[e^{s\eta}] = e^{s^2/2}$ for any s. Setting $s = t\sqrt{1-\rho^2}$:

$$\mathbb{E}_\eta\left[\exp\left(t\sqrt{1-\rho^2}\,\eta\right)\right] = \exp\left(\frac{t^2(1-\rho^2)}{2}\right). \tag{17}$$

Multiplying the two pieces, the exponents combine:

$$t\rho z - \frac{t^2}{2} + \frac{t^2(1-\rho^2)}{2} = t\rho z - \frac{t^2}{2} + \frac{t^2}{2} - \frac{t^2\rho^2}{2} = t\rho z - \frac{(t\rho)^2}{2}. \tag{18}$$

So we obtain:

$$\sum_{k=0}^{\infty} \frac{t^k}{k!}\,\mathbb{E}_\eta[\mathrm{He}_k(z')] = \exp\left((\rho t)z - \frac{(\rho t)^2}{2}\right). \tag{19}$$

The right side of (19) has exactly the form of the generating function (14) with t replaced by ρt:

$$\exp\left((\rho t)z - \frac{(\rho t)^2}{2}\right) = \sum_{k=0}^{\infty} \frac{(\rho t)^k}{k!}\,\mathrm{He}_k(z) = \sum_{k=0}^{\infty} \frac{t^k}{k!}\,\rho^k\,\mathrm{He}_k(z). \tag{20}$$

Matching coefficients of $t^k/k!$:

$$\mathbb{E}_\eta[\mathrm{He}_k(z')] = \rho^k\,\mathrm{He}_k(z). \tag{21}$$

Since $0 < \rho < 1$, we have $\rho^2 < \rho$, $\rho^3 < \rho^2$, and so on. *The linear component ($k = 1$) survives the transition best; every higher-degree component is strictly more attenuated.*

This contraction is specific to Gaussians. The proof succeeds because the Gaussian MGF ($e^{s^2/2}$) partially cancels the $e^{-t^2/2}$ in the Hermite generating function, leaving $e^{-t^2\rho^2/2}$, which is the generating function again with $t \to \rho t$. This replacement is where the ρ^k contraction originates. For non-Gaussian distributions, the transition operator still has a spectral decomposition with ordered eigenvalues (see Appendix F), but the eigenfunctions are no longer Hermite polynomials and the eigenvalues no longer decay geometrically. In particular, the first eigenfunction is monotonic but generally nonlinear, which is why monotonic (not linear) identifiability holds in the general case.

Step 2: Different Hermite components do not interact across views. Step 1 tells us what happens to each Hermite component in isolation. But to compute the correlation $\mathbb{E}[h_i(z')h_i(z)]$ for an arbitrary function h_i, we also need to know whether different components interact across views: does the degree-2 part of $h_i(z')$ correlate with the degree-1 part of $h_i(z)$?

The answer is no. We show that $\mathbb{E}[H_\alpha(z')\,H_\beta(z)] = \rho^{|\alpha|}\,\delta_{\alpha\beta}$ (Mehler's formula). The $\delta_{\alpha\beta}$ means that cross-terms vanish: the degree-1 part of $h(z')$ only correlates with the degree-1 part of $h(z)$, the degree-2 part only with degree-2, and so on. Each degree-d component contributes correlation ρ^d. The Hermite basis *diagonalizes* the transition.

In the univariate case, using the law of iterated expectations:

$$\mathbb{E}\left[\hat{h}_k(z')\,\hat{h}_j(z)\right] = \mathbb{E}_z\left[\mathbb{E}_\eta\left[\hat{h}_k(z')\right]\,\hat{h}_j(z)\right]. \tag{22}$$

By the contraction property (21), $\mathbb{E}_\eta[\hat{h}_k(z')] = \rho^k\,\hat{h}_k(z)$. Substituting:

$$\mathbb{E}\left[\hat{h}_k(z')\,\hat{h}_j(z)\right] = \rho^k\,\mathbb{E}_z\left[\hat{h}_k(z)\,\hat{h}_j(z)\right] = \rho^k\,\delta_{jk}, \tag{23}$$

where the last step uses orthonormality of the $\hat{h}_k$ under γ.

In the multivariate case, since z has independent components and the noise in the Ornstein–Uhlenbeck channel acts independently on each coordinate, the multivariate Hermite products factorize:

$$\mathbb{E}[H_\alpha(z')\,H_\beta(z)] = \prod_{j=1}^{N} \mathbb{E}\left[\hat{h}_{\alpha_j}(z'_j)\,\hat{h}_{\beta_j}(z_j)\right] = \prod_{j=1}^{N} \rho^{\alpha_j}\,\delta_{\alpha_j\beta_j} = \rho^{|\alpha|}\,\delta_{\alpha\beta}. \tag{24}$$

Step 3: Nonlinearity strictly reduces correlation. Now we can compute the total correlation of *any* function through the transition. Since the Hermite basis diagonalizes the problem, the answer is simply a weighted sum over degrees, and because ρ^d decreases with d, putting variance on higher degrees always costs correlation.

Expanding h_i in the Hermite basis (13) and applying Mehler's formula (24):

$$\mathbb{E}[h_i(z')\, h_i(z)] = \sum_{|\alpha|\geq 1} \sum_{|\beta|\geq 1} c_\alpha^{(i)} c_\beta^{(i)}\, \mathbb{E}[H_\alpha(z')\, H_\beta(z)] = \sum_{|\alpha|\geq 1} (c_\alpha^{(i)})^2\, \rho^{|\alpha|}. \tag{25}$$

Note that all cross-terms vanish ($\delta_{\alpha\beta} = 0$ for $\alpha \neq \beta$), and each surviving term contributes its squared coefficient times $\rho^{|\alpha|}$.

Grouping by degree $d = |\alpha|$ with $w_d = \sum_{|\alpha|=d} (c_\alpha^{(i)})^2$:

$$\mathbb{E}[h_i(z')\, h_i(z)] = w_1 \cdot \rho + w_2 \cdot \rho^2 + w_3 \cdot \rho^3 + \cdots \tag{26}$$

with $w_d \geq 0$ and $w_1 + w_2 + w_3 + \cdots = 1$. This is a weighted average of $\rho, \rho^2, \rho^3, \ldots$ with weights summing to 1. Since $0 < \rho < 1$, each $\rho^d \leq \rho$ with strict inequality for $d \geq 2$:

$$w_1 \cdot \rho + w_2 \cdot \rho^2 + w_3 \cdot \rho^3 + \cdots \leq w_1 \cdot \rho + w_2 \cdot \rho + w_3 \cdot \rho + \cdots = \rho \cdot (w_1 + w_2 + \cdots) = \rho. \tag{27}$$

Equality requires $w_d \cdot \rho^d = w_d \cdot \rho$ for every d, which means $w_d = 0$ for all $d \geq 2$ (since $\rho^d < \rho$ for $d \geq 2$). So all the variance is at degree 1: $w_1 = 1$. The degree-1 Hermite basis consists of the coordinate functions $\{z_1, \ldots, z_N\}$, so:

$$h_i(z) = \sum_{j=1}^{N} c_{e_j}^{(i)}\, z_j = \langle q_i, z \rangle \tag{28}$$

for some vector q_i with $\|q_i\|^2 = w_1 = 1$. That is, h_i must be a linear function.

No nonlinear map can match the correlation of a linear map. This closes both gaps from the information-theoretic argument. That argument failed because it assumed the joint distribution of $(h(z), h(z'))$ is Gaussian, which need not hold for nonlinear h. The Hermite analysis avoids this entirely: Mehler's formula computes $\mathbb{E}[h_i(z')h_i(z)]$ exactly for any $h_i \in L^2(\gamma_N)$, using only the joint Gaussianity of (z, z'), not of $(h_i(z), h_i(z'))$. The prior argument also left open the possibility that a nonlinear map could match the correlation of a linear one. The strict inequality above rules this out: any nonlinearity ($w_d > 0$ for some $d \geq 2$) strictly reduces correlation.

Step 4: Assembly. Summing the correlation bound over all N components:

$$\sum_{i=1}^{N} \mathbb{E}[h_i(z')h_i(z)] \leq \rho N, \tag{29}$$

which by (3) gives $\mathcal{L}(h) \geq 2(1-\rho)N$.

At equality, every h_i is linear: $h_i(z) = \langle q_i, z \rangle$ with $\|q_i\| = 1$. So $h(z) = Qz$ for some matrix Q with unit-norm rows. The Gaussianity constraint requires $\mathrm{Cov}(h(z)) = QQ^\top = I_N$, so Q is orthogonal. Conversely, any orthogonal Q achieves $\sum_i \mathbb{E}[\langle Q_i, z' \rangle \langle Q_i, z \rangle] = \rho N$, giving equality.

Step 5: Recovery of the transition. At the optimum, $h(z') = Qz' = \rho\, Qz + \sqrt{1-\rho^2}\, Q\eta = \rho\, h(z) + \sqrt{1-\rho^2}\, Q\eta$. Since Q is orthogonal and $\eta \sim \mathcal{N}(0, I_N)$, we have $Q\eta \sim \mathcal{N}(0, I_N)$ independently of $h(z)$. Therefore $h(z') \mid h(z) \sim \mathcal{N}(\rho\, h(z), (1-\rho^2) I_N)$, which is the same conditional as $z' \mid z$. The encoder recovers the latent variables and their dynamics up to a global rotation. $\square$

B Proof of Theorem 2 (Gaussian Uniqueness)

Theorem 2 (Gaussian Uniqueness)

Consider any world satisfying Assumptions 3.1. Suppose every minimizer of (3) with $\mathrm{Cov}(h(z)) = I_N$ is linear, $h(z) = Qz$. Then z is Gaussian.

Since the latent variables are independent, the transition operator separates, and it suffices to prove the result for a single component z_i whose stationary distribution has full support on $\mathbb{R}$.

B.1 Background

The forward result (Thm. 1) shows that the optimal encoder extracts the first few eigenfunctions of the transition operator T (Section 4): these are the functions of the latent variables with the highest temporal autocorrelation. If the first (slowest) non-constant eigenfunction is linear, the encoder recovers the latent variable itself; if it is nonlinear, the encoder recovers a nonlinear transformation of the latent variable, and linear identifiability fails. The question is therefore: for which latent variable distributions is the first eigenfunction linear?

The proof uses the Sturm–Liouville characterization of these eigenfunctions [74]. Assumption (iii) (η_i independent of z_i) implies the noise variance does not depend on the state; in continuous time, this is the constant-diffusion regime in which Sturm–Liouville machinery applies. The transition operator T has an infinitesimal generator $\mathcal{D}$ that shares the same eigenfunctions; $\mathcal{D}$ is a second-order differential operator whose structure makes the eigenvalue problem tractable (see Appendix F for the full connection to Slow Feature Analysis). For a scalar latent variable evolving under constant diffusion $K > 0$ with stationary density $p(z_i) > 0$, the generator, written in self-adjoint form, acts on test functions φ as

$$\mathcal{D}\varphi = \frac{1}{p(z_i)} \frac{d}{dz_i} \Big[K\, p(z_i)\, \varphi'(z_i) \Big]. \tag{30}$$

The eigenfunctions φ_k are properties of the latent process, not of the encoder h; the encoder inherits its structure from them.

When the score function diverges ($(\log p)'(z_i) \to \mp\infty$ as $z_i \to \pm\infty$), the operator has a purely discrete spectrum $0 < \lambda_1 < \lambda_2 < \cdots$ with eigenfunctions $\varphi_1, \varphi_2, \ldots$ forming a complete orthogonal basis of $L^2(\mathbb{R}, p\, dz)$. The eigenvalues are ordered by increasing "speed": φ_1 is the slowest non-constant feature, φ_2 the next slowest, and so on. By classical Sturm–Liouville oscillation theory, φ_k has exactly $k - 1$ interior zeros. In particular, φ_1 is always monotonic, so one always gets identifiability up to a monotonic transformation. Linear identifiability requires φ_1 to be affine, and since an affine function is monotonic (zero interior zeros), only φ_1 can be affine: no higher eigenfunction can be. (For heavy-tailed densities such as Laplace, $\mathcal{D}$ develops continuous spectrum and the discrete picture does not apply; this does not affect our result, since assuming an affine eigenfunction exists forces p to be Gaussian, which has discrete spectrum.) For the full connection between the Sturm–Liouville framework and Slow Feature Analysis, see Appendix F.

B.2 Proof

The key observation is that the operator $\mathcal{D}$ encodes the score function $(\log p)'$ of the latent variable distribution. Demanding that the first eigenfunction is affine forces the score to be linear in z_i, and a linear score uniquely characterizes the Gaussian.

Suppose $\varphi(z_i) = az_i + b$ is an eigenfunction of $\mathcal{D}$ with eigenvalue $\lambda > 0$. Since $\varphi' = a$ is constant, the operator simplifies:

$$\mathcal{D}[az_i + b] = a\, K (\log p)'(z_i). \tag{31}$$

The eigenfunction equation $\mathcal{D}\varphi = -\lambda\varphi$ requires $a\, K(\log p)'(z_i) = -\lambda(az_i + b)$, giving

$$K(\log p)'(z_i) = -\lambda z_i - \lambda b/a. \tag{32}$$

Integrating both sides:

$$\log p(z_i) = -\frac{\lambda}{2K} \left(z_i + \frac{b}{a} \right)^2 + C, \tag{33}$$

which is a Gaussian density with mean $-b/a$. No other density satisfies this equation. Under the zero-mean convention $b = 0$, and a is fixed by unit-variance normalization. □

Remark: the OU transition is a modeling choice. Gaussian marginals and Assumptions 3.1 do not by themselves force the OU transition (1). Any map m that preserves $\mathcal{N}(0, 1)$ yields a stationary transition, e.g., $m(z) = z$ for $|z| > 1$ and $m(z) = -z$ otherwise, with $\eta = 0$; composing such a map with the OU channel, $z' = \rho\, m(z) + \sqrt{1-\rho^2}\,\eta$, gives examples with $\eta \neq 0$ (we thank Siddharth Ancha for pointing this out). Thm. 1 relies on the OU form through its Hermite spectrum, and such transitions have different eigenfunctions. The OU form is, however, forced in the continuous-time setting that motivates Assumption (iii) (Sec. 3.1): a diffusion $dz = b(z)\,dt + \sigma\,dW$ with constant volatility has stationary density $p(z) \propto \exp\left(\frac{2}{\sigma^2}\int^z b(u)\,du\right)$, which is Gaussian if and only if the drift b is linear, i.e., if and only if the process is OU. The counterexamples above are discrete-time maps that no such diffusion produces over a short interval.

C Proof of Theorem 3 (Approximate Identifiability)

Theorem 3 (Approximate Identifiability)

Consider the Gaussian world (3.1.1). Let $h : \mathbb{R}^N \to \mathbb{R}^N$ be measurable with $\mathbb{E}[h(z)] = 0$, satisfying: *Approximate alignment:* $\mathcal{L}(h) \leq 2(1-\rho)\,\mathrm{tr}(\mathrm{Cov}(h(z))) + \delta$; *Approximate whitening:* $\|\mathrm{Cov}(h(z)) - I_N\|_F \leq \varepsilon$. Define $D = \delta/(2\rho(1-\rho))$. Then there exists $Q \in O(N)$ such that

$$\mathbb{E}\big[\|h(z) - Qz\|^2\big] \;\leq\; D \;+\; (\varepsilon + D)^2 \tag{34}$$

We prove the stability bound in three steps: (1) the alignment gap controls nonlinear energy, (2) the whitening constraint forces the linear part near orthogonality, and (3) orthogonality of Hermite degrees gives a clean Pythagorean decomposition.

Setup and notation. Since $z \sim \mathcal{N}(0, I_N)$, each zero-mean component h_i of the encoder admits a Hermite expansion (Appendix A.2):

$$h_i(z) \;=\; \sum_{|\alpha|\geq 1} c_\alpha^{(i)}\, H_\alpha(z), \tag{35}$$

where the constant term vanishes by the zero-mean assumption. We split this into a linear part and a nonlinear residual.

The linear part is Mz, where $M \in \mathbb{R}^{N\times N}$ is the matrix of correlations between encoder outputs and latent variables:

$$M_{ij} \;=\; \mathbb{E}\big[h_i(z)\, z_j\big]. \tag{36}$$

This extracts exactly the degree-1 Hermite coefficients, since all higher-degree terms are orthogonal to the coordinate functions z_j under the Gaussian measure.

The nonlinear residual $\nu(z) = h(z) - Mz$ contains everything at Hermite degree ≥ 2. By construction, $\mathbb{E}[\nu(z)\, z^\top] = 0$: the residual is uncorrelated with the latent variables. We define the total nonlinear energy as

$$W_{\mathrm{nl}} \;=\; \mathbb{E}\big[\|\nu(z)\|^2\big]. \tag{37}$$

Since the linear and nonlinear parts live on orthogonal Hermite subspaces, the total variance decomposes without cross terms:

$$\mathrm{tr}\big(\mathrm{Cov}(h(z))\big) \;=\; \|M\|_F^2 \;+\; W_{\mathrm{nl}}. \tag{38}$$

This Pythagorean decomposition is the key identity underlying Step 3 below.

C.1 Step 1: Alignment gap controls nonlinear energy

Recall that the alignment loss decomposes as

$$\mathcal{L}(h) \;=\; 2N \;-\; 2\sum_{i=1}^{N} \mathbb{E}[h_i(z')\, h_i(z)]\,, \tag{39}$$

where (z, z') is a pair from the OU process with autocorrelation ρ. Mehler's formula (equation (24) in Appendix A) states that Hermite polynomials diagonalize the OU conditional expectation:

$$\mathbb{E}\big[H_\alpha(z') \,\big|\, z\big] \;=\; \rho^{|\alpha|}\, H_\alpha(z), \tag{40}$$

so each Hermite degree $|\alpha|$ is an eigenspace with eigenvalue $\rho^{|\alpha|}$. Substituting the Hermite expansion of h into the cross-correlation $\mathbb{E}[h_i(z')\, h_i(z)]$ and applying this eigenstructure yields an exact decomposition over degrees:

$$\mathcal{L}(h) \;=\; 2\sum_{i=1}^{N} \sum_{|\alpha|\geq 1} (1-\rho^{|\alpha|})\, (c_\alpha^{(i)})^2. \tag{41}$$

The weight $1-\rho^{|\alpha|}$ increases with degree: higher-order components are penalized more heavily because they decorrelate faster under the OU transition.

We now split the sum by degree. At degree 1 the weight is exactly $1-\rho$. At degree $|\alpha| \geq 2$, the bound $\rho^{|\alpha|} \leq \rho^2$ (valid for $0 < \rho < 1$) gives $1-\rho^{|\alpha|} \geq 1-\rho^2 = (1-\rho)(1+\rho)$, so

$$\mathcal{L}(h) \;\geq\; 2(1-\rho)\, \|M\|_F^2 \;+\; 2(1-\rho)(1+\rho)\, W_{\mathrm{nl}}. \tag{42}$$

Substituting the Pythagorean decomposition $\|M\|_F^2 = \mathrm{tr}(\mathrm{Cov}(h(z))) - W_{\mathrm{nl}}$ from (38) and collecting terms:

$$\begin{aligned}
\mathcal{L}(h) \;&\geq\; 2(1-\rho)\big[\mathrm{tr}(\mathrm{Cov}(h(z))) - W_{\mathrm{nl}}\big] + 2(1-\rho)(1+\rho)\, W_{\mathrm{nl}} \\
&=\; 2(1-\rho)\,\mathrm{tr}(\mathrm{Cov}(h(z))) + 2(1-\rho)\big[(1+\rho)-1\big]\, W_{\mathrm{nl}} \\
&=\; 2(1-\rho)\,\mathrm{tr}(\mathrm{Cov}(h(z))) + 2\rho(1-\rho)\, W_{\mathrm{nl}}.
\end{aligned} \tag{43}$$

To extract a bound on W_{nl}, we need an upper bound on $\mathcal{L}(h)$. An encoder achieving exact alignment would satisfy $\mathcal{L}(h) = 2(1-\rho)\,\mathrm{tr}(\mathrm{Cov}(h(z)))$ (attained when h is linear). We assume the encoder is δ-approximately aligned, meaning

$$\mathcal{L}(h) \;\leq\; 2(1-\rho)\,\mathrm{tr}(\mathrm{Cov}(h(z))) + \delta. \tag{44}$$

Combining the upper and lower bounds and rearranging:

$$\boxed{W_{\mathrm{nl}} \;\leq\; \frac{\delta}{2\rho(1-\rho)}.} \tag{45}$$

The alignment gap δ directly controls how much variance the encoder can place on nonlinear Hermite components. The denominator $2\rho(1-\rho)$ is the spectral gap between degree-1 and degree-2 eigenvalues, and governs the sensitivity of this bound.

C.2 Step 2: Whitening constraint forces M near orthogonal

The covariance of $h(z)$ decomposes by Hermite orthogonality as

$$\mathrm{Cov}(h(z)) = MM^\top + N, \tag{46}$$

where $N_{ij} = \sum_{|\alpha|\geq 2} c_\alpha^{(i)}\, c_\alpha^{(j)}$ is the contribution from nonlinear terms. Since N is a Gram matrix (positive semidefinite) with $\mathrm{tr}(N) = W_{\mathrm{nl}}$, and for any PSD matrix $\|N\|_F \leq \mathrm{tr}(N)$ (all eigenvalues are non-negative, so $\sum \lambda_i^2 \leq (\sum \lambda_i)^2$), we have $\|N\|_F \leq W_{\mathrm{nl}}$.

The approximate whitening assumption gives:

$$\|MM^\top - I_N\|_F = \|\mathrm{Cov}(h(z)) - I_N - N\|_F \leq \varepsilon + W_{\mathrm{nl}}. \tag{47}$$

Let $M = QP$ be the polar decomposition with $Q \in O(N)$ and P symmetric positive semidefinite with singular values $\sigma_1, \ldots, \sigma_N$. Then $MM^\top = QP^2Q^\top$, so $\|P^2 - I_N\|_F = \|MM^\top - I_N\|_F$. Since P is PSD, each $\sigma_i \geq 0$, and $|\sigma_i - 1| \leq |\sigma_i + 1||\sigma_i - 1| = |\sigma_i^2 - 1|$. Therefore:

$$\|M - Q\|_F = \|P - I_N\|_F \leq \|P^2 - I_N\|_F \leq \varepsilon + W_{\mathrm{nl}}. \tag{48}$$

C.3 Step 3: Pythagorean decomposition

Write

$$h(z) - Qz = \underbrace{(M-Q)z}_{\text{degree 1}} + \underbrace{\nu(z)}_{\text{degree}\geq 2}. \tag{49}$$

Since $(M-Q)z$ lives entirely in Hermite degree 1 and $\nu(z)$ lives entirely in degree ≥ 2, orthogonality of Hermite degrees under γ_N gives:

$$\mathbb{E}[\|h(z) - Qz\|^2] = \mathbb{E}[\|(M-Q)z\|^2] + \mathbb{E}[\|\nu(z)\|^2]. \tag{50}$$

The cross term $\mathbb{E}[\nu(z)^\top (M-Q)z] = 0$ vanishes exactly. For the first term, since $z \sim \mathcal{N}(0, I_N)$:

$$\mathbb{E}[\|(M-Q)z\|^2] = \operatorname{tr}\big((M-Q)\,\mathbb{E}[zz^\top]\,(M-Q)^\top\big) = \|M-Q\|_F^2. \tag{51}$$

Combining with (45) and (48):

$$\mathbb{E}[\|h(z)-Qz\|^2] = \|M-Q\|_F^2 + W_{\text{nl}} \leq (\varepsilon + W_{\text{nl}})^2 + W_{\text{nl}} \leq \left(\varepsilon + \frac{\delta}{2\rho(1-\rho)}\right)^2 + \frac{\delta}{2\rho(1-\rho)}. \tag{52}$$

This completes the proof. □

Role of the Gaussian on the latent side vs. the output side. The proof uses the Gaussian assumption on z in two essential ways: the Hermite basis and Mehler's formula (which provide the spectral decomposition (41)), and the geometric eigenvalue decay ρ^d (which creates the spectral gap exploited in Step 1). By contrast, the proof uses *no* distributional assumption on $h(z)$ beyond zero mean and approximate whitening. In particular, the full distributional Gaussianity condition $h(z) \sim \mathcal{N}(0, I_N)$ of Thm. 1 can be relaxed to $\mathbb{E}[h(z)] = 0$ and $\|\mathrm{Cov}(h(z)) - I_N\|_F \leq \varepsilon$ without any change to the bound.

D Proof and Discussion of Theorem 4

Theorem 4 (Optimal Latent Planning)

Let $h(z) = Qz$ with $Q \in O(N)$, write $\hat{z} := h(z)$, and consider any finite-horizon optimal control problem whose stage and terminal costs ℓ, ℓ_T are $O(N)$-invariant in the state argument:

$$\ell(Rz, a) = \ell(z, a), \qquad \ell_T(Rz) = \ell_T(z), \qquad \forall\, R \in O(N). \tag{53}$$

Let $\hat{p}(\cdot \mid \hat{z}_{t-1}, a_t)$ denote the pushforward of p under h, and let $\hat{V}^*, \hat{a}^*_{1:T}$ denote the value function and optimal plan of the $\hat{z}$-space problem defined by replacing p with $\hat{p}$ and leaving costs unchanged. Then $\hat{V}^*\big(h(z_0)\big) = V^*(z_0)$ and $\hat{a}^*_{1:T}\big(h(z_0)\big) = a^*_{1:T}(z_0)$.

Proof. Fix any action sequence $a_{1:T} \in \mathcal{A}^T$ and initial state z_0, and let $\hat{z}_0 = h(z_0) = Qz_0$. By construction of the pushforward, the joint law of $(\hat{z}_0, \hat{z}_1, \ldots, \hat{z}_T)$ under $\hat{p}$ equals the joint law of $(Qz_0, Qz_1, \ldots, Qz_T)$ under p. Applying (6) with $R = Q$,

$$\ell(\hat{z}_t, a_{t+1}) = \ell(Qz_t, a_{t+1}) = \ell(z_t, a_{t+1}) \quad \text{a.s.}, \qquad \ell_T(\hat{z}_T) = \ell_T(z_T) \quad \text{a.s.} \tag{54}$$

Summing and taking expectations,

$$\mathbb{E}_{\hat{p}}\left[\sum_{t=1}^{T} \ell(\hat{z}_{t-1}, a_t) + \ell_T(\hat{z}_T) \,\middle|\, \hat{z}_0\right] = \mathbb{E}_{p}\left[\sum_{t=1}^{T} \ell(z_{t-1}, a_t) + \ell_T(z_T) \,\middle|\, z_0\right]. \tag{55}$$

Since this holds for every $a_{1:T}$, the minima coincide, giving $\hat{V}^*(\hat{z}_0) = V^*(z_0)$, and any minimizer of one side is a minimizer of the other. □

D.1 Worked Applications

The invariance condition (6) covers a substantial part of control applications. We spell out three.

Goal-reaching. The terminal cost $\ell_T(z) = \|z - z^*\|^2$ is $O(N)$-invariant when the state and goal rotate together: $\|Rz - Rz^*\|^2 = \|z - z^*\|^2$. A planner choosing a goal $\hat{z}^* \in \mathbb{R}^N$ in the learned latent and minimizing $\|\hat{z}_T - \hat{z}^*\|^2$ solves the true goal-reaching problem for the goal $z^* = Q^\top \hat{z}^*$. The learned goal need never be mapped back to the true latent for this to work; the planner's actions are correct as a sequence, regardless of the coordinate system they are computed in.

Linear-quadratic regulation. Take linear dynamics $z_{t+1} = Az_t + Ba_t + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$, and quadratic costs $\ell(z, a) = z^\top W z + a^\top R a$, $\ell_T(z) = z^\top W_T z$, with $W, W_T \succeq 0$ and $R \succ 0$. The pushforward dynamics in $\hat{z}$-coordinates are

$$\hat{z}_{t+1} = (QAQ^\top)\,\hat{z}_t + (QB)\,a_t + Q\varepsilon_t, \qquad Q\varepsilon_t \sim \mathcal{N}(0, Q\Sigma Q^\top), \tag{56}$$

so the LQR problem in $\hat{z}$-space has system matrices $(\hat{A}, \hat{B}, \hat{\Sigma}) = (QAQ^\top, QB, Q\Sigma Q^\top)$ and cost matrices $(\hat{W}, \hat{W}_T, R) = (QWQ^\top, QW_TQ^\top, R)$. The discrete algebraic Riccati equation

$$P = A^\top PA - A^\top PB(R + B^\top PB)^{-1}B^\top PA + W \tag{57}$$

is covariant under $(A, B, W, P) \mapsto (QAQ^\top, QB, QWQ^\top, QPQ^\top)$: if P solves the true-latent equation, then $\hat{P} = QPQ^\top$ solves the $\hat{z}$-space equation. The optimal feedback gain $K = (R + B^\top PB)^{-1}B^\top PA$ transforms as $\hat{K} = KQ^\top$, and the feedback action is

$$\hat{a}_t = -\hat{K}\hat{z}_t = -(KQ^\top)(Qz_t) = -Kz_t, \tag{58}$$

which is the true LQR action at every state. The value function is likewise $\hat{V}^*(\hat{z}) = \hat{z}^\top \hat{P}\hat{z} = z^\top Pz = V^*(z)$. LQR is therefore a special case of Thm. 4 in which the invariance is inherited from the quadratic structure and the covariance of all system matrices under the rotation.

General rotation-invariant objectives. Any cost depending on the state only through quantities preserved by rotations satisfies (6). Examples: norms $\|z\|$ and radial functions $f(\|z\|)$; inner products $\langle z, z' \rangle$ between pairs of states; quadratic forms $z^\top Wz$ with W specified in the learned coordinates (which silently corresponds to $Q^\top WQ$ in the true latent); Gaussian belief updates, whose covariance updates are covariant under rotation; and linear value functions $w^\top z$ with w specified in the learned coordinates. In each case, an engineer who writes down the cost using only these operations has unwittingly written down a cost in the equivalence class of (6), and Thm. 4 certifies that planning is unaffected by the ambiguity.

D.2 Scope: encoder versus action-conditioned dynamics

Thm. 4 concerns the encoder alone. The action-conditioned transition $\hat{p}(\hat{z}' \mid \hat{z}, a)$ must still be learned from data [9]; our theorems do not prove that it is. What they do guarantee is that the encoder does not corrupt the learning problem for the transition: for $h(z) = Qz$, the pushforward of the true transition lies in the same function class as the original (linear transitions remain linear, Gaussian transitions remain Gaussian, and any class closed under orthogonal reparametrization of the state is equally expressive in both coordinate systems). The residue of the orthogonal ambiguity is a consistent reparametrization across the encoder, the transition model, and the cost: so long as these three components are trained or specified to be mutually consistent in the learned coordinates, planning in those coordinates is planning in the true ones.

The natural next theoretical step is to prove that the action-conditioned transition itself is identifiable. This is a genuinely different problem: the encoder identifiability proved here relies on stationarity plus Gaussianity of the latent distribution, whereas transition identifiability would trade these for an excitation condition on the action sequence (the actions must explore all latent directions). Persistent excitation is the classical condition in subspace identification [68, 69], and extending LeJEPA-style identifiability to the controlled setting is the subject of ongoing work.

D.3 Connection to classical optimal control

The covariance of LQR under orthogonal reparametrization has been exploited for decades in control theory: the Kalman filter, the Riccati equation, and Lyapunov stability analysis all transform covariantly under orthogonal changes of basis, and this is why principal-component and whitening

transforms are used as preprocessing steps in classical state estimation. What is new here is not the covariance itself but its role as a *consequence of an identifiability guarantee*. In classical control, the state is assumed given; the engineer chooses a coordinate system. In the LeJEPA setting, the coordinate system is chosen by the learning procedure, and the theorem is that the choice is made consistently enough, orthogonal rather than arbitrary linear (let alone arbitrary nonlinear), that classical control tools continue to apply without modification. Latent-space planning with a learned world model has a long heritage [95, 46, 52]; Thm. 4 provides a representation-learning foundation for the classical pipeline.

E Alternative Proof via Dirichlet Energy

We provide a second, independent proof of identifiability using a different noise model and different mathematical machinery. This proof requires stronger regularity (C^1 diffeomorphism and infinitesimal noise limit) but offers complementary geometric insight and makes explicit contact with the local isometry approach used in prior nonlinear ICA [76].

E.1 Setup

Consider the alternative noise model $z' = z + \sqrt{\epsilon}\,\eta$ with $\eta \sim \mathcal{N}(0, I_N)$, define the normalized loss:

$$\mathcal{L}(h) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}\left[\|h(z + \sqrt{\epsilon}\eta) - h(z)\|^2\right]. \tag{59}$$

We require $h : \mathbb{R}^N \to \mathbb{R}^N$ to be a C^1 diffeomorphism with $h(z) \sim \mathcal{N}(0, I_N)$ when $z \sim \mathcal{N}(0, I_N)$.

E.2 Proof

Theorem 5 (Identifiability via Dirichlet energy)

Any C^1 diffeomorphism h that preserves the standard Gaussian measure and minimizes $\mathcal{L}(h)$ must satisfy $h(z) = Qz$ for some orthogonal matrix $Q \in O(N)$.

Step 1: Reduction to Dirichlet energy. By the first-order Taylor expansion $h(z+\sqrt{\epsilon}\eta) - h(z) = \sqrt{\epsilon}\, J_h(z)\,\eta + o(\sqrt{\epsilon})$, where $J_h(z)$ is the Jacobian of h at z. Squaring, dividing by ϵ, taking expectations, and using $\mathbb{E}_\eta[\|A\eta\|^2] = \|A\|_F^2$ for any matrix A:

$$\mathcal{L}(h) = \mathbb{E}_{z \sim \gamma_N}\left[\|J_h(z)\|_F^2\right], \tag{60}$$

which is the Dirichlet energy of h with respect to the Gaussian measure γ_N.

The exchange of limit and expectation is justified by dominated convergence: the C^1 regularity of h combined with the Gaussian decay of γ_N provides the necessary integrability. The full computation, including careful treatment of the remainder term, is given below.

Starting from the Taylor expansion:

$$h(z + \sqrt{\epsilon}\eta) - h(z) = \sqrt{\epsilon}\, J_h(z)\,\eta + R(z, \sqrt{\epsilon}\eta), \tag{61}$$

where $\|R(z, \sqrt{\epsilon}\eta)\| = o(\sqrt{\epsilon})$. Expanding the squared norm:

$$\|h(z+\sqrt{\epsilon}\eta) - h(z)\|^2 = \epsilon\,\|J_h(z)\eta\|^2 + 2\sqrt{\epsilon}\,(J_h(z)\eta)^\top R + \|R\|^2. \tag{62}$$

Dividing by ϵ and taking $\epsilon \to 0$, the last two terms vanish, yielding:

$$\frac{1}{\epsilon}\,\mathbb{E}_{z,\eta}\left[\|h(z+\sqrt{\epsilon}\eta) - h(z)\|^2\right] \xrightarrow{\epsilon \to 0} \mathbb{E}_{z,\eta}\left[\|J_h(z)\eta\|^2\right]. \tag{63}$$

Using $\mathbb{E}_\eta[\eta^\top M \eta] = \mathrm{tr}(M)$ for $M = J_h(z)^\top J_h(z)$:

$$\mathbb{E}_\eta\left[\|J_h(z)\eta\|^2\right] = \mathrm{tr}(J_h(z)^\top J_h(z)) = \|J_h(z)\|_F^2. \tag{64}$$

Step 2: The expected log-determinant is zero. By the change of variables formula for densities, $p(h(z))\,|\det J_h(z)| = p(z)$ for all z. Substituting the Gaussian density $p(z) = (2\pi)^{-N/2}\exp(-\|z\|^2/2)$ and noting that the normalizing constants cancel:

$$\ln|\det J_h(z)| = \tfrac{1}{2}\|h(z)\|^2 - \tfrac{1}{2}\|z\|^2. \tag{65}$$

Taking expectations and using $\mathbb{E}[\|h(z)\|^2] = \mathbb{E}[\|z\|^2] = N$ (both follow from $h(z) \sim \mathcal{N}(0, I_N)$):

$$\mathbb{E}[\ln|\det J_h(z)|] = 0. \tag{66}$$

Step 3: Lower bound via AM-GM and Jensen. Let $\sigma_1(z), \ldots, \sigma_N(z)$ be the singular values of $J_h(z)$. Then $\|J_h(z)\|_F^2 = \sum_i \sigma_i(z)^2$ and $|\det J_h(z)| = \prod_i \sigma_i(z)$. By AM-GM applied to $\sigma_1^2, \ldots, \sigma_N^2$:

$$\frac{1}{N}\sum_{i=1}^{N}\sigma_i(z)^2 \geq \left(\prod_{i=1}^{N}\sigma_i(z)^2\right)^{1/N} = |\det J_h(z)|^{2/N}, \tag{67}$$

so $\|J_h(z)\|_F^2 \geq N\,|\det J_h(z)|^{2/N}$ pointwise. Taking expectations:

$$\mathcal{L}(h) \geq N\,\mathbb{E}\Big[|\det J_h(z)|^{2/N}\Big]. \tag{68}$$

Writing $|\det J_h(z)|^{2/N} = \exp\big(\frac{2}{N}\ln|\det J_h(z)|\big)$ and applying Jensen's inequality to the strictly convex function $x \mapsto e^{(2/N)x}$:

$$\mathbb{E}\Big[|\det J_h(z)|^{2/N}\Big] \geq \exp\big(\tfrac{2}{N}\,\mathbb{E}[\ln|\det J_h(z)|]\big) = \exp(0) = 1. \tag{69}$$

Therefore $\mathcal{L}(h) \geq N$. Any orthogonal map $h(z) = Qz$ achieves this bound since $\|Q\|_F^2 = N$.

Step 4: Equality forces orthogonal Jacobian. At the minimum $\mathcal{L}(h) = N$, equality must hold in both Jensen and AM-GM. Strict convexity of the exponential forces $\ln|\det J_h(z)|$ to be constant a.e., and since its expectation is 0, we get $|\det J_h(z)| = 1$ a.e. Equality in AM-GM forces all squared singular values to be equal: $\sigma_i(z)^2 = \cdots = \sigma_N(z)^2$. Combined with $\prod_i \sigma_i = 1$, this gives $\sigma_i(z) = 1$ for all i. Thus:

$$J_h(z)^\top J_h(z) = I_N \qquad \text{for a.e. } z. \tag{70}$$

By C^1 continuity, this holds everywhere.

Step 5: Orthogonal Jacobian implies linearity (Mazur–Ulam). Since $J_h(z)^\top J_h(z) = I_N$ everywhere, h preserves the norm of tangent vectors: $\|J_h(z)v\| = \|v\|$ for all v. For any $z_1, z_2 \in \mathbb{R}^N$, integrating along the straight-line path $\gamma(t) = z_1 + t(z_2 - z_1)$:

$$h(z_2) - h(z_1) = \int_0^1 J_h(\gamma(t))(z_2 - z_1)\,dt. \tag{71}$$

Taking norms and applying the triangle inequality:

$$\|h(z_2) - h(z_1)\| \leq \int_0^1 \|J_h(\gamma(t))(z_2 - z_1)\|\,dt = \int_0^1 \|z_2 - z_1\|\,dt = \|z_2 - z_1\|. \tag{72}$$

Since h is a diffeomorphism, h^{-1} exists and is C^1 with $J_{h^{-1}}(h(z)) = J_h(z)^{-1} = J_h(z)^\top$, which is also orthogonal ($J_h J_h^\top = I_N$ for square matrices with $J_h^\top J_h = I_N$). Applying the same argument to h^{-1} yields $\|z_2 - z_1\| \leq \|h(z_2) - h(z_1)\|$. Together:

$$\|h(z_1) - h(z_2)\| = \|z_1 - z_2\| \qquad \text{for all } z_1, z_2 \in \mathbb{R}^N, \tag{73}$$

so h is a surjective isometry of $\mathbb{R}^N$. By the Mazur–Ulam theorem [96], every surjective isometry of a real normed space is affine: $h(z) = Qz + b$ with Q orthogonal and $b \in \mathbb{R}^N$. Measure preservation forces $0 = \mathbb{E}[h(z)] = Q\mathbb{E}[z] + b = b$, giving $h(z) = Qz$. □

Generalization. No step in the proof above uses the Gaussian assumption on p; the noise η need only satisfy $\mathbb{E}[\eta\eta^\top] = I_N$ (Step 1), and Steps 2–5 hold for any density preserved by h. The identity map always achieves $\mathcal{L} = N$, so the lower bound is tight regardless of p. The conclusion is that any C^1 measure-preserving diffeomorphism minimizing $\mathcal{L}$ must be an isometry $h(z) = Qz + b$ with $p(Qz + b) = p(z)$.

E.3 Comparison of the Two Proofs

The Hermite proof (Thm. 1) and the Dirichlet energy proof (Thm. E.2) establish the same conclusion through different mechanisms:

	Hermite proof	**Dirichlet energy proof**
Regularity	Measurable	C^1 diffeomorphism
Noise level	Any $\rho \in (0, 1)$	Infinitesimal ($\rho \to 1$)
Key tool	Mehler's formula	AM-GM + Jensen + Mazur–Ulam
Mechanism	Low-pass filter on degrees	Jacobian rigidity
Connection to	Harmonic analysis	Differential geometry

The Hermite proof is strictly more general (weaker assumptions, finite noise), but the Dirichlet energy proof provides geometric intuition that complements the spectral perspective: the Jacobian must be an isometry at every point, so the map cannot compress or stretch any direction anywhere.

F Prior Work: Connection to Slow Feature Analysis

Our identifiability result has a deep connection to the theory of Slow Feature Analysis (SFA) developed by Wiskott and Sejnowski [72] and extended to nonlinear blind source separation by Sprekeler et al. [74]. Sobal et al. [73] analyzed the tendency of JEPAs to focus on slow features. We spell out this connection in detail, both because it clarifies the scope of our contribution and because the SFA perspective suggests natural generalizations.

SFA and its optimization problem. SFA seeks scalar output functions $g_j(x)$ whose output signals $y_j(t) = g_j(x(t))$ vary as slowly as possible in time, as measured by the Δ-value $\Delta(y_j) = \langle \dot{y}_j^2 \rangle_t$, subject to the constraints $\langle y_j \rangle = 0$ (zero mean), $\langle y_j^2 \rangle = 1$ (unit variance), and $\langle y_i y_j \rangle = 0$ for $i < j$ (decorrelation) [72]. For discrete-time data, the Δ-value is proportional to $\langle y^2 \rangle - \langle y(t)y(t+1) \rangle$, so that for unit-variance outputs, minimizing Δ is equivalent to maximizing the temporal correlation $\langle y(t)y(t+1) \rangle$. The SFA constraints are therefore precisely whitening, and the SFA objective is precisely the alignment loss used in our setting.

Nonlinear identifiability via SFA. Sprekeler et al. [74] analyzed SFA in the setting where the input data $x(t) = F(s(t))$ are generated from statistically independent latent variables $s(t)$ via an unknown invertible nonlinear mixing function F. They showed that if the function space $\mathcal{F}$ accessible to SFA is unrestricted (a Sobolev space), then the set of achievable output signals is independent of F: for every $g \in \mathcal{F}$, the composition $g \circ F$ is also in $\mathcal{F}$, so the mixing function drops out of the analysis entirely. This is the same role played by our assumption that the encoder is expressive enough to invert the mixing function.

Eigenfunction structure and Sturm–Liouville theory. Building on Franzius et al. [97], Sprekeler et al. [74] showed that the optimal SFA functions satisfy an eigenvalue equation $\mathcal{D}g = \lambda g$ for a second-order partial differential operator $\mathcal{D}$. When the latent variables are statistically independent, the operator separates as $\mathcal{D} = \sum_\alpha \mathcal{D}_\alpha$, where each $\mathcal{D}_\alpha$ depends on a single latent variable s_α only. By separation of variables, the eigenfunctions of the full operator are products $g_{\mathbf{i}}(s) = \prod_\alpha g_{\alpha i_\alpha}(s_\alpha)$ of the eigenfunctions of the individual operators, with eigenvalues $\lambda_{\mathbf{i}} = \sum_\alpha \lambda_{\alpha i_\alpha}$. Each one-dimensional eigenvalue problem is of Sturm–Liouville type, and standard results guarantee that the eigenfunctions $g_{\alpha i}$ are oscillatory with exactly i zeros. In particular, the first non-constant eigenfunction $g_{\alpha 1}$ is monotonic and therefore invertible, yielding identifiability of each latent variable up to a monotonic transformation for *any* latent variable distribution.

The Gaussian specialization and Hermite polynomials. When the latent variables are reversible Gaussian processes (equivalently, Ornstein–Uhlenbeck processes), the conditional variance of the derivative given the current value is constant ($K_\alpha(s_\alpha) = K_\alpha$), and the latent variable density is Gaussian. The Sturm–Liouville equation then reduces to Hermite's differential equation, with eigenfunctions given by the Hermite polynomials H_i and eigenvalues $\lambda_{\alpha i} = iK_\alpha$. The first non-constant

eigenfunction is simply $g_{\alpha 1}(s_\alpha) = s_\alpha$: the identity. This is the same Hermite eigenstructure that appears in our proof via Mehler's formula. Indeed, Mehler's formula is the spectral expansion of the Green's function (transition kernel) of the operator $\mathcal{D}_\alpha$ in the Gaussian case, so the two proof routes, Sturm–Liouville eigenvalue analysis and direct Mehler expansion, access the same underlying mathematical structure from different entry points.

Sequential versus simultaneous extraction. A key difference between the approach of Sprekeler et al. [74] and ours lies in how the d output dimensions are determined. Their xSFA algorithm extracts latent variables *sequentially*: it identifies the slowest latent variable, projects out all its nonlinear transformations, and repeats. This greedy procedure, combined with the assumption that the latent variables have distinct autocorrelation rates (K_α all different), yields identifiability up to *permutation*, a stronger result than our orthogonal identifiability. However, the sequential approach is fragile in practice: it requires representing the nonlinear transformations of already-extracted latent variables (approximated by high-degree polynomial expansions), accumulates estimation errors across iterations, and degrades rapidly as the number of latent variables increases [74, Section 5].

Our approach instead optimizes all d output dimensions *simultaneously* via a joint loss (alignment plus whitening or SIGReg). Because whitening is invariant under orthogonal transformations, the simultaneous approach cannot resolve individual latent variable axes, yielding identifiability up to orthogonal rotation rather than permutation. This is the natural and complete identifiability class for the simultaneous setting.

The role of isotropic transitions. The distinction between sequential and simultaneous extraction is intimately connected to the structure of the latent transitions. When the autocorrelation parameters ρ_α differ across dimensions, the eigenvalue $\lambda_{\alpha i} = iK_\alpha$ of the i-th Hermite component of latent variable α depends on α. For the simultaneous approach to correctly identify the subspace of first harmonics (degree-1 Hermite components), all first harmonics must have smaller eigenvalues than all higher-order terms. This requires $\max_\alpha K_\alpha < 2 \min_\beta K_\beta$, or equivalently, the fastest latent variable must not be more than twice as fast as the slowest. When ρ_α are all equal (isotropic transitions), this condition is trivially satisfied, and the first d eigenfunctions are exactly the d first harmonics, yielding clean orthogonal identifiability.

We verified empirically that anisotropic transitions break the simultaneous approach: with distinct ρ_α, the encoder recovers the second Hermite polynomial of the slow latent variable in place of the first Hermite polynomial of the fast latent variable, exactly as the eigenvalue interleaving predicts. See also the anisotropy in Tab. 2. This confirms that isotropic transitions are not merely a simplifying assumption but a necessary condition for simultaneous (parallel) identifiability. The anisotropic case requires either sequential extraction à la Sprekeler et al. [74] or a modified objective such as the AnInfoNCE loss of Rusak et al. [98], which reweights the contrastive objective to account for per-dimension variance differences, at the cost of downstream accuracy.

Beyond Gaussian latents. The Sturm–Liouville perspective of Sprekeler et al. [74] suggests a generalization of our result that does not require Gaussian latents. If all latent variables are drawn i.i.d. from the *same* (arbitrary) distribution with the same transition structure, then the eigenvalue spectra of the individual operators $\mathcal{D}_\alpha$ are identical: $\lambda_{\alpha i} = \lambda_i$ for all α. All first harmonics share the same eigenvalue λ_1, all second harmonics share $\lambda_2 > \lambda_1$ (the gap is guaranteed by Sturm–Liouville theory), and the top-d eigenspace of the simultaneous problem is exactly the span of the first harmonics. The simultaneous approach then yields identifiability up to orthogonal rotation, not in the original latent variable space but in the space of first harmonics. Since the first harmonics are monotonic functions of the latent variables, this amounts to identifiability up to an orthogonal rotation composed with a shared monotonic nonlinearity applied elementwise. Only in the Gaussian case does the first harmonic equal the identity, collapsing this to linear orthogonal identifiability. A rigorous treatment of this non-Gaussian generalization is an interesting direction for future work.

Relation to Mehler's formula and maximal correlation. The Hermite expansion and temporal correlation bounds used in our proof are closely related to classical results in probability theory. Specifically, the correlation bound (equation (4)) is a consequence of the classical Hirschfeld–Gebelein–Rényi maximal correlation for jointly Gaussian variables [99, 100, 101]. Our contribution is not the bound itself but its application to identifiability via the measure-preservation constraint. As discussed above, Mehler's formula, the key technical tool in our proof, is equivalent to the spectral

	Sprekeler et al. (2014)	This work
Identifiability class	Permutation	Orthogonal
Latent variable distributions	Any independent	Gaussian (or i.i.d. factorial)
Transition structure	Distinct rates required	Isotropic required
Extraction	Sequential (greedy)	Simultaneous
Function space	Fixed polynomial kernel	Learned (neural network)
Proof technique	Sturm–Liouville / PDE	Mehler's formula / Parseval
Approximate bound	None	$D + (\varepsilon + D)^2$
Practical algorithm	xSFA (fragile, ≤ 6 latent variables)	LeJEPA / SIGReg (scalable)

Table 4: Comparison of the SFA identifiability theory of Sprekeler et al. [74] and the present work.

expansion of the transition kernel associated with the Sturm–Liouville (SL) operator of Sprekeler et al. [74] in the Gaussian case. The SL route is more general (handles arbitrary latent variable distributions) but does not directly yield the quantitative bound (Thm. 3) that Mehler's formula provides.

Practical differences: learned features versus fixed kernels. While the theoretical optimization problems coincide, the practical implementations differ substantially. Sprekeler et al. [74] use fixed polynomial kernel expansions (degree 7 for two latent variables) followed by a linear eigenvalue problem, a two-stage procedure in which the feature space is chosen a priori and never adapted. In contrast, the LeJEPA/SIGReg pipeline learns the encoder end-to-end via gradient descent on a neural network, jointly optimizing the feature representation and the alignment plus regularization objectives. This learned-feature approach avoids the numerical instabilities of high-degree polynomial expansions noted by Sprekeler et al. [74], scales to high-dimensional data, and is the standard paradigm in modern self-supervised learning. Our identifiability result thus provides theoretical grounding for an architecture and training procedure that is already in widespread use.

Connections to diffusion maps and spectral geometry. Sprekeler [102] showed that SFA is closely related to Laplacian eigenmaps and diffusion maps, with the operator $\mathcal{D}$ playing the role of a Laplace–Beltrami operator weighted by the temporal structure of the data. The eigenfunctions of $\mathcal{D}$ are the same objects studied in spectral geometry and manifold learning (see also [75]). Singer and Coifman [103] exploited a similar connection, using data-driven diffusion maps with a carefully chosen local metric to achieve nonlinear latent variable separation. This spectral-geometric perspective suggests that the identifiability phenomenon we study, encoders converging to specific linear transformations of the latents, may be an instance of a broader principle: that temporal or geometric structure on the data manifold, combined with appropriate regularization, constrains learned representations to align with the intrinsic coordinates of the generative process.

Summary of the relationship. Table 4 summarizes the key similarities and differences. The mathematical core, Hermite eigenfunctions under Gaussian measure with temporal correlations yielding identifiability, is shared. Our contributions are (a) the simultaneous formulation with its clean orthogonal identifiability class, (b) the converse result identifying the Gaussian as the *unique* latent distribution yielding linear identifiability (Thm. 2), (c) the quantitative approximate identifiability bound under whitening alone (Thm. 3), (d) the observation that isotropic transitions are necessary for the simultaneous approach, and (e) the planning equivalence (Thm. 4) connecting linear identifiability to optimal latent-space planning.

G Lean Verification

We formally verify all five theoretical results, namely the Hermite proof of Thm. 1, the Gaussian uniqueness result (Thm. 2), the Dirichlet energy proof (Appendix E), the approximate identifiability bound (Thm. 3), and the planning equivalence theorem (Thm. 4), in the Lean 4 theorem prover using the Mathlib mathematical library. The formalization compiles with zero `sorry` obligations: every logical step from axiomatized premises to stated conclusions is machine-checked.

Scope and methodology. Lean 4 verification requires every inference step to be justified by a previously established lemma or axiom. When a standard mathematical result exists in Mathlib

(e.g., `Finset.sum_lt_sum` for strict finite-sum monotonicity), we invoke it directly. When a result is standard but not yet available in Mathlib, either because the mathematical objects have not been formalized (Hermite polynomials) or because connecting available results to our specific statement forms would require substantial software engineering (Mazur–Ulam, AM-GM with uniform weights), we introduce it as a Lean `axiom` with a reference to its mathematical source. The complete reasoning chains *between* these axioms are fully verified. Table 5 gives the complete inventory. The formalization is organized into five files corresponding to the five results.

G.1 What Is Axiomatized and Why

The axiomatized results fall into five categories:

1. Hermite polynomial infrastructure. Mathlib does not yet define Hermite polynomials or their properties under the Gaussian measure. The contraction property (21), Mehler's formula (24), and the correspondence between degree-1 concentration and linearity are axiomatized for this reason. These are classical results in probability theory [83, 101]; their eventual formalization in Mathlib would allow these axioms to be discharged.

2. Sturm–Liouville spectral theory. The Gaussian uniqueness proof (Thm. 2) axiomatizes two results: that an affine score function uniquely determines a Gaussian density (standard exponential family characterization), and that a Gaussian density yields Hermite eigenfunctions with the identity as the first non-constant eigenfunction. The core algebraic step between these, namely extracting the affine score from the Sturm–Liouville eigenfunction equation via field arithmetic (including the sign argument that the score slope $-\lambda_1/K$ is strictly negative), is fully verified.

3. Standard analysis results requiring API software engineering. The AM-GM inequality with uniform weights, Jensen's inequality for $\exp$, and the Mazur–Ulam theorem are all available in Mathlib in closely related forms, but connecting them to our specific statement signatures requires nontrivial type-class and measure-theory software engineering. We axiomatize the exact statements we need and reference their Mathlib counterparts in the source code.

4. Measure-theoretic and matrix-analytic facts. The polar decomposition bound ($\|M - Q\|_F \leq \varepsilon + W_{\mathrm{nl}}$) and the Pythagorean decomposition ($\mathbb{E}[\|h(z) - Qz\|^2] = \|M - Q\|_F^2 + W_{\mathrm{nl}}$) combine matrix analysis with integration under the Gaussian measure. These are stated as axioms with their mathematical content documented in the source.

5. Trajectory pushforward for expected costs. The planning equivalence theorem (Thm. 4) axiomatizes the pushforward relation between the learned-latent and true-latent expected per-step costs: for any test function c, the expected value of $c(\hat{z}_t, a_t)$ under the pushforward dynamics starting from Qz_0 equals the expected value of $c(Qz_t, a_t)$ under the original dynamics starting from z_0. This is the measure-theoretic content of "the joint law of $(\hat{z}_0, \ldots, \hat{z}_T)$ under the pushforward equals the joint law of $(Qz_0, \ldots, Qz_T)$ under the original"; stating it at the level of expected values rather than joint distributions sidesteps the need to formalize stochastic dynamics in Lean while preserving the algebraic content of the theorem's proof.

What is verified. All reasoning chains *between* the axiomatized premises are fully machine-checked: the correlation bound and its strict equality characterization (the mathematical core of the Hermite proof), the algebraic chain from the SL eigenfunction equation to the affine score function and the biconditional assembly (the Gaussian uniqueness proof), the geometric chain from orthogonal Jacobian to linear isometry (the core of the Dirichlet proof), the spectral gap argument controlling nonlinear energy, the monotonicity and algebraic assembly of the approximate bound, and the recovery of the exact theorem as a special case. The inverse function theorem is incorporated as a structural assumption on the diffeomorphism rather than invoked as a separate lemma.

G.2 Build Information

The project compiles against Lean 4 v4.28.0 with Mathlib v4.28.0 (8,032 build targets, zero errors, zero `sorry` obligations) and is available at `https://github.com/klindtlab/lejepa-identifiability`.

H Experimental Details and Additional Results

H.1 Nonlinear Mixing Functions

The four mixing functions $g : \mathbb{R}^2 \to \mathbb{R}^2$ used in Figs. 1,3 are:

1. **Spiral:** $g(z) = R(\pi\|z\|_2)\, z$, where $R(\theta) = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$. This norm-dependent rotation is a measure-preserving diffeomorphism of $\mathcal{N}(0, I_2)$: at each point, g acts as an orthogonal rotation ($|\det J_g| = 1$), so it preserves the Gaussian density despite being highly nonlinear.

2. **Sinusoidal shear:** $g(z_1, z_2) = (z_1 + \sin(1.5\, z_2),\ z_2)$.

3. **Parabolic shear:** $g(z_1, z_2) = (z_1,\ z_2 + z_1^2)$.

4. **RealNVP coupling layer:** a single affine coupling block [85] of the form $g(z_1, z_2) = (z_1,\ z_2 \odot \exp(s(z_1)) + t(z_1))$, where s, t are small MLPs with random initialization. This is the same family used as the matched encoder in the scaling experiment (App. H.10); here we use it in 2D as an additional nonlinear diffeomorphism.

The first (spiral) map applies a rotation whose angle $\pi\|z\|_2$ depends on the distance from the origin. Points at different radii are rotated by different angles, producing a characteristic spiral structure in observation space (Figure 3, center). The map is a measure-preserving diffeomorphism of $\mathcal{N}(0, I_2)$: at each point, g acts as an orthogonal rotation ($|\det J_g| = 1$ everywhere), so it preserves the Gaussian density. This is an interesting case because it entails SIGReg is already fulfilled in input space. However, g is highly nonlinear; it entangles the radial and angular components of the latent space.

H.2 Network Architecture and Training

The encoder f is a 4-layer MLP with hidden dimension 256 and GELU activations. The architecture is:

$$\mathbb{R}^2 \xrightarrow{\text{Linear}(2,256)} \xrightarrow{\text{GELU}} \xrightarrow{\text{Linear}(256,256)} \xrightarrow{\text{GELU}} \xrightarrow{\text{Linear}(256,256)} \xrightarrow{\text{GELU}} \xrightarrow{\text{Linear}(256,2)} \mathbb{R}^2. \tag{74}$$

H.3 Training Procedure

The loss function combines two terms with a balancing coefficient $\lambda \in [0, 1]$:

$$\mathcal{L} = \lambda\, \mathcal{L}_{\text{SIG}} + (1 - \lambda)\, \mathcal{L}_{\text{inv}}. \tag{75}$$

1. **Invariance loss:** $\mathcal{L}_{\text{inv}} = \frac{1}{B} \sum_{i=1}^{B} \|f(x_i) - f(x_i')\|^2$, where (x_i, x_i') are positive pairs generated by sampling $z_i \sim \mathcal{N}(0, I_2)$, generating $z_i' = \rho z_i + \sqrt{1 - \rho^2}\eta_i$ via the Ornstein–Uhlenbeck channel, and mixing: $x_i = g(z_i)$, $x_i' = g(z_i')$.

2. **SIGReg regularization:** The SIGReg regularizer of Balestriero and LeCun [2] encourages the empirical distribution of embeddings to match the standard Gaussian by penalizing deviations of the empirical sliced characteristic function from the Gaussian target.

H.4 Hyperparameters

All experiments share the same training infrastructure and learning rate schedule: constant learning rate for the first half of training, followed by cosine decay to zero. Data is generated online (fresh samples each step, infinite data regime).

Hyperparameter	2D / Gennorm / Grid	Scaling
Optimizer	AdamW	AdamW
Learning rate	3×10^{-3}	3×10^{-3}
LR schedule	warmup (10k) + cosine (10k)	warmup (10k) + cosine (10k)
Weight decay	0	0
Batch size	256	256
Training steps	20,000	20,000
Data regime	online (infinite)	online (infinite)
Eval points	10,000 (fixed)	10,000 (fixed)
Views per sample	2	2
ρ	0.95 (2D/Gennorm), swept (Grid)	0.95
λ	10^{-3} (2D/Gennorm), swept (Grid)	10^{-6}
Encoder	MLP (spiral/banana/sinusoid) matched NVP (nvp mixing)	matched inverse-NVP [85]
Hidden dimension	256 (MLP)	—
MLP layers	4	—
NVP coupling layers	8 (2D/Gennorm)	4
Activation	GELU (MLP)	tanh (NVP)

H.5 Evaluation Metrics

After training, we evaluate the composed map $h = f \circ g$ on a fixed evaluation set of 10,000 points. All metrics are computed on GPU via PyTorch for efficiency at large N. We report:

1. **Linear R^2 (bidirectional):** We fit linear maps $\hat{z} = Az + b$ and $\hat{h} = Bh + c$ via ordinary least squares and report $R^2(z \to h)$ and $R^2(h \to z)$. Values close to 1 indicate that h is well-approximated by a linear function of z.
2. **Orthogonality error:** For the fitted linear map $\hat{Q}$, we report $\|\hat{Q}^\top \hat{Q} - I_N\|_F / \sqrt{N}$. A value near 0 indicates that $\hat{Q}$ is orthogonal, consistent with Thm. 1.
3. **Approximate bound quantities:** The covariance deviation $\varepsilon = \|\mathrm{Cov}(h(z)) - I_N\|_F$, the alignment gap $\delta = \mathcal{L}(h) - 2(1-\rho)\,\mathrm{tr}(\mathrm{Cov}(h(z)))$ (clamped to ≥ 0), and the Orthogonal recovery error $\min_{Q \in O(N)} \mathbb{E}[\|h(z) - Qz\|^2]$ solved via SVD.

H.6 Grid Search

We perform a 2D grid search over the regularization weight λ and the OU correlation ρ:

- $\lambda \in \{10^{-6},\ 10^{-5},\ 10^{-4},\ 10^{-3},\ 5 \times 10^{-3},\ 10^{-2},\ 5 \times 10^{-2},\ 10^{-1},\ 5 \times 10^{-1}\}$
- $\rho \in \{0.3,\ 0.5,\ 0.7,\ 0.8,\ 0.9,\ 0.95,\ 0.99\}$

For each of the $9 \times 7 = 63$ settings, we train 3 independent seeds, for a total of 189 runs. The extended λ range (down to 10^{-6}) covers the regime where SIGReg has negligible weight, showing that Gaussianity regularization below $\lambda = 10^{-4}$ is insufficient for identifiability regardless of ρ.

Figure 7 shows the results. Three regimes are visible. First, when Gaussianity regularization is too strong ($\lambda = 0.5$), the encoder collapses to a non-informative Gaussian representation with $R^2 \approx 0$, regardless of ρ: the regularizer overwhelms the alignment signal. Second, at low λ ($\leq 10^{-2}$), identifiability improves monotonically with ρ, consistent with the theoretical prediction that the correlation bound tightens as $\rho \to 1$. The best performance ($R^2 > 0.97$, orthogonality error ≈ 0.15) is achieved at $\lambda \in \{10^{-3}, 5 \times 10^{-3}\}$ with $\rho \in \{0.9, 0.95\}$. Third, at intermediate λ (5×10^{-2} to 10^{-1}), there is a non-monotonic interaction: performance peaks at moderate ρ and degrades at $\rho = 0.99$. This occurs because very high correlation makes the invariance loss trivially small, shifting the loss balance toward Gaussianity and pushing the system into the collapse regime.

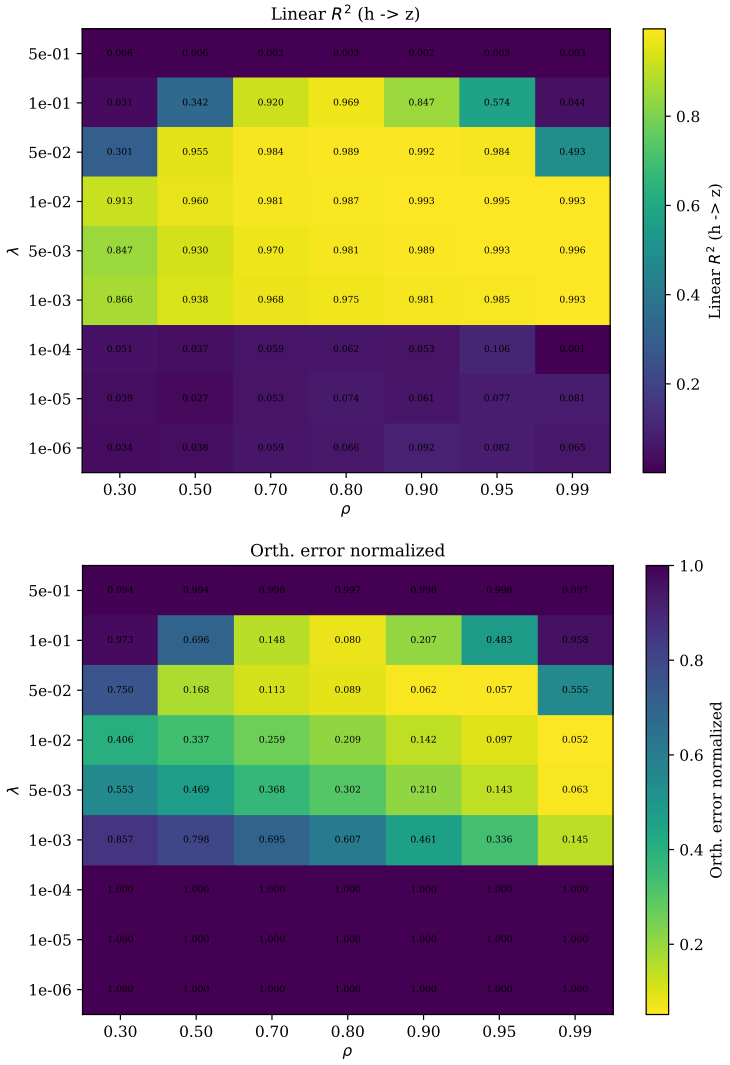


Figure 7: Grid search over the regularization weight λ and OU correlation ρ (5 seeds each, mean $\pm$ std). **Left:** Linear R^2 between true and learned latents ($\uparrow$ better). **Right:** Normalized orthogonality error of the fitted linear map ($\downarrow$ better). Identifiability requires both objectives to be active: too much Gaussianity ($\lambda = 0.5$) collapses the representation entirely ($R^2 \approx 0$), while even weak regularization ($\lambda \leq 10^{-2}$) suffices when the alignment signal is strong. At very high correlation ($\rho = 0.99$), performance degrades for moderate λ because near-identical positive pairs make the invariance loss trivially small, allowing SIGReg to dominate. The best identifiability ($R^2 > 0.96$, orthogonality error < 0.2) occurs at $\lambda \in [10^{-3}, 10^{-2}]$ with $\rho \in [0.8, 0.95]$.

H.7 Latent Variable Distribution Sweep Across Mixings

The main-text gennorm figure (Fig. 4b) shows linear recovery on the spiral mixing across latent distribution shapes $\alpha \in \{2^{-3}, \ldots, 2^5\}$ in the generalized normal family ($\alpha = 2$ Gaussian, $\alpha = 1$ Laplace, $\alpha \to \infty$ uniform). We extend this sweep to all four 2D mixings used in our paper (Figs. 8, 9).

Two patterns are robust across mixings. First, both methods peak at $\alpha = 2$, confirming Thm. 2 empirically: linear identifiability is maximal precisely at the Gaussian latents. Second, SIGReg's plateau is consistently wider than whitening's. Whitening collapses for $\alpha < 2$ (heavier-than-Gaussian tails), whereas SIGReg retains high recovery across a broad range of latent distribution

shapes. This confirms that SIGReg's Gaussianization of h is doing more than enforcing covariance: it absorbs latent-side non-Gaussianity that whitening cannot see.

The contrast between methods is most pronounced for the spiral mixing, the only one of the four that is approximately measure-preserving and therefore unconstrained by second-order statistics alone. For the banana, sinusoid, and NVP mixings, the mixing structure itself partially constrains the recovery, narrowing the gap between methods. The orthogonality error (Fig. 9) tells the same story from the geometric side: whitening achieves a minimum at $\alpha = 2$ where the second-order constraint exactly matches the latents, while SIGReg's basin is broader.

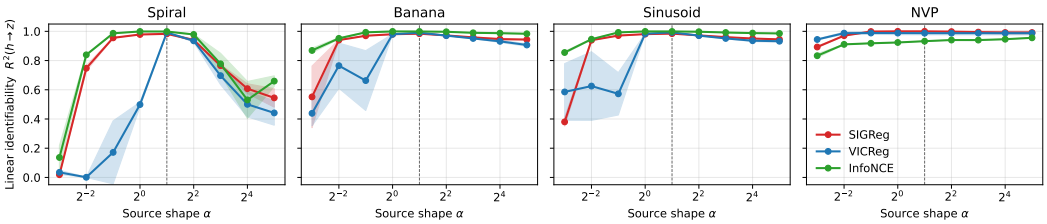


Figure 8: **Linear identifiability across the gennorm family for all four 2D mixings.** $R^2(h \to z)$ on a fixed evaluation grid as a function of latent shape α (α=1 Laplace, α=2 Gaussian). All three methods peak at $\alpha = 2$ as predicted by Thm. 2; SIGReg and InfoNCE, which constrain h beyond second moments, retain a wider plateau than VICReg for heavy-tailed latents. Mean $\pm$ std, 3 seeds.

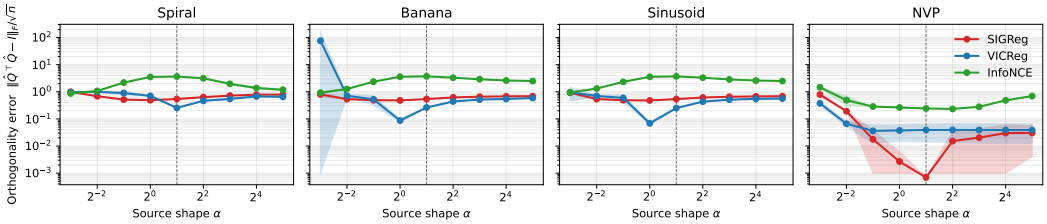


Figure 9: **Orthogonality error across the gennorm family for all four 2D mixings.** $\|\hat{Q}^\top\hat{Q} - I\|_F/\sqrt{N}$ on the same fixed grid (log scale). VICReg and SIGReg dip near $\alpha = 2$ where their constraints align with the latent distribution; InfoNCE remains roughly flat, reflecting weaker control over the linear map under fixed kernel width. Mean $\pm$ std over 3 seeds.

H.8 Approximate Bound Decomposition

Figure 10 decomposes the recovery error into its two sources: the whitening error ε and the alignment gap δ. The two quantities have distinct roles. Large ε alone does not predict poor recovery: the $\lambda = 5 \times 10^{-1}$ runs (crosses) achieve near-zero ε because SIGReg dominates, yet their recovery error is maximal because the alignment signal is absent (δ is large). Conversely, the alignment gap δ is a much stronger predictor of actual recovery error, consistent with the interpretation that the D term in the bound captures the nonlinear energy that alignment fails to suppress. This asymmetry confirms the practical takeaway from Thm. 3: approximate whitening is easy to achieve in practice, so the binding constraint on identifiability is the quality of the alignment objective.

H.9 Loss Predictivity

Figure 11 pools all trained encoders across our four experiments to examine whether the training losses are predictive of linear identifiability. We restrict the view to converged runs ($R^2 > 0.9$) to focus on the regime where the theory applies. The alignment loss is the strongest single predictor of R^2, consistent with its direct connection to the alignment gap δ in Thm. 3.

SIGReg does not vanish at the optimum. SIGReg is a goodness-of-fit statistic of Epps–Pulley type, multiplied by the batch size [2]. Under the null, i.e., for embeddings that are exactly $\mathcal{N}(0, I)$,

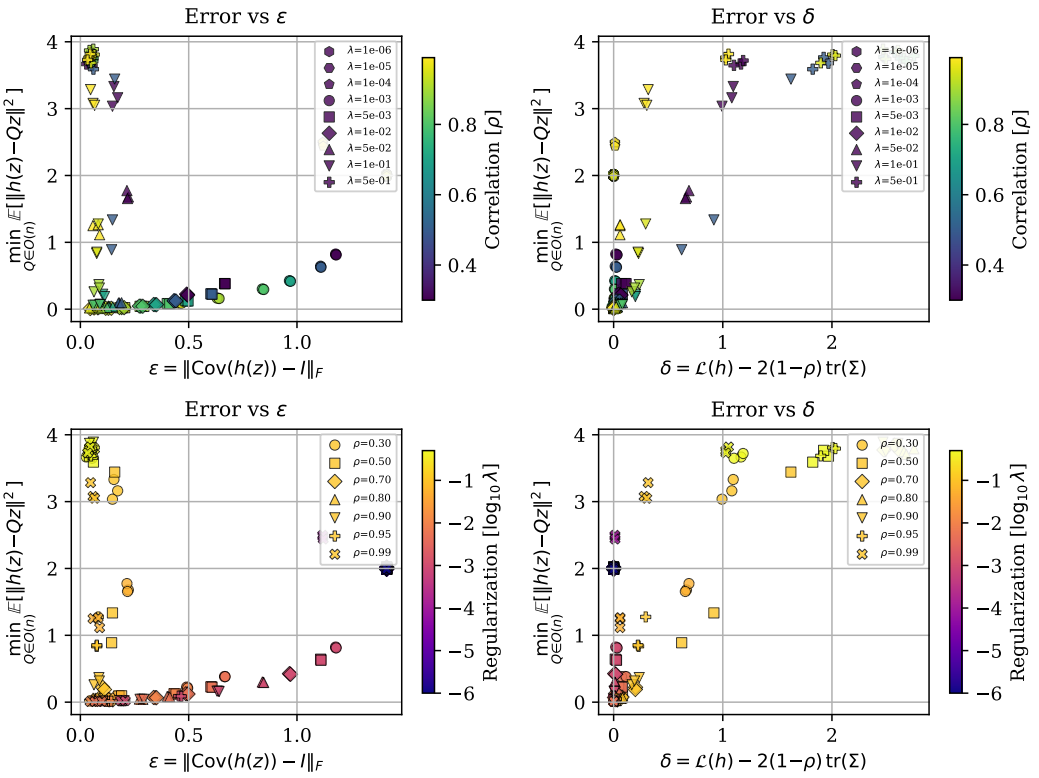


Figure 10: Decomposition of the recovery error into its two sources. **Left column:** Error vs whitening deviation ε. **Right column:** Error vs alignment gap δ. Top and bottom rows use different marker/color encodings for λ and ρ. The alignment gap δ is a much stronger predictor of recovery error than the whitening error ε, confirming that approximate whitening is the easy part; the binding constraint is alignment quality.

the bias of the estimator vanishes with the batch size while the multiplication holds its variance at order one, so the statistic converges to a constant of order one rather than to zero. On true isotropic Gaussian samples at our batch size and embedding dimension it reads 1.04; in identified runs it settles at 1.8–2.0 for all $N \geq 16$, independently of the dimension up to N=4096 (Tabs. 6, 9). A value of order one therefore indicates that the target is met up to sampling noise, not a failure to optimize, and the remaining gap is what Thm. 3 accounts for through ε.

H.10 Scaling Experiment

We test whether identifiability holds at scale by sweeping the latent dimension $N \in \{2^1, \ldots, 2^{12}\}$ with 5 seeds each; for $N > 1250$ the evaluation set grows to $8N$ points. The mixing function is a RealNVP-style coupling layer architecture (4 layers with orthogonal weight matrices) [85], and the encoder is a *matched* inverse-coupling-layer architecture with learnable weights. This architectural matching removes encoder expressivity as a confounder: any failure of identifiability is due to the optimization landscape, not the encoder's function class. For $N \leq 32$, we train $K = 3$ encoders per seed and select the one with lowest final loss; for $N > 32$, all runs converge to equivalent solutions ($K = 1$).

Tabs. 6–8 report the per-method results. The mixing difficulty is consistently nonlinear across dimensions ($R^2(x \to z) \approx 0.73$–$0.78$) for all methods, since the RealNVP mixing is shared. The three methods differ in their failure modes:

SIGReg (Tab. 6). Linear identifiability is essentially perfect at all dimensions ($R^2 > 0.999$, at N=4096 with batch size 1024). Alignment and SIGReg losses are stable across N. The only quantity that drifts is the orthogonality error $\|\hat{Q}^\top \hat{Q} - I\|_F / \sqrt{N}$, which grows from $\sim 10^{-5}$ at

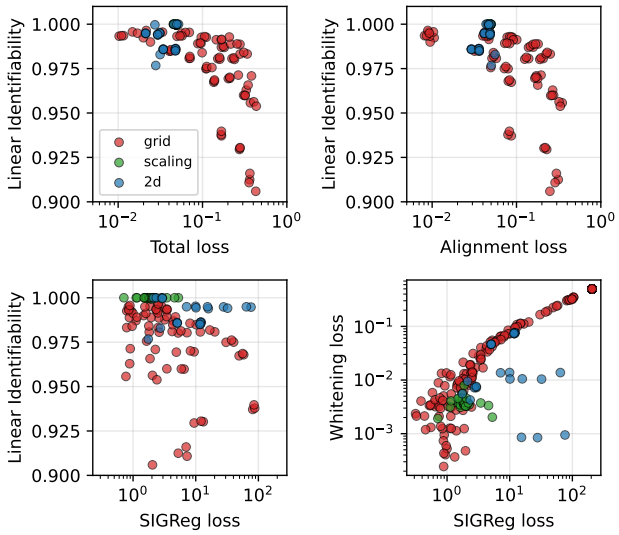


Figure 11: **Losses predict linear identifiability.** Each point is one trained encoder from one of our three experiments (2D illustrations, grid search, scaling), zoomed to the converged regime ($R^2 > 0.9$). **Top left:** Total loss (alignment + λ SIGReg) correlates with linear $R^2(h \to z)$. **Top right:** Alignment loss alone is predictive of identifiability quality. **Bottom left:** SIGReg loss vs. R^2. **Bottom right:** SIGReg and whitening losses are correlated.

N=2 to $\sim 2 \times 10^{-2}$ at N=2048. This is the slow degradation predicted by Thm. 3: at fixed alignment quality, the bound on the linear-part distortion grows with the dimension being whitened.

VICReg (Tab. 7). Numerically indistinguishable from SIGReg across the entire sweep, on both linear identifiability and orthogonality error. The whitening loss is small ($\sim 3.5 \times 10^{-3}$) and stable. Confirms that for Gaussian latents at the population optimum, second-moment whitening is sufficient (the Gaussianity assumption on $h(z)$ in Thm. 1 can be relaxed to whitening, see Sec. 5.2).

InfoNCE (Tab. 8). Qualitatively different. R^2 is non-monotone in N: high at N=16 (~ 0.9999), then collapsing to ~ 0.57–0.73 for $N \geq 64$. Orthogonality error is one to three orders of magnitude larger than for the batch-statistic methods. The InfoNCE loss column reveals the mechanism: at $N \geq 32$, the loss saturates at zero, indicating that the Gaussian kernel $\exp(-\|\cdot\|^2/\sigma^2)$ underflows for both positive and negative pairs once $\|z\|^2 \approx N$ at σ=1, killing the gradient. The non-monotone R^2 is the fingerprint of this kernel-width mismatch, not a property of the InfoNCE objective itself; per-dimension kernel tuning would presumably restore performance, at the cost of a hyperparameter sweep absent from the batch-statistic methods.

These three patterns make the main-text story precise: the practical gap between methods is not in their theoretical identifiability guarantees (Thm. 1 applies to all), but in how each method's optimization behaves at scale.

$N = 2^{12}$**.** At the default batch size of 256, all three objectives fail at N=4096 (Tab. 9, Fig. 12): the alignment loss stays at five times its value for $N \leq 2048$ and recovery drops to 0.65–0.73. This is an estimation limit rather than an identifiability one. A batch of 256 is smaller than the dimension, so the batch statistics that SIGReg and VICReg constrain are rank-deficient; raising the batch size to 1024 restores SIGReg to $R^2 = 0.9994$ (3 seeds), whereas quadrupling the number of SIGReg slices at batch size 256 has no effect. We have not determined the exact transition batch size.

H.11 Reacher Experiment

Environment and rendering. We use the DeepMind Control Suite Reacher ("hard" variant) [84] with two revolute joints: a shoulder (unbounded) and a wrist (limited to $[-2.79, 2.79]$ rad). The latent state is $z = (\theta_0, \theta_1) \in \mathbb{R}^2$ (Fig. 13). Images are rendered at 64×64 pixels via MuJoCo with EGL headless rendering. The target position is fixed at $(0.1, 0.1)$ across all experiments to eliminate

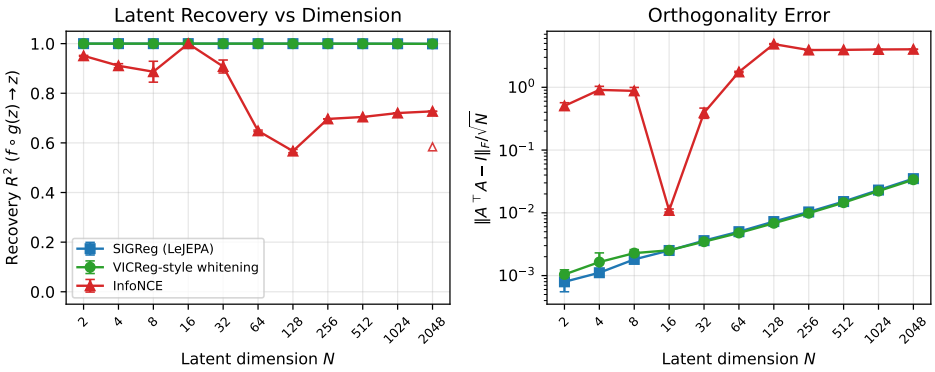


Figure 12: **Scaling with latent dimension** (5 seeds, batch size 256). **Left:** Recovery $R^2(h \to z)$; open marker: held-out recovery for InfoNCE at N=2048. **Right:** Normalized orthogonality error of the fitted linear map (log scale), which grows gradually with N for SIGReg and VICReg. N=4096 is shown in Tab. 9.

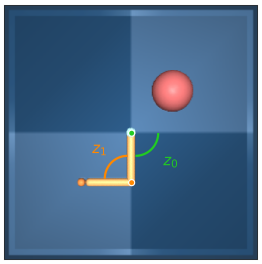


Figure 13: **DMC Reacher.** The latent state $z = (z_0, z_1)$ consists of two joint angles (shoulder and wrist) that fully determine the arm configuration. The nonlinear mixing g is the MuJoCo rendering pipeline producing 64×64 pixel observations.

it as a confound. For each experimental condition, we pre-render 100,000 image pairs and 10,000 evaluation images (i.i.d. Gaussian samples, shared across all conditions). Images are stored as uint8 and normalized per-channel at training time.

OU condition. For each $\rho \in \{0.3, 0.5, 0.7, 0.8, 0.9, 0.95, 0.99\}$, we sample $z \sim \mathcal{N}(0, I_2)$, generate z' via (1), and render. The theory's assumptions (Gaussian marginals, isotropic OU transitions, additive noise) are exactly satisfied. Wrapping beyond $\pm\pi$ occurs for $< 0.2\%$ of samples.

Trajectory condition. We use 10,000 episodes (201 steps each) from the LeWorldModel [9] dataset, collected by a trained SAC policy. For each temporal stride $\delta \in \{1, 2, 4, 8, 16, 32, 64\}$, we subsample 10 pairs per episode (100,000 total), extract the joint angles (θ_0^t, θ_1^t) and $(\theta_0^{t+\delta}, \theta_1^{t+\delta})$, and re-render with the fixed target. The marginal distributions (Figure 14) violate the Gaussian assumption: the shoulder is broad and heavy-tailed (13.6% beyond $\pm\pi$), while the wrist is nearly uniform over $[-\pi, \pi]$. Autocorrelations are anisotropic (e.g., at $\delta = 8$: $\rho_0 = 0.992$, $\rho_1 = 0.982$).

Encoder architecture. We use a CNN with BatchNorm:

$$\texttt{Conv}(3 \to 32, 4 \times 4, s{=}2) \to \texttt{BN} \to \texttt{GELU} \to \texttt{Conv}(32 \to 64) \to \texttt{BN} \to \texttt{GELU}$$
$$\to \texttt{Conv}(64 \to 128) \to \texttt{BN} \to \texttt{GELU} \to \texttt{Conv}(128 \to 256) \to \texttt{BN} \to \texttt{GELU}$$
$$\to \texttt{AvgPool}(4) \to \texttt{Flatten} \to \texttt{Linear}(256, 256) \to \texttt{BN} \to \texttt{GELU} \to \texttt{Linear}(256, 2).$$

This architecture has $\sim$1.1M parameters. BatchNorm was critical for stable training: without it, $\sim$36% of runs collapsed to zero-variance outputs. Gradient clipping ($\|\nabla\|_{\max} = 1.0$) is applied after each backward pass.

Training. We train with AdamW (lr $= 3 \times 10^{-3}$, weight decay 10^{-4}) and cosine annealing for 100 epochs, batch size 256, with SIGReg using 256 random slices. The loss is $\mathcal{L} = \lambda\mathcal{L}_{\text{SIG}} + (1-\lambda)\mathcal{L}_{\text{inv}}$.

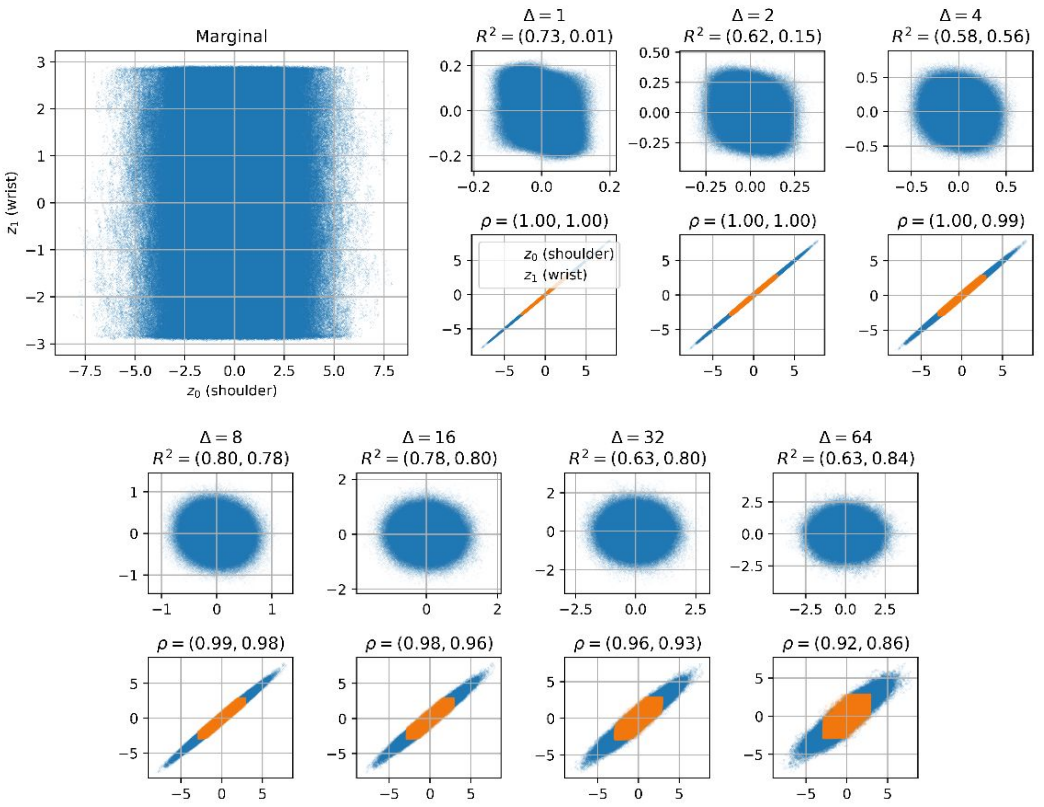


Figure 14: **Reacher trajectory latent distributions across temporal strides δ. Left:** Stationary marginal of (z_0, z_1); shoulder is broad, wrist is nearly bimodal. **Top row:** 2D transition differences $z_{t+\delta} - z_t$, with the per-dimension R^2 of the best-tuned encoder. **Bottom row:** Autocorrelation scatter z_t vs. $z_{t+\delta}$, per dimension, with Pearson ρ annotated. Small δ: $\rho \approx 1$, transition is trivial, alignment carries no signal. Large δ: ρ has dropped but the difference distribution inherits the non-Gaussian structure of the stationary marginals. Intermediate δ: both ρ is informative and the differences are effectively Gaussian, and identifiability is highest.

We sweep $\lambda \in \{10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$ with 3 seeds per setting, reporting the best λ per condition (mean $\pm$ std over seeds). The final model (after 100 epochs) is always used; no early stopping or checkpoint selection is applied.

Evaluation. We fit a linear regression from embeddings to joint angles on 10,000 training images and score on 10,000 held-out evaluation images (proper train/test split). We report bidirectional R^2, per-dimension R^2, and an R^2 diagnostic for $\sin / \cos$ targets $(\sin\theta_0, \cos\theta_0, \sin\theta_1, \cos\theta_1)$. The $\sin / \cos$ diagnostic tests whether the encoder learns a trigonometric representation of the cyclic angles; negative values indicate the 2D embedding encodes raw angles (linear in θ) rather than their trigonometric functions.

Latent distributions. Figure 14 shows the stationary joint-angle distribution and, for each stride δ, the 2D transition-difference distribution together with the per-dimension autocorrelation scatter. The stationary marginal (leftmost panel) reveals a striking asymmetry: the shoulder (z_0) is broad and mildly platykurtic, while the wrist (z_1) is nearly bimodal, reflecting a policy-induced preference for two joint-limit configurations. The transition-difference panels (top row) grow from tight, narrow clouds at small δ to fully-developed isotropic clouds at $\delta = 64$. The autocorrelation scatter (bottom row) confirms that both joints have $\rho \approx 1$ at $\delta = 1$ (the transition is nearly trivial) and decorrelate monotonically, with the wrist decorrelating faster. The per-dimension R^2 values annotated on each panel already hint at the core phenomenon examined in the next paragraph: both small and large δ are bad, and identifiability peaks in an intermediate regime.

Gaussianity and autocorrelation jointly determine identifiability. Thm. 1 requires two conditions on the data-generating process: Gaussian latent marginals and a non-trivial autocorrelation $\rho \in (0, 1)$. Thm. 3 relaxes these to approximate whitening and a non-vanishing spectral gap. The trajectory condition satisfies neither exactly, and the temporal stride δ controls how close each condition comes to being met. To disentangle these two effects, we measure the (zscored) SIGReg value of each transition-difference distribution and of their 2D joint, and plot it against the Pearson autocorrelation, colored by identifiability R^2 (Figure 15). We compare to a finite-sample floor obtained by evaluating SIGReg on iid standard Gaussian samples of matched size; this floor is approximately 1.2, so any value near 1–2 is indistinguishable from Gaussian at this sample size.

Three regimes are visible. At $\delta = 1, 2$ (top right of each panel), ρ is essentially 1 and SIGReg is two or three orders of magnitude above the floor: the transition is too trivial to carry an identifiability signal *and* the differences inherit the non-Gaussian structure of the stationary marginals, because at this stride the differences $z_{t+\delta} - z_t$ are tiny perturbations that cannot Gaussianize via any averaging. At $\delta = 64$ (far left), ρ has dropped to 0.86–0.92 and the shoulder differences are near the Gaussian floor, but the wrist differences climb again because the increments at this stride start to resolve the bimodal stationary distribution (they approach a symmetrized sum of two draws from a bimodal law). At $\delta = 8, 16$, *both* constraints are satisfied simultaneously: $\rho \in [0.96, 0.99]$ gives a meaningful spectral gap $2\rho(1 - \rho)$, and SIGReg drops to within a few times the floor for both marginals and the joint. This is precisely the regime where R^2 is maximal, as shown by the yellow points clustered in the bottom-middle region of each panel. The best OU ρ value (vertical dashed line, $\rho = 0.95$) sits near these high-R^2 points, confirming that the trajectory condition most closely approaches the theoretical assumptions at intermediate stride.

The joint panel (rightmost) tells a consistent story at slightly higher absolute values: the 2D SIGReg of the transition-difference distribution is always larger than either marginal because it additionally penalizes the cross-dimension dependence that neither marginal can see. The R^2 ordering is the same as in the per-dimension panels.

Additional results on OU and Trajectory conditions. Fig. 16 directly contrasts the two: at matched ρ, Gaussian (OU) latents achieve substantially higher R^2 than SAC-policy trajectories (right panel); an empirical instance of Thm. 2 on raw pixels. Within OU alone (left), R^2 rises monotonically in ρ, and stronger λ helps only at low ρ where the alignment signal is weak.

Fig. 17 resolves identifiability by joint. OU recovers shoulder and wrist symmetrically, consistent with the isotropy of the Gaussian OU transition (see App. F for why isotropic transitions are necessary for the simultaneous approach). Trajectory recovery is massively asymmetric: the wrist, driven by a nearly-bimodal policy-induced marginal and a faster autocorrelation, is recovered poorly at small δ and improves only once δ is large enough to provide informative temporal variation; the shoulder degrades at large δ due to wrapping beyond $\pm\pi$. This is the empirical signature predicted by the Sturm–Liouville analysis when isotropy fails.

Fig. 18 connects back to the approximate bound. **Left:** OU identifiability is robust to λ at high ρ; only at $\rho = 0.99$ with the strongest λ does SIGReg begin to dominate alignment, mirroring the collapse regime of the grid search (Fig. 7). **Right:** the fitted orthogonality error $\|\hat{Q}^\top \hat{Q} - I\|_F / \sqrt{N}$ decreases monotonically with ρ, tracking the tightening spectral gap $2\rho(1 - \rho)$ that governs Thm. 3.

H.12 Planning Experiment

The planning experiment (Sec. 6.4, Fig. 5) evaluates straight-line latent interpolation across all trained Reacher encoders (App. H.11). The qualitative panels (Fig. 5 left) compare the best Gaussian encoder (OU condition at $\rho = 0.99$) against the best Trajectory encoder (stride $\delta = 8$); the quantitative panels (middle, right) aggregate over the full set.

Setup. We construct a retrieval library of 10,000 held-out evaluation frames paired with their joint-angle labels. Given a start frame x_s and a goal frame x_g, we encode both to latents $y_s = f(x_s)$ and $y_g = f(x_g)$ and form the straight-line latent interpolant $y_t = (1 - t)\, y_s + t\, y_g$ for $t \in \{0, \frac{1}{T}, \ldots, 1\}$ with $T = 15$. Each y_t is decoded back to a frame by k-nearest-neighbor retrieval ($k = 1$, Euclidean distance) against the library's embeddings, yielding a decoded joint-space path $\{\hat{\theta}_t\}$. We evaluate this path against the ideal straight-in-joint-space trajectory connecting θ_s and θ_g.

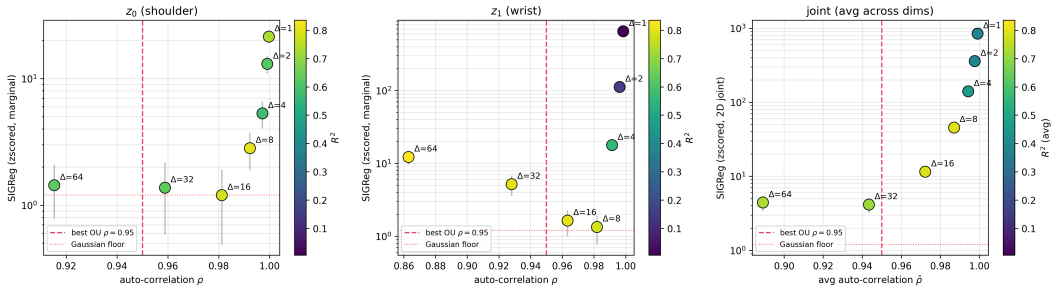


Figure 15: **Identifiability requires both approximate Gaussianity and non-trivial autocorrelation.** For each temporal stride δ, we plot SIGReg (zscored, averaged over random projections and subsamples, error bars show one standard deviation) against the Pearson autocorrelation ρ of the transition-difference distribution, colored by the corresponding identifiability R^2 (see Table 2). **Left:** shoulder (z_0, 1D marginal). **Middle:** wrist (z_1, 1D marginal). **Right:** 2D joint; x-axis is the average $\bar{\rho}$ across dimensions. Dotted red line: SIGReg floor from matched iid $\mathcal{N}(0, I)$ samples (≈ 1.2). Dashed red line: best-performing OU autocorrelation ($\rho = 0.95$). Small δ (top right): $\rho \approx 1$ yields no spectral gap, and small differences inherit the non-Gaussian stationary shape; both constraints fail. Large δ (far left): the wrist difference starts resolving the bimodal stationary, so SIGReg climbs again. Intermediate $\delta \in \{8, 16\}$: simultaneously near the Gaussian floor and in a regime of informative ρ; identifiability R^2 peaks here.

Metric. We sample $K = 30$ start–goal pairs uniformly from the eval library (shared seed across encoders) and report **path length**: $\left(\sum_t \|\hat{\theta}_{t+1} - \hat{\theta}_t\|\right) / \|\theta_g - \theta_s\|$, the θ-space arc length of the decoded path divided by the chord length. Ideal value 1 (straight chord); larger values indicate the encoder warps joint-space geometry. We also track mean orthogonal deviation from the chord as a complementary diagnostic.

Interpretation. Under an affine encoder $h(z) = Qz$, the latent straight line maps to the true straight line in joint space and the metric attains its ideal value; deviations measure how much the encoder warps joint-space geometry. Fig. 19 decomposes this geometrically.

H.13 Misspecified Latent Dimension

Setup. The true latent dimension N is fixed and the encoder output dimension M is swept. Since the matched encoder of the scaling experiment is a bijection and cannot change dimension, we use the 4-layer MLP encoder (hidden width 512 for RealNVP with N=8, 256 for the spiral with N=2), trained with SIGReg ($\lambda = 10^{-3}$, $\rho = 0.95$, batch size 256, 20,000 steps, 5 seeds) and evaluated on 20,000 held-out points. For $M \neq N$, recovery $R^2(h \to z)$ regresses the N latents on the M-dimensional embedding and encoding $R^2(z \to h)$ the reverse. The participation ratio $\mathrm{PR} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ is computed from the eigenvalues λ_i of the $M \times M$ embedding covariance C; equivalently, $\mathrm{PR} = (\mathrm{tr}\, C)^2 / (\varepsilon^2 + 2\,\mathrm{tr}\, C - M)$ with $\varepsilon = \|C - I_M\|_F$.

Results. Tab. 10 and Figs. 6, 20 give both settings. Below N, both settings behave identically: recovery is close to M/N and encoding is near-perfect, so the embedding is an isometric copy of an M-dimensional subspace of the latents, and $\mathrm{PR} = M$ exactly because SIGReg fills every output dimension. Above N the two settings differ. On RealNVP the surplus dimensions collapse, PR stays at N up to M=16, and recovery and encoding stay near 0.99; at M=32, 64 some seeds lose a latent ($\mathrm{PR} < N$, larger seed variance). On the spiral the surplus dimensions are filled with nonlinear functions of z: recovery stays high, encoding drops to 0.61–0.73, and PR keeps growing to about 4. In both settings, however, $\mathrm{PR} = M$ holds exactly up to M=N and fails from M=N+1 on, which is the basis of the estimator in Sec. 6.5.

H.14 Predictors

Setup. LeJEPA's objective (Sec. 3.2) aligns the two embeddings directly. JEPAs with a predictor instead align $h(z')$ with a learned prediction $P(h(z))$, i.e., $\mathcal{L}_{\mathrm{inv}} = \mathbb{E}\left[\|h(z') - P(h(z))\|^2\right]$, trained

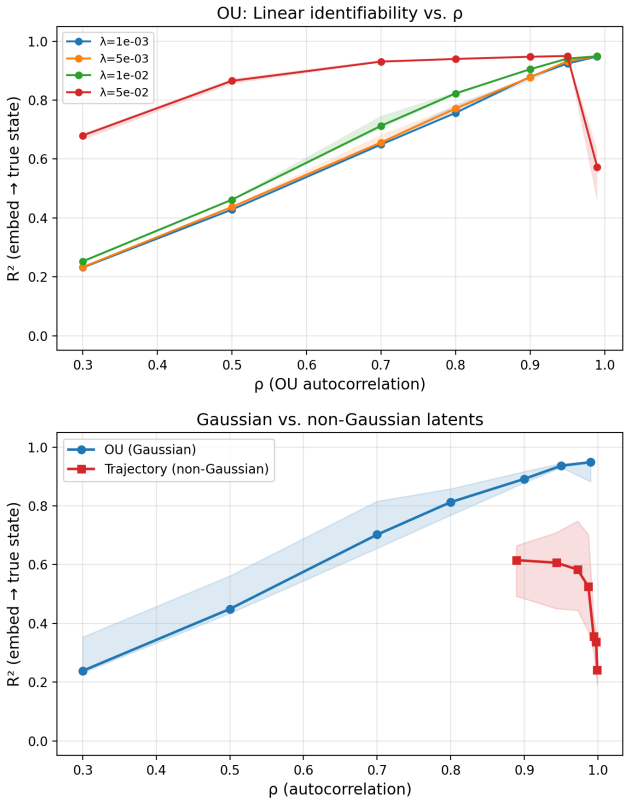


Figure 16: **Left:** OU identifiability vs. ρ for different λ values. R^2 increases monotonically, with all λ values converging at high ρ. Higher λ (stronger SIGReg) helps at low ρ where the alignment signal is weak. **Right:** Gaussian (OU) vs. trajectory data at matched autocorrelation ρ. At the same ρ, Gaussian latents achieve substantially higher R^2, directly validating the converse theorem: non-Gaussian marginals reduce identifiability.

jointly with the encoder. We compare three arms that are identical in every other respect: no predictor ($P = \mathrm{id}$, Eq. 3), a linear predictor ($P \in \mathbb{R}^{M\times M}$, no bias, random initialization), and an MLP predictor (two hidden layers of the encoder's width with GELU; 271k parameters for RealNVP). All arms use positive pairs (z, z') from the OU transition (1) with $\rho = 0.95$, the 4-layer MLP encoder (hidden width 512 for RealNVP with N=8, 256 for the spiral with N=2), SIGReg with $\lambda = 4 \times 10^{-3}$, batch size 256, 20,000 steps, and 5 seeds.

Results. All three arms reach the same recovery in both settings (Tab. 11, Fig. 21). For the linear predictor, the theory suggests a specific solution: at $h(z) = Qz$ with $Q \in O(N)$, the least-squares predictor is $\mathrm{Cov}(h(z'), h(z))\,\mathrm{Cov}(h(z))^{-1} = \rho I$, the true transition operator in the learned coordinates (Thm. 1). The learned P converges to it from a random initialization, $\|P - \rho I\|_F = 0.011$ (RealNVP) and 0.002 (spiral). We do not have a proof that the predictive objective selects $h(z) = Qz$, so this is consistency rather than a theorem. For the MLP predictor no such statement is possible: an expressive predictor can in principle absorb a nonlinear distortion of h without changing the loss. That gradient descent does not find such solutions here is an empirical observation, which a full treatment of action-conditioned predictors [9] would need to explain.

H.15 Training Dynamics

Our theorems characterize the optimum, not the trajectory that reaches it. Fig. 22 shows recovery during training for the scaling runs (SIGReg, matched encoder, batch size 256). Recovery first rises slowly and then jumps to its final value within a few hundred steps, and Tab. 12 lists the first step at which $R^2(h \to z) > 0.9$. This escape step is not monotone in N: it grows up to $N \approx 8$–

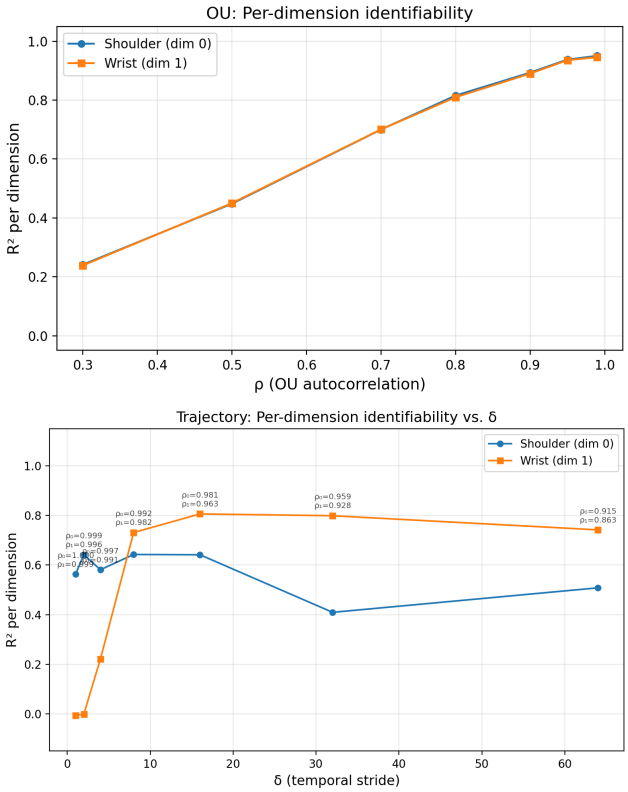


Figure 17: **Per-dimension identifiability. Left:** OU condition; shoulder and wrist are recovered symmetrically, consistent with the isotropic transition. **Right:** Trajectory condition; massive asymmetry: the wrist ($R^2 \approx 0$ at $\delta = 1$) recovers only at larger δ where temporal variation provides learning signal; the shoulder is consistently easier but degrades at large δ due to wrapping beyond $\pm\pi$. Per-dimension ρ values are annotated.

32, where it also varies most across seeds, and then shortens and becomes nearly seed-independent for $N \geq 128$. At N=4096 and batch size 256, recovery never crosses 0.9 within 20,000 steps (App. H.10). Explaining these dynamics, for instance by tracking per-coordinate recovery during training, is left to future work.

H.16 ImageNet-X Linear Probes

Encoders. LeJEPA, DINO [22], and iBOT [87] are ViT-B/16 models trained on ImageNet-1k with a shared codebase and schedule, so that the objective is the only variable among them. DINOv2 [88] and DINOv3 [30] (LVD-142M and LVD-1689M, the latter distilled from a 7B teacher) are reference points with two to three orders of magnitude more data and are not part of the matched comparison. Two controls separate the objective from the architecture and the images: the same ViT-B/16 at random initialization (*init*) and a random linear map of the raw pixels (*pixels*).

Analysis space. All analyses use 64 dimensions. For LeJEPA this is the output of the projector, where both loss terms attach and therefore the space the theory is about. For every other encoder it is the top 64 principal components of the 768-dimensional CLS embedding; since linear decodability is invariant to invertible linear maps, nothing linearly decodable within this subspace is lost. The subspace is matched in nominal dimension, not in retained information: LeJEPA's backbone has fewer effective dimensions than DINO's or iBOT's, so the projection discards proportionally more from the latter.

Probes. We probe the 17 binary factors annotated by ImageNet-X [24] on 43,603 ImageNet validation images, following Reizinger et al. [70, Fig. 3]. For each factor we split 70/30, balance both

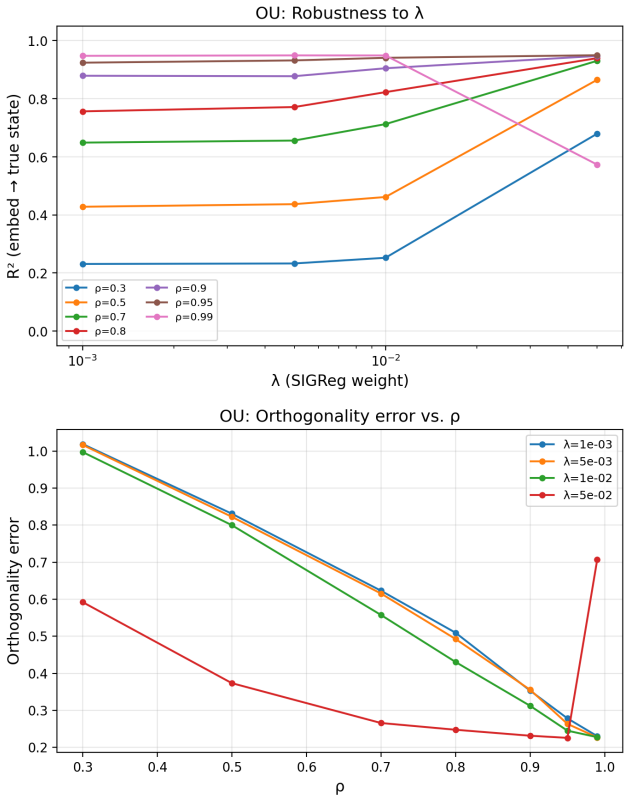


Figure 18: **Left:** λ robustness in the OU condition. For high ρ, identifiability is stable across λ; for low ρ, stronger regularization ($\lambda = 5 \times 10^{-2}$) compensates for the weak alignment signal. At $\rho = 0.99$, the highest λ degrades slightly as SIGReg begins to dominate alignment. **Right:** Orthogonality error decreases monotonically with ρ, consistent with the approximate bound (Thm. 3).

classes by subsampling, hold out 20% of the training split for early stopping, and report balanced test accuracy over 5 repetitions. Probes are full-batch AdamW (learning rate 10^{-2}, weight decay 10^{-3}, early stopping with patience 300 steps); the MLP probe has one hidden layer of width 256 and is otherwise identical to the linear probe.

Results. Tab. 13 lists all factors. Balanced sample sizes span three orders of magnitude, so per-factor comparisons are only meaningful for the nine factors with $N > 1000$, which give the main-text means (Tab. 3). LeJEPA has the highest accuracy on eight of these nine, both among the matched IN-1k encoders and against DINOv2 and DINOv3; the exception is *smaller*. Both controls stay well below every trained encoder, so the probes do not overfit. MLP probes improve on linear ones by only 0.01–0.03 for every encoder (Tab. 14); this gap is bounded by the headroom below the Bayes rate of noisy human annotations and supports no cross-model claim.

Gaussianity. Tab. 14 reports the sketched Epps–Pulley statistic of LeJEPA [2, Alg. 1] against $\mathcal{N}(0, I)$, divided by the sample size, and the participation ratio of the covariance, normalized by the dimension. The claim-bearing comparison is LeJEPA's projector output against a true Gaussian sample of the same size, since that is where SIGReg was applied: its distance is 0.0049, two orders of magnitude above the sampling floor (3×10^{-5}) but fifty times below LeJEPA's own backbone (0.25), with nearly isotropic covariance (normalized participation ratio 0.91). SIGReg is therefore met approximately, the regime of Thm. 3. For the other encoders the coordinates are principal components, which Gaussianize marginals by the central limit theorem regardless of the underlying distribution, so their statistics are not directly comparable.

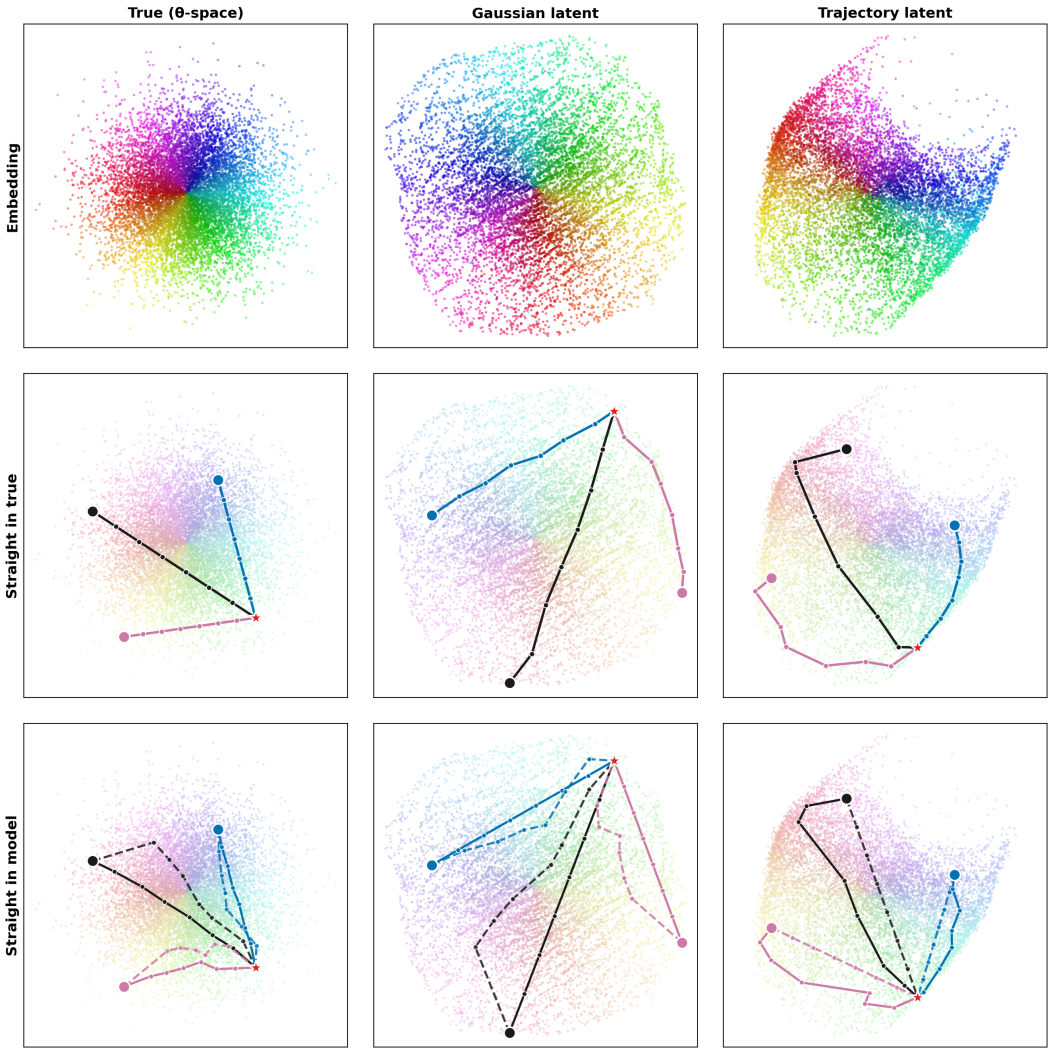


Figure 19: **Planning: embedding and straight-line paths.** Columns: true θ-space (left), Gaussian/OU encoder (center), Trajectory encoder (right). **Top row:** scatter of eval-set embeddings, colored by true θ-space polar angle. The Gaussian encoder is an approximate rotation of the true latent; the Trajectory encoder is visibly warped. **Middle row:** three example trajectories that are straight in the *true* joint space, rendered in each representation. They remain approximately straight in the Gaussian embedding but curve in the Trajectory embedding. **Bottom row:** the complementary test; straight lines in each model's latent, decoded back to joint space. Gaussian-encoder plans stay near-straight in θ-space; Trajectory-encoder plans curve. Together, the middle and bottom rows confirm Fig. 5: only linear identifiability turns straight-latent plans into straight joint-space paths.

File / Component	Lean name	Status
Hermite.lean: Thm. 1 via Hermite polynomials		
Hermite basis completeness	`mehler_summability`	axiomatized
Contraction lemma (ρ^d decay)	(folded into above)	axiomatized
Mehler's formula	`correlation_eq_spectral_sum`	axiomatized
$\rho^d \leq \rho$ for $d \geq 1$	`pow_le_self_of_pos_lt_one`	VERIFIED
$\rho^d < \rho$ for $d \geq 2$ (strict)	`pow_lt_self_of_ge_two`	VERIFIED
Pointwise bound $w_d \rho^d \leq w_d \rho$	`spectral_term_le`	VERIFIED
Correlation bound $\leq \rho$ (4)	`correlation_le_rho`	VERIFIED
Equality $\Leftrightarrow w_1 = 1$	`equality_forces_degree_one`	VERIFIED
Degree-1 $\Rightarrow$ linearity	`linear_of_degree_one`	axiomatized
Gaussianity + linear $\Rightarrow$ orthogonal	`orthogonal_of_gaussian_linear`	axiomatized
Loss lower bound $\mathcal{L} \geq 2(1-\rho)N$	`loss_lower_bound`	VERIFIED
Theorem assembly $h = Qz$	`hermite_identifiability`	VERIFIED
Uniqueness.lean: Thm. 2 (Gaussian uniqueness)		
Affine eigenfunction $\Rightarrow$ affine score	`score_affine_of_eigenfunction`	VERIFIED
Affine score $\Rightarrow$ Gaussian density	`gaussian_of_affine_score`	axiomatized
Only-if direction assembly	`gaussian_of_affine_eigenfunction`	VERIFIED
Gaussian $\Rightarrow$ Hermite eigenfunctions	`hermite_first_eigenfunction_of_gaussian`	axiomatized
Full biconditional	`gaussian_uniqueness`	VERIFIED
Zero-mean specialization	`score_pure_linear_zero_mean`	VERIFIED
Dirichlet.lean: Appendix E via Dirichlet energy		
AM-GM inequality	`amgm_sum_ge_prod_pow`	axiomatized
Jensen's inequality (exp)	`exp_mean_ge_mean_exp`	axiomatized
Mazur–Ulam theorem	`mazur_ulam`	axiomatized
Norm-preserving CLM $\Rightarrow$ isometry	`clm_isometry_of_norm_preserving`	VERIFIED
Orthogonal Jacobian $\Rightarrow$ 1-Lipschitz	`lipschitz_of_orthogonal_jacobian`	VERIFIED
Bilipschitz $\Rightarrow$ global isometry	`isometry_of_bilipschitz`	VERIFIED
$h(0) = 0 \Rightarrow$ linear isometry	(part of assembly)	VERIFIED
Theorem assembly $h = Qz$	`dirichlet_identifiability`	VERIFIED
Approx.lean: Thm. 3 (approximate bound)		
Spectral gap positivity	`spectral_gap_pos`	VERIFIED
$W_{\mathrm{nl}} \leq D$ from alignment gap	`nonlinear_energy_le_D`	VERIFIED
Polar decomposition bound	`polar_bound_axiom`	axiomatized
Pythagorean decomposition	`pythagorean_axiom`	axiomatized
$\|M - Q\|_F^2$ bound	`linear_deviation_sq_bound`	VERIFIED
Monotonicity of $(\varepsilon + t)^2 + t$	`bound_monotone`	VERIFIED
Full bound $D + (\varepsilon + D)^2$	`approximate_identifiability`	VERIFIED
Exact recovery ($\delta = \varepsilon = 0$)	`exact_recovery_special_case`	VERIFIED
Planning.lean: Thm. 4 (planning equivalence)		
Orthogonal invariance definition	`IsOrthogonalInvariant`	VERIFIED
Stage pushforward	`stage_pushforward`	axiomatized
Terminal pushforward	`terminal_pushforward`	axiomatized
Per-step stage cost equivalence	`stage_cost_equiv`	VERIFIED
Terminal cost equivalence	`terminal_cost_equiv`	VERIFIED
Total cost equivalence (main step)	`planning_equivalence`	VERIFIED
Minimizer equivalence	`minimizer_equivalence`	VERIFIED
Value equivalence	`value_equivalence`	VERIFIED

Table 5: Lean 4 verification status for each proof component. VERIFIED: machine-checked with zero `sorry`. AXIOMATIZED: stated as a Lean `axiom`; standard result not yet in Mathlib or requiring nontrivial API software engineering.

Table 6: **Scaling Experiment (SIGReg)** (mean ± std, 5 seeds). The RealNVP mixing is consistently nonlinear across dimensions ($R^2(x \to z) < 1$). The learned model nonetheless recovers the true latents at all dimensions. Training losses are stable; orthogonality error grows gradually with N. †Batch size 1024, 3 seeds (App. H.10, Tab. 9).

Latents	**Mixing difficulty**		**Linear identifiability**		**Orthogonality**	**SIGReg losses**	
N	$R^2(z \to x)$ $\pm\text{std} \times 10^{-3}$	$R^2(x \to z)$ $\pm\text{std} \times 10^{-3}$	$R^2(z \to h)$ $\pm\text{std} \times 10^{-7}$	$R^2(h \to z)$ $\pm\text{std} \times 10^{-7}$	$\|\hat{Q}^\top \hat{Q} - I\|_F/\sqrt{N}$ $\pm\text{std} \times 10^{-5}$	Align $\pm\text{std} \times 10^{-4}$	SIGReg $\pm\text{std}$
2^1	$0.871_{\pm 1.4}$	$0.782_{\pm 2.8}$	$0.99999_{\pm 3.4}$	$0.99999_{\pm 3.4}$	$0.000_{\pm 24}$	$0.0472_{\pm 15}$	$2.54_{\pm 1.4}$
2^2	$0.865_{\pm 16}$	$0.724_{\pm 24}$	$0.99999_{\pm 12}$	$0.99999_{\pm 12}$	$0.001_{\pm 12}$	$0.0475_{\pm 10}$	$2.09_{\pm 1.3}$
2^3	$0.860_{\pm 4.2}$	$0.729_{\pm 10}$	$0.99999_{\pm 9.0}$	$0.99999_{\pm 9.0}$	$0.001_{\pm 14}$	$0.0473_{\pm 13}$	$2.06_{\pm 0.78}$
2^4	$0.858_{\pm 6.9}$	$0.733_{\pm 6.0}$	$0.99998_{\pm 4.9}$	$0.99998_{\pm 4.9}$	$0.002_{\pm 18}$	$0.0476_{\pm 3.6}$	$1.76_{\pm 0.18}$
2^5	$0.851_{\pm 3.0}$	$0.738_{\pm 1.6}$	$0.99998_{\pm 7.2}$	$0.99998_{\pm 7.2}$	$0.003_{\pm 7.6}$	$0.0479_{\pm 5.0}$	$1.93_{\pm 0.17}$
2^6	$0.854_{\pm 1.5}$	$0.737_{\pm 1.4}$	$0.99996_{\pm 7.6}$	$0.99996_{\pm 7.4}$	$0.005_{\pm 10}$	$0.0484_{\pm 3.9}$	$2.03_{\pm 0.14}$
2^7	$0.853_{\pm 0.62}$	$0.739_{\pm 0.50}$	$0.99993_{\pm 3.1}$	$0.99993_{\pm 3.2}$	$0.007_{\pm 7.4}$	$0.0486_{\pm 1.2}$	$1.80_{\pm 0.09}$
2^8	$0.855_{\pm 0.30}$	$0.742_{\pm 0.48}$	$0.99988_{\pm 7.7}$	$0.99988_{\pm 7.9}$	$0.010_{\pm 5.0}$	$0.0486_{\pm 2.9}$	$1.88_{\pm 0.17}$
2^9	$0.859_{\pm 0.24}$	$0.749_{\pm 0.30}$	$0.99977_{\pm 6.5}$	$0.99977_{\pm 6.7}$	$0.015_{\pm 3.7}$	$0.0489_{\pm 2.4}$	$1.93_{\pm 0.08}$
2^{10}	$0.866_{\pm 0.10}$	$0.763_{\pm 0.17}$	$0.99956_{\pm 12}$	$0.99956_{\pm 12}$	$0.023_{\pm 4.7}$	$0.0488_{\pm 0.91}$	$1.93_{\pm 0.06}$
2^{11}	$0.869_{\pm 0.03}$	$0.768_{\pm 0.05}$	$0.99904_{\pm 8.1}$	$0.99904_{\pm 7.9}$	$0.035_{\pm 1.5}$	$0.0491_{\pm 0.48}$	$1.88_{\pm 0.07}$
2^{12} †	$0.869_{\pm 0.04}$	$0.768_{\pm 0.02}$	$0.99945_{\pm 1.7}$	$0.99944_{\pm 1.9}$	$0.026_{\pm 0.90}$	$0.0489_{\pm 0.14}$	$1.91_{\pm 0.11}$

Table 7: **Scaling Experiment (VICReg)** (mean ± std, 5 seeds). VICReg matches SIGReg on linear identifiability across all dimensions; orthogonality error grows similarly with N. The whitening loss is stable and small throughout.

Latents	**Mixing difficulty**		**Linear identifiability**		**Orthogonality**	**VICReg losses**	
N	$R^2(z \to x)$ $\pm\text{std} \times 10^{-3}$	$R^2(x \to z)$ $\pm\text{std} \times 10^{-3}$	$R^2(z \to h)$ $\pm\text{std} \times 10^{-7}$	$R^2(h \to z)$ $\pm\text{std} \times 10^{-7}$	$\|\hat{Q}^\top \hat{Q} - I\|_F/\sqrt{N}$ $\pm\text{std} \times 10^{-5}$	Align $\pm\text{std} \times 10^{-4}$	Whitening $\pm\text{std} \times 10^{-4}$
2^1	$0.871_{\pm 0.66}$	$0.782_{\pm 1.7}$	$0.99999_{\pm 8.4}$	$0.99999_{\pm 8.4}$	$0.001_{\pm 19}$	$0.0491_{\pm 24}$	$0.0022_{\pm 13}$
2^2	$0.865_{\pm 15}$	$0.726_{\pm 24}$	$0.99998_{\pm 54}$	$0.99998_{\pm 54}$	$0.001_{\pm 65}$	$0.0476_{\pm 11}$	$0.0041_{\pm 13}$
2^3	$0.869_{\pm 4.4}$	$0.731_{\pm 8.1}$	$0.99998_{\pm 4.9}$	$0.99998_{\pm 4.8}$	$0.002_{\pm 31}$	$0.0476_{\pm 13}$	$0.0045_{\pm 12}$
2^4	$0.860_{\pm 6.9}$	$0.733_{\pm 5.9}$	$0.99998_{\pm 4.7}$	$0.99998_{\pm 4.6}$	$0.002_{\pm 12}$	$0.0476_{\pm 3.6}$	$0.0036_{\pm 2.8}$
2^5	$0.855_{\pm 1.5}$	$0.736_{\pm 1.9}$	$0.99998_{\pm 9.4}$	$0.99998_{\pm 9.4}$	$0.003_{\pm 3.6}$	$0.0477_{\pm 6.7}$	$0.0035_{\pm 3.3}$
2^6	$0.853_{\pm 1.7}$	$0.737_{\pm 1.4}$	$0.99996_{\pm 8.2}$	$0.99996_{\pm 8.1}$	$0.004_{\pm 9.0}$	$0.0488_{\pm 2.1}$	$0.0035_{\pm 1.8}$
2^7	$0.853_{\pm 0.87}$	$0.739_{\pm 0.66}$	$0.99994_{\pm 7.1}$	$0.99994_{\pm 7.2}$	$0.006_{\pm 9.5}$	$0.0487_{\pm 0.99}$	$0.0035_{\pm 0.75}$
2^8	$0.855_{\pm 0.25}$	$0.742_{\pm 0.49}$	$0.99988_{\pm 7.2}$	$0.99988_{\pm 7.2}$	$0.009_{\pm 4.1}$	$0.0486_{\pm 2.9}$	$0.0035_{\pm 0.39}$
2^9	$0.858_{\pm 0.23}$	$0.749_{\pm 0.30}$	$0.99978_{\pm 7.2}$	$0.99978_{\pm 6.9}$	$0.014_{\pm 3.7}$	$0.0489_{\pm 2.4}$	$0.0035_{\pm 0.29}$
2^{10}	$0.866_{\pm 0.10}$	$0.763_{\pm 0.17}$	$0.99958_{\pm 11}$	$0.99958_{\pm 11}$	$0.022_{\pm 4.5}$	$0.0488_{\pm 0.91}$	$0.0035_{\pm 0.06}$
2^{11}	$0.869_{\pm 0.03}$	$0.768_{\pm 0.05}$	$0.99909_{\pm 7.5}$	$0.99909_{\pm 7.5}$	$0.033_{\pm 1.3}$	$0.0490_{\pm 0.47}$	$0.0035_{\pm 0.11}$

Table 8: **Scaling Experiment (InfoNCE)** (mean ± std, 5 seeds). InfoNCE matches the batch-statistic methods at low N but degrades at scale under a fixed Gaussian kernel width. The InfoNCE loss column shows the saturation directly: at $N \geq 32$, both positive- and negative-pair similarities $\exp(-\| \cdot \|^2/\sigma^2)$ underflow to zero, killing the gradient. The non-monotone R^2 across N (note the spike at N=16) is the fingerprint of this kernel-width mismatch rather than a property of the InfoNCE objective itself.

Latents	**Mixing difficulty**		**Linear identifiability**		**Orthogonality**	**InfoNCE losses**	
N	$R^2(z \to x)$ $\pm\text{std} \times 10^{-3}$	$R^2(x \to z)$ $\pm\text{std} \times 10^{-3}$	$R^2(z \to h)$ $\pm\text{std} \times 10^{-3}$	$R^2(h \to z)$ $\pm\text{std} \times 10^{-3}$	$\|\hat{Q}^\top \hat{Q} - I\|_F/\sqrt{N}$ $\pm\text{std}$	Align $\pm\text{std} \times 10^{-3}$	InfoNCE $\pm\text{std}$
2^1	$0.870_{\pm 0.92}$	$0.781_{\pm 1.4}$	$0.96030_{\pm 2.4}$	$0.95096_{\pm 1.6}$	$0.505_{\pm 0.06}$	$0.0653_{\pm 5.1}$	$4.529_{\pm 0.03}$
2^2	$0.855_{\pm 11}$	$0.730_{\pm 23}$	$0.93874_{\pm 8.1}$	$0.91087_{\pm 8.2}$	$0.908_{\pm 0.12}$	$0.0866_{\pm 5.0}$	$3.557_{\pm 0.03}$
2^3	$0.870_{\pm 2.0}$	$0.724_{\pm 12}$	$0.92503_{\pm 25}$	$0.88681_{\pm 42}$	$0.876_{\pm 0.11}$	$0.0883_{\pm 11}$	$1.811_{\pm 0.05}$
2^4	$0.858_{\pm 6.4}$	$0.736_{\pm 6.8}$	$0.99988_{\pm 0.01}$	$0.99988_{\pm 0.01}$	$0.010_{\pm 0.00}$	$0.0489_{\pm 1.2}$	$0.156_{\pm 0.01}$
2^5	$0.853_{\pm 2.9}$	$0.737_{\pm 2.8}$	$0.92206_{\pm 18}$	$0.90780_{\pm 26}$	$0.393_{\pm 0.07}$	$0.0738_{\pm 6.5}$	$0.000_{\pm 0.00}$
2^6	$0.853_{\pm 1.7}$	$0.737_{\pm 1.4}$	$0.77351_{\pm 0.82}$	$0.64849_{\pm 3.1}$	$1.756_{\pm 0.04}$	$0.2123_{\pm 3.5}$	$0.000_{\pm 0.00}$
2^7	$0.853_{\pm 0.87}$	$0.739_{\pm 0.66}$	$0.77232_{\pm 5.2}$	$0.56695_{\pm 6.6}$	$4.875_{\pm 0.08}$	$0.3803_{\pm 6.2}$	$0.000_{\pm 0.00}$
2^8	$0.855_{\pm 0.25}$	$0.742_{\pm 0.49}$	$0.83727_{\pm 0.51}$	$0.69658_{\pm 0.49}$	$3.922_{\pm 0.04}$	$0.2519_{\pm 2.5}$	$0.000_{\pm 0.00}$
2^9	$0.858_{\pm 0.23}$	$0.749_{\pm 0.30}$	$0.84131_{\pm 0.15}$	$0.70439_{\pm 0.26}$	$3.946_{\pm 0.00}$	$0.2507_{\pm 0.95}$	$0.000_{\pm 0.00}$
2^{10}	$0.866_{\pm 0.10}$	$0.763_{\pm 0.17}$	$0.84972_{\pm 0.18}$	$0.72024_{\pm 0.20}$	$4.003_{\pm 0.00}$	$0.2507_{\pm 1.1}$	$0.000_{\pm 0.00}$
2^{11}	$0.869_{\pm 0.03}$	$0.768_{\pm 0.05}$	$0.85343_{\pm 0.07}$	$0.72714_{\pm 0.10}$	$4.028_{\pm 0.00}$	$0.2514_{\pm 0.84}$	$0.000_{\pm 0.00}$

Table 9: **Scaling diagnostic at N=4096.** Recovery, orthogonality error, and final alignment loss (mean ± std where more than one seed).

Objective	Batch	Slices	Seeds	$R^2(h \to z)$	$\|\hat{Q}^\top \hat{Q} - I\|_F / \sqrt{N}$	Align
SIGReg	256	256	5	0.663$\pm$0.001	2.51$\pm$0.15	0.272
SIGReg	256	1024	1	0.663	2.27	0.270
SIGReg	1024	256	3	**0.99945**$\pm$2e-7	0.027	0.049
SIGReg	1024	1024	1	**0.99945**	0.027	0.049
VICReg	256	–	5	0.652$\pm$0.002	2.09$\pm$0.09	0.273
InfoNCE	256	–	5	0.727$\pm$0.000	4.03$\pm$0.00	0.252

Table 10: **Misspecified latent dimension** (mean over 5 seeds; std $\leq$ 0.01 unless given). PR: participation ratio of $\mathrm{Cov}(h)$. The last M with PR $\approx M$ (bold) equals the true N in both settings.

RealNVP, N=8				**Spiral,** N=2			
M	$R^2(h \to z)$	$R^2(z \to h)$	PR	M	$R^2(h \to z)$	$R^2(z \to h)$	PR
2	0.242	0.971	2.00	1	0.492	0.985	1.00
4	0.491	0.982	4.00	**2**	0.985	0.985	**2.00**
6	0.740	0.986	6.00	3	0.933	0.728	2.83
8	0.987	0.987	**7.99**	4	0.927	0.637	3.49
12	0.988	0.988	7.99	8	0.976	0.613	3.98
16	0.989	0.989	7.99	16	0.998	0.651	3.80
32	0.949$\pm$0.06	0.988	7.19$\pm$0.75				
64	0.977$\pm$0.02	0.986	7.71$\pm$0.39				

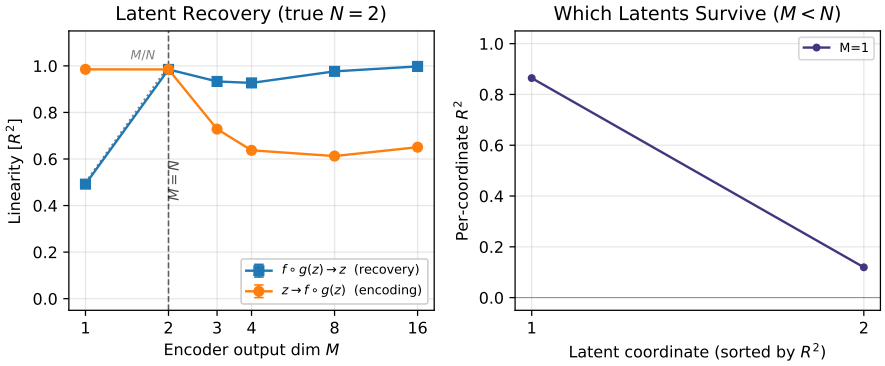


Figure 20: **Misspecified latent dimension, spiral mixing** (true N=2, 5 seeds; compare Fig. 6). **Left:** Recovery tracks M/N below N and saturates above; unlike RealNVP, encoding $R^2(z \to h)$ drops for $M > N$ because the surplus dimensions carry nonlinear functions of z. **Right:** At M=1, the encoder keeps one latent variable and discards the other.

Table 11: **Predictors.** Recovery $R^2(h \to z)$, mean ± std over 5 seeds.

Predictor	RealNVP (N=8)	Spiral (N=2)
None (Eq. 3)	0.989$\pm$0.001	0.989$\pm$0.000
Linear	0.987$\pm$0.001	0.989$\pm$0.000
MLP	0.988$\pm$0.001	0.988$\pm$0.000

Table 12: **Escape step**, the first evaluation step with $R^2(h \to z) > 0.9$ (median and range over 5 seeds; SIGReg scaling runs, batch size 256).

N	2	4	8	16	32	64	128	256	512	1024	2048
median	500	800	1500	1500	1500	1000	700	600	600	600	1000
range	200–600	600–1500	900–3500	700–2000	800–3000	700–3000	600–800	600	500–600	600	1000

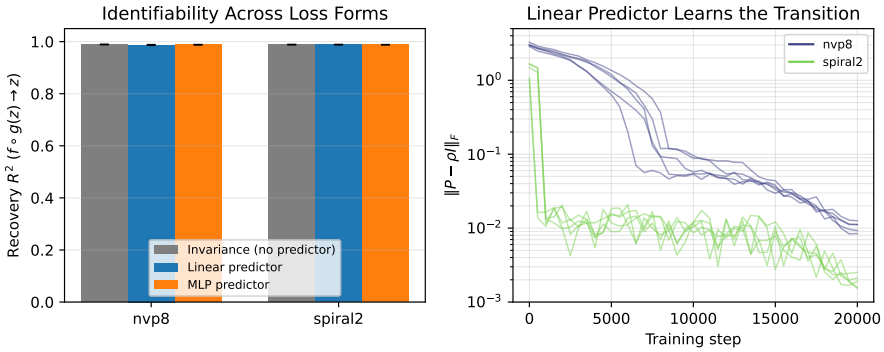


Figure 21: **Predictors. Left:** Recovery for the three arms (mean $\pm$ std, 5 seeds). **Right:** Distance of the linear predictor from the true transition operator, $\|P - \rho I\|_F$, during training (lines are seeds).

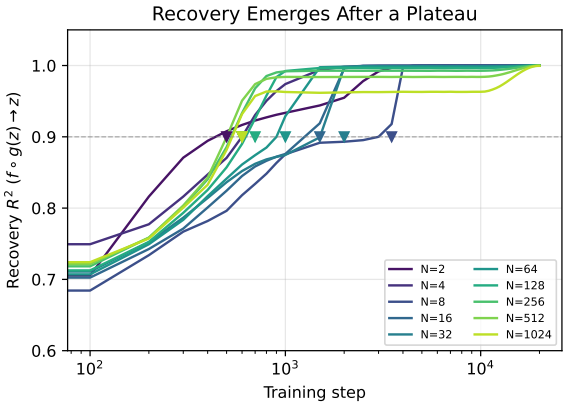


Figure 22: **Recovery during training** (SIGReg scaling runs, seed 0, $N \leq 1024$). Triangles mark the escape step, where $R^2(h \to z)$ first exceeds 0.9 (dashed line).

Table 13: **ImageNet-X linear probe accuracy, all 17 factors** (balanced accuracy, mean over 5 repetitions; chance 0.5). N: balanced sample size. Bold: above chance at a Bonferroni-corrected $p < 0.05$ over all factors and models. Factors above the line ($N > 1000$) enter the main-text mean.

		IN-1k, matched			reference		controls	
Factor	N	LeJEPA	DINO	iBOT	DINOv2	DINOv3	init	pixels
class	43500	**0.865**	**0.666**	**0.741**	**0.856**	**0.813**	**0.620**	**0.585**
shape	1276	**0.737**	0.590	**0.661**	**0.709**	**0.657**	0.552	0.575
subcategory	1118	**0.720**	**0.604**	**0.648**	**0.682**	0.614	0.554	0.502
pattern	11354	**0.674**	**0.593**	**0.618**	**0.658**	**0.653**	0.536	0.536
smaller	2738	**0.651**	**0.618**	**0.663**	**0.627**	**0.692**	0.529	0.538
partial view	1208	**0.629**	0.543	0.592	0.595	0.606	0.507	0.506
color	11642	**0.616**	**0.571**	**0.580**	**0.599**	**0.606**	0.544	0.526
pose	28404	**0.605**	**0.559**	**0.583**	**0.598**	**0.589**	0.530	0.516
background	27898	**0.594**	**0.559**	**0.570**	**0.576**	**0.578**	**0.532**	0.517
darker	230	0.809	0.714	0.751	0.643	0.620	0.700	0.520
person blocking	112	0.800	0.647	**0.771**	0.671	0.712	0.482	0.588
texture	542	**0.664**	0.619	0.643	**0.698**	0.625	0.548	0.521
larger	284	0.614	0.551	0.549	0.493	0.558	0.528	0.493
style	84	0.600	0.562	0.638	0.500	0.608	0.438	0.500
brighter	84	0.592	0.523	0.485	0.492	0.523	0.562	0.562
multiple objects	76	0.555	0.445	0.627	0.491	0.536	0.473	0.491
object blocking	152	0.552	0.565	0.435	0.574	0.513	0.491	0.474
mean ($N > 1000$)		**0.677**	0.589	0.628	0.656	0.645	0.545	0.533
mean (all 17)		**0.663**	0.584	0.621	0.615	0.618	0.537	0.526

Table 14: **ImageNet-X probe summary and Gaussianity diagnostics** in the 64-dimensional analysis space. EP: sketched Epps–Pulley distance to $\mathcal{N}(0, I)$ divided by the sample size (lower is more Gaussian). PR$/K$: participation ratio of the covariance divided by the dimension (1 for isotropic). For LeJEPA the coordinates are the projector output; for all other rows, principal components.

	Linear	MLP	Gap	EP	PR$/K$	Excess kurtosis
LeJEPA (projector)	0.663	0.684	0.021	0.0049	0.91	0.39
LeJEPA backbone (768-d)	–	–	–	0.2534	0.08	0.38
DINO	0.584	0.617	0.033	0.5330	0.99	0.04
iBOT	0.621	0.632	0.011	0.6083	0.98	0.03
DINOv2	0.615	0.645	0.030	0.4712	0.92	0.82
DINOv3	0.618	0.643	0.025	0.0033	0.94	0.65
init	0.537	0.554	0.017	0.5311	0.06	1.18
pixels	0.526	0.540	0.014	0.5874	0.14	0.99
$\mathcal{N}(0, I)$ sample	–	–	–	0.00003	1.00	0.00

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: Experiments are tailored to assess the validity of each of the theoretical results.

   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: The paper includes a limitations section which highlights the key assumptions and shortcomings of the theory.

   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: Yes, full proofs + Lean verification in appendix.

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: Yes, all experimental details necessary for a complete reproduction are included in the appendix.

   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 Secs. 6.1–6.4 and the Lean proofs is public at `https://github.com/klindtlab/lejepa-identifiability`; the RL trajectories are public on HuggingFace [9]. The Sec. 6.5 experiments are fully specified in App. H and use public data and checkpoints [24].

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: Yes, all specified and explained in appendix and code.

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: Yes, means, standard deviations and the number of seeds are reported in each experiment. The appendix includes a detailed study of the RL trajectory statistics.

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: Yes, the slurm scripts in the code indicate resources. All synthetic and RL experiments were run on single V100 GPUs and are quick to reproduce; ImageNet-X features were extracted from pretrained checkpoints on a single Colab GPU.

   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 does not violate any of the Ethics Guidelines.

   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: [Yes]

    Justification: The work is mostly theoretical but includes a brief motivation for identifiable world models that can contribute to robust, interpretable and fair decision making outside of the training data.

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 poses no such risks.

    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: All existing assets are cited and used for non-commercial research under their respective licenses: the RL trajectories [9], the ImageNet validation set with ImageNet-X annotations [24], and pretrained LeJEPA, DINO, iBOT, DINOv2 and DINOv3 checkpoints [2, 22, 87, 88, 30].

    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: [Yes]

    Justification: Code and Lean proofs are released with documentation at `https://github.com/klindtlab/lejepa-identifiability`.

    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: Neither of the two was used in this paper.

    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: No such research included.

    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 theoretical innovation of this project does not use LLMs.

    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.