REPRESENTATION SELECTION IN CAPACITY-LIMITED LEJEPA

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised Joint-Embedding Predictive Architectures (JEPAs) learn from correlated observations of the same system. Existing LeJEPA theory shows that, when the representation dimension matches the number of latent variables and all latent variables evolve at the same rate, the training objective can linearly recover the latent variables. In practical models, however, the representation dimension is often smaller, and there is still no complete theory of how limited representation capacity is allocated. We study latent-variable selection in LeJEPA under limited representation capacity and analyze how its distributional constraint affects this selection. We show that, when latent variables evolve at different rates, the model ranks both the latent variables and nonlinear functions of them. As a result, higher-order functions of slower latent variables can replace faster latent variables in the learned representation. LeJEPA's distributional constraint can rule out simple higher-order polynomial solutions, but it can still allow multiple representation coordinates to depend on the same slow latent variable. This source reuse can leave out information that remains predictable from the current observation. Based on this analysis, we introduce an additional conditional-mean constraint to reduce this information loss. Under the Gaussian model, the two constraints yield linear invariant structure, and alignment selects the leading linear subspace when the cutoff is strict. Our experiments validate the predicted selection boundary and the effect of the proposed correction.

1 INTRODUCTION

Modern learning systems often need to compress complex high-dimensional observations into lower-dimensional internal representations for prediction and downstream decision making. Because manual annotation is costly, an important class of self-supervised learning methods learns from correlated observations of the same system (Jing & Tian, 2021), such as different transformations of the same image (Chen et al., 2020; He et al., 2020) or adjacent frames in a video (Assran et al., 2025). Joint-Embedding Predictive Architectures (JEPAs) learn representations by comparing correlated observations and capturing the predictable relationships between them (Assran et al., 2023; 2022; Baevski et al., 2022). LeJEPA further constrains the representation distribution to prevent representation collapse (Balestriero & LeCun, 2025). Existing theory shows that, in an analytically tractable Gaussian model, LeJEPA can linearly recover the latent structure generating the observations when the representation dimension matches the number of latent variables and all latent variables evolve at the same rate (Klindt et al., 2026).

In practical models, however, representation capacity is often limited. When the representation dimension is smaller than the number of latent variables, the model must discard part of the latent information. This turns the problem from full recovery of the latent structure into a choice of which information to retain under limited capacity. Intuitively, LeJEPA favors representations that remain close across correlated observations, which points to a preference for more stable variables. However, the encoder can also represent functions of the latent variables, so the competing candidates are not limited to the latent variables themselves. Existing LeJEPA theory has already shown that higher-order functions of slow latent variables can replace faster latent variables, and it explicitly identifies dimensionality mismatch as an open problem (Klindt et al., 2026). Therefore, when the representation dimension is smaller than the number of latent variables, there is still no systematic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

characterization of what information is retained under LeJEPA's isotropic Gaussian distributional constraint.

The answer depends on the rates at which the latent variables evolve. Under LeJEPA's isotropic Gaussian distributional constraint, when all latent variables evolve at the same rate, a finite representation remains within the linear latent space, but the objective cannot distinguish among equally optimal m-dimensional linear latent subspaces. Different rates break this tie and favor slower latent variables. When the rate differences are large, however, nonlinear functions of slow latent variables can also enter the selected representation, and multiple representation coordinates may depend on the same slow latent variable. This occurs because LeJEPA favors representations that remain similar across correlated observations, and higher-order functions of slow latent variables can satisfy this preference better than linear functions of faster latent variables. To address this concentration of information under limited representation capacity, we further use the predictive relationship between adjacent observations to constrain the representation. This additional constraint encourages the limited representation dimensions to correspond to distinct latent directions, and together with the Gaussian distributional constraint it yields linear invariant structure, after which alignment selects the leading linear subspace when the cutoff is strict. Figure 1 summarizes this progression.

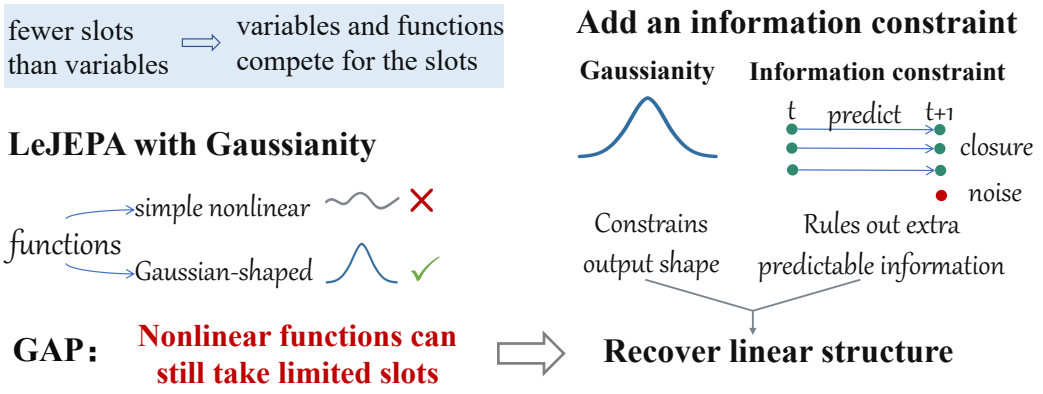

Figure 1: Finite representation capacity and linear recovery. Variables and their functions compete for the available coordinates. Gaussianity excludes non-affine polynomial competitors but permits source reuse; closure then yields linear invariant structure, and alignment selects the leading subspace.

The main contributions of this paper are as follows.

1. We characterize the selection rule under limited representation capacity. We analyze both equal-rate and different-rate settings and derive error bounds for near-optimal representations.
2. We characterize the role and limitation of LeJEPA's isotropic Gaussian distributional constraint. We prove that it rules out simple higher-order polynomial solutions, but can still allow multiple representation coordinates to depend on the same slow latent variable.
3. To resolve this remaining ambiguity, we introduce a conditional-mean constraint on the prediction residual. Together with the Gaussian distributional constraint, it recovers linear latent structure, after which alignment selects the target linear latent subspace under the stated conditions.

2 Background and Motivation

Joint-embedding methods learn predictable representations from correlated observations, using contrastive mechanisms, architectural design, or statistical constraints to avoid collapse (Chen et al., 2020; Grill et al., 2020; Zbontar et al., 2021). JEPA predicts in representation space, and LeJEPA additionally encourages an isotropic Gaussian output law (Assran et al., 2023; Balestriero & LeCun, 2025). Existing LeJEPA theory establishes linear recovery in a Gaussian latent model when representation and latent dimensions match and transition rates are equal (Klindt et al., 2026). Limited capacity requires a selection rule for the information retained.

Alignment ranks latent functions by predictability, bringing linear coordinates and nonlinear functions into the same competition (Wiskott & Sejnowski, 2002; Balestriero & LeCun, 2022; Noé & Nüske, 2013). Equal transition rates favor linear subspaces without selecting a unique one. Unequal rates rank latent variables and can favor nonlinear functions of a slow source. For example, correlations 0.8 and 0.6 give the quadratic mode of the first variable a higher score, $0.8^2 = 0.64$, than the second variable. Existing LeJEPA theory identifies this competition and leaves limited representation capacity open (Klindt et al., 2026). Extending recovery theory to this regime hinges on characterizing which functions occupy the available coordinates.

Gaussianity constrains output shape while permitting independent Gaussian coordinates to draw all their randomness from one latent variable. Thus, even mutually independent outputs can reuse one source while other latent variables remain unrepresented. Identifying their sources calls for structure beyond the marginal output law. Related identifiability results use temporal dependence, auxiliary variables, or multiple views for this purpose (Khemakhem et al., 2020; Hyvärinen & Morioka, 2016; Gresele et al., 2019). Here, we examine how the dynamics of correlated observations constrain source allocation.

Invariant subspaces in dynamical systems describe sets of observable functions whose span is preserved by prediction (Mezić, 2005; Rowley et al., 2009; Williams et al., 2015). This suggests requiring the selected representation span to remain closed under the transition operator. Concretely, after the best linear prediction within the representation, the current observation should contain no information that predicts the residual mean. This condition connects the observable pair to the dynamics of the selected functions. We study how it combines with Gaussianity to identify linear structure and when this recovery persists under approximate constraints.

3 SETUP

We study a stationary Gaussian latent state $Z \sim \mathcal{N}(0, I_d)$ and its successor after one transition

$$Z' = RZ + (I_d - R^2)^{1/2}\eta, \qquad R = \mathrm{diag}(\rho_1, \dots, \rho_d), \qquad 1 > \rho_1 \geq \cdots \geq \rho_d > 0, \tag{1}$$

where $\eta \sim \mathcal{N}(0, I_d)$ is independent of Z. Stationarity gives $Z' \sim \mathcal{N}(0, I_d)$. Each ρ_i is the autocorrelation over one transition of Z_i, so larger rates correspond to slower, more predictable latent variables.

The learner observes $X = g(Z)$ and $X' = g(Z')$. We assume that X and Z generate the same sigma-algebra up to null sets, as holds when g has a measurable inverse on a set of full measure. This observation model preserving latent information isolates the allocation performed by the representation. An encoder f produces $h = f \circ g = (h_1, \dots, h_m)^\top : \mathbb{R}^d \to \mathbb{R}^m$, with $1 \leq m < d$. In this regime of limited capacity, the objective allocates m representation coordinates across information supplied by d latent variables.

At the population level, we center and whiten the representation,

$$\mathbb{E}[h(Z)] = 0, \qquad \mathbb{E}[h(Z)h(Z)^\top] = I_m. \tag{2}$$

These constraints fix the second-order statistics. LeJEPA's full distributional constraint additionally requires $h(Z) \sim \mathcal{N}(0, I_m)$, fixing the entire marginal output law.

For a square-integrable scalar function q, define the transition operator by $(Tq)(z) = \mathbb{E}[q(Z') \mid Z = z]$, the conditional mean of the function at the next state. Write $\langle u, v \rangle = \mathbb{E}[u(Z)v(Z)]$. Under (2), the population alignment loss and retained predictability satisfy

$$L(h) := \mathbb{E}\|h(Z') - h(Z)\|_2^2 = 2m - 2C(h), \qquad C(h) := \sum_{j=1}^{m} \langle h_j, Th_j \rangle. \tag{3}$$

Thus minimizing alignment maximizes the total predictability retained by the representation coordinates.

We ask which m-dimensional function space this objective selects and when it corresponds to linear latent structure. A linear representation has the form $h(Z) = AZ$, with $A \in \mathbb{R}^{m \times d}$, and standardization requires $AA^\top = I_m$. Recovery is defined at the subspace level, up to orthogonal output rotation. With the rates ordered as in (1), write $Z_{1:m} = (Z_1, \dots, Z_m)^\top$ and $\mathcal{S}_m = \mathrm{span}\{Z_1, \dots, Z_m\}$ for the leading linear latent subspace. A strict cutoff singles out $\mathcal{S}_m$, while tied rates admit a family of equally ranked subspaces.

4 CAPACITY-LIMITED SELECTION

We first characterize how alignment allocates $m < d$ standardized coordinates across latent functions, then examine the effect of imposing a Gaussian output law.

4.1 SELECTION OVER THE TRANSITION SPECTRUM

For the Gaussian transition in Eq. (1), the normalized tensor-product Hermite functions $\{\psi_\alpha\}_{\alpha\in\mathbb{N}_0^d}$ diagonalize the transition operator.

$$T\psi_\alpha = \lambda_\alpha \psi_\alpha, \qquad \lambda_\alpha = \prod_{i=1}^{d} \rho_i^{\alpha_i}. \tag{4}$$

Here λ_α is the transition eigenvalue of the Hermite mode ψ_α. Because the Hermite modes are standardized, λ_α also equals the autocorrelation over one transition of ψ_α, and therefore serves as its predictability score. In particular, $\rho_i = \lambda_{e_i}$ is the score of the linear coordinate function Z_i, while the normalized quadratic Hermite function $H_2(Z_i) = \frac{Z_i^2-1}{\sqrt{2}}$ has score ρ_i^2. Order the nonconstant eigenpairs, with multiplicity, as $(\varphi_k, \lambda_k)_{k\geq 1}$, where $\lambda_1 \geq \lambda_2 \geq \cdots > 0$, and write $\Phi_m = (\varphi_1, \ldots, \varphi_m)^\top$. We call $\Gamma_m := \lambda_m - \lambda_{m+1}$ the spectral cutoff gap; it measures how well the m-th selected mode is separated from the next candidate.

Theorem 1 (Capacity-limited spectral selection)**.** *Under the standardization constraints in Eq.* (2)*,*

$$\sup_h C(h) = \sum_{k=1}^{m} \lambda_k, \qquad \inf_h L(h) = 2m - 2\sum_{k=1}^{m} \lambda_k.$$

If $\Gamma_m > 0$, every optimum has the form $h = Q\Phi_m$ for $Q \in O(m)$. Moreover, suppose h has excess alignment loss $L(h) - \inf L \leq \tau$. Define its off target spectral energy by

$$q := \sum_{j=1}^{m} \left\| P_{\mathrm{span}(\Phi_m)^\perp} h_j \right\|_{L^2}^2 .$$

Then

$$q \leq \frac{\tau}{2\Gamma_m}, \qquad \inf_{Q\in O(m)} \|h - Q\Phi_m\|_{L^2}^2 \leq \frac{\tau}{\Gamma_m}. \tag{5}$$

If $\lambda_m = \lambda_{m+1}$, every mode with eigenvalue strictly above the cutoff is included, while the remaining representation dimensions may be chosen within the tied eigenspace.

Proof sketch. Expand each representation coordinate as $h_j = \sum_k a_{jk}\varphi_k$ and let $p_k := \sum_j a_{jk}^2$ be the total projection energy assigned to mode φ_k. Standardization together with Bessel's inequality and Parseval's identity gives $0 \leq p_k \leq 1$ and $\sum_k p_k = m$, while $C(h) = \sum_k \lambda_k p_k$. The optimization reduces to a capacity allocation problem. The available m units of projection mass are assigned to the modes with the largest eigenvalues. Any mass q placed below a strict cutoff loses at least $\Gamma_m q$ in correlation, which yields the first bound. Within the leading eigenspace, the coefficient matrix is close to row-orthonormal; replacing it by its nearest orthogonal polar factor gives the Procrustes bound. See Appendix B for the proof and tie case.

Theorem 1 characterizes capacity allocation by ranking predictable functions of the latent variables. The leading linear latent subspace $\mathrm{span}\{Z_1, \ldots, Z_m\}$ is optimal exactly when

$$\rho_m \geq \max\{\rho_{m+1}, \rho_1^2\}, \tag{6}$$

and under strict inequality it is the unique optimal function subspace up to output rotation. Here ρ_{m+1} is the best omitted linear competitor, while ρ_1^2 is the largest possible score of any higher-order Hermite mode. Hence $H_2(Z_1)$ is selected before Z_m whenever $\rho_1^2 > \rho_m$.

If $R = \rho I_d$, all linear modes have score ρ, while modes of higher order have score at most $\rho^2 < \rho$. Every optimum is therefore an m-dimensional linear latent subspace, with equal rates leaving its orientation undetermined. Unequal rates rank the candidates, and we next examine how Gaussianity restricts this competition.

4.2 Gaussianity and Source Allocation

Exact Gaussianity removes the simplest polynomial competitors of higher order. A nonconstant polynomial of a Gaussian vector can itself be Gaussian only if it is affine, so the quadratic takeover above is infeasible under $h(Z) \sim \mathcal{N}(0, I_m)$. The remaining question concerns latent source allocation, namely which latent variables supply the randomness in the representation coordinates. Define the correlation of the leading linear latent subspace by $C_{\mathrm{lin}} := C(Z_{1:m}) = \sum_{i=1}^{m} \rho_i$. We use W_2 to denote the 2-Wasserstein distance between probability laws.

Theorem 2 (Gaussian latent source reuse)**.** *Fix $2 \le m < d$ and $1 > \rho_2 \ge \cdots \ge \rho_d > 0$.*

(a) There exist a measurable map $H : \mathbb{R} \to \mathbb{R}^m$, depending only on Z_1, and a threshold $\rho_* < 1$ such that for every $\rho_1 > \rho_*$,

$$H(Z_1) \sim \mathcal{N}(0, I_m), \qquad C(H) > C_{\mathrm{lin}}.$$

Thus all m Gaussian representation coordinates can reuse one sufficiently slow latent variable while outperforming the leading linear baseline.

(b) Fix a model and an exact latent source reuse map with advantage $a_0 := C(H) - C_{\mathrm{lin}} > 0$. For every $\delta > 0$, there exists a finite width, globally Lipschitz rectified linear unit (ReLU) network $h_\delta : \mathbb{R} \to \mathbb{R}^m$ with one hidden layer such that

$$\mathbb{E}h_\delta = 0, \qquad \mathrm{Cov}(h_\delta) = I_m, \qquad W_2(\mathrm{Law}(h_\delta(Z_1)), \mathcal{N}(0, I_m)) \le \delta,$$

$$C(h_\delta) \ge C_{\mathrm{lin}} + \frac{a_0}{2}, \qquad \sum_{j=1}^{m} \mathrm{Var}(\mathbb{E}[h_{\delta,j} \mid Z_1]) = m.$$

The final equality means that all output variance is still supplied by the single latent variable Z_1.

Proof sketch. Let $U = \Phi(Z_1)$, which is uniform on $(0, 1)$. Split disjoint binary digits of U among m new uniform variables and map each through Φ^{-1}. The resulting coordinates are independent standard Gaussians, although every coordinate remains measurable with respect to Z_1. Their Hermite expansions imply

$$C(H) \to m \qquad \text{as} \qquad \rho_1 \to 1^-,$$

whereas $C_{\mathrm{lin}} \to 1 + \sum_{i=2}^{m} \rho_i < m$. Hence the latent source reuse representation has a strictly positive advantage when Z_1 is sufficiently persistent. For part (b), approximate the fixed measurable map by a continuous piecewise linear function, represent that function with one ReLU layer, and apply population centering and whitening. The resulting L^2 approximation controls W_2 under the explicit coupling and preserves a positive correlation margin. The construction and approximation metric are described in Appendix C.

Theorem 2 shows that independent Gaussian outputs can share a single latent source, motivating a constraint on how the selected functions evolve across observations.

5 Transition Closure and Linear Recovery

We use the relation between correlated observations to constrain the dynamics of the selected representation and establish exact and stable linear recovery.

5.1 Zero Conditional Mean and Exact Recovery

For a standardized representation, define $B_h := \mathbb{E}[h(Z')h(Z)^\top]$ and $R_h := h(Z') - B_h h(Z)$. Because $\mathrm{Cov}(h(Z)) = I_m$, $B_h h(Z)$ is the population least-squares linear predictor of the next representation from the current representation. We impose the zero conditional mean (ZCM) condition

$$\mathbb{E}[R_h \mid X] = 0. \tag{7}$$

It states that, after removing the best linear prediction available within the representation space, the full current observation contains no remaining information that can systematically predict the residual mean.

Condition (7) is stated on the observation, so we read it on the operator the observation defines. Let $(Kq)(x) := \mathbb{E}[q(X') \mid X = x]$ on $L^2(P_X)$; reversibility of (Z, Z') makes (X, X') exchangeable,

so K is a self-adjoint contraction. With $V_h := \mathrm{span}\{h_1, \ldots, h_m\} \subset L^2(P_X)$, P_h the orthogonal projector onto V_h, $\bar{R}_h := \mathbb{E}[R_h \mid X] = Kh - B_h h$ and $\epsilon_h := \|\bar{R}_h\|^2_{L^2}$, the defect ϵ_h measures the predictable residual energy not captured by the linear predictor within the representation. We write $\|\cdot\|_{HS}$ for the Hilbert-Schmidt norm and $[K, P_h] := KP_h - P_h K$.

Proposition 1 (Transition-closure defect)**.**

$$\epsilon_h = \|(I - P_h)KP_h\|^2_{HS} = \frac{1}{2}\|[K, P_h]\|^2_{HS}. \tag{8}$$

Consequently, ZCM holds exactly if and only if $KV_h \subseteq V_h$. Under the assumption of Section 3 that observations preserve latent information, K is unitarily equivalent to T, so the statement reads on the latent transition operator.

Proof sketch. Standardization makes $h_1, \ldots, h_m$ orthonormal, and $(B_h)_{jk} = \langle Kh_j, h_k \rangle$ gives $P_h K h = B_h h$, so $\bar{R}_h = (I - P_h)Kh$ is exactly the transition energy leaking from V_h into $V_h^\perp$. Block form of the self-adjoint K on $V_h \oplus V_h^\perp$ gives the commutator identity; Appendix D.1 also records the compression relating K to T.

Equation (7) uses only the observable pair (X, X') and requires R_h to be orthogonal in expectation to every square-integrable function of X. It therefore enforces transition closure whether or not the observation determines the latent state; full observability is what makes the closed subspace closed for the latent dynamics rather than only for their observable compression. The defect is also directly trainable. Over square-integrable critics c, the variational identity $\sup_c\{2\mathbb{E}\langle R_h, c(X)\rangle - \mathbb{E}\|c(X)\|^2\} = \epsilon_h$ turns (7) into a penalty that a restricted critic class can only underestimate, and the experiments below train with this penalty rather than only measure it afterwards. Combined with Gaussianity, it yields the following recovery result.

Theorem 3 (Exact linear recovery and target selection)**.** *Assume the model and standardization conditions of Section 3.*

(i) Structural identification. If $h(Z) \sim \mathcal{N}(0, I_m)$ and $\epsilon_h = 0$, then there exists $A \in \mathbb{R}^{m \times d}$ such that, almost surely (a.s.),

$$h(Z) = AZ, \qquad AA^\top = I_m, \qquad AR = B_h A. \tag{9}$$

Equivalently, $\mathrm{row}(A)$ is an R-invariant linear latent subspace. Conversely, every row-orthonormal linear map whose row space is R-invariant satisfies exact Gaussianity and ZCM.

(ii) Target selection. If, in addition, $\rho_m > \rho_{m+1}$, then minimizing alignment over this exact feasible family gives precisely $h(Z) = QZ_{1:m}$ for $Q \in O(m)$. If there is a tie at the linear cutoff, the corresponding tied family of invariant subspaces remains optimal.

Proof sketch. Exact ZCM gives $Th = B_h h$. Under the Gaussian transition considered here, B_h is symmetric positive definite for every fixed standardized representation. An orthogonal rotation of the outputs therefore diagonalizes B_h and makes each rotated coordinate a transition eigenfunction with a positive eigenvalue. A fixed positive transition eigenvalue contains only finitely many Hermite modes, so each rotated coordinate is a finite polynomial. Exact Gaussianity then invokes polynomial Gaussian rigidity, forcing every coordinate to be affine. Centering removes the constant term and yields $h = AZ$. Substitution gives $AR = B_h A$, hence row space invariance. Once the feasible family has been reduced to row-orthonormal invariant linear subspaces, alignment becomes the finite-dimensional problem of maximizing $\mathrm{tr}(ARA^\top)$, which selects the leading m latent directions under a strict cutoff. Appendix D gives the proof and a fourth moment strengthening.

Theorem 3 separates structural identification by Gaussianity and closure from target selection by alignment. Training satisfies these constraints approximately, motivating the stability question below.

5.2 Stable Recovery under Approximate Constraints

To characterize stable recovery, we track nonlinear energy, constraint errors, and the predictability retained by every representation direction.

Define

$$M_h := \mathbb{E}[h(Z)Z^\top], \qquad s(h) := m - \|M_h\|^2_F = \|P_{\geq 2}h\|^2_{L^2},$$

where $P_{\geq 2}$ projects onto Hermite modes of total degree at least two. Thus $s(h)$ measures nonlinear representation energy and vanishes exactly when a standardized representation lies entirely in the linear latent space. We write $D_{\mathrm{lin}}(h) := \inf_{A:AA^\top = I_m} \|h - AZ\|_{L^2}^2$ for the squared distance to the nearest standardized linear representation. We measure Gaussianity error by $W_2(\mathrm{Law}(h(Z)), \mathcal{N}(0, I_m)) \leq \delta$; Appendix E.3 gives the same bound with this term replaced by the fourth-moment discrepancy $\Delta_4 := \|\mathbb{E}[h^{\otimes 4}] - \mathbb{E}[G^{\otimes 4}]\|_F$, $G \sim \mathcal{N}(0, I_m)$, which is the form the experiments measure. A uniform predictability floor is written

$$B_h \succeq bI_m, \qquad b > 0,$$

meaning that every representation direction retains a nonvanishing amount of predictability over one transition; equivalently $b_{\min}(h) := \lambda_{\min}(B_h) \geq b$.

Theorem 4 (Stable recovery under approximate constraints)**.** *(a) Necessity of nondegeneracy. For every fixed model in Eq.* (1)*, there exists a standardized sequence h_n such that*

$$h_n(Z) \sim \mathcal{N}(0, I_m), \qquad \epsilon_{h_n} \to 0, \qquad B_{h_n} \to 0, \qquad s(h_n) \to m, \tag{10}$$

and $D_{\mathrm{lin}}(h_n) \to 2m$. Thus exact Gaussianity and vanishing ZCM error do not imply near linearity when predictability is allowed to vanish.

(b) Stable near linearity. Fix d, m, R and $b > 0$. If a standardized representation satisfies

$$B_h \succeq bI_m, \qquad \epsilon_h \leq \epsilon < b^2/8, \qquad W_2(\mathrm{Law}(h(Z)), \mathcal{N}(0, I_m)) \leq \delta, \tag{11}$$

then

$$s(h) \leq S_G(\delta, \epsilon; d, m, R, b), \qquad D_{\mathrm{lin}}(h) \leq 2s(h), \tag{12}$$

where the explicit function S_G is given in Appendix E, $S_G \to 0$ as $(\delta, \epsilon) \to (0, 0)$, and, for fixed model parameters, $s(h) = O(\delta^{1/2} + \epsilon^{1/4})$.

Proof sketch. For part (a), repeatedly recode one uniformized slow latent variable at increasing frequency. The output law remains exactly Gaussian, but the resulting functions become weakly orthogonal to every fixed low-frequency component. Compactness of T then makes their predictable part vanish, so both B_h and the ZCM residual go to zero while the representation loses all first-order linear energy. For part (b), diagonalize B_h. The lower bound $B_h \succeq bI_m$ keeps the resulting approximate transition eigenvalues away from the zero accumulation point, so small ZCM error confines the outputs to a finite spectral region. Finite polynomial truncation controls the higher moments in this region. Wasserstein proximity then transfers Gaussian fourth moments to the truncated coordinates, and quantitative fourth-order rigidity bounds their nonlinear energy. Projection onto the linear latent space yields a nearby row-orthonormal linear map. See Appendix E for the constants and proof.

The uniform predictability floor is what makes small distributional and closure errors informative about linear structure. Part (a) establishes its necessity, and part (b) gives the quantitative recovery guarantee. Alignment itself supplies the floor. Compression of the self-adjoint transition operator gives $\beta_j \leq \lambda_j$ for the eigenvalues of B_h, and $C(h) = \sum_j \beta_j$, so $b_{\min}(h) \geq C(h) - \sum_{j<m} \lambda_j$; maximizing alignment raises this bound. Appendix E.6 records it with a further sufficient condition.

6 Experiments

The main training experiments use the shared setting $\rho = (0.98, 0.20, 0.10, 0.05)$ with $d = 4$ and $m = 2$, under which the theoretical target is the leading two-dimensional linear latent subspace. Its total one-step predictability is $C_{\mathrm{lin}} = 1.18$, and the weaker target direction has one-step predictability 0.20. We measure recovery by D_{target}, the squared distance to the target subspace after optimal orthogonal alignment, which ranges over $[0, 4]$ when $m = 2$. Full protocols and per-seed results are reported in Appendix F.

6.1 Spectral Selection

Theorem 1 predicts that under limited capacity alignment selects the most predictable functions across the full transition spectrum. We test this with $d = 3$ and $m = 2$. We fix $\rho_1 = 0.8$ and $\rho_3 = 0.3$, so the strongest higher-order competitor has predictability $\rho_1^2 = 0.64$, and sweep ρ_2 across this boundary. On the left side of the boundary the second learned direction consistently selects $H_2(Z_1)$; once ρ_2 exceeds 0.64 it switches to the linear variable Z_2 (Figure 2A). All strictly separated sweep points agree with the theoretical ordering.

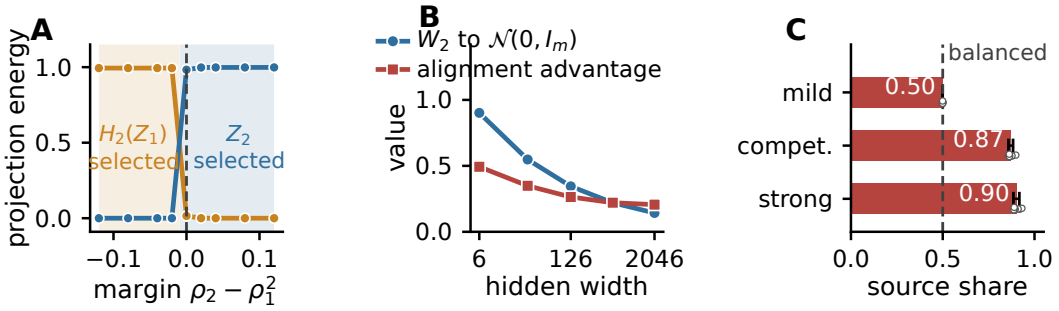

Figure 2: Selection and allocation. (A) Projection energy of the two candidates for the second representation direction across the spectral margin, under whitening alone; the dashed line marks the theoretical tie. (B) A one-hidden-layer ReLU network approaching the single-source construction as width grows. (C) Share of representation variance explained by a single latent variable on the mild, competitive and strongly competitive spectra; bars are means over eight initializations, whiskers one standard deviation, and the dashed line the balanced two-source value.

6.2 Source Allocation

Adding Gaussian regularization to the boundary experiment keeps the learned representation on the linear directions across the sweep, showing that the Gaussian constraint removes this simple higher-order competitor.

We approximate the single-source Gaussian construction of Theorem 2 with one-hidden-layer ReLU networks of output dimension two, bounding the W_2 distance between the two output laws. As width increases from 6 to 2046 this distance decreases from about 0.90 to 0.14, while the alignment advantage over the leading linear baseline remains positive throughout (Figure 2B). Finite ReLU networks reach the construction.

We next ask whether the same ambiguity appears under ordinary Gaussian-regularized training, using the same MLP on three spectra that widen the advantage of the slowest source Z_1: mild $(0.80, 0.52, 0.30)$, competitive $(0.98, 0.25, 0.02)$, and strongly competitive $(0.98, 0.10, 0.05)$. A balanced two-source target puts this share near one half. Figure 2C shows about 50% in the mild setting, rising to about 87% and 90% in the competitive and strongly competitive settings. As the predictability advantage of the slowest source grows, representation capacity becomes increasingly concentrated on Z_1.

6.3 Linear Recovery

Adding transition closure to the Gaussian constraint should recover the leading linear latent subspace (Theorem 3). We compare Gaussian-only and Gaussian+closure training under the shared setting, with the same data and encoder in both cases, against the target $\mathrm{span}\{Z_1, Z_2\}$.

The fraction of representation variation explained by a single latent variable drops from about 87% to about 50%, showing that capacity is redistributed across the two leading sources, and the distance to the target falls from about 2.36 to nearly zero. The recovered total one-step predictability is close to C_{lin} and $b_{\min}$ is close to 0.20, so the recovered subspace matches the target both geometrically and dynamically.

The same pattern appears with pixel observations. Adding closure reduces the target distance from about 1.48 to about 0.15, while more than 90% of the representation energy lies in linear latent directions. Together, the two comparisons in Figure 3A support the two steps in Theorem 3: Gaussianity with transition closure recovers the linear dynamical structure, and alignment selects the most predictable linear directions within it, and the same conclusion holds across the range of Gaussian weights we sweep.

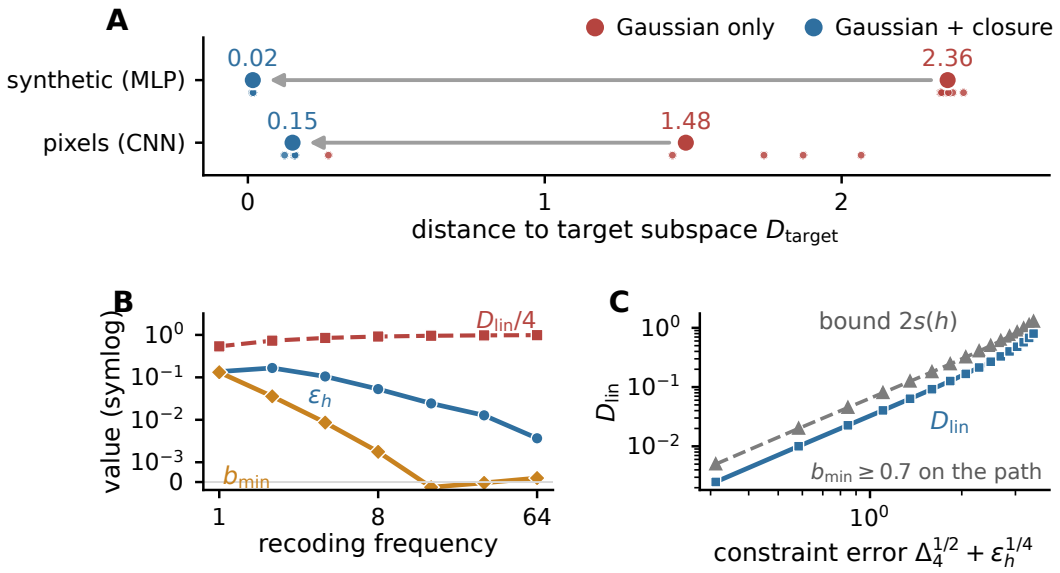

Figure 3: Recovery and its stability. (A) Distance to the target subspace under the Gaussian objective and after adding closure, on the synthetic model (eight initializations) and on pixels (five); small dots are initializations and large dots their means. (B) As the recoding frequency grows, ϵ_h and $b_{\min}$ fall to within noise of zero on a symmetric logarithmic axis, while D_{lin} approaches its maximum $2m = 4$. (C) Along a path with $b_{\min} \geq 0.7$, D_{lin} falls with the constraint error and stays below the bound $2s(h)$.

6.4 Stable Recovery

Training satisfies the two constraints only approximately, the regime covered by Theorem 4. We first illustrate Theorem 4(a) with a sequence of representations that remains exactly Gaussian while encoding the same latent source at progressively higher frequencies, which makes the representation less predictable from one step to the next. In Figure 3B the closure error eventually approaches zero as one-step predictability also approaches zero, while D_{lin} increases toward its maximum $2m = 4$. Exact Gaussianity and a small closure error can therefore coexist with a representation that is far from linear.

With a positive predictability floor the picture changes, and part (b) of the same theorem applies. We use $\rho = (0.9, 0.7, 0.3)$ and start from a two-dimensional representation containing a quadratic nonlinear component, which is gradually reduced; along the entire path $b_{\min} \geq 0.7$. Figure 3C shows that as Δ_4 and the closure error decrease, D_{lin} approaches zero while remaining within the theoretical bound.

Theorem 4 extends Theorem 3 to this approximate regime, and the recovery we observe persists under several departures from the exact model assumptions. The closure term costs a constant factor in compute, and giving that factor to the Gaussian objective alone does not recover the target.

7 Conclusion

With fewer representation coordinates than latent variables, alignment allocates capacity across the full transition spectrum. Nonlinear functions of a slow source can therefore compete with faster linear directions. Gaussianity excludes polynomial competitors while allowing multiple coordinates to reuse one source. Under the Gaussian model, the conditional mean constraint supplies transition closure and, together with Gaussianity, yields linear invariant structure. Alignment selects the leading linear subspace under a strict cutoff, and a uniform predictability floor makes near linearity stable under approximate constraints. Neural training reproduces the spectral boundary and recovers the linear structure once the closure term is included, and controlled constructions isolate source reuse and nondegenerate predictability.

AI USE STATEMENT

AI tools were used solely to assist with language polishing and code organization. We developed the research content and wrote the manuscript, and we take full responsibility for its content and accuracy.

REPRODUCIBILITY STATEMENT

Every theoretical claim is stated together with its assumptions in Sections 3–5 and proved in Appendices A–E. Appendix F gives the data-generating processes, encoder architectures, optimizer settings and training budgets for every experiment, defines each reported diagnostic, and reports per-seed results for all comparisons summarized in Section 6.

REFERENCES

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked Siamese networks for label-efficient learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pp. 456–473. Springer, 2022. doi: 10.1007/978-3-031-19821-2_26. URL https://doi.org/10.1007/978-3-031-19821-2_26.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 15619–15629. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://doi.org/10.1109/CVPR52729.2023.01499.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *CoRR*, abs/2506.09985, 2025. doi: 10.48550/ARXIV.2506.09985. URL https://doi.org/10.48550/arXiv.2506.09985.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/aa56c74513a5e35768a11f4e82dd7ffb-Abstract-Conference.html.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *CoRR*, abs/2511.08544, 2025. doi: 10.48550/ARXIV.2511.08544. URL https://doi.org/10.48550/arXiv.2511.08544.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *The Tenth International Conference on Learning*

Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete Rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In Amir Globerson and Ricardo Silva (eds.), *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pp. 217–227. AUAI Press, 2019. URL http://proceedings.mlr.press/v115/gresele20a.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Hermanni Hälvä and Aapo Hyvärinen. Hidden markov nonlinear ICA: unsupervised learning from nonstationary time series. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pp. 939–948. AUAI Press, 2020. URL http://proceedings.mlr.press/v124/halva20a.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3765–3773, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/d305281faf947ca7acade9ad5c8c818c-Abstract.html.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):4037–4058, 2021. doi: 10.1109/TPAMI.2020.2992393. URL https://doi.org/10.1109/TPAMI.2020.2992393.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020. URL http://proceedings.mlr.press/v108/khemakhem20a.html.

David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan M. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse

coding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=EbIDjBynYJ8.

David A. Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model? *CoRR*, abs/2605.26379, 2026. doi: 10.48550/ARXIV.2605.26379. URL https://doi.org/10.48550/arXiv.2605.26379.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo K. Rohde. Generalized sliced wasserstein distances. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 261–272, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f0935e4cd5920aa6c7c996a5ee53a70f-Abstract.html.

Milan Korda and Igor Mezić. On convergence of extended dynamic mode decomposition to the koopman operator. *J. Nonlinear Sci.*, 28(2):687–710, 2018. doi: 10.1007/S00332-017-9423-0. URL https://doi.org/10.1007/s00332-017-9423-0.

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 Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. URL http://proceedings.mlr.press/v97/locatello19a.html.

Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1-3):309–325, August 2005. ISSN 1573-269X. doi: 10.1007/s11071-005-2824-x.

Frank Noé and Feliks Nüske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Model. Simul.*, 11(2):635–655, 2013. doi: 10.1137/110858616. URL https://doi.org/10.1137/110858616.

Clarence W. Rowley, Igor Mezić, Shervin Bagheri, Philipp Schlatter, and Dan S. Henningson. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, November 2009. ISSN 1469-7645. doi: 10.1017/S0022112009992059.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020. URL http://proceedings.mlr.press/v119/wang20k.html.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *J. Nonlinear Sci.*, 25(6):1307–1346, 2015. doi: 10.1007/S00332-015-9258-5. URL https://doi.org/10.1007/s00332-015-9258-5.

Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Comput.*, 14(4):715–770, 2002. doi: 10.1162/089976602317318938. URL https://doi.org/10.1162/089976602317318938.

Hao Wu and Frank Noé. Variational approach for learning markov processes from time series data. *J. Nonlinear Sci.*, 30(1):23–66, 2020. doi: 10.1007/S00332-019-09567-Y. URL https://doi.org/10.1007/s00332-019-09567-y.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL http://proceedings.mlr.press/v139/zbontar21a.html.

A Mathematical Foundations

A.1 Hermite Basis

Let He_k be the probabilists' Hermite polynomial defined by

$$e^{tx-t^2/2} = \sum_{k\geq 0} \mathrm{He}_k(x) t^k/k!.$$

For $G \sim \mathcal{N}(0,1)$, comparison of coefficients in $\mathbb{E}[e^{sG-s^2/2}e^{tG-t^2/2}] = e^{st}$ gives $\mathbb{E}[\mathrm{He}_p(G)\,\mathrm{He}_q(G)] = p!\,\mathbf{1}\{p=q\}$. Hence

$$\psi_\alpha(z) = \prod_{i=1}^{d} \mathrm{He}_{\alpha_i}(z_i)/\sqrt{\alpha_i!}$$

is orthonormal in $L^2(\gamma_d)$, where $\gamma_d = \mathcal{N}(0, I_d)$. Completeness follows from Gaussian polynomial density. One direct argument observes that if $u \in L^2(\gamma_d)$ is orthogonal to all polynomials, its Gaussian weighted Laplace transform is entire with every Taylor coefficient equal to zero. Fourier uniqueness then implies $u = 0$.

A.2 Transition Spectrum

For one coordinate $Y = \rho z + \sqrt{1-\rho^2}\eta$,

$$\mathbb{E}[e^{tY-t^2/2} \mid z] = e^{\rho t z - \rho^2 t^2/2}.$$

Coefficient comparison yields $T\,\mathrm{He}_k = \rho^k\,\mathrm{He}_k$. Independence across coordinates gives Eq. (4). Writing $w_i = -\log\rho_i$ and $E_\alpha = \sum_i w_i\alpha_i$, we have $\lambda_\alpha = e^{-E_\alpha}$. Since $E_\alpha \geq (-\log\rho_1)|\alpha|$, only finitely many nonconstant eigenvalues exceed any fixed positive threshold. Thus T is compact, self-adjoint, and strictly positive on the centered subspace, while its eigenvalues accumulate at zero.

The one-dimensional Hermite product identity

$$\mathrm{He}_p\,\mathrm{He}_q = \sum_{r=0}^{\min(p,q)} r!\binom{p}{r}\binom{q}{r}\mathrm{He}_{p+q-2r}$$

has nonnegative coefficients. Taking tensor products shows that if $\psi_\alpha\psi_\beta = \sum_\gamma M_{\alpha\beta\gamma}\psi_\gamma$, then $M_{\alpha\beta\gamma} \geq 0$, and every nonzero term satisfies $E_\gamma \leq E_\alpha + E_\beta$. This weighted energy subadditivity is used in the fourth-order rigidity proof.

A.3 Function Spaces and the Weighted Generator

The positive weighted Ornstein-Uhlenbeck generator is

$$\mathcal{A} = -\sum_{i=1}^{d} w_i(\partial_{ii} - z_i\partial_i), \qquad \mathcal{A}\psi_\alpha = E_\alpha\psi_\alpha, \quad T = e^{-\mathcal{A}}.$$

Its operator domain is $D(\mathcal{A}) = \{u = \sum c_\alpha\psi_\alpha : \sum E_\alpha^2 c_\alpha^2 < \infty\}$ and its quadratic form domain is $D(\mathcal{E}) = \{u : \sum E_\alpha c_\alpha^2 < \infty\} = W^{1,2}(\gamma_d)$. On this form domain,

$$\mathcal{E}(u,u) = \sum_i w_i\|\partial_i u\|_2^2.$$

Finite Hermite truncations converge in the form norm. All higher derivative and product manipulations used in the fourth-order lemma below are applied only after a finite spectral truncation, so ordinary Gaussian integration by parts is justified term by term.

A.4 Projection Identities

For a standardized h, its coordinates are orthonormal in L^2. Since $(B_h)_{jk} = \langle Th_j, h_k \rangle$,

$$P_h T h = B_h h, \qquad \bar{R}_h = (I - P_h) T h.$$

Therefore

$$\epsilon_h = \|(I - P_h) T P_h\|_{\mathrm{HS}}^2 = \mathrm{tr}(P_h T^2) - \mathrm{tr}[(P_h T P_h)^2].$$

The full projection and commutator calculation is given in Appendix D.1. Finally, for $v \neq 0$, $u = v^\top h$ is centered and

$$v^\top B_h v = \langle u, Tu \rangle = \sum_{\alpha \neq 0} \lambda_\alpha |\langle u, \psi_\alpha \rangle|^2 > 0.$$

Thus every fixed standardized representation has $0 \prec B_h \preceq \rho_1 I$. Stable recovery over a family further uses the uniform lower bound $B_h \succeq bI$ used over a class in Theorem 4.

B Spectral Selection

B.1 Proof of Theorem 1

Expand $h_j = \sum_{k \geq 1} a_{jk} \varphi_k$ and set $p_k = \sum_j a_{jk}^2$. The quantity p_k is the total representation energy assigned to mode φ_k. Bessel gives $p_k \leq 1$, nonnegativity is immediate, and Parseval gives $\sum_k p_k = m$. Hence

$$C(h) = \sum_k \lambda_k p_k \leq \sum_{k=1}^{m} \lambda_k.$$

The upper bound is attained by $h = \Phi_m$. Let $q = \sum_{k>m} p_k = \sum_{k \leq m} (1 - p_k)$. Then

$$\begin{aligned} \sum_{k \leq m} \lambda_k - C(h) &= \sum_{k \leq m} \lambda_k (1 - p_k) - \sum_{k > m} \lambda_k p_k \\ &\geq (\lambda_m - \lambda_{m+1}) q. \end{aligned}$$

If the cutoff gap is positive, optimality gives $q = 0$ and the representation span equals $\mathrm{span}(\Phi_m)$. Standardization then forces the coefficient matrix to be orthogonal.

For a near-optimal h, $L(h) - L^* = 2(C^* - C)$ and therefore $q \leq \tau / (2\Gamma_m)$. Write $h = M\Phi_m + r$ with $r \perp \mathrm{span}(\Phi_m)$, and whitening gives $MM^\top + \mathbb{E}[rr^\top] = I_m$ and $\|r\|_2^2 = q$. If $M = U \,\mathrm{diag}(s_i) V^\top$, take $Q = UV^\top$. Since $0 \leq s_i \leq 1$,

$$\|M - Q\|_F^2 = \sum_i (1 - s_i)^2 \leq \sum_i (1 - s_i^2) = q.$$

Orthogonality of the two components yields $\|h - Q\Phi_m\|_2^2 \leq 2q$, completing Eq. (5).

B.2 Ties and the Procrustes Identity

When $\lambda_m = \lambda_{m+1} = c$, rewrite the objective deficit as

$$\sum_{\lambda_k > c} (\lambda_k - c)(1 - p_k) + \sum_{\lambda_k < c} (c - \lambda_k) p_k.$$

Each term is nonnegative. Equality forces all modes above c into the representation span, excludes all modes below c, and leaves the remaining dimension inside the finite eigenspace at c. This proves the stated family of tied optima.

Lemma B.1 (Procrustes identity)**.** *For any row-orthonormal target vector Ξ of m functions, set $N = \mathbb{E}[h\Xi^\top]$. Expanding the squared distance and maximizing the cross term over $Q \in O(m)$ gives*

$$\inf_Q \|h - Q\Xi\|_2^2 = 2m - 2\|N\|_* \leq 2m.$$

Here $\|N\|_*$ is the nuclear norm, the sum of singular values.

Proof. Expand the square and maximize $\mathrm{tr}(Q^\top N)$ over $Q \in O(m)$. The singular value decomposition gives the maximum $\|N\|_*$. □

B.3 Linear Boundary and Capacity Interpretation

Every nonlinear Hermite mode has total degree at least two and therefore eigenvalue at most ρ_1^2, attained by $H_2(Z_1)$. Outside the first m linear directions, the largest competitor is $\max\{\rho_{m+1}, \rho_1^2\}$. This proves Eq. (6) with a non-strict inequality for existence of the specified linear optimum and a strict inequality for uniqueness of its coordinate space.

Remark B.2 (Meaning of representation capacity). Capacity counts the m mutually orthogonal function coordinates whose predictability contributes simultaneously to the objective. Whitening fixes their total spectral mass at m, so the allocation problem is defined in the space of functions.

The same question – which of the available directions a limited number of coordinates actually covers – has been studied empirically for contrastive objectives, where the learned representation can settle into a strict subspace of the available dimensions (Jing et al., 2022). Theorem 1 gives the counterpart of that observation for a transition spectrum: the directions that survive are the most predictable functions, and the cutoff gap Γ_m controls how sharply they are separated from the rest.

C Gaussian Source Reuse

C.1 Polynomial Gaussian Rigidity

Lemma C.1 (Polynomial Gaussian rigidity)**.** *A polynomial of a nondegenerate Gaussian vector has a Gaussian distribution only if it is affine.*

Proof. Let a real polynomial $P = P_r + P_{<r}$ have highest degree $r \geq 2$. Choose a sphere direction where P_r is nonzero and a sign σ such that σP_r is positive on a cone. Along this cone and for sufficiently large radius s, $\sigma P(su) \geq cs^r/2$. Taking $t = \sigma t_0$ with $t_0 > 0$, the moment generating integral has the lower bound

$$\mathbb{E}e^{tP(Z)} \geq C \int_{s_0}^{\infty} \exp\{t_0 c s^r/2 - s^2/2\} s^{d-1}\, ds.$$

It diverges for every $t_0 > 0$ when $r > 2$, and for sufficiently large t_0 when $r = 2$. A Gaussian random variable has a finite moment-generating function (MGF) for every real t, so P must have degree at most one. □

C.2 Exact Source Reuse

We fix the map H throughout the variation in ρ_1. Let $U = \Phi(Z_1)$. Outside the dyadic set, write its binary digits as $b_k(U)$. Split the digits cyclically across m subsequences,

$$U_j = \sum_{\ell \geq 1} 2^{-\ell} b_{m(\ell-1)+j}(U), \qquad H_j = \Phi^{-1}(U_j).$$

Every finite set of the original bits is independent Bernoulli$(1/2)$, so the U_j are independent uniforms and $H(Z_1) \sim \mathcal{N}(0, I_m)$. Input dyadic points and the events $U_j \in \{0, 1\}$ have measure zero and arbitrary finite assignments there preserve the law.

For each fixed coordinate, write $H_j = \sum_{k \geq 1} a_{jk} \psi_k(Z_1)$ with $\sum_k a_{jk}^2 = 1$. Then

$$C_{\rho_1}(H) = \sum_{j=1}^{m} \sum_{k \geq 1} \rho_1^k a_{jk}^2 \to m \quad \text{as } \rho_1 \uparrow 1$$

by dominated convergence. The leading linear baseline tends to $1 + \sum_{i=2}^{m} \rho_i < m$, so the same fixed digit splitting map has a strict advantage for every ρ_1 above some threshold below one.

The construction above is a statement about what a constraint on the output law leaves open. Such constraints are a standard ingredient of self-supervised objectives and do shape the geometry of the representation (Wang & Isola, 2020), but fixing the marginal law does not by itself pin down the latent variables; without further structure the unsupervised recovery of independent factors is not identifiable at all (Locatello et al., 2019). Theorem 2 locates the specific gap that remains once the output law is exactly Gaussian.

C.3 Finite-ReLU Approximation

Fix any exact source reuse map H with advantage $a_0 > 0$. Approximate H in L^2 by a bounded step function on a finite partition of the scalar input. Replace every jump by a narrow linear ramp. A scalar-input vector-valued continuous piecewise linear function has an exact ReLU representation with one hidden layer of the form

$$F(z) = c + az + \sum_{r=1}^{J} v_r (z - t_r)_+.$$

Choosing the partition fine and the ramps narrow makes $\|F - H\|_2$ arbitrarily small.

Center F and let Σ_F be its covariance. If $\|F - H\|_2 \le r$, then $\|\Sigma_F - I\|_{\mathrm{op}} \le 2r + r^2$. Here $\|\cdot\|_{\mathrm{op}}$ is the operator norm. For small enough r, define

$$h_\delta = \Sigma_F^{-1/2}(F - \mathbb{E}F).$$

This changes only the final affine layer and preserves finite width and global Lipschitzness. Matrix square-root continuity gives $\|h_\delta - H\|_2 \to 0$. Coupling the two outputs on the same Z_1 yields $W_2 \le \|h_\delta - H\|_2$. Since T is a contraction and both representations have bounded L^2 norm, $|C(h_\delta) - C(H)| \to 0$. Taking the approximation sufficiently accurate establishes the $a_0/2$ margin in Theorem 2. Every coordinate remains measurable with respect to Z_1, so $\mathbb{E}[h_{\delta,j} \mid Z_1] = h_{\delta,j}$. Population whitening therefore gives

$$\sum_j \mathrm{Var}(\mathbb{E}[h_{\delta,j} \mid Z_1]) = \sum_j \mathrm{Var}(h_{\delta,j}) = m.$$

C.4 Metric and Complexity Scope

The finite network result is an approximation in W_2, controlled by the explicit L^2 coupling above. For every requested accuracy, a finite width and finite Lipschitz constant suffice; these may increase as the accuracy improves. This metric permits a network with scalar input to approximate the multivariate Gaussian law while preserving dependence on the same latent source. The exact Gaussian law belongs to the measurable construction, and the finite networks approach it in the stated metric.

D Exact Transition Identifiability

D.1 Proof of Proposition 1

Since h is a function of X, the residual satisfies $\mathbb{E}[R_h \mid X] = Kh - B_h h = \bar{R}_h$. Reversibility of (Z, Z') makes (X, X') exchangeable, so K is self-adjoint on $L^2(P_X)$, and it is a contraction by Jensen. The coordinates of h form an orthonormal basis of V_h, so $(B_h)_{jk} = \langle Kh_j, h_k \rangle$ and $P_h K h = B_h h$. Hence $\bar{R}_h = (I - P_h)Kh$ and

$$\epsilon_h = \sum_j \|(I - P_h)Kh_j\|_2^2 = \|(I - P_h)KP_h\|_{\mathrm{HS}}^2.$$

On $V_h \oplus V_h^\perp$, write

$$K = \begin{pmatrix} B & C^* \\ C & D \end{pmatrix}, \qquad P_h = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \qquad [K, P_h] = \begin{pmatrix} 0 & -C^* \\ C & 0 \end{pmatrix}.$$

Thus $\|[K, P_h]\|_{\mathrm{HS}}^2 = 2\|C\|_{\mathrm{HS}}^2 = 2\epsilon_h$. The defect vanishes exactly when $C = 0$, equivalently $KV_h \subseteq V_h$.

For the relation to the latent operator, let $U : L^2(P_X) \to \mathcal{H}_X := L^2(\sigma(X)) \subseteq L^2(P_Z)$ be the isometry $(Uq)(Z) = q(X)$ and let Π_X project $L^2(P_Z)$ onto $\mathcal{H}_X$. For $q \in L^2(P_X)$ and any $w \in \mathcal{H}_X$, the tower property gives $\langle UKq, w \rangle = \mathbb{E}[q(X')w(X)] = \langle TUq, w \rangle$, so $UK = \Pi_X TU$: the observable operator is the latent operator compressed to observable functions. If $\sigma(X) = \sigma(Z)$ up to null sets then $\Pi_X = I$ and $UK = TU$, so the two are unitarily equivalent and the statement reads on T.

D.2 Closure Under Partial Observation

Without observation equivalence the two readings separate, and by exactly one term. Identify P_h with the projector onto $UV_h \subseteq \mathcal{H}_X$ and define the latent defect of the same observable representation by $\epsilon_h^Z := \sum_j \|(I - P_h)T(Uh_j)\|_2^2$. Decomposing along $\mathcal{H}_X$ and its complement, and using $UK = \Pi_X TU$ together with $UV_h \subseteq \mathcal{H}_X$,

$$(I - P_h)T(Uh_j) = (I - \Pi_X)T(Uh_j) + U(I - P_h)Kh_j,$$

an orthogonal sum, so

$$\epsilon_h^Z = \epsilon_h + \sum_{j=1}^{m} \|(I - \Pi_X)T(Uh_j)\|_2^2. \tag{13}$$

Exact observable closure therefore does not imply latent closure: the gap is precisely the predictable transition content that leaves the observable function space, which no diagnostic computed from X can see. Under observation equivalence the second term vanishes and the two defects coincide. Theorem 3 is stated under that assumption; Eq. (13) says what survives without it, namely closure for the observable dynamics.

D.3 Exact Linear Recovery

Appendix A.4 gives $B_h \succ 0$. Diagonalize B_h as $UB_hU^\top = \mathrm{diag}(\beta_1, \ldots, \beta_m)$ with every $\beta_j > 0$, and let $u = Uh$. Exact ZCM gives $Tu_j = \beta_j u_j$. Taking inner products with the Hermite basis yields

$$(\lambda_\alpha - \beta_j)\langle u_j, \psi_\alpha \rangle = 0.$$

Since only finitely many λ_α can equal a fixed positive β_j, each u_j is a finite polynomial. The Gaussian output law is preserved by U, so Lemma C.1 forces every u_j to be affine. Centering makes it linear. Therefore $h = AZ$, standardization gives $AA^\top = I$, and $Th = B_h h$ gives $AR = B_h A$.

Conversely, suppose $h = AZ$, $AA^\top = I$, and the row space is R-invariant. With $P = A^\top A$, invariance means $RP = PR$. Then $B_h = ARA^\top$ and $B_h A = ARP = AR$, which gives ZCM. Since $AA^\top = I_m$, the Gaussian vector AZ has law $\mathcal{N}(0, I_m)$.

Proposition 1 reads exact ZCM as invariance of V_h under the transition operator, and searching for a function space the operator maps into itself is also what data-driven approximations of that operator do, whether by fitting a dictionary and reading off its leading modes (Korda & Mezić, 2018) or by optimizing a variational score over candidate feature spaces (Wu & Noé, 2020). What Theorem 3 adds is that combining invariance with an exactly Gaussian output law forces the invariant space to be linear in the latent variables. The temporal dependence doing this work is the same structure that identifies latent variables from time series in nonlinear independent component analysis (Hälvä & Hyvärinen, 2020; Klindt et al., 2021).

D.4 Fourth-Order Rigidity

The next estimate converts transition eigenfunction structure and a Gaussian fourth moment into linearity.

Lemma D.1 (Fourth-order rigidity)**.** *If F is a centered, unit-variance finite polynomial satisfying $\mathcal{A}F = aF$ with $a > 0$, then*

$$\mathbb{E}F^4 - 3 \geq \frac{6}{d}\|P_{\geq 2}F\|_2^4. \tag{14}$$

Proof. Let a centered, unit-variance finite polynomial F satisfy $\mathcal{A}F = aF$ with $a > 0$, and define

$$\Gamma(F, G) = \sum_i w_i \partial_i F\, \partial_i G, \qquad \Gamma(F) = \Gamma(F, F), \qquad s_F = \|P_{\geq 2}F\|_2^2.$$

Gaussian integration by parts gives $\mathbb{E}[F\mathcal{A}G] = \mathbb{E}\Gamma(F, G)$, while the weighted generator product rule gives

$$\mathcal{A}(FG) = F\mathcal{A}G + G\mathcal{A}F - 2\Gamma(F, G).$$

Set $G = F^2 - 1$. Since $\mathbb{E}\Gamma(F) = \langle F, \mathcal{A}F\rangle = a$,

$$\Gamma(F) - a = (a - \mathcal{A}/2)G. \tag{15}$$

Every nonconstant Hermite mode of G has weighted energy $\nu \in (0, 2a]$. This follows from the nonnegative Hermite product coefficients and the energy subadditivity property in Appendix A.2. Decomposing $G = \sum_{\nu>0} G_\nu$ into orthogonal energy levels therefore yields

$$\operatorname{Var}\Gamma(F) = \sum_{\nu>0}(a - \nu/2)^2\|G_\nu\|_2^2 \le a\sum_{\nu>0}(a - \nu/2)\|G_\nu\|_2^2.$$

Moreover, integration by parts with F^3 gives $3\mathbb{E}[F^2\Gamma(F)] = a\mathbb{E}F^4$. Hence

$$\langle \Gamma(F) - a, F^2 - 1\rangle = \frac{a}{3}(\mathbb{E}F^4 - 3),$$

and Eq. (15) gives the upper bound

$$\operatorname{Var}\Gamma(F) \le \frac{a^2}{3}(\mathbb{E}F^4 - 3). \tag{16}$$

For the reverse bound, use the normalized second Hermite mode $\psi_{2e_i} = (Z_i^2 - 1)/\sqrt{2}$ and put $q_i = \langle \Gamma(F), \psi_{2e_i}\rangle$. Two Gaussian integrations by parts give

$$q_i = \sqrt{2}\sum_j w_j \left(\|\partial_{ij}F\|_2^2 + \langle \partial_j F, \partial_{iij}F\rangle\right).$$

If the derivatives are nonzero, $\partial_j F$ and $\partial_{iij}F$ lie in weighted energy spaces $a - w_j$ and $a - w_j - 2w_i$, respectively. These levels differ because $w_i > 0$, so the cross term vanishes. Thus

$$q_i = \sqrt{2}\sum_j w_j\|\partial_{ij}F\|_2^2 \ge 0.$$

Expanding $F = \sum_\alpha c_\alpha\psi_\alpha$ on the level $E_\alpha = a$ and using $\partial_i\psi_\alpha = \sqrt{\alpha_i}\psi_{\alpha - e_i}$ gives

$$\sum_{i,j} w_j\|\partial_{ij}F\|_2^2 = \sum_\alpha c_\alpha^2 E_\alpha(|\alpha| - 1) \ge a s_F.$$

The ψ_{2e_i} are orthonormal, so Bessel and finite-dimensional Cauchy-Schwarz imply

$$\operatorname{Var}\Gamma(F) \ge \sum_i q_i^2 \ge \frac{1}{d}\left(\sum_i q_i\right)^2 \ge \frac{2a^2}{d}s_F^2. \tag{17}$$

Combining Eqs. (16) and (17) proves Eq. (14). All operations involve finite polynomials. □

Corollary D.2 (Recovery from Gaussian fourth moments)**.** *Under the model and standardization assumptions, exact ZCM and $\mathbb{E}[(v^\top h)^4] = 3$ for every unit vector v imply the linear invariant structure in Theorem 3(i).*

Proof. After exact ZCM, the rotated outputs in the proof of Theorem 3 are finite weighted transition eigenpolynomials. If every unit output direction has Gaussian fourth moment, the property survives the output rotation and each rotated coordinate has $\mathbb{E}F^4 = 3$. Equation (14) then forces every higher-order Hermite coefficient to vanish. This gives the fourth-order strengthening of exact linear recovery. The transition-eigenfunction structure carries this step, since it is what upgrades a second- and fourth moment match into full linearity.

□

D.5 Alignment Selection within Linear Invariant Subspaces

For a row-orthonormal linear map A, let $P = A^\top A$. Then

$$C(AZ) = \operatorname{tr}(PR) = \sum_i \rho_i P_{ii}, \qquad 0 \le P_{ii} \le 1, \quad \sum_i P_{ii} = m.$$

This finite-dimensional mass allocation problem is maximized by the first m coordinates. If $\rho_m > \rho_{m+1}$, equality uniquely fixes the target projection, while repeated eigenvalues at the cutoff leave the corresponding invariant family. Output bases inside a fixed row space differ by an orthogonal rotation.

E Stable Recovery

E.1 High-Frequency Escape

Take any measurable digit splitting $H : (0,1) \to \mathbb{R}^m$ with $H(U) \sim \mathcal{N}(0, I_m)$ for $U \sim \mathrm{Unif}(0,1)$, and define

$$h^{(n)}(Z) = H(\{n\Phi(Z_1)\}),$$

where $\{\cdot\}$ is fractional part. The inner variable remains uniform, so every output law is exactly standard Gaussian. For a continuous test function a on $[0,1]$,

$$\int_0^1 a(u) H_j(\{nu\})\, du = \int_0^1 H_j(v) \left[\frac{1}{n}\sum_{k=0}^{n-1} a\Big(\frac{k+v}{n}\Big)\right] dv \to 0.$$

Density extends this to all L^2 tests, so $h_j^{(n)} \rightharpoonup 0$. Compactness of T turns weak convergence into $\|Th_j^{(n)}\|_2 \to 0$, hence $\epsilon_{h^{(n)}} \to 0$ and $B_{h^{(n)}} \to 0$. The first chaos coefficient $M_n = \mathbb{E}[h^{(n)} Z^\top]$ also tends to zero, so $s(h^{(n)}) = m - \|M_n\|_F^2 \to m$. By the rectangular form of Lemma B.1, the minimum squared distance over row-orthonormal linear maps tends to $2m$. The sequence isolates the loss of uniform predictability excluded by Theorem 4(b).

E.2 Approximate Eigenfunction Reduction

Assume $B_h \succeq bI_m$ and $\epsilon_h \le \epsilon < b^2/8$. Choose $U \in O(m)$ so that $UB_hU^\top = \mathrm{diag}(\beta_1, \dots, \beta_m)$, and set $u = Uh$. Then

$$b \le \beta_j \le \rho_1, \qquad \epsilon_j := \|(T - \beta_j)u_j\|_2^2, \qquad \sum_j \epsilon_j = \epsilon_h \le \epsilon.$$

This is an exact diagonalization of the finite matrix B_h, available for any spectrum of that matrix.

Let P_b be the orthogonal projection onto nonconstant transition modes with $\lambda_\alpha \ge b/2$. This projection has finite rank. Put $p_j = P_b u_j$ and $r_j = \|u_j - p_j\|_2$. On every discarded mode, $|\lambda_\alpha - \beta_j| \ge b/2$, so Parseval gives

$$r_j^2 \le \frac{4\epsilon_j}{b^2}, \qquad \sum_j r_j^2 \le q_{\mathrm{tail}} := \frac{4\epsilon}{b^2} < \frac{1}{2}. \tag{18}$$

Hence $\|p_j\|_2^2 = 1 - r_j^2 > 1/2$. Define the centered unit-variance finite polynomial $F_j = p_j/\|p_j\|_2$. Then

$$\|u_j - F_j\|_2^2 = 2\left(1 - \sqrt{1 - r_j^2}\right) \le 2r_j^2. \tag{19}$$

Writing $w_- = -\log \rho_1$ and $E_{\max} = \log(2/b)$, every retained mode has $E_\alpha = -\log \lambda_\alpha \le E_{\max}$ and therefore total degree at most

$$K = \left\lfloor \frac{E_{\max}}{w_-} \right\rfloor. \tag{20}$$

To convert the transition residual to a generator residual, note that $-\log x$ is $2/b$-Lipschitz on $[b/2, \rho_1]$. Therefore

$$\|(\mathcal{A} + \log \beta_j)F_j\|_2 \le \frac{2}{b\|p_j\|_2}\|(T - \beta_j)p_j\|_2 \le \frac{2\sqrt{2}}{b}\sqrt{\epsilon_j}.$$

Let $a_j = \langle F_j, \mathcal{A}F_j\rangle$ and $t_j = \|(\mathcal{A} - a_j)F_j\|_2$. The scalar a_j minimizes $\|\mathcal{A}F_j - cF_j\|_2$ over $c \in \mathbb{R}$, so

$$\sum_j t_j^2 \le \frac{8\epsilon}{b^2}. \tag{21}$$

It is convenient to define $v = \sqrt{8\epsilon/(mb^2)}$. Equations (18) and (21) and Cauchy-Schwarz imply

$$\frac{1}{m}\sum_j t_j \le v, \qquad \frac{1}{m}\sum_j t_j^2 \le v^2, \qquad \frac{1}{m}\sum_j \|u_j - F_j\|_2 \le v.$$

Thus the lower bound on B_h keeps the retained spectrum away from the zero accumulation point and reduces each output to a controlled finite polynomial region, for any configuration of the nonzero Hermite eigenvalues.

E.3 Quantitative Rigidity and Explicit Stability Function

We first state the finite polynomial estimate used coordinatewise. Let F be centered with unit variance, supported on $|\alpha| \le K$ and $E_\alpha \le E_{\max}$. Set

$$a = \langle F, \mathcal{A}F\rangle, \qquad e = \mathcal{A}F - aF, \qquad t = \|e\|_2, \qquad \kappa_+(F) = \max\{\mathbb{E}F^4 - 3, 0\}.$$

Define the explicit finite-dimensional moment constant

$$M_{d,K} = \left(\sum_{1\le|\alpha|\le K} \|\psi_\alpha\|_6^2\right)^{1/2}, \qquad M = M_{d,K},$$

so every centered polynomial of degree at most K satisfies $\|F\|_6, \|F\|_4 \le M\|F\|_2$. Further set

$$D_K = K - 1 + K^2\Big(\sum_i w_i\Big)\Big(\sum_i w_i^{-1}\Big), \tag{22}$$

$$C_1 = \frac{2E_{\max}}{3}M^3 + 3E_{\max}M^2(M^2+1), \qquad C_2 = 2M^4. \tag{23}$$

Lemma E.1 (Quantitative rigidity on a finite spectral region)**.** *For the centered unit-variance polynomial and constants defined above,*

$$\boxed{s_F \le \frac{D_K}{w_-}t + \sqrt{\frac{d}{6}\kappa_+(F) + \frac{d}{2w_-^2}(C_1 t + C_2 t^2)}.} \tag{24}$$

Proof. With $G = F^2 - 1$ and $H = (a - \mathcal{A}/2)G$, the generator product rule gives

$$\Gamma(F) - a = H + Fe, \qquad \langle H, G\rangle = \frac{a}{3}(\mathbb{E}F^4 - 3) - \frac{2}{3}\mathbb{E}[F^3 e].$$

The nonnegative Hermite product coefficients imply that the negative spectral part of the multiplier $a - \mathcal{A}/2$ satisfies $\|H_-\|_2 \le M^2 t$. Together with $\|G\|_2 \le M^2 + 1$, $|\mathbb{E}F^3 e| \le M^3 t$, and $\|Fe\|_2 \le M^2 t$, this yields

$$\operatorname{Var}\Gamma(F) \le \frac{a^2}{3}\kappa_+(F) + C_1 t + C_2 t^2. \tag{25}$$

For the lower side, the second Hermite coefficients of $\Gamma(F)$ obey the same integration-by-parts identity as in Appendix D.4. Commuting derivatives through the approximate generator equation gives $|\langle \partial_j F, \partial_{iij} F\rangle| \le K^2 t/w_i$, while direct Hermite expansion gives

$$\sum_{i,j} w_j \|\partial_{ij} F\|_2^2 \ge a s_F - (K-1)t.$$

Combining these estimates with Bessel and Cauchy-Schwarz yields

$$a s_F \le D_K t + \sqrt{d\,\operatorname{Var}\Gamma(F)/2}.$$

Since $a \ge w_- > 0$, substituting Eq. (25) proves Eq. (24). The proof works inside a finite coefficient space, using only the retained spectral region.

□

E.3.1 Transferring distributional and fourth moment error

We now make the stability function in Theorem 4 explicit. Define

$$\mathfrak{D}(x,y) = x^3 + x^2 y + xy^2 + y^3, \qquad q_{\text{tail}} = \frac{4\epsilon}{b^2}, \qquad v = \sqrt{\frac{8\epsilon}{mb^2}}.$$

For the W_2 route, let

$$D_G = \mathfrak{D}(M, 15^{1/6}), \qquad \Xi_G = D_G(\delta + v).$$

Indeed, if two coupled scalar variables U, V have finite sixth moments, factorization of $U^4 - V^4$ followed by Cauchy-Schwarz gives

$$|\mathbb{E}U^4 - \mathbb{E}V^4| \le \|U - V\|_2 \, \mathfrak{D}(\|U\|_6, \|V\|_6).$$

Orthogonal rotation preserves vector W_2, and coordinate projection is 1-Lipschitz. Thus $W_2(\mathrm{Law}(u_j), \mathcal{N}(0,1)) \le \delta$. Apply the moment inequality after finite polynomial truncation, where the sixth moments are bounded. Since a standard Gaussian has sixth norm $15^{1/6}$, coupling F_j first to u_j and then optimally to a Gaussian gives $m^{-1}\sum_j \kappa_+(F_j) \le \Xi_G$ using the controlled sixth moments of the truncated polynomials.

For the fourth moment route, assume $\sup_{\|a\|=1} \|a^\top h\|_6 \le H_6$ and $\sup_{\|a\|=1} |\mathbb{E}(a^\top h)^4 - 3| \le \delta_4$. Set

$$D_4 = \mathfrak{D}(M, H_6), \qquad \Xi_4 = \delta_4 + D_4 v.$$

The same coupling inequality between u_j and F_j gives $m^{-1}\sum_j \kappa_+(F_j) \le \Xi_4$.

E.3.2 Aggregating coordinate bounds

For either admissible value of Ξ, define

$$\mathcal{S}(\Xi, \epsilon; d, m, R, b) = \min\left\{ m, 4q_{\mathrm{tail}} + \frac{2mD_K}{w_-} v + 2m\sqrt{\frac{d}{6}\Xi + \frac{d}{2w_-^2}(C_1 v + C_2 v^2)} \right\}. \tag{26}$$

The bound in Theorem 4(b) and its fourth moment extension are

$$\boxed{S_G = \mathcal{S}(\Xi_G, \epsilon; d, m, R, b), \qquad S_4 = \mathcal{S}(\Xi_4, \epsilon; d, m, R, b).} \tag{27}$$

To see this, apply Eq. (24) to every F_j, use $s(u_j) \le 2\|u_j - F_j\|_2^2 + 2s(F_j)$, sum over j, and apply concavity of the square root together with the average bounds from Appendix E.2. This is exactly Eq. (26). Output rotation preserves $s(h)$, and the trivial $s(h) \le m$ supplies the outer minimum.

For fixed d, m, R, b, we have $q_{\mathrm{tail}} = O(\epsilon)$, $v = O(\epsilon^{1/2})$, and $\Xi_G = O(\delta + \epsilon^{1/2})$. Equation (26) therefore gives $S_G = O(\delta^{1/2} + \epsilon^{1/4})$. The fourth-order route is identical with δ_4 in place of δ. These rates control nonlinear energy and hence squared distance.

E.4 Near Linearity

Corollary E.2 (Near linearity)**.** *For a standardized representation with nonlinear energy $s(h)$,*

$$\inf_{AA^\top = I_m} \|h - AZ\|_2^2 = 2m - 2\|M_h\|_* \le 2s(h).$$

In particular, the preceding bounds on $s(h)$ imply stable near linearity.

Proof. Summing the coordinatewise rigidity bounds gives $s(h) \le S$. Let $M_h = \mathbb{E}[hZ^\top]$. Since $M_h M_h^\top \preceq I$ and $m - \|M_h\|_F^2 = s(h)$, the polar factor A of M_h can be completed to a row-orthonormal matrix. Exactly as in the proof of Theorem 1,

$$\|M_h - A\|_F^2 \le s(h), \qquad \|h - AZ\|_2^2 \le 2s(h).$$

This proves near linearity. The rate is stated for squared distance, and the corresponding unsquared L^2 rate is $O(\delta^{1/4} + \epsilon^{1/8})$.

□

E.5 A Supplementary Bound for Target Selection

This consequence combines a nonlinear energy bound with a separate alignment assumption.

Corollary E.3 (Target selection from alignment)**.** *Suppose $s(h) \le S$, $\Delta = \rho_m - \rho_{m+1} > 0$, and $C(h) \ge C_{\mathrm{lin}} - \xi$ for $\xi \ge 0$. Then*

$$\inf_{Q \in O(m)} \|h - QZ_{1:m}\|_2^2 \le \min\left\{ 2m,\ 2\left[\frac{\xi}{\Delta} + \left(1 + \frac{(\rho_1^2 - \rho_m)_+}{\Delta}\right) S \right] \right\}. \tag{28}$$

Proof. Write $h = M_h Z + P_{\geq 2} h$ and $p_i = \|M_h e_i\|^2$. Standardization gives $0 \leq p_i \leq 1$ and $\sum_i p_i = m - s(h)$. Set $O = \sum_{i>m} p_i$. Since T preserves Hermite degree and its nonlinear spectrum is bounded by ρ_1^2,

$$\begin{aligned} C(h) &\leq \sum_i \rho_i p_i + \rho_1^2 s(h) \\ &\leq C_{\mathrm{lin}} - \rho_m \sum_{i \leq m} (1 - p_i) + \rho_{m+1} O + \rho_1^2 s(h) \\ &= C_{\mathrm{lin}} - \Delta O + (\rho_1^2 - \rho_m) s(h). \end{aligned}$$

Combining this with the alignment assumption bounds O by $\xi/\Delta + (\rho_1^2 - \rho_m)_+ s(h)/\Delta$. The missing linear mass in the target is $O + s(h)$. Lemma B.1 bounds the squared target distance by twice that mass, giving Eq. (28). □

E.6 Sufficient Conditions for Uniform Predictability

A uniform floor can be imposed on the representation family or obtained from another constraint. These are independent sufficient routes. If every normalized $u \in V_h$ satisfies $\langle u, \mathcal{A} u \rangle \leq K_0$, Jensen's inequality applied to its spectral weights gives $\langle u, Tu \rangle \geq e^{-K_0}$, hence $B_h \succeq e^{-K_0} I_m$.

Alignment supplies another route when its lower bound is sufficiently large. If $\beta_1 \geq \cdots \geq \beta_m$ are the eigenvalues of B_h, compression of the self-adjoint operator gives $\beta_j \leq \lambda_j$. Since $C(h) = \sum_j \beta_j$,

$$\lambda_{\min}(B_h) \geq C(h) - \sum_{j=1}^{m-1} \lambda_j.$$

A uniform positive lower bound on the right-hand side therefore supplies the required floor. For $m = 2$, $C(h) \geq C_{\mathrm{lin}} - \xi$ implies $\lambda_{\min}(B_h) \geq \rho_2 - \xi$. For larger m, the bound uses the full transition spectrum, which includes nonlinear modes.

F Experimental Protocols and Full Results

F.1 Evaluation and Training Protocols

All trained experiments use fresh pairs from the stationary Gaussian transition model and independent evaluation samples. Comparisons retain the same encoder architecture and training budget within each experiment. Tables report means and sample standard deviations across initializations. These summarize optimization variability conditional on the evaluation sample. Controlled constructions use exact spectral identities, Gaussian quadrature, and Monte Carlo integration where needed.

Representations are centered and symmetrically whitened at evaluation. For MLPs, the statistics are estimated from pooled current and future outputs. For pixels, they are fitted on the first half of the current evaluation frames and applied to current and future frames. Pixel probes are trained on the first half and evaluated on the second. This common normalization aligns the diagnostics with the population representation in Section 3.

For a standardized target vector Φ, we evaluate the squared distance invariant under rotation

$$D_{\mathrm{target}} = 2m - 2\|\widehat{\mathbb{E}}[h\Phi^\top]\|_*.$$

Here D_{runner} uses the same formula with Φ spanning the two linear directions with the second largest sum of transition rates. Empirical cross moment estimates can give slightly negative values near zero, since the population normalization identities hold only approximately in the sample. For linear recovery over all latent directions, the corresponding population quantity is $D_{\mathrm{lin}} = 2m - 2\|M_h\|_*$, as in Corollary E.2. Projection energy onto a normalized scalar feature is $E_\phi = \|\widehat{\mathbb{E}}[h\phi]\|^2$, and energy in linear modes is $E_{\mathrm{lin}} = \|\widehat{\mathbb{E}}[hZ^\top]\|_F^2$. We also report $C = \widehat{\mathbb{E}}[h^\top h']$ and the smallest eigenvalue $\widehat{b}_{\min}$ of the symmetrized empirical transition matrix. Gaussian fit is measured by sliced W_2 over 256 random directions against standard normal quantiles, a projection-based surrogate for the Wasserstein distance in the sense of Kolouri et al. (2019). These empirical quantities use standardized outputs on evaluation samples.

Table 1: Training and evaluation budgets. Batch sizes count transition pairs. MLPs use Adam and CNNs use AdamW, with learning rate 10^{-3} and cosine decay.

Experiment	Batch	Steps	Initializations	Evaluation pairs
Spectral boundary	4096	6000	10	200000
Synthetic closure	4096	6000	8	100000
Pixel allocation	256	12000	5	40000
Pixel closure, $\lambda = 0.60$	256	12000	5	40000

F.2 Distributional and Closure Objectives

Training with whitening alone centers and whitens the pooled current and future outputs in each batch and minimizes their mean squared difference. A covariance ridge of $10^{-5}I$ stabilizes the whitening. Training with Gaussian regularization acts on raw encoder outputs f through

$$\widehat{L} = (1-\lambda)\widehat{\mathbb{E}}\|f(X') - f(X)\|^2 + \lambda \widehat{G}_{\mathrm{EP}}(f).$$

Writing N for the number of pooled outputs, K for the number of projection directions, and a_j for random unit directions, the sliced Epps-Pulley statistic is

$$\widehat{G}_{\mathrm{EP}}(f) = \frac{2N}{K}\sum_{j=1}^{K} \mathrm{Trap}_{[0,3]}\left[\left|\frac{1}{N}\sum_{n=1}^{N} e^{ita_j^\top f_n} - e^{-t^2/2}\right|^2 e^{-t^2/2}\right].$$

Here $\mathrm{Trap}_{[0,3]}$ denotes trapezoidal quadrature on $[0,3]$ using 17 equally spaced nodes, with 64 projection directions for MLPs and 128 for CNNs. The factor N is included in the statistic, so weights are specified together with the batch size. Relative to the statistic divided by N, the ratio of Gaussian and alignment coefficients is $\lambda N/(1-\lambda)$. We choose weights within each task; the spectral and allocation experiments use $\lambda = 0.05$, and the main pixel closure comparison uses $\lambda = 0.60$.

For closure training, let $y = f(X)$ and $y' = f(X')$. A regularized batch linear predictor gives

$$\widehat{B} = \widehat{\mathbb{E}}[y'y^\top]\big(\widehat{\mathbb{E}}[yy^\top] + 10^{-6}I\big)^{-1}, \qquad r = y' - \widehat{B}y.$$

We hold $\widehat{B}$ fixed when differentiating the closure term. A critic $c_\theta(X)$ estimates predictable residual energy through

$$\widehat{J}(\theta) = 2\widehat{\mathbb{E}}\langle r, c_\theta(X)\rangle - \widehat{\mathbb{E}}\|c_\theta(X)\|^2.$$

Completing the square gives $\widehat{J}(\theta) = \epsilon_h - \|c_\theta - \bar{R}_h\|^2$ in population, so the supremum over square-integrable critics equals the conditional mean residual energy ϵ_h, and the supremum over any restricted critic class equals ϵ_h minus that class's squared approximation error for $\bar{R}_h$. A finite critic therefore never overstates the defect, and a small measured value can be attributed to three separable sources: a genuinely small defect, a critic class too small to represent $\bar{R}_h$, or an incompletely optimized critic. The sweeps below vary the second and third and evaluate the first with an independent critic. Let λ_Z denote the closure weight. The critic takes three ascent steps per encoder update, with the encoder then minimizing $\widehat{L} + \lambda_Z \max\{\widehat{J}, 0\}$. Synthetic critics have two hidden Gaussian error linear unit (GELU) layers of width 128. Pixel critics use the same CNN architecture as the encoder and receive the current image. The fixed closure weight is $\lambda_Z = 2$.

The pixel objective controls use the same CNN, optimizer, and budget. Closure alone adds the same residual critic to alignment after batch whitening, with $\lambda_Z = 2$. For the two alternative second-order controls, let $\widehat{\Sigma}$ be the sample covariance of pooled raw outputs. The variance and covariance control uses the corresponding terms of VICReg (Bardes et al., 2022), with $G_{\mathrm{VC}} = 25m^{-1}\sum_j[1 - \sqrt{\widehat{\Sigma}_{jj} + 10^{-4}}]_+ + m^{-1}\sum_{i\neq j}\widehat{\Sigma}_{ij}^2$. Here $[u]_+ = \max\{u, 0\}$. The orthogonality control uses $G_{\mathrm{orth}} = \|\widehat{\Sigma} - I_m\|_F^2$. Each replaces $\widehat{G}_{\mathrm{EP}}$ in $\widehat{L}$ with weight $\lambda = 0.60$. The VICReg control isolates these two regularization terms with the shared encoder and weighting, without an additional projection head. Its role is a controlled comparison of constraints.

Evaluation separately measures the conditional mean residual of standardized representations. A ridge predictor based on latent tensor Hermite features of total degree at most four is fitted on one

half of the evaluation sample, and the dual expression is evaluated on the other; the two halves are exchanged and averaged. Tables denote the resulting energy by $\widehat{e}_h$ and its ratio to residual energy by $\widehat{e}_{\mathrm{rel}}$. This diagnostic using latent features provides a common reference across training objectives. Estimates using random features of observations are examined separately below. Signed estimates near zero are retained.

F.3 Spectral Selection

The latent dimension is three and the representation dimension is two. We fix $\rho_1 = 0.8$, $\rho_3 = 0.3$, and vary ρ_2 from 0.52 to 0.76 across the quadratic boundary. Observations are

$$X = A[Z; Z^{\odot 3}/\sqrt{15}],$$

where A is a fixed invertible mixing matrix. Inclusion of the linear coordinates makes the map injective. The MLP has three hidden GELU layers of width 256. Table 2 gives the projection energies, selection frequencies, and correlation. At a strict cutoff, selection agrees with the larger transition eigenvalue. At the tie, the population optimum includes the full family described in Theorem 1.

Table 2: Spectral selection across the boundary. Values are seed means and sample standard deviations; P_{quad} is the selection frequency.

Objective	ρ_2	n	E_{Z_2}	$E_{H_2(Z_1)}$	P_{quad}	C
Gaussian	0.52	10	0.993 ± 0.000	0.000 ± 0.000	0.000	1.316 ± 0.000
Gaussian	0.56	10	0.994 ± 0.000	0.000 ± 0.000	0.000	1.356 ± 0.000
Gaussian	0.60	10	0.994 ± 0.000	0.000 ± 0.000	0.000	1.396 ± 0.000
Gaussian	0.62	10	0.995 ± 0.000	0.000 ± 0.000	0.000	1.417 ± 0.000
Gaussian	0.64	10	0.995 ± 0.000	0.000 ± 0.000	0.000	1.437 ± 0.000
Gaussian	0.66	10	0.995 ± 0.000	0.000 ± 0.000	0.000	1.457 ± 0.000
Gaussian	0.68	10	0.996 ± 0.000	0.000 ± 0.000	0.000	1.477 ± 0.000
Gaussian	0.72	10	0.996 ± 0.000	0.000 ± 0.000	0.000	1.517 ± 0.000
Gaussian	0.76	10	0.997 ± 0.000	0.000 ± 0.000	0.000	1.558 ± 0.000
Whitening	0.52	10	0.000 ± 0.000	0.994 ± 0.000	1.000	1.437 ± 0.000
Whitening	0.56	10	0.000 ± 0.000	0.994 ± 0.000	1.000	1.437 ± 0.000
Whitening	0.60	10	0.000 ± 0.000	0.994 ± 0.000	1.000	1.437 ± 0.000
Whitening	0.62	10	0.000 ± 0.000	0.994 ± 0.000	1.000	1.437 ± 0.000
Whitening	0.64	10	0.983 ± 0.023	0.015 ± 0.023	0.000	1.438 ± 0.000
Whitening	0.66	10	0.998 ± 0.000	0.000 ± 0.000	0.000	1.458 ± 0.000
Whitening	0.68	10	0.998 ± 0.000	0.000 ± 0.000	0.000	1.478 ± 0.000
Whitening	0.72	10	0.998 ± 0.000	0.000 ± 0.000	0.000	1.518 ± 0.000
Whitening	0.76	10	0.998 ± 0.000	0.000 ± 0.000	0.000	1.559 ± 0.000

F.4 Finite Networks and Source Allocation

We realize the fixed construction using a single source at $\rho = (0.98, 0.10, 0.05)$ and $m = 2$. Increasing the resolution of its piecewise linear approximation produces ReLU networks with one hidden layer of widths 6, 30, 126, 510, and 2046. Each level uses 400000 evaluation pairs. The Gaussian coupling estimate compares the finite network with a numerical digit splitter using 26 bits under the same scalar input. All output coordinates remain functions of Z_1, so population whitening fixes the unnormalized variance supplied by a single source at m.

Table 4 reports the approximation and correlation quantities. As network width grows, the coupling upper estimate of W_2 decreases while the correlation advantage over the leading linear baseline remains positive (Figure 2B). All output variance is still supplied by one source, illustrating the finite network approximation in Theorem 2(b).

Source concentration under free training. The construction above fixes a single source by design. We therefore ask whether training allocates capacity the same way when nothing forces it to. Three spectra share the Gaussian objective, encoder, and budget, and differ only in how much predictability advantage the slowest source holds. The single-source explained variance $S_{\max}$ rises from 0.499 on the mild spectrum $(0.80, 0.52, 0.30)$ to 0.871 on the competitive spectrum $(0.98, 0.25, 0.02)$ and 0.901 on the strongly competitive spectrum $(0.98, 0.10, 0.05)$, while the maximum squared correlation between output coordinates stays below 0.08 throughout (Figure 2C). A balanced two-source linear target has $S_{\max} \approx 0.5$, so the two competitive spectra place roughly nine tenths of the representation variance on one latent variable while keeping the coordinates nearly uncorrelated. This is the solution described by Theorem 2, reached by optimization rather than by construction.

Held-out source decomposition. Because $S_{\max}$ uses a one-dimensional probe, we check separately that the residual variance is not simply a second source that the probe cannot see. Quantile boundaries, bin-wise means, and the fallback global mean are estimated on the first half of the evaluation sample and the explained variance is computed on the second. On the strongly competitive and competitive spectra, Z_1 alone explains about 90% and 87% of representation variance under a 100-bin probe. A held-out 50×50 joint grid over (Z_1, Z_2) explains about 99%, while the sum of the two single-source additive projections explains about 91% and 88%. The gap between the joint and additive fits, about 8 and 11 percentage points, is non-additive Z_1–Z_2 interaction rather than an independent second source. Refining the one-dimensional probe from 100 to 400 quantile bins moves the explained variance by about 0.3 percentage points in both cases, so the estimate is not resolution limited. Table 3 gives the per-spectrum values.

Table 3: Held-out source decomposition under Gaussian regularization. Bin boundaries, cell means, and the fallback mean are estimated on the first half of the evaluation sample; explained variance is computed on the second. The interaction column is the joint fit minus the additive sum.

Spectrum	Z_1, 100 bins	Z_1, 400 bins	(Z_1, Z_2) joint	additive sum	interaction	Runs
$(0.98, 0.10, 0.05)$	0.901 ± 0.017	0.905 ± 0.019	0.985 ± 0.004	0.906 ± 0.016	0.079 ± 0.019	8
$(0.98, 0.25, 0.02)$	0.871 ± 0.013	0.873 ± 0.014	0.989 ± 0.003	0.880 ± 0.013	0.109 ± 0.015	8

Table 4: Finite ReLU approximation of a fixed map using a single source. The variance fraction supplied by the source is one at every width.

Width	Coupling W_2 estimate	Sliced W_2	$C - C_{\mathrm{lin}}$	C
6	0.902	0.446	0.493	1.573
30	0.548	0.219	0.349	1.429
126	0.345	0.121	0.264	1.344
510	0.220	0.069	0.221	1.301
2046	0.143	0.042	0.206	1.286

F.5 Allocation from Pixel Observations

Four independent Gaussian autoregressive factors control a soft disk. For $u = \tanh Z$, its center, radius, and intensity are $(c_x, c_y) = (32 + 14u_1, 32 + 14u_2)$, $r = 8 + 4u_3$, and $v = 0.6 + 0.35u_4$. At pixel centers (p, q), the image is

$$X_{pq} = v \ \mathrm{sigmoid}\left(\frac{r - \sqrt{(p - c_x)^2 + (q - c_y)^2}}{0.7} \right).$$

The three spectra are $(0.85, 0.85, 0.85, 0.85)$, $(0.92, 0.86, 0.80, 0.74)$, and $(0.98, 0.90, 0.65, 0.40)$. We compare $m = 2$ with the full capacity reference $m = 4$. The encoder has four convolutions with stride two with channel widths $(32, 64, 128, 128)$, GroupNorm and GELU, followed by a flattened width 256 head. AdamW uses weight decay 10^{-4}.

Affine ridge probes assess each latent factor. Concentration on a single source is the explained variance on the test sample

$$S_i = 1 - \frac{\widehat{\mathbb{E}}_{\mathrm{test}} \|h - q_i(Z_i)\|^2}{\widehat{\mathbb{E}}_{\mathrm{test}} \|h - \bar{h}_{\mathrm{test}}\|^2}, \qquad S_{\max} = \max_i S_i,$$

where q_i uses representation means in 100 training quantile bins. The maximum is taken within each initialization before aggregation. Tables 5 and 6 include whitening, Gaussian, and joint objectives. Figure 4 shows that, at equal rates, the full capacity reference $m = 4$ recovers all four factors well. As rates separate, recovery of faster factors decreases and source concentration rises. The $m = 2$ representation primarily retains position and concentrates further on the slow source under strong anisotropy. These trends extend the allocation mechanism of Section 4 to convolutional learning from pixels. At equal rates, the selected subspace of dimension two reflects the encoder and renderer within the family of equally optimal population directions.

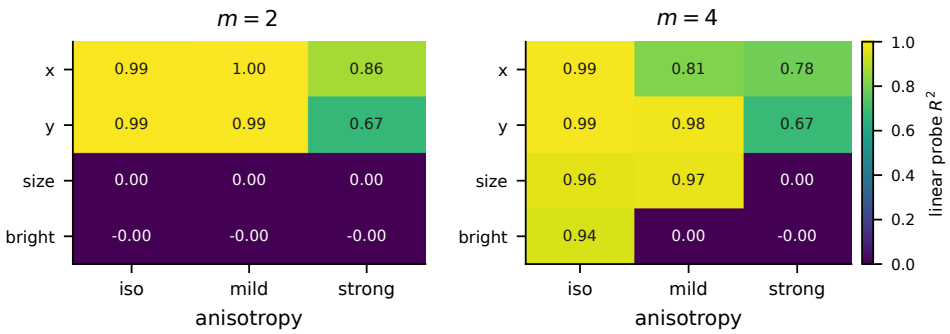

Figure 4: Mean factor recovery under Gaussian regularization for two representation dimensions. Table 5 reports the corresponding standard deviations and all training objectives.

Table 5: Factor recovery from pixels across capacities and transition regimes. Every entry is the mean and sample standard deviation of test R^2 over five initializations.

m	Regime	Objective	x	y	Size	Brightness
2	iso	Gaussian	0.992 ± 0.000	0.991 ± 0.000	0.000 ± 0.000	-0.000 ± 0.000
2	iso	Joint	0.992 ± 0.000	0.991 ± 0.001	0.000 ± 0.000	-0.000 ± 0.000
2	iso	Whitening	0.999 ± 0.000	0.999 ± 0.000	0.000 ± 0.000	-0.000 ± 0.000
2	mild	Gaussian	0.995 ± 0.000	0.990 ± 0.000	0.000 ± 0.000	-0.000 ± 0.000
2	mild	Joint	0.995 ± 0.000	0.990 ± 0.000	0.000 ± 0.000	-0.000 ± 0.000
2	mild	Whitening	0.999 ± 0.000	0.999 ± 0.000	0.000 ± 0.000	-0.000 ± 0.000
2	strong	Gaussian	0.864 ± 0.121	0.665 ± 0.294	0.000 ± 0.000	-0.000 ± 0.000
2	strong	Joint	0.901 ± 0.109	0.753 ± 0.269	0.000 ± 0.000	-0.000 ± 0.000
2	strong	Whitening	0.999 ± 0.000	0.000 ± 0.000	-0.000 ± 0.000	-0.000 ± 0.000
4	iso	Gaussian	0.988 ± 0.001	0.987 ± 0.001	0.960 ± 0.012	0.941 ± 0.017
4	iso	Joint	0.988 ± 0.001	0.988 ± 0.001	0.964 ± 0.009	0.945 ± 0.022
4	iso	Whitening	0.999 ± 0.000	0.998 ± 0.000	0.997 ± 0.001	0.001 ± 0.001
4	mild	Gaussian	0.811 ± 0.001	0.977 ± 0.000	0.968 ± 0.001	0.001 ± 0.001
4	mild	Joint	0.822 ± 0.001	0.977 ± 0.000	0.968 ± 0.001	0.001 ± 0.001
4	mild	Whitening	0.999 ± 0.000	0.998 ± 0.000	0.200 ± 0.446	-0.000 ± 0.000
4	strong	Gaussian	0.785 ± 0.137	0.675 ± 0.362	0.000 ± 0.000	-0.000 ± 0.000
4	strong	Joint	0.771 ± 0.188	0.557 ± 0.372	-0.000 ± 0.000	-0.000 ± 0.000
4	strong	Whitening	1.000 ± 0.000	0.995 ± 0.001	0.000 ± 0.000	-0.000 ± 0.000

Table 6: Pixel allocation diagnostics. The condition with full capacity serves as a reference. SW is sliced W_2.

m	Regime	Objective	$S_{\max}$	D_{target}	$\widehat{\beta}_{\min}$	SW
2	iso	Gaussian	0.507 ± 0.000	0.060 ± 0.001	0.834 ± 0.000	0.077 ± 0.001
2	iso	Joint	0.507 ± 0.000	0.060 ± 0.001	0.834 ± 0.000	0.077 ± 0.001
2	iso	Whitening	0.506 ± 0.000	0.062 ± 0.001	0.833 ± 0.000	0.027 ± 0.000
2	mild	Gaussian	0.506 ± 0.000	0.059 ± 0.001	0.845 ± 0.000	0.070 ± 0.001
2	mild	Joint	0.506 ± 0.000	0.059 ± 0.001	0.845 ± 0.000	0.070 ± 0.001
2	mild	Whitening	0.507 ± 0.000	0.061 ± 0.001	0.843 ± 0.000	0.028 ± 0.001
2	strong	Gaussian	0.617 ± 0.100	0.579 ± 0.472	0.905 ± 0.019	0.065 ± 0.005
2	strong	Joint	0.586 ± 0.091	0.429 ± 0.429	0.899 ± 0.014	0.063 ± 0.003
2	strong	Whitening	0.995 ± 0.000	1.978 ± 0.005	0.939 ± 0.002	0.454 ± 0.006
4	iso	Gaussian	0.250 ± 0.000	0.166 ± 0.029	0.829 ± 0.002	0.087 ± 0.006
4	iso	Joint	0.250 ± 0.000	0.157 ± 0.031	0.829 ± 0.002	0.084 ± 0.005
4	iso	Whitening	0.254 ± 0.000	2.033 ± 0.025	0.703 ± 0.001	0.145 ± 0.003
4	mild	Gaussian	0.485 ± 0.000	2.297 ± 0.004	0.788 ± 0.000	0.149 ± 0.004
4	mild	Joint	0.488 ± 0.000	2.281 ± 0.004	0.788 ± 0.000	0.152 ± 0.003
4	mild	Whitening	0.500 ± 0.003	3.620 ± 0.871	0.773 ± 0.012	0.281 ± 0.037
4	strong	Gaussian	0.650 ± 0.029	4.715 ± 0.728	0.862 ± 0.011	0.158 ± 0.001
4	strong	Joint	0.662 ± 0.062	4.938 ± 0.791	0.861 ± 0.009	0.161 ± 0.005
4	strong	Whitening	0.736 ± 0.001	4.040 ± 0.003	0.883 ± 0.000	0.335 ± 0.007

F.6 Closure Interventions and Spectral Ordering

The synthetic comparison uses spectra $(0.80, 0.52, 0.30)$, $(0.98, 0.25, 0.02)$, and $(0.98, 0.10, 0.05)$, with $d = 3, m = 2$. Their leading linear correlations are 1.32, 1.23, and 1.08. We keep the cubic observations, encoder, Gaussian weight, and training budget fixed within each comparison. Tables 7 and 8 report the closure variants. The adaptive variant updates the nonnegative closure multiplier

toward a residual level of 0.01. The variant with two stages switches the penalty off once that level is reached and reactivates it if the residual rises. The variant using latent features replaces the observation critic with a cross-fitted Hermite predictor. These variants describe alternative optimization procedures for the same conditional mean objective.

Table 7: Synthetic closure interventions and alternative implementations of the same conditional-mean objective. Control is $(0.80, 0.52, 0.30)$, Anisotropic is $(0.98, 0.25, 0.02)$, Strong is $(0.98, 0.10, 0.05)$. Joint denotes Gaussian regularization with the observation-based closure term at a fixed weight.

Spectrum	Objective	λ_Z	n	D_{target}	C	$\widehat{\beta}_{\min}$
Control	Gaussian	–	8	0.001 ± 0.000	1.314 ± 0.000	0.515 ± 0.000
Control	Joint	2	8	0.000 ± 0.000	1.315 ± 0.000	0.515 ± 0.000
Control	Joint	10	8	-0.001 ± 0.000	1.315 ± 0.000	0.515 ± 0.000
Control	Adaptive	2	8	0.001 ± 0.000	1.314 ± 0.000	0.515 ± 0.000
Control	Latent features	2	8	0.001 ± 0.000	1.314 ± 0.000	0.515 ± 0.000
Control	Two stage	2	8	0.001 ± 0.000	1.314 ± 0.000	0.515 ± 0.000
Anisotropic	Gaussian	–	8	2.344 ± 0.035	1.683 ± 0.011	0.825 ± 0.009
Anisotropic	Joint	2	8	0.028 ± 0.026	1.226 ± 0.010	0.247 ± 0.011
Anisotropic	Joint	10	8	0.016 ± 0.000	1.223 ± 0.000	0.243 ± 0.000
Anisotropic	Adaptive	2	8	0.040 ± 0.004	1.231 ± 0.002	0.255 ± 0.003
Anisotropic	Latent features	2	8	1.416 ± 0.777	1.608 ± 0.179	0.693 ± 0.202
Anisotropic	Two stage	2	8	0.087 ± 0.077	1.249 ± 0.027	0.273 ± 0.028
Strong	Gaussian	–	8	2.441 ± 0.057	1.683 ± 0.026	0.823 ± 0.020
Strong	Joint	2	8	0.032 ± 0.001	1.073 ± 0.000	0.093 ± 0.000
Strong	Joint	10	8	0.030 ± 0.001	1.073 ± 0.000	0.093 ± 0.000
Strong	Adaptive	2	8	0.049 ± 0.001	1.081 ± 0.001	0.104 ± 0.002
Strong	Latent features	2	8	1.956 ± 0.253	1.711 ± 0.038	0.795 ± 0.039
Strong	Two stage	2	8	0.057 ± 0.007	1.086 ± 0.005	0.108 ± 0.004

Table 8: Companion diagnostics for the synthetic interventions. E_{lin} is energy in linear modes, $\widehat{\epsilon}_h$ is the closure diagnostic using latent features, and SW denotes sliced W_2.

Spectrum	Objective	λ_Z	E_{lin}	$\widehat{\epsilon}_h$	SW
Control	Gaussian	–	1.999 ± 0.000	-0.0004 ± 0.0002	0.039 ± 0.001
Control	Joint	2	2.000 ± 0.000	-0.0005 ± 0.0002	0.036 ± 0.000
Control	Joint	10	2.001 ± 0.000	-0.0006 ± 0.0002	0.031 ± 0.000
Control	Adaptive	2	1.999 ± 0.000	-0.0004 ± 0.0002	0.038 ± 0.001
Control	Latent features	2	1.999 ± 0.000	-0.0004 ± 0.0002	0.039 ± 0.001
Control	Two stage	2	1.999 ± 0.000	-0.0004 ± 0.0002	0.038 ± 0.001
Anisotropic	Gaussian	–	0.478 ± 0.020	0.1092 ± 0.0056	0.073 ± 0.004
Anisotropic	Joint	2	1.973 ± 0.026	0.0016 ± 0.0063	0.066 ± 0.002
Anisotropic	Joint	10	1.985 ± 0.000	-0.0006 ± 0.0003	0.062 ± 0.001
Anisotropic	Adaptive	2	1.960 ± 0.004	0.0093 ± 0.0024	0.066 ± 0.002
Anisotropic	Latent features	2	1.033 ± 0.448	0.0639 ± 0.0368	0.100 ± 0.023
Anisotropic	Two stage	2	1.916 ± 0.073	0.0139 ± 0.0146	0.068 ± 0.003
Strong	Gaussian	–	0.446 ± 0.034	0.0993 ± 0.0096	0.082 ± 0.009
Strong	Joint	2	1.968 ± 0.001	-0.0008 ± 0.0003	0.081 ± 0.002
Strong	Joint	10	1.971 ± 0.001	-0.0008 ± 0.0003	0.078 ± 0.002
Strong	Adaptive	2	1.952 ± 0.001	0.0076 ± 0.0023	0.081 ± 0.002
Strong	Latent features	2	0.789 ± 0.142	0.0487 ± 0.0150	0.114 ± 0.009
Strong	Two stage	2	1.944 ± 0.007	0.0054 ± 0.0042	0.082 ± 0.005

To examine spectral dependence, we vary the slowest rate and the linear cutoff gap. Figure 5 and Table 9 compare the fixed joint objective with Gaussian regularization. For target selection, we keep the shared rate set $(0.98, 0.20, 0.10, 0.05)$ and permute which latent coordinate carries each rate, with $d = 4$, $m = 2$, and eight initializations per assignment. Figure 5 and Table 10 show that the joint objective moves the recovered directions with the spectrum rather than with the coordinate index: all four assignments give a target distance between 0.016 and 0.026, against 2.35 to 2.39 for Gaussian regularization alone.

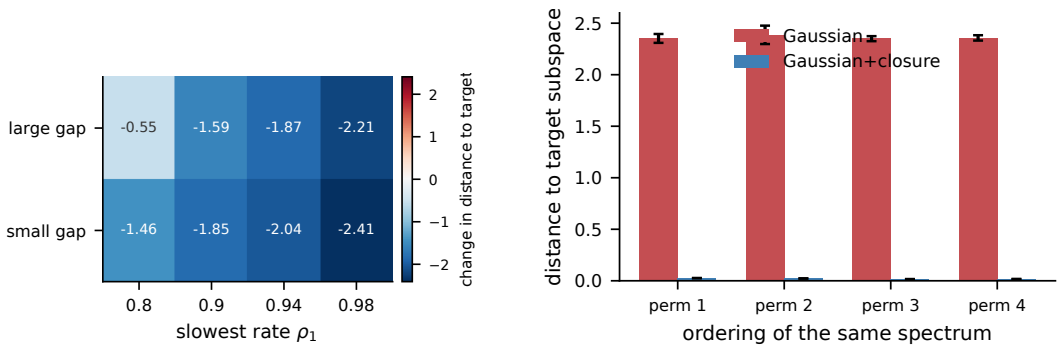

Figure 5: Left. Change in target distance after adding closure across slowest rates and linear gaps; negative values indicate improvement. Right. Target recovery under four assignments of the same rate set to the latent coordinates, with one standard deviation over eight initializations.

Table 9: Where the closure intervention changes the outcome. The slowest rate ρ_1 and the gap between the second linear rate and the strongest higher-order competitor are varied independently. The adaptive row sets the closure weight by an augmented Lagrangian targeting a fixed residual level rather than fixing it.

ρ_1	Rank gap	Objective	n	D_{target}	$\widehat{\beta}_{\min}$
0.80	0.05	Gaussian	8	1.472 ± 0.010	0.454 ± 0.003
0.80	0.05	Joint	8	0.017 ± 0.000	0.093 ± 0.000
0.80	0.05	Joint, adaptive	8	0.050 ± 0.005	0.105 ± 0.002
0.80	0.28	Gaussian	8	0.558 ± 0.432	0.394 ± 0.078
0.80	0.28	Joint	8	0.005 ± 0.000	0.293 ± 0.000
0.80	0.28	Joint, adaptive	8	0.008 ± 0.003	0.293 ± 0.001
0.90	0.05	Gaussian	8	1.873 ± 0.006	0.644 ± 0.002
0.90	0.05	Joint	8	0.019 ± 0.001	0.093 ± 0.000
0.90	0.05	Joint, adaptive	8	0.041 ± 0.003	0.104 ± 0.002
0.90	0.28	Gaussian	8	1.601 ± 0.006	0.651 ± 0.002
0.90	0.28	Joint	8	0.007 ± 0.000	0.293 ± 0.000
0.90	0.28	Joint, adaptive	8	0.033 ± 0.013	0.302 ± 0.004
0.94	0.05	Gaussian	8	2.064 ± 0.007	0.724 ± 0.001
0.94	0.05	Joint	8	0.021 ± 0.001	0.093 ± 0.000
0.94	0.05	Joint, adaptive	8	0.041 ± 0.003	0.104 ± 0.002
0.94	0.28	Gaussian	8	1.883 ± 0.006	0.745 ± 0.002
0.94	0.28	Joint	8	0.009 ± 0.000	0.293 ± 0.000
0.94	0.28	Joint, adaptive	8	0.035 ± 0.010	0.304 ± 0.004
0.98	0.05	Gaussian	8	2.441 ± 0.057	0.823 ± 0.020
0.98	0.05	Joint	8	0.032 ± 0.001	0.093 ± 0.000
0.98	0.05	Joint, adaptive	8	0.049 ± 0.001	0.104 ± 0.002
0.98	0.28	Gaussian	8	2.304 ± 0.034	0.829 ± 0.006
0.98	0.28	Joint	8	0.094 ± 0.222	0.317 ± 0.068
0.98	0.28	Joint, adaptive	8	0.039 ± 0.005	0.305 ± 0.003

Table 10: Recovery under permuted spectral orderings. The same four transition rates are assigned to different latent coordinates; the target follows the spectrum rather than the coordinate index.

Latent rate assignment	Objective	n	D_{target}	C	$\widehat{\beta}_{\min}$
$(0.05, 0.20, 0.10, 0.98)$	Gaussian	8	2.352 ± 0.043	1.685 ± 0.046	0.831 ± 0.026
$(0.05, 0.20, 0.10, 0.98)$	Joint	8	0.026 ± 0.001	1.173 ± 0.000	0.194 ± 0.000
$(0.10, 0.20, 0.98, 0.05)$	Gaussian	8	2.387 ± 0.089	1.668 ± 0.003	0.816 ± 0.002
$(0.10, 0.20, 0.98, 0.05)$	Joint	8	0.022 ± 0.000	1.174 ± 0.000	0.195 ± 0.000
$(0.20, 0.98, 0.10, 0.05)$	Gaussian	8	2.350 ± 0.024	1.665 ± 0.002	0.816 ± 0.001
$(0.20, 0.98, 0.10, 0.05)$	Joint	8	0.016 ± 0.001	1.171 ± 0.000	0.192 ± 0.000
$(0.98, 0.20, 0.10, 0.05)$	Gaussian	8	2.357 ± 0.026	1.668 ± 0.004	0.818 ± 0.002
$(0.98, 0.20, 0.10, 0.05)$	Joint	8	0.017 ± 0.001	1.173 ± 0.000	0.194 ± 0.000

Recovery across a family of spectra. The main comparison fixes one spectrum, so we repeat it over a grid. We sweep $\rho_1 \in \{0.95, 0.98, 0.99\}$ and $\rho_2 \in \{0.10, 0.20, 0.30\}$, setting $\rho_3 = \rho_2/2$ and $\rho_4 = \rho_2/4$, with eight initializations in each of the nine cells. Table 11 reports the result. Gaussian regularization alone leaves a target distance between 1.95 and 2.90 in every cell, while adding closure brings it to between 0.004 and 0.037. The recovered correlation tracks $C_{\text{lin}} = \rho_1 + \rho_2$ and the minimum transition eigenvalue tracks ρ_2, both to within 0.01, and source concentration returns to

the balanced value 0.5. The margin over Gaussian regularization widens as ρ_1 increases and as ρ_2 decreases, which is the direction in which Theorem 2 predicts a larger reuse advantage. No single cell carries the comparison.

Table 11: Recovery across a family of synthetic transition spectra, with $\rho_3 = \rho_2/2$ and $\rho_4 = \rho_2/4$. The joint objective is Gaussian regularization with closure. Entries are means and sample standard deviations over eight initializations.

ρ_1	ρ_2	C_{lin}	D_{target} Gaussian	D_{target} joint	C joint	$\widehat{\beta}_{\mathrm{min}}$ joint	S_{max} joint	Runs
0.95	0.10	1.05	2.146 ± 0.045	0.019 ± 0.002	1.044 ± 0.000	0.095 ± 0.000	0.498 ± 0.000	8
0.95	0.20	1.15	2.041 ± 0.008	0.009 ± 0.001	1.144 ± 0.000	0.194 ± 0.000	0.499 ± 0.000	8
0.95	0.30	1.25	1.948 ± 0.010	0.004 ± 0.000	1.244 ± 0.000	0.294 ± 0.000	0.499 ± 0.000	8
0.98	0.10	1.08	2.433 ± 0.049	0.028 ± 0.002	1.074 ± 0.000	0.095 ± 0.000	0.494 ± 0.000	8
0.98	0.20	1.18	2.357 ± 0.026	0.017 ± 0.001	1.173 ± 0.000	0.194 ± 0.000	0.495 ± 0.000	8
0.98	0.30	1.28	2.281 ± 0.012	0.010 ± 0.001	1.274 ± 0.000	0.294 ± 0.000	0.496 ± 0.000	8
0.99	0.10	1.09	2.902 ± 0.163	0.037 ± 0.003	1.084 ± 0.000	0.095 ± 0.000	0.491 ± 0.000	8
0.99	0.20	1.19	2.781 ± 0.113	0.024 ± 0.002	1.184 ± 0.000	0.194 ± 0.000	0.493 ± 0.000	8
0.99	0.30	1.29	2.587 ± 0.093	0.016 ± 0.001	1.284 ± 0.000	0.294 ± 0.000	0.494 ± 0.000	8

The pixel intervention retains the renderer and CNN above and uses $\rho = (0.98, 0.20, 0.10, 0.05)$, with $m = 2$ and $C_{\mathrm{lin}} = 1.18$. The larger separation between the slowest factor and the remaining factors gives the competitor using a single source a positive advantage over the linear target. We sweep the Gaussian weight over $\lambda \in \{0.05, 0.30, 0.45, 0.60, 0.75, 0.85\}$ with the closure weight fixed at $\lambda_Z = 2$, five initializations in every cell. Tables 12 and 13 report the sweep and Figure 6 plots it. Except at $\lambda = 0.05$, where the distributional constraint is too weak for either objective, adding closure holds the target distance between 0.12 and 0.33 across the whole range, keeps source concentration near the balanced value 0.5, and keeps linear energy above 0.86. The minimum transition eigenvalue stays near the target value 0.20. Recovery therefore does not depend on a particular Gaussian weight.

Table 12: Pixel closure intervention across Gaussian weights at $\rho = (0.98, 0.20, 0.10, 0.05)$, $m = 2$, and $C_{\mathrm{lin}} = 1.18$. Every cell uses five initializations and the joint objective uses a fixed closure weight $\lambda_Z = 2$.

λ	Objective	λ_Z	n	D_{target}	C	$\widehat{\beta}_{\mathrm{min}}$
0.05	Gaussian	–	5	2.238 ± 0.005	1.844 ± 0.002	0.901 ± 0.003
0.05	Joint	2	5	2.226 ± 0.005	1.849 ± 0.000	0.904 ± 0.001
0.30	Gaussian	–	5	2.299 ± 0.122	1.670 ± 0.060	0.793 ± 0.035
0.30	Joint	2	5	0.116 ± 0.010	1.175 ± 0.006	0.196 ± 0.006
0.45	Gaussian	–	5	2.248 ± 0.062	1.599 ± 0.007	0.765 ± 0.011
0.45	Joint	2	5	0.119 ± 0.007	1.166 ± 0.002	0.187 ± 0.002
0.60	Gaussian	–	5	1.475 ± 0.712	1.422 ± 0.146	0.614 ± 0.211
0.60	Joint	2	5	0.150 ± 0.015	1.161 ± 0.002	0.185 ± 0.002
0.75	Gaussian	–	5	0.865 ± 0.661	1.117 ± 0.105	0.318 ± 0.192
0.75	Joint	2	5	0.209 ± 0.073	1.150 ± 0.011	0.180 ± 0.011
0.85	Gaussian	–	5	1.022 ± 0.468	0.905 ± 0.073	0.207 ± 0.064
0.85	Joint	2	5	0.328 ± 0.232	1.137 ± 0.019	0.171 ± 0.019

Table 13: Companion diagnostics for the pixel Gaussian weight sweep. E_{lin}/m is the linear energy fraction.

λ	Objective	n	SW	$\widehat{\epsilon}_h$	E_{lin}/m	S_{max}
0.05	Gaussian	5	0.194 ± 0.006	0.0038 ± 0.0005	0.379 ± 0.002	0.998 ± 0.000
0.05	Joint	5	0.193 ± 0.004	0.0030 ± 0.0006	0.384 ± 0.002	0.998 ± 0.000
0.30	Gaussian	5	0.109 ± 0.016	0.1264 ± 0.0552	0.279 ± 0.057	0.881 ± 0.054
0.30	Joint	5	0.098 ± 0.003	-0.0014 ± 0.0010	0.960 ± 0.005	0.512 ± 0.006
0.45	Gaussian	5	0.091 ± 0.005	0.1690 ± 0.0160	0.265 ± 0.007	0.823 ± 0.007
0.45	Joint	5	0.096 ± 0.008	-0.0014 ± 0.0011	0.957 ± 0.003	0.506 ± 0.001
0.60	Gaussian	5	0.088 ± 0.013	0.1540 ± 0.0674	0.468 ± 0.240	0.704 ± 0.110
0.60	Joint	5	0.086 ± 0.015	-0.0014 ± 0.0010	0.942 ± 0.007	0.505 ± 0.001
0.75	Gaussian	5	0.083 ± 0.012	0.0846 ± 0.0743	0.652 ± 0.244	0.515 ± 0.093
0.75	Joint	5	0.079 ± 0.014	-0.0008 ± 0.0022	0.914 ± 0.035	0.503 ± 0.003
0.85	Gaussian	5	0.084 ± 0.005	0.0843 ± 0.0335	0.572 ± 0.175	0.409 ± 0.042
0.85	Joint	5	0.077 ± 0.008	0.0003 ± 0.0029	0.864 ± 0.096	0.500 ± 0.003

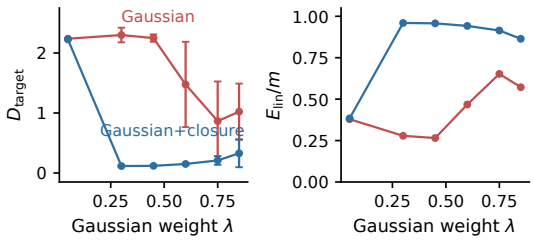

Figure 6: Pixel recovery across Gaussian weights at fixed closure weight $\lambda_Z = 2$. Left, squared distance to the target subspace; right, the linear energy fraction. Error bars show one sample standard deviation over five initializations.

We add four objective controls to the fixed $\lambda = 0.60$ pixel comparison, with five initializations each. Tables 14, 15, and 16 report all six objectives. Whitening and closure alone yield nearly identical target distances of about 1.98, with closure diagnostics near zero and source concentration near one. These measurements are consistent with selection of a nonlinear invariant subspace concentrated on the slow source. Such subspaces are allowed by transition closure, as illustrated by $\mathrm{span}\{Z_1, H_2(Z_1)\}$. The variance and covariance control and the orthogonality control exhibit similar concentration and large target distances. Across these controls, substantial Gaussian error accompanies small closure error, showing why the distributional condition in Theorem 3 is essential.

The Gaussian and joint objectives have similar sliced Gaussian errors at $\lambda = 0.60$, while closure reduces target distance from 1.475 to 0.150 and brings correlation to 1.161, close to $C_{\text{lin}} = 1.18$. The joint representation also has higher linear energy, 0.942 against 0.468, and lower source concentration, 0.505 against 0.704. Its minimum transition eigenvalue is approximately 0.185, close to the target value $\rho_2 = 0.20$. This decrease reflects recovery of the faster target direction. Together, the objective controls support the roles of Gaussianity, transition closure, and alignment in structural identification and target selection.

Table 14: Pixel objective controls at $\lambda = 0.60$ and $\lambda_Z = 2$, five initializations each.

Objective	n	D_{target}	C	$\widehat{\beta}_{\min}$
Whitening	5	1.978 ± 0.003	1.919 ± 0.002	0.939 ± 0.002
Closure only	5	1.978 ± 0.003	1.919 ± 0.001	0.939 ± 0.002
Variance/covariance	5	2.010 ± 0.003	1.932 ± 0.000	0.951 ± 0.000
Orthogonality	5	2.039 ± 0.004	1.932 ± 0.000	0.953 ± 0.000
Gaussian	5	1.475 ± 0.712	1.422 ± 0.146	0.614 ± 0.211
Joint	5	0.150 ± 0.015	1.161 ± 0.002	0.185 ± 0.002

Table 15: Distribution, closure, and allocation diagnostics for the pixel objective controls. SW is the sliced Gaussian error; $\widehat{\epsilon}_h$ is the closure diagnostic estimated from observations, retained with sign near zero.

Objective	n	SW	$\widehat{\epsilon}_h$	E_{lin}/m	$S_{\max}$
Whitening	5	0.453 ± 0.003	-0.0002 ± 0.0015	0.502 ± 0.000	0.995 ± 0.000
Closure only	5	0.454 ± 0.003	-0.0002 ± 0.0015	0.502 ± 0.000	0.995 ± 0.000
Variance/covariance	5	0.373 ± 0.003	-0.0011 ± 0.0009	0.483 ± 0.002	1.000 ± 0.000
Orthogonality	5	0.372 ± 0.002	-0.0012 ± 0.0009	0.468 ± 0.003	0.999 ± 0.000
Gaussian	5	0.088 ± 0.013	0.1540 ± 0.0674	0.468 ± 0.240	0.704 ± 0.110
Joint	5	0.086 ± 0.015	-0.0014 ± 0.0010	0.942 ± 0.007	0.505 ± 0.001

Table 16: Linear factor probes for the pixel objective controls. Entries are test R^2 for the two target factors.

Objective	n	x	y
Whitening	5	0.999 ± 0.000	0.000 ± 0.000
Closure only	5	0.999 ± 0.000	0.000 ± 0.000
Variance/covariance	5	0.960 ± 0.003	0.000 ± 0.000
Orthogonality	5	0.931 ± 0.005	0.000 ± 0.000
Gaussian	5	0.651 ± 0.152	0.281 ± 0.332
Joint	5	0.974 ± 0.007	0.915 ± 0.022

Table 17 gives the linear factor probes across the same sweep. At $\lambda = 0.30$ closure reduces target distance from 2.299 to 0.116 and brings correlation to 1.175, close to the linear reference, with the second target factor reaching a test R^2 of 0.934. At $\lambda = 0.05$ both objectives retain substantial target error alongside a small closure diagnostic, which is the pattern Theorem 3 predicts when only the dynamical constraint is active.

At $\lambda = 0.30$, closure reduces target distance from 2.299 to 0.116, while correlation reaches 1.175, close to the linear reference. The second target factor attains a test probe R^2 of 0.934 at this weight. Recovery therefore occurs at two tested Gaussian weights. At $\lambda = 0.05$, both objectives retain substantial target error alongside a small closure diagnostic. This pattern is consistent with the joint conditions of Theorem 3, which require both the distributional and dynamical constraints. The supplementary results at $\lambda = 0.45$ show the same recovery direction, with their reported counts determining the available precision. The weight comparisons support the observed mechanism at the tested settings.

Table 17: Linear factor probes across the pixel Gaussian weight sweep. Entries are test R^2.

λ	Objective	n	x	y
0.05	Gaussian	5	0.753 ± 0.006	0.000 ± 0.000
0.05	Joint	5	0.764 ± 0.004	0.000 ± 0.000
0.30	Gaussian	5	0.542 ± 0.114	0.015 ± 0.005
0.30	Joint	5	0.993 ± 0.001	0.934 ± 0.011
0.45	Gaussian	5	0.495 ± 0.009	0.035 ± 0.011
0.45	Joint	5	0.986 ± 0.004	0.934 ± 0.009
0.60	Gaussian	5	0.651 ± 0.152	0.281 ± 0.332
0.60	Joint	5	0.974 ± 0.007	0.915 ± 0.022
0.75	Gaussian	5	0.716 ± 0.166	0.592 ± 0.326
0.75	Joint	5	0.962 ± 0.008	0.870 ± 0.080
0.85	Gaussian	5	0.580 ± 0.171	0.570 ± 0.186
0.85	Joint	5	0.953 ± 0.007	0.777 ± 0.196

F.7 Conditional Mean Calibration

For an explicit linear mixture $h = (\cos\theta Z_1 + \sin\theta Z_2, Z_3)$, Proposition 1 gives

$$\epsilon_h = \cos^2\theta \sin^2\theta (\rho_1 - \rho_2)^2.$$

We compare this value with a latent Hermite diagnostic and a cross-fitted observation-based estimator using 1024 random features. Figure 7 shows the response across the three synthetic spectra. Both estimates track the growing defect on this controlled family. Finite-feature estimates quantify residual predictability within their fitted function class, providing a calibration reference for the empirical closure diagnostic.

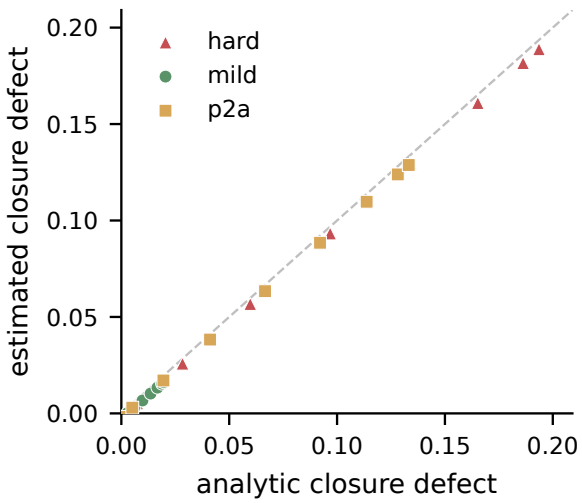

Figure 7: Conditional mean estimates against the analytic closure defect. Filled markers use latent Hermite features; open markers use random features of observations. The diagonal is exact agreement.

F.8 Controlled Stability and Predictability

For $\rho = (0.9, 0.7, 0.3)$, consider

$$h_\alpha = (\sqrt{1-\alpha^2}Z_1 + \alpha H_2(Z_1), Z_2), \qquad 0 \le \alpha \le 0.8.$$

Exact Hermite identities and Gauss-Hermite quadrature with 80 nodes give

$$s(h_\alpha) = \alpha^2, \qquad D_{\mathrm{lin}}(h_\alpha) = 2(1-\sqrt{1-\alpha^2}),$$
$$\epsilon_{h_\alpha} = \alpha^2(1-\alpha^2)(0.9-0.9^2)^2, \qquad \Delta_4 = 24\alpha^2 - 12\alpha^4.$$

Here $\Delta_4 = \|\mathbb{E}[h_\alpha^{\otimes 4}] - \mathbb{E}[G^{\otimes 4}]\|_F$ for $G \sim \mathcal{N}(0, I_2)$. The smallest transition eigenvalue is at least 0.7 throughout. Figure 8A uses $\Delta_4^{1/2} + \epsilon_h^{1/4}$, consistent with the fourth moment extension in Appendix E.3. Coupling to (Z_1, Z_2) also gives

$$W_2(\mathrm{Law}(h_\alpha), \mathcal{N}(0, I_2)) \le \sqrt{D_{\mathrm{lin}}(h_\alpha)},$$

so Gaussian error tends to zero along the same path. As the nonlinear component decreases, nonlinear energy and exact best linear distance decrease with the fourth moment and closure error scale (Figure 8A). The uniform lower bound on predictability makes this path a controlled illustration of Theorem 4(b). Figure 8A distinguishes the exact minimum distance from its bound $2s(h)$.

The high-frequency construction in Appendix E.1 preserves the exact Gaussian law. Numerical evaluation uses a approximation using 26 bits at frequencies $1, 2, 4, 8, 16, 32, 64$. We estimate $\|Th\|^2$ through the cross-product of conditionally independent successors and compute linear distance from the cross moment matrix. At high frequencies, the closure residual decreases together with the minimum transition eigenvalue, while best linear distance approaches its maximum (Figure 8B,C). This illustrates Theorem 4(a). Exact Gaussianity and a small closure residual coexist with nonlinear structure as predictability vanishes.

A complementary resonance construction illustrates Lemma D.1. For $\rho = (0.8, 0.64)$, the scalar function $F_t = \sqrt{1-t^2}Z_2 + tH_2(Z_1)$ is a transition eigenfunction with eigenvalue 0.64. Its span is exactly closed for every t, while $s(F_t) = t^2$ and $\mathbb{E}F_t^4 - 3 = 12t^4$. Gaussian fourth moment matching selects the linear endpoint, illustrating the structural contribution of distributional control within a closed span.

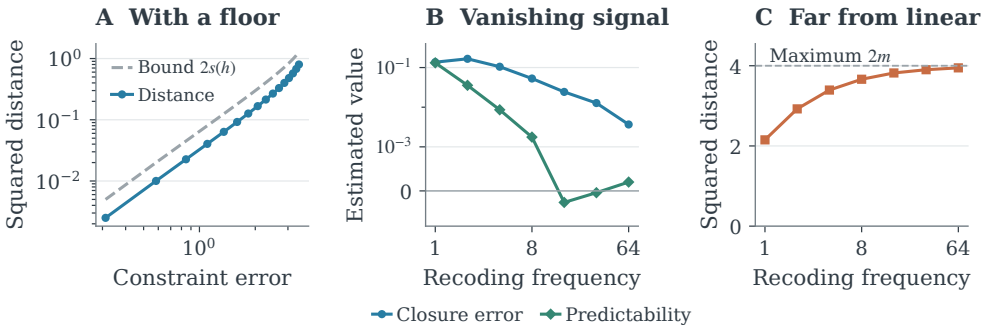

Figure 8: Controlled stability and predictability. (A) The nonlinear energy bound $2s(h)$ and exact best linear squared distance along a mixture path with a uniform predictability floor. The error scale uses fourth moment discrepancy Δ_4 and closure error ϵ_h. (B) High-frequency recoding reduces closure error and predictability; a symmetric logarithmic axis retains signed empirical estimates near zero. (C) The same exactly Gaussian construction approaches the maximum squared linear distance $2m = 4$.

Table 18: Exact quantities along the controlled mixture path. $\lambda_{\min}(B_h) = 0.7$ throughout; the distance column gives the exact minimum, and $2s(h)$ is its separate upper bound.

α	Δ_4	ϵ_h	$s(h)$	D_{lin}	$2s(h)$
0.00	0.0000	0.000000	0.0000	0.0000	0.0000
0.05	0.0599	0.000020	0.0025	0.0025	0.0050
0.10	0.2388	0.000080	0.0100	0.0100	0.0200
0.15	0.5339	0.000178	0.0225	0.0226	0.0450
0.20	0.9408	0.000311	0.0400	0.0404	0.0800
0.25	1.4531	0.000475	0.0625	0.0635	0.1250
0.30	2.0628	0.000663	0.0900	0.0921	0.1800
0.35	2.7599	0.000871	0.1225	0.1265	0.2450
0.40	3.5328	0.001089	0.1600	0.1670	0.3200
0.45	4.3679	0.001308	0.2025	0.2139	0.4050
0.50	5.2500	0.001519	0.2500	0.2679	0.5000
0.55	6.1619	0.001709	0.3025	0.3297	0.6050
0.60	7.0848	0.001866	0.3600	0.4000	0.7200
0.65	7.9979	0.001976	0.4225	0.4801	0.8450
0.70	8.8788	0.002024	0.4900	0.5717	0.9800
0.75	9.7031	0.001993	0.5625	0.6771	1.1250
0.80	10.4448	0.001866	0.6400	0.8000	1.2800

Table 19: Gaussian recoding at high frequency. Exact Gaussianity is a property of the construction; transition and distance entries are numerical estimates.

Frequency	$\widehat{\epsilon}_h$	$\widehat{\beta}_{\min}$	$s(h)$	D_{lin}
1	0.13667	0.13171	1.14596	2.15171
2	0.16621	0.03565	1.71071	2.92428
4	0.10495	0.00860	1.90811	3.39375
8	0.05294	0.00176	1.97165	3.66324
16	0.02432	-0.00026	1.99190	3.82002
32	0.01273	-0.00004	1.99755	3.90102
64	0.00367	0.00020	1.99930	3.94727

F.9 Stability Across Trained Runs

The two paths above are controlled constructions in which the constraint error is set by hand. To check that the same relation holds for representations reached by optimization, we pool the 136 trained runs on the shared spectrum and compare the combined constraint error $\Delta_4^{1/2} + \widehat{\epsilon}_h^{1/4}$ with the exact distance to the nearest linear representation, $D_{\text{lin}} = 2m - 2\|M_h\|_*$. On a logarithmic scale the two are correlated at 0.92. Splitting the runs by combined error, the lowest quartile has median $D_{\text{lin}} = 0.017$ and the highest quartile has median 1.993, close to the value attained by a nonlinear invariant subspace. Trained runs with small constraint error therefore sit near the linear

region, as Theorem 4(b) requires, and this holds across objectives and spectra rather than along a single designed path.

Runs that degenerate supply the matching evidence for Theorem 4(a). On the reversible chains below, the variance-covariance and orthogonality controls produce 16 runs whose closure defect is near zero, about 0.003, but whose predictability floor collapses to a median of 0.039 and 0.015 respectively; their median distance to the two leading transition modes is 1.85 and 1.88. Whitening on the same chains reaches a closure defect of 0.0000 with a median $\widehat{b}_{\min}$ of 0.46 and a distance of 0.0005. Small closure error alone therefore does not imply proximity to a linear representation; the predictability floor is what separates the two groups, which is the hypothesis Theorem 4(a) isolates.

F.10 Cost of the Closure Term

The closure term adds a critic and three critic ascent steps per encoder update. Measured on the same device, precision and batch size, with a warm-up and repeated timings, it multiplies wall-clock per update by 1.33 on the synthetic setting, where the critic is a small multilayer perceptron, and by 3.3 on pixels, where the critic matches the encoder architecture; trainable parameters grow by 1.13 and peak memory by 1.17 on the synthetic setting. To check that the improvement is not bought with that extra compute, we retrain the Gaussian objective alone on pixels at two, three and five times its original budget, with the cosine schedule covering each extended horizon. Target distance does not improve but degrades monotonically, from 1.475 at the original budget to 1.896, 2.110 and 2.196, while source concentration rises from 0.704 to 0.815; the joint objective reaches 0.150 at the original budget. Additional compute therefore drives the Gaussian objective further into the single-source solution of Theorem 2 rather than towards the target.

F.11 Robustness of the Modelling Assumptions

Perturbed assumptions. The synthetic model assumes a diagonal transition, exactly Gaussian latent variables, noiseless observation, and full observation. We relax each in turn on the spectrum $(0.98, 0.10, 0.05)$, holding the encoder, objective weights, and budget fixed, with eight initializations per cell. Table 20 reports the result. Rotating the transition by 0.2, 0.4, and 0.8 radians leaves the joint target distance at 0.029, 0.026, and 0.024, against an unperturbed reference of 0.032. Observation noise of standard deviation 0.05 and 0.15 gives 0.037 and 0.069, and a uniform latent law gives 0.097. Under a Laplace latent law the joint objective succeeds on five of eight initializations and fails on the remaining three, giving 0.93 ± 1.07; the heavier tails interact with the Gaussian term rather than with closure. Recovery also survives hiding one, two, or three latent coordinates from the observation, and breaks only when the fourth is hidden, at which point the target itself is no longer identifiable from the observation and both objectives coincide at 1.50. Mild departures from the modelling assumptions therefore preserve the mechanism.

General reversible dynamics. The theory is stated for reversible transition operators, of which the Gaussian autoregressive model is one instance. We replace the latent dynamics with a product of two independent reversible Markov chains, each built as $P = D^{-1}W$ from a random symmetric weight matrix on a discrete state space, so the transition spectrum is computed numerically rather than given in closed form. The slower chain has second eigenvalue 0.458. Sweeping the leading eigenvalue λ_1 of the second chain across that value, whitening places the second representation direction on the slow chain's second mode for $\lambda_1 \leq 0.42$ and on the second chain's leading mode for $\lambda_1 \geq 0.46$, with projection energy above 0.997 on the selected mode in every one of the eight initializations (Table 21). The switch occurs where the numerical spectrum crosses, not at any value supplied to the training procedure. Table 22 reports objectives against a factorised target taking one leading eigenmode from each chain: whitening and closure alone concentrate on a single chain and leave a distance of 1.99, Gaussian regularization reaches 0.99, and adding closure reaches 0.51. The same ordering holds on the other two stationary laws we tried, with the joint objective reaching 0.37–0.44 under a Gaussian-shaped stationary law and 0.087–0.089 under a skewed one. Selection by transition spectrum and the direction of the closure intervention therefore extend beyond the Gaussian autoregressive setting. We note that on a non-Gaussian stationary law the Gaussian term is itself misspecified, so its role here is weaker than in the setting the theory assumes; the factorised target is correspondingly only partly reached.

Partial observation. When the observation does not determine the latent state, the closure defect an observer can estimate and the defect defined on the latent variables need not agree. On a small construction in which one latent coordinate is folded out of view, the observable estimate is 0.0002 to 0.0018 while the latent one is 0.0048 to 0.0163, so an observer underestimates by a factor of three to nine. The gap appears only when the hidden content is predictable from what remains visible: under a symmetric folding, for which the visible view is itself closed, the two estimates agree to within 0.0005. Non-injective observation therefore does not by itself create a blind spot.

Table 20: Assumption robustness on the synthetic spectrum $(0.98, 0.10, 0.05)$. Each block perturbs one modelling assumption while holding the encoder, objective weights, and budget fixed. Runs that did not produce a finite evaluation are excluded and the remaining count is given.

Perturbation	Level	D_{target} Gaussian	D_{target} joint	$\widehat{\beta}_{\min}$ joint	Runs
unperturbed reference	0.0	2.435 ± 0.024	0.032 ± 0.001	0.094 ± 0.000	8
transition rotation (rad)	0.2	2.374 ± 0.105	0.029 ± 0.000	0.094 ± 0.000	8
transition rotation (rad)	0.4	2.492 ± 0.325	0.026 ± 0.001	0.094 ± 0.000	8
transition rotation (rad)	0.8	2.482 ± 0.066	0.024 ± 0.001	0.094 ± 0.000	8
observation noise s.d.	0.05	2.363 ± 0.055	0.037 ± 0.001	0.094 ± 0.000	8
observation noise s.d.	0.15	2.171 ± 0.022	0.069 ± 0.001	0.093 ± 0.000	8
observation noise s.d.	0.4	1.792 ± 0.040	0.259 ± 0.001	0.085 ± 0.000	8
latent law	uniform	2.430 ± 0.174	0.097 ± 0.001	0.119 ± 0.000	8
latent law	laplace	2.239 ± 0.050	0.927 ± 1.071	0.354 ± 0.345	8
hidden latent coordinates	1	2.485 ± 0.069	0.048 ± 0.007	0.093 ± 0.000	8
hidden latent coordinates	2	2.513 ± 0.123	0.057 ± 0.011	0.091 ± 0.001	8
hidden latent coordinates	3	2.311 ± 0.092	0.164 ± 0.010	0.087 ± 0.000	8
hidden latent coordinates	4	1.500 ± 0.022	1.497 ± 0.025	0.106 ± 0.001	8

Table 21: Mode selection under a product of two reversible Markov chains. The slower chain has second eigenvalue 0.458; the competing mode is the leading eigenmode of the second chain. E denotes projection energy of the second representation direction.

λ_1 (second chain)	$E_{\mathrm{slow},2}$ wh.	$E_{\mathrm{fast},1}$ wh.	$E_{\mathrm{slow},2}$ Gauss.	$E_{\mathrm{fast},1}$ Gauss.	Runs
0.30	1.001 ± 0.001	0.000 ± 0.000	0.465 ± 0.406	0.481 ± 0.385	8
0.38	1.000 ± 0.001	0.000 ± 0.000	0.131 ± 0.276	0.797 ± 0.258	8
0.42	1.000 ± 0.001	0.000 ± 0.000	0.090 ± 0.224	0.833 ± 0.205	8
0.46	0.001 ± 0.001	0.999 ± 0.001	0.070 ± 0.185	0.849 ± 0.167	8
0.50	0.000 ± 0.000	1.000 ± 0.001	0.055 ± 0.151	0.859 ± 0.132	8
0.58	0.000 ± 0.000	1.000 ± 0.001	0.031 ± 0.087	0.872 ± 0.070	8
0.70	0.000 ± 0.000	1.000 ± 0.001	0.013 ± 0.037	0.865 ± 0.025	8

Table 22: Objectives on a product of two reversible chains with bimodal stationary law. $D_{\mathrm{factorised}}$ targets one leading eigenmode from each chain; D_{top} targets the two leading transition modes overall. Objectives evaluated only against D_{top} are marked with a dash. Runs pool the swept values of λ_1.

Objective	$D_{\mathrm{factorised}}$	D_{top}	$\widehat{\beta}_{\min}$	$\widehat{\epsilon}_h$	Runs
whitening	1.990 ± 0.007	-0.003 ± 0.000	0.457 ± 0.000	0.0000 ± 0.0000	32
closure only	1.987 ± 0.008	-0.003 ± 0.000	0.457 ± 0.000	0.0000 ± 0.0000	32
Gaussian	0.988 ± 0.746	1.144 ± 0.792	0.419 ± 0.106	0.0993 ± 0.0218	48
Gaussian+closure	0.515 ± 0.422	1.677 ± 0.764	0.366 ± 0.137	0.0827 ± 0.0159	48
VICReg	—	1.815 ± 0.181	0.040 ± 0.017	0.0033 ± 0.0037	8
orthogonality	—	1.847 ± 0.136	0.020 ± 0.014	0.0028 ± 0.0034	8