BREAKING THE ROTATION BARRIER IN JEPA WORLD MODELS WITH ANISOTROPIC PREDICTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint-Embedding Predictive Architectures (JEPAs) learn world models by predicting in representation space, but their predictor can absorb any rotation of that space without changing the loss, leaving physical factors mixed across coordinates and hard to monitor or control individually. Recovering one factor per coordinate requires the right subspace, which contains the factors, and the right basis, which gives each factor its own coordinate. We break this rotation barrier with a minimal change, replacing the dense predictor by an anisotropic predictor (**AP-JEPA**), and prove that its loss splits exactly into two terms: a rotation-invariant ***dense-predictor term***, which selects the subspace, and a ***cross-talk term***, which selects the basis and aligns coordinates with factors that evolve on distinct timescales. However, the encoder can produce features that are more predictable than genuine factors, such as the square of a slow factor, so the selected subspace may miss some factors. We characterize when two existing regularizers, whitening and SIGReg, rule out such features, and propose a new regularizer, PairSIGReg, which constrains consecutive representations jointly and recovers the factor subspace for every spectrum of distinct timescales. Across six dynamical and image benchmarks, AP-JEPA with PairSIGReg raises the average mean correlation coefficient (MCC) from 63.6% for the strongest baseline to 77.8%, while also achieving the highest average linear readout (90.2%). Beyond training from scratch, AP-JEPA also applies to existing JEPAs trained with dense predictors: since the cross-talk term depends only on the basis, minimizing it alone yields a closed-form change of coordinates that raises the single-coordinate R^2 of released I-JEPA and V-JEPA models from about 10% to 81% without information loss.

1 INTRODUCTION

Joint-Embedding Predictive Architectures (JEPAs) learn representations by predicting in embedding space rather than pixel space (LeCun, 2022). The approach has proven effective across images, video, audio and point clouds (Assran et al., 2023; Bardes et al., 2024; Fei et al., 2024; Saito et al., 2025), and in world models that plan in embedding space (Assran et al., 2025; Zhou et al., 2025; Sobal et al., 2025; Maes et al., 2026). LeJEPA recently put its training on a provable footing with an isotropic Gaussian regularizer, SIGReg (Balestriero & LeCun, 2025). These objectives share a blind spot: they do not change when the representation space is rotated. For representations y, y' of consecutive observations, a dense linear predictor W gives the prediction loss L_{dense}, and any rotation Q of the representation is absorbed into the predictor:

$$L_{\text{dense}} = \min_{W} \mathbb{E}\|y' - Wy\|^2, \qquad \left\|Qy' - (QWQ^\top)\,Qy\right\|^2 = \left\|y' - Wy\right\|^2. \tag{1}$$

The isotropic regularizers that prevent collapse are unchanged by Q as well, so the learned coordinates are fixed only up to an arbitrary rotation, as proven for LeJEPA and for action-conditioned world models (Klindt et al., 2026; Zhang et al., 2026), and each coordinate can mix several physical factors. We call this symmetry the *rotation barrier*.

The rotation barrier leaves JEPA representations non-identifiable. Identifiability, the recovery of the true latent factors up to permutation and sign, is a basic theoretical requirement in representation learning (Locatello et al., 2019; Khemakhem et al., 2020; Zimmermann et al., 2021): without it, the objective cannot distinguish the true factors from their mixtures. The consequence is also practical:

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.

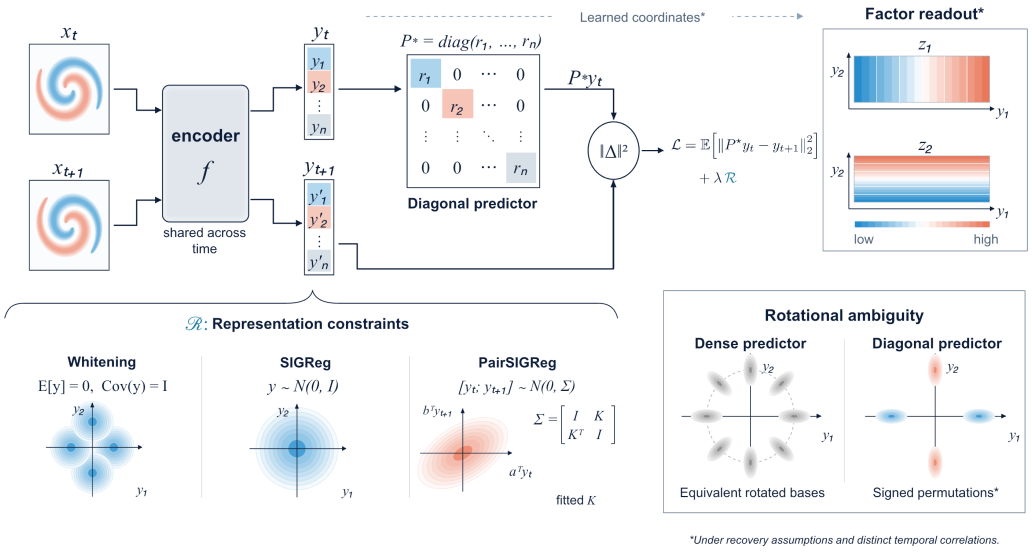


Figure 1: **Anisotropic prediction and representation constraints.** AP-JEPA fits one coefficient per coordinate. When these coefficients differ, the cross-talk term selects the basis, while the regularizer constrains which features can enter the subspace. Factor recovery requires the factor subspace and distinct temporal correlations under the stated assumptions.

when physical factors are mixed across coordinates, monitoring or controlling any one of them requires a decoder trained on labels, which need not transfer between models because the rotation is arbitrary, whereas a factor with its own coordinate can be calibrated from a handful of labels.

Recovering one factor per coordinate requires two things: the representation must contain the factors (the right *subspace*), and each coordinate must carry a single factor (the right *basis*). With a shoulder and a wrist angle, the coordinates (shoulder, wrist) and (shoulder+wrist, shoulder−wrist) span the same subspace and give the same full linear readout, but only the first can be read one coordinate at a time; the rotation barrier leaves this choice of basis arbitrary. Fixing the basis requires an objective that prefers some coordinate systems over others, and the data already contain such a signal: physical factors typically evolve at different rates. A factor with lag-one correlation ρ_i is best predicted by ρ_i times itself, whereas a coordinate that mixes factors of different rates must settle for a single compromise coefficient; slow feature analysis and temporal source separation build on this cue (Wiskott & Sejnowski, 2002; Hyvärinen & Morioka, 2017; Yao et al., 2022). A dense predictor, however, sidesteps the compromise through its cross-coordinate weights. We break the barrier with a minimal change that removes them: AP-JEPA replaces the dense predictor by an anisotropic one (Figure 1), a diagonal matrix with one coefficient per coordinate,

$$L_{\mathrm{AP}} = \min_{P=\mathrm{diag}(p_1,\ldots,p_n)} \mathbb{E}\|y' - Py\|^2, \tag{2}$$

where each optimal p_j has a closed-form least-squares solution from minibatch moments, requiring neither labels nor extra parameters. Since $QPQ^\top$ is in general not diagonal, a rotation can no longer be absorbed. Under whitening and reversible dynamics, we prove that L_{AP} splits exactly into two terms,

$$L_{\mathrm{AP}} = \underbrace{L_{\mathrm{dense}}}_{\text{dense-predictor term}} + \underbrace{\textstyle\sum_{j\neq l} \mathbb{E}[y_j y'_l]^2}_{\text{cross-talk term}}. \tag{3}$$

The ***dense-predictor term*** is L_{dense} from equation 1, the loss of the best dense linear predictor: it is rotation invariant and depends only on the subspace. The ***cross-talk term*** measures how much each coordinate predicts the future of another, and depends on the basis. For two factors with lag-one correlations ρ_1, ρ_2 and coordinates rotated by θ from the factor axes,

$$\textstyle\sum_{j\neq l} \mathbb{E}[y_j y'_l]^2 = \frac{1}{2}(\rho_1 - \rho_2)^2 \sin^2(2\theta), \tag{4}$$

which vanishes only at the factor axes, and everywhere when the timescales coincide. AP thus selects the basis whenever factors evolve on distinct timescales. The two terms do not compete: a

rotation within the subspace removes the cross-talk without changing the dense-predictor term, so AP attains the same minimum predictive risk as a dense linear predictor.

However, the encoder can produce features that are more predictable than genuine factors. The square of a slow factor with $\rho_1 = 0.9$ has lag correlation $\rho_1^2 = 0.81$, above that of a fast factor with $\rho_2 = 0.5$, so the selected subspace may contain this harmonic in place of the fast factor, a known failure of nonlinear slow features (Wiskott, 2003; Sprekeler, 2011). The regularizer decides which features are admissible. We show that whitening recovers the factor subspace exactly when $\rho_{\max}^2 < \rho_{\min}$, and that SIGReg excludes the bare harmonic and extends this to a larger, but still bounded, regime of second-order stability. We then propose PairSIGReg, which requires consecutive representations to be jointly Gaussian with a fitted cross-time covariance. Combined with AP, it recovers the factors for every spectrum of distinct timescales, provided the constraints hold exactly and the encoder preserves information; under approximate constraints, the error degrades gracefully.

The split gives two ways to use AP-JEPA. Trained from scratch with PairSIGReg, it raises the average mean correlation coefficient (MCC) over six dynamical and image benchmarks from 63.6% for the strongest baseline to 77.8%, with the highest average linear readout (90.2%). When the expressive dense predictor of a practical world model is kept instead, a linear change of coordinates leaves the dense-predictor term unchanged, so minimizing the cross-talk term alone has a closed-form solution; without retraining, it raises the single-coordinate R^2 of released I-JEPA and V-JEPA models (Assran et al., 2023; Bardes et al., 2024) from about 10% to 81%.

Our contributions are:

- We identify the rotation barrier in JEPAs and propose AP-JEPA, which makes the predictor anisotropic and provably restores identifiability up to permutation and sign.
- We characterize the conditions under which existing regularizers, whitening and SIGReg, recover the factor subspace, and propose a complementary regularizer, PairSIGReg, that provides a global factor-recovery guarantee for every set of distinct timescales.
- Extensive experiments on dynamical and image benchmarks show that AP-JEPA recovers factors in individual coordinates while preserving the information needed for linear readout.

2 Related Work

JEPA world models fix the subspace, not the basis. JEPAs predict in embedding space (LeCun, 2022), training image and video encoders (Assran et al., 2023; Bardes et al., 2024) and world models that plan in latent space (Assran et al., 2025; Zhou et al., 2025; Sobal et al., 2025; Maes et al., 2026). Their anti-collapse mechanisms, a stop-gradient predictor, variance penalties, whitening or SIGReg's isotropic Gaussian target (Grill et al., 2020; Bardes et al., 2022; Ermolov et al., 2021; Balestriero & LeCun, 2025), are all rotation invariant. Theory explains *which* features JEPAs keep, predictive and slow ones (Littwin et al., 2024; Sobal et al., 2022), but LeJEPA recovers Gaussian latents only up to a rotation, also with actions (Klindt et al., 2026; Zhang et al., 2026). Off-diagonal predictor weights, which transfer features between neurons (Wen & Li, 2022), absorb this rotation; AP-JEPA removes them, keeping the encoder, regularizer and minimum predictive risk.

Breaking the rotation. Without inductive bias, factors are not identifiable (Locatello et al., 2019). Nonlinear ICA exploits temporal structure but needs non-Gaussian or nonstationary sources (Hyvärinen & Morioka, 2017; Yao et al., 2022); contrastive learning stops at a linear or orthogonal map (Zimmermann et al., 2021; Roeder et al., 2021) unless positive pairs vary factors by different amounts (Rusak et al., 2025). Recent JEPAs impose structure: a Hamiltonian geometry (Alvarez, 2026), a progression subspace (Thil et al., 2026) or a prescribed Gaussian-process kernel per channel (Wei, 2026). AP-JEPA prescribes no kernel, geometry or subspace: it learns the timescales from data and reaches signed permutations for Gaussian factors, where isotropic objectives cannot.

Slow features and spectral methods. Slow feature analysis ranks features by slowness but can prefer a harmonic of a slow factor to a fast one (Wiskott & Sejnowski, 2002; Sprekeler, 2011). Neural spectral methods face the same rotation symmetry and break it with a mechanism designed by hand for a known operator, such as a Laplacian: masked gradients, a sequential game or prescribed per-dimension weights (Pfau et al., 2020; Gemp et al., 2021; Wang et al., 2021). AP-JEPA

trains a nonlinear encoder with an asymmetry learned from data, whose coefficients converge to the timescale spectrum, and shows which regularizer keeps harmonics out of the learned span.

3 PRELIMINARIES

A latent state $z \in \mathbb{R}^n$ evolves to z' and is observed through an unknown injective map g. The JEPA encoder f yields $y = f(g(z))$ and $y' = f(g(z'))$. We analyze the induced map $h := f \circ g$ from states to representations; training optimizes f.

Dynamics. Following Klindt et al. (2026), we analyze independent stationary Gaussian factors,

$$z \sim \mathcal{N}(0, I_n), \qquad z'_i = \rho_i z_i + \sqrt{1-\rho_i^2}\,\eta_i, \qquad \eta \sim \mathcal{N}(0, I_n) \perp z. \tag{5}$$

Each factor has a standard Gaussian marginal; larger lag-one correlation $\rho_i \in (0, 1)$ means slower variation. We order $\rho_1 \geq \cdots \geq \rho_n$ only for notation, not as input to the learner.

Anti-collapse constraints. At the population level we compare three nested constraint sets:

$$\mathcal{H}_{\text{pair}} \subseteq \mathcal{H} = \big\{h : h_{\#}\mathcal{N}(0, I_n) = \mathcal{N}(0, I_n)\big\} \subseteq \mathcal{C}_2 = \big\{h : \mathbb{E}[h(z)] = 0,\ \mathrm{Cov}(h(z)) = I_n\big\}, \tag{6}$$

where $h_{\#}$ denotes the law of $h(z)$. They are the exact zero sets of three regularizers: whitening ($\mathcal{C}_2$, the second-moment relaxation of $\mathcal{H}$), LeJEPA's SIGReg ($\mathcal{H}$; Balestriero & LeCun, 2025) and PairSIGReg (§4.4), whose zero set $\mathcal{H}_{\text{pair}}$ additionally requires $(h(z), h(z'))$ to be jointly Gaussian. The three sets agree on linear maps, which are exactly $h(z) = Qz$ with $Q \in \mathrm{O}(n)$, but a nonlinear feature can be whitened without being Gaussian, and Gaussian at each time without being jointly Gaussian across time. We call h *information preserving* if z is a.s. a measurable function of $h(z)$.

Coordinate-level recovery means $h(z) = D\Pi z$, with permutation Π and $D = \mathrm{diag}(\pm 1)$, and we call such an h an *axis map*. No objective that treats coordinates alike can do better, and equal-correlation factors are identifiable only up to rotations within their joint subspace.

4 ANISOTROPIC PREDICTOR JEPA

4.1 AN ANISOTROPIC PREDICTOR

AP-JEPA keeps the JEPA encoder and replaces the dense predictor W by a diagonal one, $P = \mathrm{diag}(p_1, \ldots, p_n)$, which predicts each coordinate from its own past value instead of from a mixture of all coordinates. The prediction loss of equation 2 then splits into n one-dimensional least-squares problems, $\min_{p_j} \mathbb{E}[(y'_j - p_j y_j)^2]$, each solved in closed form by the regression coefficient r_j. This needs no assumption on the encoder or the data beyond $\mathbb{E}[y_j^2] > 0$, so the optimal predictor is available exactly at every training step. Substituting it gives

$$L_{\text{AP}}(h) = \sum_{j=1}^{n} \mathbb{E}\big[(y'_j - r_j y_j)^2\big] = \sum_{j=1}^{n} \Big(\mathbb{E}[y_j'^2] - \frac{\mathbb{E}[y_j y'_j]^2}{\mathbb{E}[y_j^2]}\Big), \qquad r_j = \frac{\mathbb{E}[y_j y'_j]}{\mathbb{E}[y_j^2]}. \tag{7}$$

Each minibatch estimates every r_j from its moments at cost $O(Bn)$. Since r_j minimizes the minibatch loss, the gradient through it vanishes (envelope theorem), so backpropagating through the closed form is the same as holding P fixed. The prediction loss alone admits collapsed solutions such as a constant y, so the encoder minimizes $L_{\text{AP}} + \lambda\mathcal{R}$ for an anti-collapse regularizer $\mathcal{R}$ whose zero set is one of the classes in equation 6; the next subsections ask which one is needed. On all three classes y and y' have zero mean and identity covariance, so $r_j = \mathbb{E}[y_j y'_j]$ and $L_{\text{AP}} = \sum_j (1 - r_j^2)$. The coefficients break the rotation symmetry only when they differ: a shared scalar or the identity is also diagonal and, like a dense predictor, absorbs any rotation, $(h, P) \mapsto (Qh, QPQ^\top)$. Our analysis replaces $\lambda\mathcal{R}$ by its exact zero set and studies $\min_{h \in \mathcal{C}} L_{\text{AP}}(h)$ (Appendix B).

4.2 SUBSPACE AND BASIS: AN EXACT DECOMPOSITION

For such h, let $K := \mathbb{E}[y\, y'^\top]$, with entries $K_{jl} = \mathbb{E}[y_j y'_l]$; it is symmetric because equation 5 is reversible. The diagonal predictor leaves $1 - K_{jj}^2$ per coordinate and the dense predictor, $W = K^\top$,

Table 1: **Regularizer hierarchy under a shared anisotropic predictor.** "Second-order" means stability along all curves in $\mathcal{H}$ through the axis map, not global optimality.

Regularizer	Constrains	Global recovery	Beyond that regime
Whitening $\mathcal{C}_2$	second moments	iff $\rho_1^2 < \rho_n$	a harmonic replaces z_n
SIGReg $\mathcal{H}$	law of y	if $\rho_1^2 < \rho_n$	second-order iff $\rho_2/\rho_1 < \tau_*$ (n=2, Thm. 4)
PairSIGReg $\mathcal{H}_{\text{pair}}$	joint law of (y, y')	every spectrum	—

leaves $n - \|K\|_F^2$ in total, so subtracting gives equation 3:

$$L_{\text{AP}} = \underbrace{n - \|K\|_F^2}_{\text{dense-predictor term}} + \underbrace{\textstyle\sum_{j \neq l} K_{jl}^2}_{\text{cross-talk term}}. \tag{8}$$

The dense-predictor term measures how predictable the span of y is and is rotation invariant, since $y \mapsto Qy$ maps K to $QKQ^\top$. The cross-talk term measures how much of it crosses between coordinates, which a diagonal predictor cannot use; it vanishes exactly when K is diagonal.

Theorem 1 (Subspace–basis decomposition)**.** *Let $\mathcal{C} \subseteq \mathcal{C}_2$ be closed under $h \mapsto Qh$, as $\mathcal{C}_2$, $\mathcal{H}$ and $\mathcal{H}_{\text{pair}}$ are. Then h minimizes L_{AP} over $\mathcal{C}$ iff (a) h minimizes the dense-predictor term over $\mathcal{C}$ (the* subspace condition*) and (b) $K(h)$ is diagonal (the* basis condition*). If K has distinct eigenvalues, (b) fixes the coordinates within the span up to signed permutation, and $\min_{\mathcal{C}} L_{\text{AP}} = \min_{\mathcal{C}} L_{\text{dense}}$.*

The subspace condition is decided by the regularizer, and the basis condition by the predictor, which meets it whenever the factors have distinct lag correlations:

Proposition 1 (Linear maps: the basis step)**.** *Under equation 5, minimize equation 7 over linear induced maps $h_Q(z) = Qz$, $Q \in \mathrm{O}(n)$, and diagonal predictors. For pairwise distinct ρ_i, the global minimizers are exactly $Q = D\Pi$, where D is a diagonal sign matrix and Π a permutation, with $p_j^\star = r_j = \rho_{\pi(j)}$: the coefficients recover the lag-one correlation spectrum. Tied correlations leave orthogonal freedom within the tied subspaces.*

In the factor basis $K = \Lambda$ is diagonal, while any other basis mixes the ρ_i into off-diagonal entries, as in equation 4. The same step also realigns a learned representation: whitening and rotating it into the eigenbasis of its symmetrized lag-one covariance minimizes L_{AP} over its whitening linear maps without changing its span or linear readout (Proposition 7); §5.1 applies it to released JEPAs.

4.3 Which regularizer selects the subspace

When training from scratch, the predictor handles the basis condition, and recovery hinges on the subspace condition. It fails when a nonlinear encoder produces features more predictable than the factors: the regularizer decides which functions of z are admissible, and the loss keeps the most predictable ones. Predictability is measured by the transition operator $(T\varphi)(z) := \mathbb{E}[\varphi(z') \mid z]$: the dense-predictor term equals $n - \sum_i \mu_i^2$, where μ_i are the eigenvalues of T compressed to the span of h; by interlacing $\mu_i \leq \kappa_i$, the eigenvalues of T, so the best span is the top admissible eigenspace. Under equation 5 the eigenfunctions are the Hermite polynomials He_k, with eigenvalues $\lambda_k = \prod_i \rho_i^{k_i}$ (Mehler's formula). The factors are the degree-one modes, with eigenvalues ρ_i; the strongest competitor is the harmonic $\mathrm{He}_2(z_1) = (z_1^2 - 1)/\sqrt{2}$, with eigenvalue ρ_1^2. Stronger constraints remove more competitors (Table 1); whitening and SIGReg do so only for some spectra.

Whitening: a narrow regime. It admits all whitened features; the subspace condition is $\rho_1^2 < \rho_n$.

Proposition 2 (Whitening: recovery iff the factors win outright)**.** *Assume equation 5 and pairwise distinct correlations. Over $h \in \mathcal{C}_2$, signed permutations are the unique global minimizers of L_{AP} if and only if $\rho_1^2 < \rho_n$, i.e. every factor outranks the strongest nonlinear harmonic.*

In the *dangerous* regime $\rho_1^2 \geq \rho_n$ the harmonic enters the best span in place of the fastest factor; with $\rho_1 = 0.95$, avoiding it already requires $\rho_i > 0.9025$ for all i.

SIGReg: the same global regime, a larger local one. Marginal matching excludes the bare harmonic (excess kurtosis 12), and since $\mathcal{H} \subset \mathcal{C}_2$ it inherits the global guarantee when $\rho_1^2 < \rho_n$. It cannot exclude measure-preserving *deformations* of the harmonic, which keep every marginal Gaussian

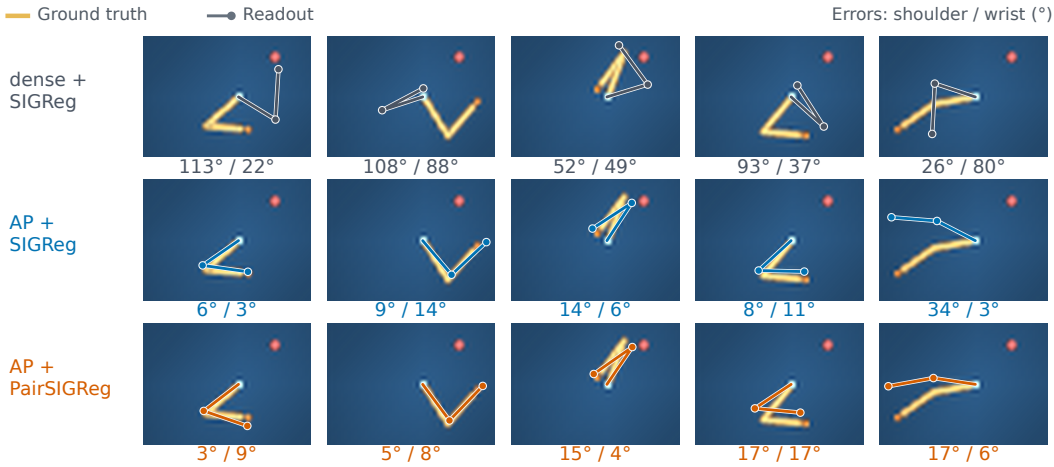


Figure 2: **Low-dimensional Reacher factor readout.** Five held-out frames; frozen readouts decode shoulder and wrist angles, with errors below each frame and the rendered arm as ground truth.

while turning the fast coordinate toward $\mathrm{He}_m(z_1)$, $m \geq 2$. Theorem 4 gives the second variation along all such curves in every dimension; for $n = 2$ the axis map is stable iff $\rho_2/\rho_1 < \tau_*(\rho)$, an explicit threshold. This window contains the whitening regime but is finite and shrinks at fine stride: for $\rho_i = e^{-\mu_i \delta}$, $\delta \to 0$, stability is lost once $\mu_2 > 4\mu_1$. Neither covers every spectrum.

4.4 PairSIGReg: Joint Gaussianity Across Time

SIGReg requires each embedding y to be Gaussian. PairSIGReg additionally requires the consecutive pair (y, y') to be jointly Gaussian, applying the sliced test to the stacked pair $u = [y; y'] \in \mathbb{R}^{2n}$:

$$\mathcal{R}_{\mathrm{Pair}}(f) = \sum_{w \in \{y, y'\}} \left(\|\mathbb{E}w\|_2^2 + \|\mathrm{Cov}(w) - I_n\|_F^2 \right) + \mathbb{E}_{v,t} \left| \phi_v(t) - \phi_{\mathrm{G},v}(t) \right|^2, \tag{9}$$

where $\phi_v(t) = \mathbb{E} \exp(itv^\top(u - \mathbb{E}u))$ and $\phi_{\mathrm{G},v}$ is the characteristic function of $\mathcal{N}(0, v^\top \mathrm{Cov}(u)v)$. With direction and frequency measures of full support, the exact zero set is $\mathcal{H}_{\mathrm{pair}}$. The cross-time covariance is fitted, not prescribed, so the basis still comes only from the predictor; minibatch moments and random slices implement it at SIGReg's cost.

Theorem 2 (Global identification with PairSIGReg)**.** *Under equation 5, let $h = f \circ g$ be information preserving with $\mathcal{R}_{\mathrm{Pair}}(f) = 0$. Then $h(z) = Qz$ a.s. for some $Q \in \mathrm{O}(n)$. For pairwise distinct ρ_i, the minimizers of equation 7 over such maps are exactly $h(z) = D\Pi z$, with $p_j^\star = \rho_{\pi(j)}$.*

Joint Gaussianity confines every admissible feature to the degree-one span: it makes $\mathbb{E}[h(z') \mid h(z)]$ linear in $h(z)$, so the span of h is invariant under T and consists of Hermite polynomials, and a polynomial of a Gaussian vector is itself Gaussian only if it is linear. The subspace condition therefore holds for every spectrum, and the basis condition follows from Proposition 1. When the constraints hold only approximately, recovery degrades gracefully (Proposition 8): the error of each coordinate is bounded by the linear-readout deficit, the cross-talk and the spectral gap of K.

5 Experiments

We test three claims of Section 4: an anisotropic predictor places each factor on its own coordinate without losing linear readout (§5.1), its closed-form basis step does so for released JEPAs (§5.1), and recovery depends on the regularizer and on distinct timescales (§5.2–5.3).

Metrics. On held-out data, MCC averages the absolute Pearson correlations between factors and coordinates after one-to-one Hungarian matching, testing whether each factor has its own coordinate. Full R^2 averages over factors the R^2 of a ridge probe on all coordinates, testing linear readability in any basis; high full R^2 with low MCC thus marks the rotation barrier. **Datasets.** We train on OU-Spiral, a two-factor Ornstein–Uhlenbeck process seen through an invertible spiral map; Reacher and Cartpole from DeepMind Control (Tassa et al., 2018), paired by a fixed frame stride; and MPI3D

Table 2: **Factor recovery on benchmarks with native transitions.** MCC and full R^2 (%), mean $\pm$ SD over five seeds; the average weights benchmarks equally. Bold: best; underline: second.

	OU-Spiral		Reacher		Cartpole		Average	
Method	MCC	R^2	MCC	R^2	MCC	R^2	MCC	R^2
BYOL	81.9 ± 11.0	80.8 ± 9.8	60.6 ± 3.5	67.5 ± 4.0	48.3 ± 2.3	64.3 ± 2.8	63.6	70.9
Barlow Twins	94.4 ± 2.4	94.5 ± 0.7	36.2 ± 1.8	45.8 ± 3.9	40.5 ± 6.8	67.2 ± 1.7	57.0	69.2
VICReg	94.4 ± 2.2	92.8 ± 2.7	54.1 ± 2.8	64.8 ± 1.0	40.7 ± 4.0	66.2 ± 2.7	63.1	74.6
SimDINO	76.7 ± 10.3	85.3 ± 15.0	48.9 ± 7.2	60.4 ± 7.0	49.2 ± 4.9	77.5 ± 4.8	58.2	74.4
LeJEPA	52.0 ± 5.5	45.0 ± 1.2	34.7 ± 3.0	58.3 ± 0.7	35.6 ± 3.1	77.2 ± 1.3	40.8	60.2
dense+SIGReg	91.3 ± 3.7	$\underline{94.9} \pm 1.4$	56.2 ± 5.4	$\mathbf{95.9} \pm 0.1$	56.3 ± 4.6	$\mathbf{86.7} \pm 0.4$	67.9	**92.5**
AP+SIGReg	$\underline{96.6} \pm 0.8$	94.1 ± 1.7	$\underline{88.8} \pm 8.3$	92.8 ± 4.1	$\underline{66.8} \pm 5.7$	$\underline{86.2} \pm 0.7$	$\underline{84.1}$	91.0
AP+PairSIGReg	$\mathbf{97.5} \pm 0.4$	$\mathbf{96.2} \pm 0.3$	$\mathbf{94.7} \pm 1.9$	$\underline{95.7} \pm 0.3$	$\mathbf{69.9} \pm 3.4$	85.4 ± 0.2	**87.4**	$\underline{92.4}$

Table 3: **Factor recovery on image datasets with constructed transitions.** Format as in Table 2; MPI3D averages its real, toy and realistic versions.

	MPI3D (3-var. avg.)		Falcor3D		Isaac3D		Average	
Method	MCC	R^2	MCC	R^2	MCC	R^2	MCC	R^2
BYOL	34.9 ± 1.6	65.0 ± 2.7	85.8 ± 3.9	87.1 ± 0.2	$\underline{70.1} \pm 4.4$	75.1 ± 1.6	$\underline{63.6}$	75.7
Barlow Twins	29.4 ± 2.6	51.1 ± 4.3	76.8 ± 2.8	86.1 ± 0.2	60.8 ± 4.6	69.3 ± 5.8	55.6	68.8
VICReg	6.4 ± 0.4	1.5 ± 0.5	82.1 ± 3.4	85.7 ± 0.2	66.6 ± 2.1	72.9 ± 0.8	51.7	53.4
SimDINO	27.4 ± 2.2	32.7 ± 3.5	61.2 ± 6.4	77.3 ± 3.1	62.6 ± 6.5	$\underline{81.4} \pm 2.8$	50.4	63.8
LeJEPA	$\underline{36.4} \pm 2.3$	$\underline{82.6} \pm 2.0$	45.6 ± 3.3	89.0 ± 0.1	30.6 ± 1.3	52.6 ± 2.3	37.5	74.8
dense+SIGReg	31.4 ± 2.6	66.0 ± 8.9	83.7 ± 2.1	93.1 ± 0.4	47.3 ± 3.9	$\mathbf{83.6} \pm 5.4$	54.1	$\underline{80.9}$
AP+SIGReg	22.5 ± 4.9	40.8 ± 11.3	$\mathbf{91.5} \pm 7.4$	$\mathbf{94.3} \pm 0.2$	69.1 ± 4.8	80.0 ± 0.5	61.0	71.7
AP+PairSIGReg	$\mathbf{40.4} \pm 1.9$	$\mathbf{89.9} \pm 0.4$	$\underline{91.2} \pm 4.4$	$\underline{93.8} \pm 0.2$	$\mathbf{72.9} \pm 6.3$	80.2 ± 1.1	**68.1**	**88.0**

(Gondal et al., 2019), Falcor3D and Isaac3D (Nie et al., 2020), whose transitions we construct by stepping the factors. For released I-JEPA (Assran et al., 2023) and V-JEPA (Bardes et al., 2024), we also report top-1 accuracy on Kinetics-400 (K400) (Kay et al., 2017) and Something-Something-v2 (SSv2) (Goyal et al., 2017). **Baselines.** We compare AP+SIGReg and AP+PairSIGReg with LeJEPA (Balestriero & LeCun, 2025) and dense+SIGReg, which differ from AP only in the predictor (identity and dense linear), and with VICReg (Bardes et al., 2022), BYOL (Grill et al., 2020), Barlow Twins (Zbontar et al., 2021) and SimDINO (Wu et al., 2025), whose two views are the temporal pair; within a benchmark all methods share the encoder and pairs. Appendix C details all datasets, baselines and metrics, and Appendix D.3 gives full results.

5.1 Main results

Coordinate recovery and linear readout. Figure 2 shows what is at stake: on a held-out Reacher episode, dense+SIGReg decodes the joints from all coordinates as well as AP-JEPA, yet its low-dimensional readouts miss them. Tables 2 and 3 show that this gap is general. On Reacher, Cartpole and Isaac3D, dense+SIGReg attains the highest full R^2 but only 47–56% MCC: the factors are present, merely spread over rotated coordinates. The anisotropic predictor removes this rotation. On the native-transition benchmarks, AP+PairSIGReg attains the highest MCC at essentially the same full R^2 as dense+SIGReg (e.g., 94.7% versus 56.2% MCC on Reacher), so the gain lies in how the information is arranged rather than in how much of it is kept; on the image datasets, it leads in both metrics on average. AP+SIGReg comes close on most benchmarks, but since SIGReg excludes non-factor features only in a limited regime (§4), it can lose a factor, as it does on MPI3D. MPI3D is hard for every method: its two arm axes decorrelate at nearly the same rate, so, as the theory predicts, MCC stays at about 40% or below. Even so, AP+PairSIGReg leads both metrics on each of the three MPI3D versions and keeps every factor reliably readable.

Post-hoc recovery in pretrained JEPAs. By Theorem 1, the AP loss selects subspace and basis through separate terms, so a frozen representation can be realigned without retraining: minimizing the cross-talk term alone has a closed form (Proposition 7(i)), which whitens the features and rotates them into the eigenbasis of their symmetrized lag-one covariance. We fit this post-hoc AP map on

Table 4: **Post-hoc coordinate recovery in released JEPAs.** Values in %; frozen ViT-H/14 and ViT-L/16 features; AP rows apply the post-hoc AP map; top-1 uses a linear probe on pooled features.

		MPI3D			K400	SSv2
		MCC	single R^2	full R^2	top-1	top-1
I-JEPA	native	28.3	8.4	91.5	49.9	20.0
	AP	**90.9** +62.6	**82.6** +74.2	91.3 −0.2	48.2 −1.7	18.6 −1.4
V-JEPA	native	32.3	10.7	89.2	54.3	48.4
	AP	**89.3** +57.0	**79.8** +69.1	89.0 −0.2	54.4 +0.1	46.5 −1.9

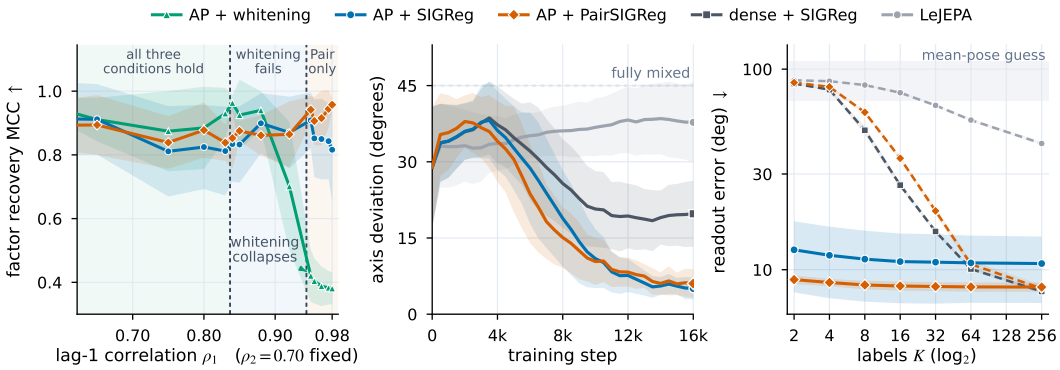


Figure 3: **Mechanism diagnostics. (a)** Recovery against ρ_1 at $\rho_2 = 0.70$; shading marks theoretical boundaries under the stated assumptions (SIGReg's extension is second-order). **(b)** Axis alignment on OU-Spiral. **(c)** Shoulder error with K labels: pair calibration (solid), full decoder (dashed).

unlabelled MPI3D walk pairs and apply it to released I-JEPA and V-JEPA features (Table 4). Native features show the rotation barrier: full R^2 is near 90%, yet MCC is about 30% and single-coordinate R^2, the R^2 of an affine fit of each factor on its matched coordinate, is only about 10%. Post-hoc AP raises MCC to about 90% and single-coordinate R^2 to about 80%, while full R^2 and K400 and SSv2 top-1 barely move (at most 0.3 and 1.9 points), as the map is invertible.

5.2 Mechanism Analysis

Figure 3(a) tests the regularizer's role in Theorem 1, keeping non-factor features out of the learned subspace. With the predictor fixed to AP, the slow factor's correlation ρ_1 grows until its square, a non-factor feature, becomes more predictable than the fast factor. The three regularizers behave as the theory predicts in each region: all recover the factors while the whitening condition holds; beyond it, whitening admits this harmonic and collapses to about 40% MCC, while SIGReg and PairSIGReg stay above 80%; and where only the PairSIGReg guarantee holds, PairSIGReg is best.

Figure 3(b) follows the predictor's role, fixing the basis within the subspace, over training on OU-Spiral. All models start about 30° from the factor axes. The AP variants rotate onto them, ending within 6°, whereas dense+SIGReg, which reaches the same full R^2 (Table 2) but pays nothing for a rotation, stalls near 20°, and LeJEPA drifts toward the fully mixed basis.

Figure 3(c) turns to few-label calibration of the Reacher shoulder angle. AP places the angle on its own coordinate pair, which is found without labels, so K labels only calibrate its direction and zero; in all ten AP runs, this label-free choice is the pair that all 4096 labels would select (Appendix D.5.1). With $K = 2$, AP+PairSIGReg reads the shoulder to within 8.9°. Other models spread the angle over many coordinates and need a full decoder, which reaches 8° for dense+SIGReg at $K = 256$ but needs more than 64 labels to match two-label calibration, and LeJEPA stays above 40°.

5.3 Ablation Study

Table 5 isolates the anisotropic predictor on the two benchmarks with distinct timescales. Predictors with shared or cross-coordinate coefficients already make the factors linearly readable (full R^2 of

Table 5: **Predictor ablation** with distinct timescales. Mean±SD (%); small numbers: anisotropic minus dense predictor, same regularizer. Bold/underline: first/second; ⋆: full method.

		Reacher-hard		OU-Spiral	
Regularizer	Predictor	MCC	full R^2	MCC	full R^2
SIGReg	identity (LeJEPA)	34.7±3.0	58.3±0.7	52.0±5.5	45.0±1.2
	shared scalar	49.3±4.8	93.4±1.0	90.7±3.7	94.1±1.8
	dense	56.2±5.4	**95.9**±0.1	91.3±3.7	94.9±1.4
	anisotropic	<u>88.8</u>±8.3 +32.6	92.8±4.1 −3.1	<u>96.6</u>±0.8 +5.3	94.1±1.7 −0.8
PairSIGReg	identity	31.2±3.4	54.2±0.3	76.4±10.0	82.1±3.1
	shared scalar	50.3±4.3	92.4±0.5	88.1±5.6	<u>96.2</u>±1.0
	dense	55.5±3.0	<u>95.8</u>±0.3	89.1±5.7	**96.7**±0.4
	anisotropic ⋆	**94.7**±1.9 +39.2	95.7±0.3 −0.1	**97.5**±0.4 +8.4	<u>96.2</u>±0.3 −0.5

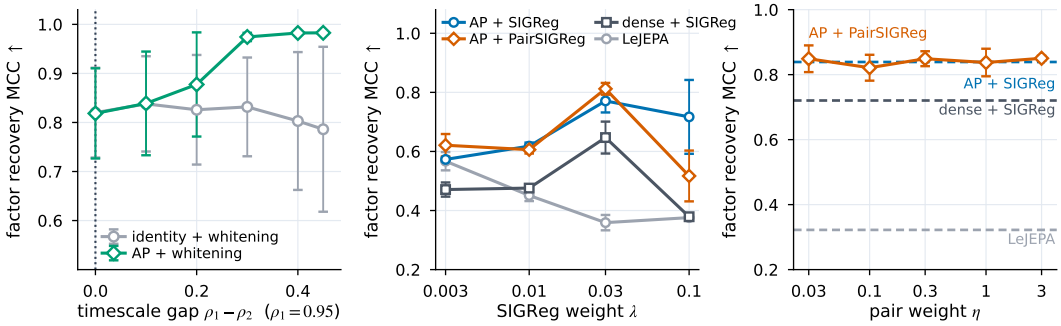


Figure 4: **Ablation trends** on controlled OU data. **(a)** Timescale gap under whitening. **(b)** SIGReg weight at $\rho = (0.97, 0.50)$. **(c)** Pair weight at $\rho = (0.95, 0.60)$. Details: Appendices D.1 and D.6.

93–96% on Reacher-hard, against 58.3% for LeJEPA), but they absorb any rotation and leave MCC at 49–56%. Unequal coefficients are what recover the coordinates: they add 32.6 and 39.2 MCC points on Reacher-hard and 5.3 and 8.4 on OU-Spiral under SIGReg and PairSIGReg, while full R^2 moves by at most 3.1 points. PairSIGReg cannot replace this step: with any other predictor it reaches only 31–56% MCC on Reacher-hard and 76–89% on OU-Spiral. Conversely, it lifts AP from 88.8% to 94.7% MCC on Reacher-hard (SD 8.3 → 1.9) but no other predictor by over 1 point. Appendix D.3 shows the gain in five of six settings; the exception is MPI3D-real, with tied timescales, where PairSIGReg restores a lost factor.

Figure 4 tests the predictor's condition and the two regularization weights on controlled OU data. Panel (a) fixes the regularizer (whitening) and $\rho_1 = 0.95$ and lowers ρ_2. Without a timescale gap the two predictors tie at about 82% MCC; as the gap widens, AP climbs to 98% while the identity predictor stays near 80%, as equation 4 predicts. Panels (b) and (c) sweep λ at $\rho = (0.97, 0.50)$, where whitening fails, and η at $\rho = (0.95, 0.60)$. Every arm varies with λ, yet both AP variants stay above dense+SIGReg and LeJEPA at every λ. Over a 100-fold η range, AP+PairSIGReg stays within 82–85% MCC, on par with AP+SIGReg and above dense+SIGReg (72%) and LeJEPA (32%).

6 Conclusion

JEPA objectives fix which information a representation keeps but, being rotation invariant, not how it is laid out across coordinates. We showed that a minimal change, an anisotropic predictor, breaks this rotation barrier, as its loss splits exactly into a dense-predictor term, which selects the subspace, and a cross-talk term, which selects the basis. Because a nonlinear encoder can produce features more predictable than the factors, the regularizer matters too: whitening and SIGReg exclude such features only in limited regimes, whereas PairSIGReg recovers the factors for every set of distinct timescales. AP-JEPA recovers factors in individual coordinates without losing linear readout, and a closed-form version exposes factors in released I-JEPA and V-JEPA models. Our guarantees are population-level and rely on distinct timescales, without which the predictor offers no advantage. Next steps are action-conditioned and non-reversible dynamics, and large JEPA world models trained with AP.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work to assist with language editing and presentation, the organization and revision of experimental code, and the implementation of numerical sanity checks related to the theoretical analysis. In particular, AI assistance was used to improve the clarity of the exposition, restructure analysis scripts, and examine whether theoretical predictions were consistent with controlled numerical cases. The authors determined the research questions, methodology, experimental design, and conclusions. All AI-assisted text, code, and numerical analyses were reviewed, rerun where applicable, and verified by the entire author team. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work involves no human subjects or new data collection. Models are trained on synthetic, simulated and public image datasets without people; Kinetics-400 and Something-Something-v2, which show people, are used only as standard action-recognition benchmarks with their official labels.

REPRODUCIBILITY STATEMENT

Assumptions, the population problems, and complete proofs are given in Appendix B, including approximate recovery (Appendix B.8) and coordinate recovery in existing representations (Appendix B.7). Experimental protocols, architectures, hyperparameter selection, and seeds are documented in Appendix C. The experimental code and scripts used to produce the reported results will be made publicly available.

REFERENCES

Robert Jenkinson Alvarez. Beyond isotropy in JEPAs: Hamiltonian geometry and symplectic prediction. arXiv preprint arXiv:2605.20107, 2026. URL https://arxiv.org/abs/2605.20107.

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

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew 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. arXiv preprint arXiv:2506.09985, 2025. URL https://arxiv.org/abs/2506.09985.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. arXiv preprint arXiv:2511.08544, 2025. URL https://arxiv.org/abs/2511.08544.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2. Featured Certification.

A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997. doi: 10.1109/78.554307.

Tobias Blaschke, Pietro Berkes, and Laurenz Wiskott. What is the relation between slow feature analysis and independent component analysis? *Neural Computation*, 18(10):2495–2508, October 2006. doi: 10.1162/neco.2006.18.10.2495. URL https://doi.org/10.1162/neco.2006.18.10.2495.

Simon Buchholz and Bernhard Schölkopf. Robustness of nonlinear representation learning. arXiv preprint arXiv:2503.15355, 2025. URL https://arxiv.org/abs/2503.15355.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001. URL https://doi.org/10.1137/0707001.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3015–3024. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ermolov21a.html.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-JEPA: Joint-embedding predictive architecture can listen. arXiv preprint arXiv:2311.15830, 2024. URL https://arxiv.org/abs/2311.15830.

Ian Gemp, Brian McWilliams, Claire Vernade, and Thore Graepel. EigenGame: PCA as a Nash equilibrium. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NzTU59SYbNq.

Diego Gomez, Michael Bowling, and Marlos C. Machado. Proper Laplacian representation learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 28911–28934, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/7c88ed03238db64cef6f23bd0d6b5d11-Paper-Conference.pdf.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d97d404b6119214e4a7018391195240a-Paper.pdf.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" video database for learning and evaluating visual common sense. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5843–5851, 2017. doi: 10.1109/ICCV.2017.622.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, April 2025. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*,

pp. 460–469. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/hyvarinen17a.html.

Markus Karmann, Shile Li, Christian Internò, Bruno Andreis, David Klindt, Randall Balestriero, Jindong Gu, Philip Torr, Qi Zhang, Peng-Tao Jiang, Hao Zhang, Bo Li, and Onay Urfalioglu. MotionJEPA: Preventing temporal feature collapse by capturing visual changes in latent space. arXiv preprint arXiv:2609.23881, 2026. URL https://arxiv.org/abs/2609.23881.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017. URL https://arxiv.org/abs/1705.06950.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model? arXiv preprint arXiv:2605.26379, 2026. URL https://arxiv.org/abs/2605.26379.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Position paper, version 0.9.2.

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How JEPA avoids noisy features: The implicit bias of deep linear self distillation networks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 91300–91336. Curran Associates, Inc., 2024. doi: 10.52202/079017-2897. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a600f0f740605205133553cb74a1c107-Paper-Conference.pdf.

Chang Liu, Fei Suo, Yanzhou Jin, Zeyu Ping, Yusuke Iwasawa, Yutaka Matsuo, and Yaonan Zhu. Temporally centered SIGReg improves LeWorldModel representations for robot policy learning. arXiv preprint arXiv:2607.26924, 2026. URL https://arxiv.org/abs/2607.26924.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. VAMPnets for deep learning of molecular kinetics. *Nature Communications*, 9(1):5, January 2018. doi: 10.1038/s41467-017-02388-1. URL https://doi.org/10.1038/s41467-017-02388-1.

L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72(23):3634–3637, Jun 1994. doi: 10.1103/PhysRevLett.72.3634. URL https://link.aps.org/doi/10.1103/PhysRevLett.72.3634.

Weili Nie, Tero Karras, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit Patel, and Animashree Anandkumar. Semi-supervised StyleGAN for disentanglement learning. In Hal Daumé, III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7360–7369. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/nie20a.html.

Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics*, 139(1):015102, July 2013. doi: 10.1063/1.4811489. URL https://doi.org/10.1063/1.4811489.

David Pfau, Stig Petersen, Ashish Agarwal, David G. T. Barrett, and Kimberly L. Stachenfeld. Spectral inference networks: Unifying deep and spectral learning. arXiv preprint arXiv:1806.02215, 2020. URL https://arxiv.org/abs/1806.02215.

Patrik Reizinger, Alice Bizeul, Attila Juhos, Julia E. Vogt, Randall Balestriero, Wieland Brendel, and David Klindt. Cross-entropy is all you need to invert the data generating process. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 38186–38216, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5ee7ed60a7e8169012224dec5fe0d27f-Paper-Conference.pdf.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9030–9039. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/roeder21a.html.

Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. InfoNCE: Identifying the gap between theory and practice. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4159–4167. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/rusak25a.html.

Ayumu Saito, Prachi Kudeshia, and Jiju Poovvancheri. Point-JEPA: A joint embedding predictive architecture for self-supervised learning on point cloud. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7348–7357, 2025. doi: 10.1109/WACV61041.2025.00714.

Uladzislau Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 43905–43941. Curran Associates, Inc., 2025. doi: 10.52202/085713-1465. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3e7cf447f21cd11c846463affefce665-Paper-Conference.pdf.

Vlad Sobal, Jyothir S V, Siddhartha Jalagam, Nicolas Carion, Kyunghyun Cho, and Yann LeCun. Joint embedding predictive architectures focus on slow features. arXiv preprint arXiv:2211.10831, 2022. URL https://arxiv.org/abs/2211.10831.

Henning Sprekeler. On the relation of slow feature analysis and Laplacian eigenmaps. *Neural Computation*, 23(12):3287–3302, December 2011. doi: 10.1162/NECO_a_00214. URL https://doi.org/10.1162/NECO_a_00214.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite. arXiv preprint arXiv:1801.00690, 2018. URL https://arxiv.org/abs/1801.00690.

Lucas Thil, Jesse Read, Rim Kaddah, and Guillaume Doquet. Subspace-decomposed JEPAs: Disentangling progression and content in latent world models. arXiv preprint arXiv:2605.31111, 2026. URL https://arxiv.org/abs/2605.31111.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning*

Research, pp. 10268–10278. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tian21a.html.

L. Tong, R.-w. Liu, V.C. Soon, and Y.-F. Huang. Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, 38(5):499–509, 1991. doi: 10.1109/31.76486.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16451–16467. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8929c70f8d710e412d38da624b21c3c8-Paper.pdf.

Kaixin Wang, Kuangqi Zhou, Qixin Zhang, Jie Shao, Bryan Hooi, and Jiashi Feng. Towards better Laplacian representation in reinforcement learning with generalized graph drawing. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11003–11012. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21ae.html.

Yuan-Hao Wei. LeJEPA-FGP: Factorized Gaussian process regularization for identifiable joint-embedding world models. *Preprints*, June 2026. doi: 10.20944/preprints202606.1157.v1. URL https://doi.org/10.20944/preprints202606.1157.v1.

Zixin Wen and Yuanzhi Li. The mechanism of prediction head in non-contrastive self-supervised learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24794–24809. Curran Associates, Inc., 2022. doi: 10.52202/068431-1798. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d276b0a087efdd2404f3295b26c24c1-Paper-Conference.pdf.

Laurenz Wiskott. Slow feature analysis: A theoretical analysis of optimal free responses. *Neural Computation*, 15(9):2147–2177, 2003. doi: 10.1162/089976603322297331.

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

Ziyang Wu, Jingyuan Zhang, Druv Pai, XuDong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma. Simplifying DINO via coding rate regularization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68036–68059. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wu25ar.html.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26492–26503. Curran Associates, Inc., 2022. doi: 10.52202/068431-1921. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a938292feb86b94ebe3e6200ff7786ef-Paper-Conference.pdf.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, June 2015. doi: 10.1093/biomet/asv008. URL https://doi.org/10.1093/biomet/asv008.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane 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*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. arXiv preprint arXiv:2607.22430, 2026. URL https://arxiv.org/abs/2607.22430.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zimmermann21a.html.

APPENDIX CONTENTS

A Extended Related Work **17**

B Theoretical Results and Proofs **18**

B.1 Setting, Notation, and the Population Problems 18

B.2 The Subspace–Basis Decomposition (Theorem 1) 19

B.3 Linear Axis Recovery (Proposition 1) 20

B.4 Whitening and Nonlinear Modes (Proposition 2) 21

B.5 Marginal Gaussian Matching: Global Regime and Local Stability 21

B.6 Global Identification with PairSIGReg (Theorem 2) 25

B.7 Basis Selection in a Fixed Representation 26

B.8 Approximate Identifiability 27

C Experimental Setup **28**

C.1 Benchmarks and Pair Construction 28

C.2 Models and Training 29

C.3 Metrics and Probe Evaluation 30

C.4 Frozen-Representation Transformation 30

C.5 Calibration and Overlay Evaluation 31

D Additional Experimental Results **32**

D.1 Basis Selection and Temporal Separation 32

D.2 Nonlinear Feature Competition and Regularization 33

D.3 Benchmark Results and Matched Ablations 33

D.4 Post-Hoc Alignment of a Fixed Representation 35

D.5 Few-Label Calibration and Overlay Results 37

D.6 Regularization Weights 38

D.7 Timescales and the Approximate-Identifiability Bound 39

D.8 Timescale Gap and Coordinate Recovery on Real Dynamics 40

A Extended Related Work

This section expands the related work of Section 2. We use the distinction of Theorem 1 between the *subspace* a representation spans and the *basis* in which it is expressed. For each line of work we state what it contributes, what fixes its basis, if anything, and how AP-JEPA differs from it.

JEPA world models fix the subspace, not the basis. JEPAs predict in representation space instead of reconstructing observations (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024). They now serve as world models that plan in embedding space: V-JEPA 2 post-trains an action-conditioned predictor on a pretrained video encoder (Assran et al., 2025), DINO-WM learns dynamics on frozen DINOv2 features (Zhou et al., 2025), Sobal et al. (2025) plan from reward-free offline data, and LeWorldModel trains end to end from pixels with SIGReg (Maes et al., 2026), an alternative to reconstruction-based agents such as DreamerV3 (Hafner et al., 2025). Collapse is prevented by a stop-gradient predictor (Grill et al., 2020), redundancy reduction (Zbontar et al., 2021; Bardes et al., 2022), whitening (Ermolov et al., 2021) or matching an isotropic Gaussian (Balestriero & LeCun, 2025). All of these are invariant to rotations of the representation, and so are the usual evaluations by planning success and linear probes. Analyses of JEPAs therefore concern which features are kept and how variance is shared among them. Deep linear JEPAs prioritize features with large regression coefficients over high-variance but noisy ones (Littwin et al., 2024). JEPA world models favour slow features, which makes them fail under static distractors (Sobal et al., 2022); MotionJEPA counteracts this bias by also predicting embeddings of frame differences (Karmann et al., 2026), and temporally centered SIGReg restores the variance of the fast residual that plain SIGReg suppresses (Liu et al., 2026). These works shape the subspace, and AP-JEPA is complementary to them: it leaves the subspace to the regularizer and fixes the basis within it. In non-contrastive learning, the predictor's eigenbasis aligns with that of the representation correlation (Tian et al., 2021), and its off-diagonal entries transfer features between neurons (Wen & Li, 2022). Theorem 1 shows that these off-diagonal entries are exactly what absorbs the cross-talk term. A diagonal predictor removes them and exposes this term to the loss, and over rotation-closed constraint classes its minimum equals that of a dense predictor, so fixing the basis costs no predictive risk.

Identifiability of JEPA world models. Our analysis builds on two recent results. Klindt et al. (2026) prove that LeJEPA recovers the latent state of a world with stationary additive-noise dynamics up to an orthogonal map, that the Gaussian is the only latent law for which this holds, and that the guarantee degrades gracefully under approximation. Zhang et al. (2026) extend recovery up to an orthogonal map to action-conditioned world models, with a transition error that scales inversely with a spectral separation margin. We adopt their Gaussian setting and close the orthogonal gap that both leave open: with an anisotropic predictor, recovery holds up to signed permutation (Proposition 1, Theorem 2), with an approximate version in Proposition 8. LeJEPA-FGP (Wei, 2026) pursues the same guarantee by a different route: it replaces SIGReg's target with a factorized Gaussian-process law in which each channel follows its own prescribed kernel, so that the regularizer carries the basis. The kernels, and with them the timescales of the factors and their assignment to channels, must therefore be specified before training. In AP-JEPA the basis comes from the predictor instead: its coefficients are fitted from minibatch moments and converge to the lag correlations (Proposition 1), and PairSIGReg fits the cross-time covariance rather than prescribing it. Because the basis step lives in the predictor, it also applies to JEPAs already trained with dense predictors (Section 5.1), and we characterize which regularizers select the factor subspace (Section 4.3). LeJEPA-FGP is validated on synthetic GP data, whereas we evaluate on dynamical and image benchmarks. Two further JEPA variants have other goals. HamJEPA predicts with a prescribed symplectic leapfrog map to improve neighbourhood geometry and does not address factor recovery (Alvarez, 2026). SD-JEPA reserves a low-dimensional subspace for task progression, shaped by a temporal triplet loss (Thil et al., 2026), while its content subspace keeps SIGReg's rotation invariance, so factors within it stay mixed.

Identifiability beyond JEPAs. Without inductive bias, factors cannot be recovered from independent samples (Locatello et al., 2019). Nonlinear ICA supplies the bias through auxiliary variables or temporal structure and identifies factors up to permutation, but requires non-Gaussian or non-stationary sources (Hyvärinen & Morioka, 2017; Yao et al., 2022), which excludes the stationary Gaussian factors studied here. Contrastive objectives recover factors up to an orthogonal or linear map (Zimmermann et al., 2021; Roeder et al., 2021), or a block of content variables when aug-

mentations change only the style (von Kügelgen et al., 2021), and supervised cross-entropy training yields linear identifiability (Reizinger et al., 2025). AnInfoNCE breaks the rotation in contrastive learning when augmentations perturb factors by different amounts, which it models with a trainable scaling in the critic (Rusak et al., 2025). AP-JEPA differs in both the source of the asymmetry and the objective: it relies on the timescales of temporal pairs rather than on augmentation strengths, and it uses a non-contrastive predictive loss without negatives, in which isotropic regularizers provably leave the rotation free (Klindt et al., 2026). Finally, Proposition 8 parallels robustness results for nonlinear representation learning (Buchholz & Schölkopf, 2025): recovery degrades gracefully when the constraints hold only approximately, at a rate set by the measurable spectral gap of K.

Slow features and spectral methods. Slow feature analysis extracts the slowest-varying features (Wiskott & Sejnowski, 2002). Its nonlinear version fails when a harmonic of a slow source outranks a fast source (Wiskott, 2003; Sprekeler, 2011); Section 4.3 shows that the same competitor threatens any predictive objective with a flexible encoder. AMUSE, SOBI and tICA separate sources with distinct autocorrelations by diagonalizing lagged covariances of fixed observations or features (Tong et al., 1991; Molgedey & Schuster, 1994; Belouchrani et al., 1997; Pérez-Hernández et al., 2013); our closed-form map for frozen JEPAs coincides with the AMUSE rotation (Appendix B.7). VAMPnets learn slow coordinates with neural networks (Mardt et al., 2018). Neural eigenfunction methods face the same symmetry and break it with a mechanism designed into the objective: masked gradients (Pfau et al., 2020), a sequential game (Gemp et al., 2021), or, for Laplacian representations, prescribed per-dimension weights and constrained formulations (Wang et al., 2021; Gomez et al., 2024). These mechanisms target a known operator and fix the asymmetry by hand. AP-JEPA differs in two respects. First, its asymmetry is learned: the coefficients are fitted in closed form from minibatch moments, need no ordering or weight schedule, and converge to the timescale spectrum (Proposition 1). Second, it trains a nonlinear encoder end to end, so the same loss must also select the span. This raises a question that linear methods on fixed features never face, whether the harmonic that defeats nonlinear SFA enters the span; the regularizer hierarchy of Section 4.3 characterizes when it is excluded, and PairSIGReg excludes it for every spectrum (Theorem 2).

B Theoretical Results and Proofs

B.1 Setting, Notation, and the Population Problems

Throughout this appendix $Z \sim \gamma_n := \mathcal{N}(0, I_n)$ and $Z' = \Lambda Z + (I_n - \Lambda^2)^{1/2}\eta$, where $\Lambda = \mathrm{diag}(\rho)$ with $\rho_i \in (0,1)$ and $\eta \sim \gamma_n$ is independent of Z, as in equation 5. The pair (Z, Z') is stationary and exchangeable, $(Z, Z') \overset{d}{=} (Z', Z)$. We write $\rho_{\max}$ and $\rho_{\min}$ for the largest and smallest correlation and, when the correlations are pairwise distinct, label them so that $\rho_1 > \cdots > \rho_n$.

Most of the analysis rests on the spectral structure of the transition operator $(T\varphi)(z) := \mathbb{E}[\varphi(Z') \mid Z = z]$ on $L^2(\gamma_n)$. Let He_k, $k \in \mathbb{N}_0^n$, be the normalized probabilists' Hermite polynomials, an orthonormal basis of $L^2(\gamma_n)$, and let $|k| = \sum_i k_i$. Mehler's formula states that

$$T\mathrm{He}_k = \lambda_k \mathrm{He}_k, \qquad \lambda_k = \prod_{i=1}^{n} \rho_i^{k_i} \le \rho_{\max}^{|k|}. \tag{10}$$

Hence $\langle \varphi, T\psi \rangle = \mathbb{E}[\varphi(Z)\psi(Z')]$, and T is self-adjoint, positive and Hilbert–Schmidt, since $\sum_k \lambda_k^2 = \prod_i (1-\rho_i^2)^{-1} < \infty$. On the mean-zero subspace $L_0^2(\gamma_n)$ its eigenvalues are the λ_k with $k \neq 0$. Every eigenspace with eigenvalue $\kappa > 0$ is finite-dimensional and spanned by Hermite polynomials, because $\lambda_k = \kappa$ forces $|k| \le \log\kappa / \log\rho_{\max}$. We list these eigenvalues with multiplicity as $\kappa_1 \ge \kappa_2 \ge \cdots$. The factors $z_i = \mathrm{He}_{e_i}$ are the degree-one eigenfunctions, with eigenvalues ρ_i, whereas every mode of degree $|k| \ge 2$ has eigenvalue at most $\rho_{\max}^2$, with equality for $\mathrm{He}_{2e_{i^*}}$ when $\rho_{i^*} = \rho_{\max}$.

An *induced map* is a map $h = f \circ g : \mathbb{R}^n \to \mathbb{R}^n$ with components in $L^2(\gamma_n)$, and its lag-one scores are $R_j(h) := \langle h_j, Th_j \rangle = \mathbb{E}[h_j(Z)h_j(Z')]$. The three constraint classes of the paper are nested,

$$\begin{aligned} \mathcal{C}_2 = \{h : \mathbb{E}h(Z) = 0,\ \mathrm{Cov}(h(Z)) = I_n\} \ &\supseteq\ \mathcal{H} = \{h : h_{\#}\gamma_n = \gamma_n\} \\ &\supseteq\ \mathcal{H}_{\mathrm{pair}} = \{h \in \mathcal{H} : (h(Z), h(Z')) \text{ jointly Gaussian}\}, \end{aligned}$$

and for $h \in \mathcal{C}_2$ the components form an orthonormal family in $L^2_0(\gamma_n)$. We call h *information preserving* if $\sigma(h(Z)) = \sigma(Z)$ up to γ_n-null sets, that is, $Z = \psi(h(Z))$ a.s. for some measurable ψ. Since g is injective, this holds whenever f is injective on the image of g with a measurable inverse.

For $h \in \mathcal{C}_2$ the coordinates of $h(Z)$ and of $h(Z')$ have unit variance, so the least-squares coefficient of coordinate j is $r_j(h) = R_j(h)$, and the optimized predictive term of equation 7 becomes

$$L_{\mathrm{pred}}(h) := \min_{P \text{ diagonal}} \mathbb{E}\|h(Z') - Ph(Z)\|^2 = \sum_{j=1}^{n} \big(1 - R_j(h)^2\big), \tag{11}$$

which is the loss L_{AP} of the main text. Our results concern the hard-constrained problem

$$(\mathrm{P}_{\mathcal{C}}) \qquad \min_{h \in \mathcal{C}} L_{\mathrm{pred}}(h), \qquad \mathcal{C} \in \{\mathcal{C}_2, \mathcal{H}, \mathcal{H}_{\mathrm{pair}}\},$$

in which the regularizer is replaced by its exact zero set. A signed permutation $h(z) = D\Pi z$ lies in $\mathcal{H}_{\mathrm{pair}}$, is information preserving and has $R_j = \rho_{\pi(j)}$, and we say that $(\mathrm{P}_{\mathcal{C}})$ *recovers the factors* if its global minimizers are exactly the signed permutations. All statements in this appendix are about population objectives.

For the local analysis it is convenient to also consider, for fixed weights $w_1, \dots, w_n > 0$, the linear functional $J_w(h) := \sum_j w_j R_j(h)$. On $\mathcal{C}_2$, maximizing J_w is the same as minimizing the weighted slowness $L_w(h) := \sum_j w_j \mathbb{E}[(h_j(Z') - h_j(Z))^2] = 2\sum_j w_j(1 - R_j(h))$. The deployed objective equation 11 is not of this form, but at the axis map its second variation equals -2 times that of J_ρ (Corollary 2), so results for J_w transfer to it.

B.2 The Subspace–Basis Decomposition (Theorem 1)

For $h \in \mathcal{C}_2$ let $K(h) := \mathbb{E}[h(Z)h(Z')^\top]$, so that $K_{jl} = \langle h_j, Th_l \rangle$. Exchangeability of (Z, Z') makes K symmetric, and since the h_j are orthonormal, K is the matrix of the compression $P_V T P_V$ of T to $V = \mathrm{span}\{h_1, \dots, h_n\}$ in the basis (h_j).

Proof of Theorem 1. Since $R_j = K_{jj}$, equation 11 gives $L_{\mathrm{pred}} = n - \sum_j K_{jj}^2 = n - \|K\|_F^2 + \sum_{j \neq l} K_{jl}^2$. For a dense predictor W, both time points are whitened, so $\mathbb{E}\|h(Z') - Wh(Z)\|^2 = n - 2\,\mathrm{Tr}(WK) + \mathrm{Tr}(WW^\top)$; this is minimized by $W = K$, with minimum $n - \|K\|_F^2$. Replacing h by Qh with $Q \in \mathrm{O}(n)$ replaces K by $QKQ^\top$ and leaves $\|K\|_F$ unchanged; indeed $\|K\|_F^2 = \mathrm{Tr}((P_V T P_V)^2)$ depends only on V.

Now let $\mathcal{C}$ be closed under rotations and put $m := \inf_{\mathcal{C}}(n - \|K\|_F^2)$. For every $h \in \mathcal{C}$ we have $L_{\mathrm{pred}}(h) \geq n - \|K(h)\|_F^2 \geq m$, and choosing Q that diagonalizes $K(h)$ gives $Qh \in \mathcal{C}$ with $L_{\mathrm{pred}}(Qh) = n - \|K(h)\|_F^2$. Hence $\inf_{\mathcal{C}} L_{\mathrm{pred}} = m$. A map satisfying (a) and (b) attains m. Conversely, if $L_{\mathrm{pred}}(h) = m$, the chain $m \leq n - \|K\|_F^2 \leq L_{\mathrm{pred}}(h) = m$ forces both equalities, which are (a) and a vanishing cross-talk term, that is, (b). Finally, if $K(h)$ and $K(Qh) = QK(h)Q^\top$ are both diagonal and $K(h)$ has distinct eigenvalues, then the rows of Q are eigenvectors of the diagonal matrix $K(h)$, hence signed unit vectors, and $Q = D\Pi$. □

Condition (a) compares spans, and the comparison reduces to the spectrum of T. We use two classical facts about compact self-adjoint operators.

Lemma 1 (Ky Fan bounds)**.** *For every orthonormal family $\varphi_1, \dots, \varphi_l$ in $L^2_0(\gamma_n)$,*

$$\sum_{j=1}^{l} \langle \varphi_j, T\varphi_j \rangle \leq \sum_{i=1}^{l} \kappa_i, \qquad \sum_{j=1}^{l} \langle \varphi_j, T\varphi_j \rangle^2 \leq \sum_{i=1}^{l} \kappa_i^2.$$

If $\kappa_l > \kappa_{l+1}$, equality in the first bound holds if and only if $\mathrm{span}\{\varphi_j\}$ is the eigenspace of $\kappa_1, \dots, \kappa_l$, and equality in the second holds if and only if, in addition, every φ_j is an eigenfunction of T.

Proof. With $V = \mathrm{span}\{\varphi_j\}$, the first sum is $\mathrm{Tr}(P_V T P_V)$, and the first bound together with its equality case under the gap is Ky Fan's maximum principle. For the second, Cauchy–Schwarz gives $\langle \varphi, T\varphi \rangle^2 \leq \|T\varphi\|^2 = \langle \varphi, T^2\varphi \rangle$ for a unit vector φ, with equality if and only if $T\varphi \parallel \varphi$. Since all κ_i are positive, T^2 has eigenvalues κ_i^2 in the same order and the same eigenspaces, and Ky Fan's principle for T^2 gives the bound and its equality case. □

Lemma 2 (Interlacing). *Let $V \subset L_0^2(\gamma_n)$ have dimension n and let $\mu_1 \geq \cdots \geq \mu_n$ be the eigenvalues of $P_V T P_V$ on V. Then $0 \leq \mu_i \leq \kappa_i$ for every i, so the dense-predictor term satisfies $n - \sum_i \mu_i^2 \geq n - \sum_{i \leq n} \kappa_i^2$. If $\kappa_n > \kappa_{n+1}$, equality holds if and only if V is the eigenspace of $\kappa_1, \ldots, \kappa_n$.*

Proof. Positivity of T gives $P_V T P_V \succeq 0$, and the Courant–Fischer characterization $\mu_i = \max_{U \subseteq V, \dim U = i} \min_{u \in U, \|u\|=1} \langle u, Tu \rangle$ gives $\mu_i \leq \kappa_i$. Equality in the sum of squares therefore forces $\mu_i = \kappa_i$ for every i, so $\mathrm{Tr}(P_V T P_V) = \sum_{i \leq n} \kappa_i$, and Lemma 1 identifies V under the gap. Conversely, the top eigenspace has $\mu_i = \kappa_i$. □

Theorem 1 thus reduces factor recovery to two questions: whether the factor span is the unique optimal span over $\mathcal{C}$, and whether its lag spectrum is simple. The second holds whenever the ρ_i are distinct. The first depends on the constraint class. Over $\mathcal{C}_2$ it is decided by Lemma 2 (Appendix B.4); over $\mathcal{H}$ we give a global sufficient condition and a complete second-order analysis (Appendix B.5); and for information-preserving maps in $\mathcal{H}_{\mathrm{pair}}$ it follows from a rigidity property of the joint Gaussian law (Appendix B.6).

B.3 Linear Axis Recovery (Proposition 1)

We first record what the constraints mean for linear maps.

Lemma 3 (Linear maps). *A linear map $h(z) = Az$ lies in $\mathcal{C}_2$ if and only if $A \in \mathrm{O}(n)$. In that case $h \in \mathcal{H}_{\mathrm{pair}}$, h is information preserving, $\mathbb{E}[h(Z)h(Z')^\top] = A\Lambda A^\top$, and $R_j(h) = \sum_i \rho_i A_{ji}^2$.*

Proof. $\mathrm{Cov}(AZ) = AA^\top$, which equals I_n if and only if A is orthogonal. Then (AZ, AZ') is a linear image of the Gaussian vector (Z, Z'), A is invertible, and $\mathbb{E}[AZ(AZ')^\top] = A\,\mathbb{E}[ZZ'^\top]A^\top = A\Lambda A^\top$, whose diagonal gives R_j. □

Proof of Proposition 1. The class $\{h_Q : Q \in \mathrm{O}(n)\}$ is closed under rotations, every regularizer of §4.1 vanishes on it, and all its members span the factors (Lemma 3). Requirement (a) of Theorem 1 therefore holds for every Q, and the minimizers are exactly the Q for which $K(h_Q) = Q\Lambda Q^\top$ is diagonal. When the ρ_i are pairwise distinct, the eigenvectors of Λ are $\pm e_i$, so $Q = D\Pi$ and $R_j = K_{jj} = \rho_{\pi(j)}$, which is also the optimal coefficient $p_j^\star$. When correlations are tied, $Q\Lambda Q^\top$ is diagonal exactly when Q permutes the eigenspaces of Λ and acts orthogonally within each tied one.

The same conclusion can be reached without the decomposition. The matrix $S_{ji} = Q_{ji}^2$ is doubly stochastic, so Jensen's inequality gives $\sum_j R_j^2 = \sum_j (\sum_i S_{ji}\rho_i)^2 \leq \sum_{j,i} S_{ji}\rho_i^2 = \sum_i \rho_i^2$, with equality if and only if every row of S is supported on a block of equal correlations. □

In two dimensions the off-diagonal entry of $Q\Lambda Q^\top$ for a rotation by θ is $\frac{1}{2}(\rho_1 - \rho_2)\sin 2\theta$, and summing its square over the two off-diagonal positions gives equation 4.

The deployed loss treats all coordinates symmetrically and recovers the factors only up to permutation. A fixed weighted objective also fixes the order.

Theorem 3 (Axis selection for a fixed weighted objective). *Let $\rho_1 > \cdots > \rho_n$ and $w_1 > \cdots > w_n > 0$. Over $h(z) = Qz$, $Q \in \mathrm{O}(n)$, the minimizers of L_w are exactly $Q = D$, that is, the map that sends the j-th slowest factor to coordinate j, up to signs. In two dimensions, with $Q(\theta)$ the rotation by θ, $L_w(\theta) = 2\big(w_1(1-\rho_1) + w_2(1-\rho_2) + G\sin^2\theta\big)$ with $G = (\rho_1 - \rho_2)(w_1 - w_2)$.*

Proof. With $S_{ji} = Q_{ji}^2$ as above, $\frac{1}{2}L_w = \sum_j w_j(1 - R_j) = \sum_{j,i} w_j(1-\rho_i)S_{ji}$ is a linear function of S. Over the Birkhoff polytope, which contains every such S, its minimum is attained on the face spanned by the minimizing permutation matrices. Since both (w_j) and $(1-\rho_i)$ are strictly monotone, the strict rearrangement inequality shows that the only minimizing permutation pairs the largest weight with the smallest $1-\rho_i$, and so on; this is the identity. The minimizing face is therefore the single vertex $S = I$, which forces $Q = D$. In two dimensions, $\frac{1}{2}L_w = w_1[\cos^2\theta(1-\rho_1) + \sin^2\theta(1-\rho_2)] + w_2[\sin^2\theta(1-\rho_1) + \cos^2\theta(1-\rho_2)]$, which rearranges to the stated expression. □

B.4 Whitening and Nonlinear Modes (Proposition 2)

Over $\mathcal{C}_2$ the encoder may use any whitened nonlinear feature, and the score of such a feature is an average of Hermite eigenvalues.

Lemma 4 (Scores of nonlinear features)**.** *Let $\varphi \in L^2_0(\gamma_n)$ have unit variance and Hermite expansion $\varphi = \sum_{k\neq 0} c_k \mathrm{He}_k$. Then $\sum_k c_k^2 = 1$ and $\langle \varphi, T\varphi\rangle = \sum_k c_k^2 \lambda_k$ is a convex combination of Hermite eigenvalues. Among modes of degree one the eigenvalues are the ρ_i, and among modes of degree at least two the largest eigenvalue is $\rho^2_{\max}$.*

Proof. This follows from orthonormality and equation 10; for $|k| \geq 2$, $\lambda_k \leq \rho^{|k|}_{\max} \leq \rho^2_{\max}$, with equality at $k = 2e_{i^*}$. □

Proof of Proposition 2. Let $h \in \mathcal{C}_2$, so that $h_1, \ldots, h_n$ are orthonormal in $L^2_0(\gamma_n)$. Suppose first that $\rho_1^2 < \rho_n$. By Lemma 4 every mode of degree at least two has eigenvalue at most $\rho_1^2 < \rho_n$, so $\kappa_i = \rho_i$ for $i \leq n$, the top-n eigenspace is $\mathrm{span}\{z_1, \ldots, z_n\}$, and $\kappa_n = \rho_n > \kappa_{n+1} = \rho_1^2$. Lemma 2 then shows that the dense-predictor term is minimized over $\mathcal{C}_2$ exactly by the families that span the factors, namely the orthogonal maps $z \mapsto Qz$; this is requirement (a). Theorem 1 and Proposition 1 leave exactly the signed permutations.

Conversely, suppose that $\rho_1^2 \geq \rho_n$. The family $(z_1, \ldots, z_{n-1}, \mathrm{He}_2(z_1))$ lies in $\mathcal{C}_2$ and has $\sum_j R_j^2 = \sum_{i<n} \rho_i^2 + \rho_1^4 \geq \sum_i \rho_i^2$, so its loss is at most that of any signed permutation. If $\rho_1^2 > \rho_n$ the inequality is strict and signed permutations are not minimizers; if $\rho_1^2 = \rho_n$ this nonlinear map attains the same value, so signed permutations cannot be the unique minimizers. □

The competing family is not available to SIGReg, because its last coordinate is not Gaussian.

Lemma 5 (The bare harmonic is whitened but not Gaussian)**.** $\mathrm{He}_2(z_1) = (z_1^2 - 1)/\sqrt{2}$ *has mean zero, unit variance, score ρ_1^2 and excess kurtosis 12. Hence the competing family above lies in $\mathcal{C}_2 \setminus \mathcal{H}$.*

Proof. $\mathbb{E}[(z^2-1)^4] = 105 - 4\cdot 15 + 6\cdot 3 - 4\cdot 1 + 1 = 60$ and $\mathrm{Var}(z^2-1) = 2$, so the kurtosis is $60/4 = 15$. □

B.5 Marginal Gaussian Matching: Global Regime and Local Stability

This subsection studies $(\mathrm{P}_{\mathcal{H}})$ and the weighted problems $\max_{\mathcal{H}} J_w$. Inside the whitening regime the global statements follow from the inclusion $\mathcal{H} \subseteq \mathcal{C}_2$. Beyond it we characterize, in every dimension, when the axis map is a second-order local optimum.

Corollary 1 (Global recovery in the whitening regime)**.** *If the ρ_i are pairwise distinct and $\rho_1^2 < \rho_n$, then $(\mathrm{P}_{\mathcal{H}})$ recovers the factors. For weights $w_1 > \cdots > w_n > 0$ the maximizers of J_w over $\mathcal{C}_2$, and hence over $\mathcal{H}$, are exactly the maps $z \mapsto Dz$.*

Proof. $\mathcal{H} \subseteq \mathcal{C}_2$ contains every signed permutation, so the first claim is Proposition 2. For the second, the first bound of Lemma 1 gives $S_l := \sum_{j\leq l} R_j \leq \sum_{i\leq l} \rho_i =: P_l$ for every l, because $\kappa_i = \rho_i$ for $i \leq n$. Abel summation with $u_l = w_l - w_{l+1} > 0$ ($w_{n+1} := 0$) gives $J_w(h) = \sum_l u_l S_l \leq \sum_l u_l P_l = \sum_j w_j \rho_j$. Equality forces $S_l = P_l$ for all l; at $l = n$ and under the gap $\rho_n > \rho_1^2$ this places h in the degree-one span, where Theorem 3 applies. □

The Abel-summation argument also shows what global optimality over $\mathcal{H}$ would require in general: for a fixed h, $J_w(h) \leq \sum_j w_j \rho_j$ holds for every strictly decreasing w if and only if $S_l \leq P_l$ for every l. The case $l = n$ is the transport inequality $\mathbb{E}\|h(Z') - h(Z)\|^2 \geq \mathbb{E}\|Z' - Z\|^2$. Over $\mathcal{C}_2$ the family $(z_1, \mathrm{He}_2(z_1))$ violates it at $l = 2$ as soon as $\rho_1^2 > \rho_2$; over $\mathcal{H}$ the results below show that it fails whenever the axis map is second-order unstable.

B.5.1 CURVES IN $\mathcal{H}$ AND THE SECOND VARIATION

For a vector field b write $\delta b := \sum_i (z_i b_i - \partial_i b_i)$, so that $\mathbb{E}[b \cdot \nabla\varphi] = \mathbb{E}[\varphi\, \delta b]$ for smooth compactly supported φ. In the Hermite basis $\delta = \sum_i a_i^\dagger$ with $a_i^\dagger \mathrm{He}_k = \sqrt{k_i + 1}\, \mathrm{He}_{k+e_i}$, because $(z_i - \partial_i)\mathrm{He}_k = \sqrt{k_i+1}\, \mathrm{He}_{k+e_i}$.

Lemma 6 (Flows of divergence-free fields)**.** *Let b be a C^1 field with $\delta b = 0$ whose flow Ψ_s is defined for all s outside a γ_n-null invariant set. Then every Ψ_s preserves γ_n, so $h_s := \Psi_s \in \mathcal{H}$, and $U_s\varphi := \varphi \circ \Psi_s$ is a strongly continuous unitary group on $L^2(\gamma_n)$. If b_j and $b \cdot \nabla b_j$ lie in $L^2(\gamma_n)$ for all j, then $s \mapsto h_s = (U_s z_j)_j$ is C^2 into $L^2(\gamma_n)^n$ with $\dot h_0 = b$.*

Proof. $\delta b = 0$ is equivalent to $\mathrm{div}(\gamma_n b) = 0$, so Liouville's theorem gives $(\Psi_s)_\# \gamma_n = \gamma_n$. Measure preservation gives $\|U_s\varphi\| = \|\varphi\|$ and $U_s U_t = U_{s+t}$, and strong continuity follows from continuity of the flow and density of continuous bounded functions. For $\varphi = z_j$, $U_{s+\epsilon} z_j - U_s z_j = \int_s^{s+\epsilon} U_r b_j\, dr$ as an L^2-valued Bochner integral whose integrand is continuous in r, so the curve is C^1 with derivative $U_s b_j$; applying the same identity to b_j, whose derivative along the flow is $b \cdot \nabla b_j \in L^2$, gives C^2. □

Proposition 3 (First and second variation)**.** *Let $s \mapsto h_s$ be a curve in $\mathcal{H}$ that is C^1 into $L^2(\gamma_n)^n$ with $h_0 = \mathrm{id}$, and let $v := \dot h_0$. Then (i) $\langle z_j, v_j\rangle = 0$ for every j; (ii) $\mathbb{E}[\nabla\varphi \cdot v] = 0$ for every φ in the Gaussian Sobolev space $W^{1,2}(\gamma_n)$, in particular for every polynomial; (iii) $\frac{d}{ds} J_w(h_s)|_0 = 0$; and (iv) if the curve is C^2,*

$$\tfrac{1}{2}\, \tfrac{d^2}{ds^2} J_w(h_s)\big|_0 = Q_w(v) := \sum_{j=1}^n w_j \big(\langle v_j, T v_j\rangle - \rho_j \|v_j\|^2\big).$$

Proof. (i) $\|h_{s,j}\|^2 = 1$ for all s; differentiate at 0. (ii) For $\varphi \in C_c^\infty$, $\mathbb{E}[\varphi(h_s)] = \mathbb{E}[\varphi]$ for all s, and a second-order Taylor bound with bounded Hessian gives $\mathbb{E}[\varphi(h_s)] - \mathbb{E}[\varphi] = \mathbb{E}[\nabla\varphi \cdot (h_s - h_0)] + O(\|h_s - h_0\|^2)$; dividing by s yields $\mathbb{E}[\nabla\varphi \cdot v] = 0$. The functional $\varphi \mapsto \mathbb{E}[\nabla\varphi \cdot v]$ is continuous on $W^{1,2}(\gamma_n)$, in which C_c^∞ is dense, and polynomials belong to $W^{1,2}(\gamma_n)$. (iii) $\frac{d}{ds}\langle h_{s,j}, T h_{s,j}\rangle|_0 = 2\langle v_j, T z_j\rangle = 2\rho_j \langle v_j, z_j\rangle = 0$ by (i). (iv) Since T is bounded and self-adjoint, $\frac{1}{2}\frac{d^2}{ds^2}\langle h_{s,j}, T h_{s,j}\rangle|_0 = \langle v_j, T v_j\rangle + \rho_j \langle a_j, z_j\rangle$ with $a := \ddot h_0$, and differentiating $\|h_{s,j}\|^2 = 1$ twice gives $\langle z_j, a_j\rangle = -\|v_j\|^2$. □

We call $v \in L^2(\gamma_n)^n$ *admissible* if it satisfies (i) and (ii), so the velocity of every C^1 curve in $\mathcal{H}$ through the axis map is admissible. The second variation Q_w decouples over Hermite degrees, which reduces its sign to a family of finite-dimensional problems.

B.5.2 BLOCK DECOMPOSITION AND THE SECOND-ORDER CRITERION

Write $\beta_k^{(j)} := \langle v_j, \mathrm{He}_k\rangle$. The map $(j,k) \mapsto (t,i) = (k + e_j, j)$ is a bijection from index pairs onto $\{(t,i) : t \in \mathbb{N}_0^n,\ i \in I_t\}$ with $I_t := \mathrm{supp}(t)$. For each t collect $c_i^t := \beta^{(i)}_{t-e_i}$, $i \in I_t$, and set

$$s_i(t) := w_i\Big(\frac{\lambda_t}{\rho_i} - \rho_i\Big), \qquad V_t := \Big\{c \in \mathbb{R}^{I_t} : \sum_{i \in I_t} \sqrt{t_i}\, c_i = 0\Big\}, \qquad Q_{w,t}(c) := \sum_{i \in I_t} s_i(t)\, c_i^2.$$

Lemma 7 (Block decomposition)**.** *v satisfies (ii) of Proposition 3 if and only if $c^t \in V_t$ for every t, and then $Q_w(v) = \sum_t Q_{w,t}(c^t)$, the series converging absolutely. If $|I_t| = 1$, then $V_t = \{0\}$; in particular (ii) already implies (i), and every nonzero admissible v has $c^t \neq 0$ for some t with $|I_t| \geq 2$.*

Proof. Testing (ii) with $\varphi = \mathrm{He}_t$ and using $\partial_i \mathrm{He}_t = \sqrt{t_i}\mathrm{He}_{t-e_i}$ gives $\sum_{i \in I_t} \sqrt{t_i}\, \beta^{(i)}_{t-e_i} = 0$; conversely these identities give (ii) for all polynomials and hence, by density, for all of $W^{1,2}(\gamma_n)$. Since $\lambda_{t-e_i} = \lambda_t/\rho_i$, the j-th term of Q_w equals $w_j \sum_k (\beta_k^{(j)})^2 (\lambda_k - \rho_j)$, and regrouping the absolutely convergent double series by t gives the block sum. For $t = m e_i$ the only constraint is $\sqrt{m}\, c_i = 0$; the case $m = 2$ is $\beta_{e_i}^{(i)} = \langle v_i, z_i\rangle = 0$. □

Lemma 8 (At most one positive coefficient)**.** *Let the ρ_i be pairwise distinct and $|I_t| \geq 2$. At most one $i \in I_t$ has $s_i(t) \geq 0$, and $s_i(t) < 0$ for all i once $\lambda_t < \rho_{\min}^2$. If all coefficients are negative, or if one is zero and the others negative, $Q_{w,t}$ is negative on $V_t \setminus \{0\}$. If exactly one is positive, then $Q_{w,t}$ takes a positive value on V_t iff $\sum_{i\in I_t} t_i/s_i(t) < 0$, is negative on $V_t \setminus \{0\}$ iff $\sum_{i\in I_t} t_i/s_i(t) > 0$, and is negative semidefinite with a null direction if the sum vanishes.*

Proof. With $r_i = -\log \rho_i > 0$, $s_i(t) \geq 0$ iff $\sum_l t_l r_l \leq 2r_i$. If this held for two indices $i \neq i'$ in I_t, then $\sum_l t_l r_l \leq 2\min(r_i, r_{i'}) \leq r_i + r_{i'} \leq \sum_l t_l r_l$, forcing $r_i = r_{i'}$, which distinctness excludes. The second claim is immediate. If $s_q = 0$ and all other coefficients are negative, then $Q_{w,t}(c) = \sum_{i\neq q} s_i c_i^2 < 0$ unless $c_i = 0$ for all $i \neq q$, in which case the constraint forces $c_q = 0$. For the test, let $s_p > 0 > s_i$ $(i \neq p)$ and $a_i = \sqrt{t_i}$. Eliminating c_p on V_t gives $Q_{w,t} = s_p(\sum_{i\neq p} a_i c_i)^2/a_p^2 - \sum_{i\neq p} |s_i| c_i^2$, whose supremum over c with $\sum_{i\neq p} |s_i| c_i^2 = 1$ is attained and equals $s_p a_p^{-2} \sum_{i\neq p} a_i^2/|s_i| - 1$ by Cauchy–Schwarz. This is positive, negative or zero exactly when $\sum_i a_i^2/s_i$ is negative, positive or zero. □

Lemma 9 (Every block direction is the velocity of a curve in $\mathcal{H}$)**.** *Let $|I_t| \geq 2$, $c \in V_t$, and let $v^{t,c}$ be the polynomial field with $v_i = c_i \mathrm{He}_{t-e_i}$ for $i \in I_t$ and $v_i = 0$ otherwise. For every $\epsilon > 0$ there is a smooth compactly supported field b with $\delta b = 0$ and $\|b - v^{t,c}\|_{L^2(\gamma_n)} < \epsilon$. Its flow is global, lies in $\mathcal{H}$, and is a C^2 curve with velocity b.*

Proof. Fix $p \in I_t$ and define an antisymmetric polynomial matrix by $\pi_{ip} := t_p^{-1/2} c_i \mathrm{He}_{t-e_i-e_p} = -\pi_{pi}$ for $i \in I_t \setminus \{p\}$, all other entries zero; $t - e_i - e_p \geq 0$ because $i, p \in I_t$. For $i \neq p$, $\sum_j a_j^\dagger \pi_{ij} = a_p^\dagger \pi_{ip} = c_i \mathrm{He}_{t-e_i}$, and $\sum_j a_j^\dagger \pi_{pj} = -t_p^{-1/2} \sum_{j\neq p} \sqrt{t_j}\, c_j \mathrm{He}_{t-e_p} = c_p \mathrm{He}_{t-e_p}$ because $c \in V_t$. Hence $v_i = \sum_j a_j^\dagger \pi_{ij}$ for all i. Since $\gamma_n(z_j - \partial_j)\pi = -\partial_j(\gamma_n \pi)$, the matrix $A := -\gamma_n \pi$ is antisymmetric and satisfies $\gamma_n v_i = \sum_j \partial_j A_{ij}$. Let $\chi \in C_c^\infty$ equal 1 on the unit ball, $\chi_R(z) = \chi(z/R)$, and $b_i^R := \gamma_n^{-1} \sum_j \partial_j(\chi_R A_{ij}) = \chi_R v_i - \sum_j (\partial_j \chi_R)\pi_{ij}$. Antisymmetry gives $\mathrm{div}(\gamma_n b^R) = \sum_{i,j} \partial_i \partial_j (\chi_R A_{ij}) = 0$, i.e. $\delta b^R = 0$, and b^R is smooth with compact support, so its flow is global and Lemma 6 applies. Finally $\|b^R - v\| \leq \|(1-\chi_R)v\| + R^{-1}\|\nabla\chi\|_\infty \sum_{i,j} \|\pi_{ij}\| \to 0$. □

Theorem 4 (Second-order characterization of the axis map over $\mathcal{H}$)**.** *Let $\rho_1 > \cdots > \rho_n$ and $w_1, \ldots, w_n > 0$, in any dimension n. Say that the* block condition (B_w) *holds if $Q_{w,t}(c) < 0$ for every t with $|I_t| \geq 2$ and every $c \in V_t \setminus \{0\}$.*

(a) If (B_w) holds, then for every C^2 curve h_s in $\mathcal{H}$ with $h_0 = \mathrm{id}$ and $\dot{h}_0 \neq 0$, $J_w(h_s) < J_w(\mathrm{id})$ for all sufficiently small $s \neq 0$.

(b) If $Q_{w,t}(c) > 0$ for some t and $c \in V_t$, there is a smooth compactly supported field with $\delta b = 0$ whose flow $h_s \in \mathcal{H}$ satisfies $J_w(h_s) > J_w(\mathrm{id})$ for all sufficiently small $s \neq 0$. The axis map is then neither a local nor a global maximizer of J_w over $\mathcal{H}$.

(c) Only finitely many blocks can violate (B_w), and each is decided by Lemma 8.

Proof. (a) By Proposition 3, $J_w(h_s) = J_w(\mathrm{id}) + Q_w(v)s^2 + o(s^2)$ with $v = \dot{h}_0$ admissible, and by Lemma 7 $Q_w(v) = \sum_t Q_{w,t}(c^t) < 0$, since every term is nonpositive and at least one block with $|I_t| \geq 2$ is nonzero. (b) By Lemma 9 choose b with $\|b - v^{t,c}\|$ small; Q_w is a bounded quadratic form, so $Q_w(b) > 0$ because $Q_w(v^{t,c}) = Q_{w,t}(c) > 0$. Along the flow of b, $J_w(h_s) = J_w(\mathrm{id}) + Q_w(b)s^2 + o(s^2)$. (c) is Lemma 8: blocks with $\lambda_t < \rho_{\min}^2$ are negative definite, and $\lambda_t \leq \rho_{\max}^{|t|}$. □

If (B_w) fails but no block has a strictly positive direction, the second variation is degenerate and the theorem makes no claim. The result is local: outside the whitening regime, (B_w) does not exclude a distant map with a larger value of J_w.

Corollary 2 (Deployed objective)**.** *Along every C^2 curve h_s in $\mathcal{H}$ through the axis map with velocity v, $L_{\mathrm{pred}}(h_s) = L_{\mathrm{pred}}(\mathrm{id}) - 2Q_\rho(v)s^2 + o(s^2)$. Hence Theorem 4 applies to $(\mathrm{P}_\mathcal{H})$ with $w = \rho$, block coefficients $s_i(t) = \lambda_t - \rho_i^2$, and "maximizer of J_w" replaced by "minimizer of L_{pred}".*

Proof. By Proposition 3, $R_j(h_s) = \rho_j + \frac{1}{2}R_j''(0)s^2 + o(s^2)$ with $\frac{1}{2}R_j''(0) = \langle v_j, Tv_j\rangle - \rho_j\|v_j\|^2$, so $1 - R_j^2 = 1 - \rho_j^2 - \rho_j R_j''(0)s^2 + o(s^2)$; summing over j gives $-2Q_\rho(v)s^2$. With $w_i = \rho_i$, $s_i(t) = \lambda_t - \rho_i^2$. □

Rotation blocks $t = e_i + e_l$ have $s_i = w_i(\rho_l - \rho_i)$ and $s_l = w_l(\rho_i - \rho_l)$ on $c_i + c_l = 0$, so $Q_{w,t} = c_i^2(\rho_i - \rho_l)(w_l - w_i)$: they are negative iff the heavier weight sits on the slower factor, which recovers Theorem 3, and for the deployed objective ($w = \rho$) they are always negative, which recovers Proposition 1.

B.5.3 The obstruction tower in two dimensions

For $n = 2$ the criterion is explicit. Blocks with $t_1, t_2 \geq 1$ have $\lambda_t \leq \rho_1\rho_2 < \rho_1^2$, so $s_1(t) < 0$, while $s_2(t) > 0$ iff $t_2 = 1$ and $\rho_1^{t_1} > \rho_2$. The only blocks that can fail are therefore $t = (m, 1)$ with $\rho_1^m > \rho_2$: $m = 1$ is the rotation block, and $m \geq 2$ occurs only in the dangerous regime $\rho_1^2 > \rho_2$. Their positive directions are realized exactly, without truncation, by explicit flows.

Lemma 10 (Tower fields). *For $m \geq 2$ let $b^{(m)} = (b_1, b_2)$ with $b_1 = -m^{-1/2}\mathrm{He}_{(m-1,1)}$ and $b_2 = \mathrm{He}_{(m,0)}$. Then $\delta b^{(m)} = 0$ and $\gamma_2 b^{(m)} = (\partial_2\Phi_m, -\partial_1\Phi_m)$ with stream function $\Phi_m = m^{-1/2}\gamma_2\mathrm{He}_{m-1}(z_1)$. The flow conserves Φ_m, is global off the null set $\{\mathrm{He}_{m-1}(z_1) = 0\}$, and $h_s = \Psi_s$ is a C^2 curve in $\mathcal{H}$ with velocity $b^{(m)}$.*

Proof. $a_1^\dagger b_1 = -\mathrm{He}_{(m,1)}$ and $a_2^\dagger b_2 = \mathrm{He}_{(m,1)}$. The identity $(\gamma\mathrm{He}_k)' = -\sqrt{k+1}\,\gamma\mathrm{He}_{k+1}$ gives the stream function. On a level set $\Phi_m = \ell \neq 0$, $|\mathrm{He}_{m-1}(z_1)|e^{-|z|^2/2}$ is bounded below, which bounds z_1 and then z_2; trajectories stay in this compact set and exist for all time. Lemma 6 applies because $b^{(m)}$ is polynomial. □

Proposition 4 (The tower thresholds). *Let $n = 2$ and $\rho_1 > \rho_2$. For $m \geq 2$ with $\rho_1^m > \rho_2$,*

$$Q_w\Big(b^{(m)}\Big) = \frac{w_1}{m}\big(\rho_1^{m-1}\rho_2 - \rho_1\big) + w_2\big(\rho_1^m - \rho_2\big) > 0 \iff \frac{w_2}{w_1} > \tau_m(\rho) := \frac{\rho_1\big(1 - \rho_1^{m-2}\rho_2\big)}{m\big(\rho_1^m - \rho_2\big)}.$$

Let $M := \max\{m : \rho_1^m > \rho_2\}$ and $\tau_*(\rho) := \min_{2\leq m\leq M}\tau_m(\rho)$, with $\tau_* := +\infty$ if $\rho_1^2 \leq \rho_2$. Then (B_w) holds iff $w_1 > w_2$ and $w_2/w_1 < \tau_*(\rho)$, and a strictly positive direction exists iff $w_2 > w_1$ or $w_2/w_1 > \tau_*(\rho)$. For the deployed objective, (B_ρ) holds iff $\rho_2/\rho_1 < \tau_*(\rho)$, and the axis map is second-order unstable iff $\rho_2/\rho_1 > \tau_*(\rho)$.

Proof. b_1 is the single mode $\mathrm{He}_{(m-1,1)}$ with squared norm $1/m$ and eigenvalue $\rho_1^{m-1}\rho_2$, and b_2 is $\mathrm{He}_{(m,0)}$ with norm one and eigenvalue ρ_1^m, which gives $Q_w(b^{(m)})$. The block $t = (m, 1)$ has $V_t = \{\sqrt{m}\,c_1 + c_2 = 0\}$, a line spanned by the coefficients of $b^{(m)}$, so its sign is that of $Q_w(b^{(m)})$; when $\rho_1^m = \rho_2$ the block is negative definite because $s_2 = 0 > s_1$. The rotation block is negative iff $w_1 > w_2$. Combining with the list of candidate blocks gives both claims, and $w = \rho$ gives the deployed case, whose rotation block is always negative. □

The field $b^{(2)}$ moves the second coordinate towards the harmonic $\mathrm{He}_2(z_1)$ of Lemma 5, while a compensating change of the first coordinate keeps the law exactly Gaussian. SIGReg therefore excludes the harmonic itself but not this deformation of the axis map. The binding threshold need not come from the lowest degree: at $\rho = (0.97, 0.50)$, for example, $\tau_2 = 0.550$, $\tau_3 = 0.404$ and $\tau_* = \tau_9 = 0.247$. The deployed ratio $\rho_2/\rho_1 = 0.515$ passes the degree-two test but not the degree-three test, so the axis map of the deployed objective is second-order unstable there. At the OU–Spiral spectrum $\rho = (0.95, 0.70)$ of §5, $\rho_2/\rho_1 = 0.737 > \tau_* = \tau_3 = 0.674$; with $\rho_2 = 0.70$ fixed, the condition $\rho_2/\rho_1 < \tau_*$ fails once ρ_1 exceeds about 0.944.

Proposition 5 (Fine-stride limit). *Write $\rho_i = e^{-\mu_i\delta}$ with $\mu_2 > \mu_1 > 0$ and $L := \mu_2/\mu_1$, independent of δ. The dangerous regime $\rho_1^2 \geq \rho_2$ is $L \geq 2$, and the admissible rungs are the integers $2 \leq m < L$, which exist iff $L > 2$. As $\delta \to 0$ at fixed L,*

$$\tau_m(\rho) \to f_L(m) := \frac{L + m - 2}{m(L - m)}, \qquad \tau_*(\rho) \to F(L) := \min_{2\leq m<L} f_L(m) \quad (F := +\infty \text{ if } L \leq 2),$$

where F is nonincreasing, $F(L) \geq 4/L$, and $F(L) = (3 + 2\sqrt{2} + o(1))/L$ as $L \to \infty$ (real minimizer $m^* = \sqrt{2(L-1)(L-2)} - (L-2)$). Consequently: (i) for fixed weights with $w_2/w_1 = \kappa < 1$, if $F(L) < \kappa$ the axis map of J_w is second-order unstable for all sufficiently small δ; this holds for every L above some $L^*(\kappa) \geq 4/\kappa$, with $L^*(\kappa) \leq (3 + 2\sqrt{2} + o(1))/\kappa$ as $\kappa \downarrow 0$; (ii) for the deployed objective, whose ratio $\rho_2/\rho_1 \to 1$, the axis map is second-order stable for all small δ if $L \leq 4$ (for every δ if $L = 4$) and unstable for all small δ if $L > 4$.

Proof. $\rho_1^m > \rho_2$ iff $m\mu_1 < \mu_2$. Expanding numerator and denominator of τ_m to first order in δ gives $f_L(m)$; convergence is uniform on the finite rung set. Each $f_L(m)$ decreases in L and the rung set grows, so F is nonincreasing. The bounds use $L + m - 2 \geq L$ and $m(L-m) \leq L^2/4$; the stationarity condition of f_L in real m is $m^2 + 2m(L-2) - L(L-2) = 0$, and $m = \alpha L$ with $\alpha = \sqrt{2} - 1$ gives the constant $3 + 2\sqrt{2}$. (i) follows from $\tau_* \to F(L) < \kappa$. For (ii), if $L > 4$ then $F(L) \leq f_L(2) = L/(2(L-2)) < 1$, so eventually $\rho_2/\rho_1 > \tau_*$. If $L < 4$, either no rung exists ($L \leq 2$), or only $m = 2$ with $f_L(2) \geq 3/2$ ($2 < L \leq 3$), or $m \in \{2, 3\}$ with $f_L(2) > 1$ and $f_L(3) = (L+1)/(3(L-3)) \geq 5/3$ ($3 < L < 4$); hence $F(L) > 1$ and eventually $\rho_2/\rho_1 < \tau_*$. At $L = 4$, with $r = \rho_1$ and $\rho_2 = r^4$, directly $\tau_2 = (1 + r^2)/(2r) > 1$ and $\tau_3 = (1 + r^2 + r^4 + r + r^3)/(3r^2) > 1 > r^3 = \rho_2/\rho_1$ for every $r \in (0, 1)$. □

At coarse stride the picture reverses. If $\rho_1 \leq \frac{1}{2}$, then $\tau_m > 1$ for every admissible m: for $m = 2$ this is equivalent to $\rho_1(1 - 2\rho_1) + \rho_2(2 - \rho_1) > 0$, and for $m \geq 3$ it follows from $\tau_m \geq (1 - \rho_1^{m-2}\rho_2)/(m\rho_1^{m-1}) \geq \frac{7}{8} \cdot \frac{4}{3}$, since $\rho_2 < \rho_1^2 \leq \frac{1}{4}$ and $m\rho_1^{m-1} \leq \frac{3}{4}$. Every strictly decreasing weighting, and the deployed objective, then satisfies the block condition however large L is; for instance $\rho = (0.5, 0.5^{20})$ has $L = 20$ and $\tau_* = \tau_2 = 1.000003$. Because τ_m and τ_* depend only on the spectrum, the regime of a trained model can be read off its fitted coefficients $\widehat{r}_j$.

B.6 Global Identification with PairSIGReg (Theorem 2)

We keep the notation of §B.1 and write $h = f \circ g$, $Y = h(Z)$, $Y' = h(Z')$ and $U = [Y; Y'] \in \mathbb{R}^{2n}$. The proof has three steps: the exact zero set of the pair penalty is $\mathcal{H}_{\text{pair}}$; joint Gaussianity together with information preservation forces h to be linear; and on linear maps Proposition 1 applies.

Lemma 11 (Zero set)**.** *Suppose the direction and frequency measures in equation 9 have full support on $\mathbb{S}^{2n-1} \times \mathbb{R}$. Then $\mathcal{R}_{\text{Pair}}(f) = 0$ if and only if $h \in \mathcal{H}_{\text{pair}}$.*

Proof. Characteristic functions are continuous in (v, t), so a vanishing integral of the squared discrepancy forces $\mathbb{E}\exp(itv^\top(U - \mathbb{E}U)) = \exp(-\frac{1}{2}t^2 v^\top \text{Cov}(U)v)$ for every (v, t). Every one-dimensional projection of $U - \mathbb{E}U$ is then Gaussian, and by the Cramér–Wold theorem U is jointly Gaussian. The moment terms vanish iff $\mathbb{E}Y = \mathbb{E}Y' = 0$ and $\text{Cov}(Y) = \text{Cov}(Y') = I_n$, so $Y \sim \gamma_n$ and $h \in \mathcal{H}$. Conversely, every $h \in \mathcal{H}_{\text{pair}}$ makes each term zero. □

The cross-time block $\mathbb{E}[Y'Y^\top]$ is not prescribed but fitted, so the penalty itself prefers no coordinate system. In practice finitely many random slices and quadrature frequencies approximate this zero set; Theorem 2 concerns the exact zero set.

The key consequence of joint Gaussianity is that the span of h becomes an invariant subspace of T.

Lemma 12 (Invariant-subspace identity)**.** *Let $h \in \mathcal{H}_{\text{pair}}$ be information preserving and $K = \mathbb{E}[Y'Y^\top]$. Then K is symmetric positive definite and $Th = Kh$ a.s.*

Proof. (Z, Z') is exchangeable, hence so is (Y, Y'), and $K = K^\top$. For $a \neq 0$ the feature $\varphi = a^\top h$ is centered with variance $\|a\|^2 > 0$, and by equation 10 $a^\top Ka = \langle \varphi, T\varphi \rangle = \sum_{k \neq 0} \widehat{\varphi}_k^2 \lambda_k > 0$. Joint Gaussianity and $\text{Cov}(Y) = I_n$ give $\mathbb{E}[Y' \mid Y] = KY$. Information preservation makes $\sigma(Y) = \sigma(Z)$ up to null sets, so $(Th)(Z) = \mathbb{E}[h(Z') \mid Z] = \mathbb{E}[Y' \mid Y] = Kh(Z)$. □

Invariant subspaces of T with positive eigenvalues consist of polynomials, and a polynomial of a Gaussian vector can itself be Gaussian only if it is affine.

Lemma 13 (Gaussian polynomial rigidity)**.** *If $p : \mathbb{R}^n \to \mathbb{R}$ is a polynomial and $p(Z)$ has a nondegenerate Gaussian law, then p is affine.*

Proof. Suppose $\deg p = d \geq 2$ with leading homogeneous part $p_d \neq 0$, pick u with $p_d(u) \neq 0$ and let $\sigma = \operatorname{sign} p_d(u)$. On a narrow cone around u, $\sigma p(x) \geq c\|x\|^d$ for $\|x\|$ large and some $c > 0$, so $\mathbb{E}e^{s\sigma p(Z)} = \infty$ for every $s > 0$ if $d > 2$ and for every $s > 1/(2c)$ if $d = 2$. A Gaussian variable has finite exponential moments of every order. □

Proposition 6 (Pair Gaussianity forces an orthogonal induced map)**.** *If $h \in \mathcal{H}_{\text{pair}}$ is information preserving, then $h(Z) = QZ$ a.s. for some $Q \in \mathrm{O}(n)$.*

Proof. Diagonalize $K = V\operatorname{diag}(\kappa_1, \ldots, \kappa_n)V^\top$ and set $q = V^\top h$. Lemma 12 gives $Tq_j = \kappa_j q_j$ with $\kappa_j > 0$, so q_j lies in a finite-dimensional Hermite eigenspace (§B.1) and is a polynomial. Since $q(Z) \sim \gamma_n$, Lemma 13 makes every q_j affine: $q(Z) = AZ + c$. The mean gives $c = 0$ and the covariance gives $AA^\top = I_n$, so $h(Z) = VAZ$ with VA orthogonal. □

Proof of Theorem 2. By Lemma 11 and Proposition 6, every information-preserving feasible induced map is $h(z) = Qz$ with $Q \in \mathrm{O}(n)$, and conversely every orthogonal map is feasible and information preserving (Lemma 3). On this class the optimized predictive term is $n - \sum_j R_j^2$, and Proposition 1 shows that for pairwise distinct ρ_i its minimizers are exactly $Q = D\Pi$, with optimal coefficients $p_j^\star = R_j = \rho_{\pi(j)}$. With tied correlations the minimizers additionally include orthogonal transformations within tied blocks. □

Information preservation enters the proof at one point: it upgrades the Gaussian-regression identity $\mathbb{E}[(Th)(Z) \mid h(Z)] = Kh(Z)$ to $Th = Kh$. Like the other results of this section, the theorem characterizes the population minimizers over induced maps, with an encoder class that contains $f = D\Pi \circ g^{-1}$ on the image of g.

B.7 Basis Selection in a Fixed Representation

Proposition 1 optimizes a coordinate system and never uses how the representation was produced. Here the admissible class is the set of whitening linear maps of a fixed representation. Let $e \in \mathbb{R}^d$ be a frozen representation observed on stationary pairs (e, e') with mean μ and covariance $\Sigma \succ 0$ (in practice, the covariance restricted to its stable subspace), and let

$$C_s := \tfrac{1}{2}\,\Sigma^{-1/2}\big(\mathbb{E}[(e-\mu)(e'-\mu)^\top] + \mathbb{E}[(e'-\mu)(e-\mu)^\top]\big)\Sigma^{-1/2}.$$

A map $y = A(e - \mu)$ satisfies $\mathrm{Cov}(y) = I_d$ iff $A = Q\Sigma^{-1/2}$ with $Q \in \mathrm{O}(d)$; write q_j for the rows of Q.

Proposition 7 (Closed form, identifiability, and information invariance)**.**

(i) Over whitening maps, $L_{\text{pred}}(A) = \sum_j \big(1 - (q_j^\top C_s q_j)^2\big) \geq d - \mathrm{Tr}(C_s^2)$, with equality iff every q_j is an eigenvector of C_s. The minimizers are therefore the eigenbases of C_s, unique up to signed permutation when C_s has distinct eigenvalues.

(ii) If $d = n$, $e = Mz + \mu$ with M invertible, and z follows equation 5 with pairwise distinct ρ_i, every minimizer gives $y = D\Pi z$.

(iii) For every invertible A and every target u with $0 < \mathrm{Var}(u) < \infty$, the population least-squares R^2 of u on $y = A(e - \mu)$ equals that on e. Consequently, for every coordinate j, $R^2(u \mid y_j) \leq R^2(u \mid e)$.

Proof. (i) $\mathbb{E}[y_j y_j'] = q_j^\top \Sigma^{-1/2}\mathbb{E}[(e-\mu)(e'-\mu)^\top]\Sigma^{-1/2} q_j = q_j^\top C_s q_j$, because a quadratic form only sees the symmetric part of its matrix, and y_j has unit variance at both times, so the coordinate-wise optimum leaves $1 - (q_j^\top C_s q_j)^2$. By Cauchy–Schwarz $(q_j^\top C_s q_j)^2 \leq \|C_s q_j\|^2 = q_j^\top C_s^2 q_j$, with equality iff $C_s q_j \parallel q_j$, and $\sum_j q_j^\top C_s^2 q_j = \mathrm{Tr}(C_s^2)$. (ii) $\Sigma = MM^\top$, so $W := \Sigma^{-1/2}M$ is orthogonal and $C_s = W\Lambda W^\top$, whose eigenvectors are the columns of W because the ρ_i are distinct. Hence $Q = D\Pi W^\top$ and $y = QWz = D\Pi z$. (iii) The linear spans of $\{1, y_1, \ldots, y_d\}$ and $\{1, e_1, \ldots, e_d\}$ coincide, so the L^2 projections of u onto them coincide; regressing on the single coordinate y_j projects onto a subspace of that span. □

Part (ii) extends Proposition 1 from orthogonal to arbitrary invertible linear mixing, since whitening turns any such mixing into a rotation. Part (iii) is the precondition used in §5.1: the map cannot raise the single-coordinate readout of a factor above its full-dimensional readout, which can be measured before the map is fitted. A high full-dimensional readout is thus necessary for recovery, and the lag spectrum of C_s decides the rest: factors with distinct autocorrelations occupy distinct eigenvectors, whereas factors with equal autocorrelations share an eigenspace, within which the basis is free. During training the same criterion also determines which information the encoder retains, which a map of a fixed representation cannot change.

The closed form in Part (i) coincides with the rotation computed by AMUSE, a classical second-order blind source separation method that diagonalizes a whitened, symmetrized lagged covariance (Tong et al., 1991; Molgedey & Schuster, 1994); linear slow feature analysis leads to the same eigenproblem (Blaschke et al., 2006). We do not start from this criterion: here the rotation follows from the cross-talk term of Theorem 1, as the minimizer of the AP loss over whitening maps. The agreement is doubly useful. The long and successful use of AMUSE in signal processing supports the post-hoc map in practice, and our derivation adds guarantees that the separation criterion does not state by itself: the map minimizes the same predictive loss that is used for training, it identifies the factors (Part (ii)), and it cannot lose linearly decodable information (Part (iii)), which yields the measurable precondition used in §5.1 for released JEPAs.

B.8 Approximate Identifiability

The preceding results assume that the constraints hold exactly. In practice the regularizer is satisfied only approximately and the encoder need not be exactly information preserving. In the spirit of the approximate-identifiability result of Klindt et al. (2026), we show that coordinate recovery degrades gracefully in terms of one measurable quantity, the linear-readout deficit. The bound holds for every whitened encoder, whatever the regularizer; the regularizer matters only through how small it makes the deficit and how far it keeps the lag spectrum of the remaining directions from the factor correlations.

Let (Z, Z') follow equation 5 with $\rho_1 > \cdots > \rho_n$, and let $h : \mathbb{R}^n \to \mathbb{R}^d$, $d \geq n$, have components in $L^2(\gamma_n)$ with $\mathbb{E}h(Z) = 0$ and $\mathrm{Cov}(h(Z)) = I_d$. Write $Y = h(Z)$, $Y' = h(Z')$, $K = \mathbb{E}[YY'^\top]$, which is symmetric by exchangeability, and $B = \mathbb{E}[YZ^\top] \in \mathbb{R}^{d\times n}$. Since $\mathrm{Cov}(Y) = I_d$, the population R^2 of Z_i regressed on all of Y is $\|Be_i\|^2$, and

$$\delta := \textstyle\sum_{i=1}^n \big(1 - R^2(Z_i \mid Y)\big) = n - \|B\|_F^2$$

is the full-dimensional linear-readout deficit, the quantity that the probe R^2 of §5 estimates. It is invariant under invertible linear maps of Y (Proposition 7(iii)). Assume $\delta < 1$. Let $B = RS$ be the polar decomposition, with $R^\top R = I_n$ and $S = (B^\top B)^{1/2}$, let $U = \mathrm{range}(R)$ and $P_\perp = I_d - RR^\top$, and set

$$K_0 := R\Lambda R^\top + P_\perp K P_\perp.$$

The spectrum of K_0 consists of $\rho_1, \ldots, \rho_n$, with unit eigenvectors $q_i := Re_i$, together with the spectrum of $N := P_\perp K P_\perp$ on $U^\perp$, which is the lag matrix of the directions that carry no linear information about the factors. Let g_i be the distance from ρ_i to the rest of the spectrum of K_0, that is, the smaller of $\min_{k\neq i} |\rho_i - \rho_k|$ and the distance from ρ_i to the spectrum of N. For orthonormal directions $v_1, \ldots, v_d$, the coordinates are $u_j = v_j^\top Y$, the fitted coefficients are $r_j = v_j^\top K v_j$ (equation 7), and the cross-talk is $\eta = \sum_{j\neq l} (v_j^\top K v_l)^2$. By Theorem 1, $\eta = 0$ exactly when the v_j are eigenvectors of K, which is the optimal basis for this span (Proposition 7(i)).

Proposition 8 (Approximate recovery)**.** *In this setting:*

(i) $\|K - K_0\|_{\mathrm{op}} \leq \varepsilon(\delta) := (2\rho_1 + \rho_1^2)\,\delta + \rho_1^2\sqrt{\delta}\,\mathbb{1}[d > n]$.

(ii) Suppose $g_i > 0$. Let $j(i)$ index the coordinate whose coefficient r_j has the same rank among $r_1, \ldots, r_d$ as ρ_i has in the spectrum of K_0, and let $s_i := 2(\|K - K_0\|_{\mathrm{op}} + \sqrt{\eta})/g_i$. Then

$$\big|\mathrm{corr}(u_{j(i)}, Z_i)\big| \;\geq\; \sqrt{1 - \min(1, s_i^2)} - \delta, \qquad |r_{j(i)} - \rho_i| \leq \|K - K_0\|_{\mathrm{op}} + \sqrt{\eta}.$$

In particular, if $d = n$ and $\delta = \eta = 0$, then every coordinate is $\pm$ one factor and $r_{j(i)} = \rho_i$, as in Theorem 2.

Proof. Let $\epsilon := Y - BZ$. Because $\mathrm{Cov}(Z) = I_n$, BZ is the L^2 projection of Y onto the factors, so $\mathbb{E}\epsilon = 0$ and $\mathbb{E}[\epsilon Z^\top] = 0$. Every component of ϵ therefore lies in the closed span of $\{\mathrm{He}_k : |k| \geq 2\}$, where T has norm at most ρ_1^2 by equation 10. Since $\mathbb{E}[Z' \mid Z] = \Lambda Z$ and (Z, Z') is exchangeable, $\mathbb{E}[\epsilon Z'^\top] = \mathbb{E}[\epsilon Z^\top]\Lambda = 0$ and similarly $\mathbb{E}[Z\epsilon'^\top] = 0$, which gives

$$K = B\Lambda B^\top + K_\epsilon, \quad K_\epsilon := \mathbb{E}[\epsilon\epsilon'^\top], \qquad I_d = \mathrm{Cov}(Y) = BB^\top + \mathrm{Cov}(\epsilon),$$

with $a^\top K_\epsilon b = \langle a^\top\epsilon, T\, b^\top\epsilon\rangle$ for fixed $a, b \in \mathbb{R}^d$.

We now bound the two parts. Let $\sigma_1, \ldots, \sigma_n$ be the singular values of B. Since $\mathrm{Cov}(\epsilon) \succeq 0$ we have $\sigma_i \leq 1$, and $\sum_i(1 - \sigma_i^2) = \delta$, so $\sigma_i \geq \sqrt{1-\delta}$ and $\|I - S\| \leq 1 - \sqrt{1-\delta} \leq \delta$. Next, $P_U\mathrm{Cov}(\epsilon)P_U = R(I - S^2)R^\top$ has norm at most δ, and $P_\perp\mathrm{Cov}(\epsilon)P_\perp = P_\perp$. For unit a, b the Cauchy–Schwarz inequality gives $|a^\top P_U K_\epsilon P_U b| \leq \rho_1^2\delta$ and $|a^\top P_U K_\epsilon P_\perp b| \leq \rho_1^2\sqrt{\delta}$. Because $P_\perp B = 0$, in the decomposition $U \oplus U^\perp$

$$K - K_0 = \begin{pmatrix} R(S\Lambda S - \Lambda)R^\top + P_U K_\epsilon P_U & P_U K_\epsilon P_\perp \\ P_\perp K_\epsilon P_U & 0 \end{pmatrix}.$$

Here $S\Lambda S - \Lambda = (S - I)\Lambda S + \Lambda(S - I)$ has norm at most $2\rho_1\delta$. A symmetric block matrix with a zero lower-right block has norm at most the norm of its upper-left block plus the norm of its off-diagonal block, which proves (i). The off-diagonal block is absent when $d = n$.

For (ii), let $\widehat{\Sigma} := \sum_j r_j v_j v_j^\top$. Then $\|K - \widehat{\Sigma}\|_{\mathrm{op}} \leq \|K - \widehat{\Sigma}\|_F = \sqrt{\eta}$, so $\|\widehat{\Sigma} - K_0\|_{\mathrm{op}} \leq \|K - K_0\|_{\mathrm{op}} + \sqrt{\eta}$. The matrix $\widehat{\Sigma}$ has eigenvectors v_j and eigenvalues r_j, and q_i is an eigenvector of K_0 whose eigenvalue ρ_i is separated from the rest of its spectrum by g_i. The Davis–Kahan theorem in the form of Yu et al. (2015, Cor. 1) gives $\sin\angle(v_{j(i)}, q_i) \leq s_i$, and Weyl's inequality gives the bound on $r_{j(i)}$. Finally, $u_{j(i)}$ and Z_i have unit variance, so $\mathrm{corr}(u_{j(i)}, Z_i) = v_{j(i)}^\top Be_i = v_{j(i)}^\top q_i - v_{j(i)}^\top R(I - S)e_i$. The first term has absolute value $\cos\angle(v_{j(i)}, q_i)$, and the second is at most $\|I - S\| \leq \delta$. □

Under the assumptions of Theorem 2, $\delta = 0$, and at its minimizers $\eta = 0$, so the proposition recovers its conclusion. The bound needs neither PairSIGReg nor information preservation. For $d = n$ it degrades linearly in δ, and with extra dimensions the leakage between U and $U^\perp$ adds a $\sqrt{\delta}$ term. The gap g_i has two parts. The first is the spacing of the factor correlations. The second is the distance to the lag spectrum N of the retained non-factor directions, and this is where the regularizer hierarchy reappears as a margin: a retained harmonic $\mathrm{He}_2(z_1)$ enters N at ρ_1^2, so the fastest factor keeps a gap only if $|\rho_n - \rho_1^2|$ is large (Proposition 2).

The bound extends in three routine directions. If a block I of factors shares one correlation, for example the $(\cos, \sin)$ pair of a circular angle, the block form of the Davis–Kahan theorem (Davis & Kahan, 1970) bounds the largest principal angle between the corresponding coordinate span and $\mathrm{span}\{q_i : i \in I\}$ by $2(\|K - K_0\|_{\mathrm{op}} + \sqrt{\eta})/g_I$, where g_I is the distance from the block to the rest of the spectrum. The basis inside the block remains free. This residual $\mathrm{O}(2)$ freedom is the sign and offset that the Reacher shoulder calibration fits (Appendix C.5). With K estimated from N pairs, add $\|\widehat{K} - K\|_{\mathrm{op}} = O_P(N^{-1/2})$ to the numerator. If $\mathrm{Cov}(Y) = I + E_w$ with $\|E_w\| \leq \gamma < 1$, apply the proposition to $\mathrm{Cov}(Y)^{-1/2}Y$: δ is unchanged and each coordinate moves by $O(\gamma)$.

Part (ii) is stated in terms of $\|K - K_0\|_{\mathrm{op}}$ itself, which can be computed on labelled evaluation data, and this step uses only the Davis–Kahan and Weyl inequalities, not equation 5; the worst-case bound $\varepsilon(\delta)$ of Part (i) is needed only when K_0 is not available. Appendix D.7 evaluates the bound on Reacher-hard. An exact population check on Hermite-feature encoders (degrees one to three, $n \leq 3$, $d \leq n + 4$, several thousand random and adversarially optimized maps) found no violation of (i), (ii) or the tied-block bound.

C Experimental Setup

C.1 Benchmarks and Pair Construction

Every benchmark has known ground-truth factors, which are used only to score the frozen representation. Images are fed at 64×64, and angular factors are scored as $(\cos, \sin)$ pairs whose two columns are averaged into one factor (Appendix C.3).

OU-Spiral. The latent $z_t \in \mathbb{R}^2$ is an Ornstein–Uhlenbeck process with $(\rho_1, \rho_2) = (0.95, 0.70)$, observed through the fixed invertible spiral map $x = (r\cos\theta, r\sin\theta)$ with $r = \|z\|$ and $\theta = \mathrm{atan2}(z_2, z_1) + 1.5r$. It is a vector-observation benchmark with native pairs (x_t, x_{t+1}), scored on 10,000 held-out samples.

Synthetic OU pixels. The three synthetic pixel settings of Table 8 render the two-joint Reacher arm from a fixed camera, with its shoulder and wrist angles set to a two-factor OU latent with standard normal marginals and the stated lag-one autocorrelations; the target is pinned, so each frame is a deterministic function of the two factors. Each setting has 100,000 training pairs (x_t, x_{t+1}). The first, $\rho = (0.70, 0.60)$, satisfies the whitening condition $\rho_1^2 < \rho_2$; the other two violate it.

DeepMind Control. Reacher-hard and Cartpole swing-up (Tassa et al., 2018) are rendered from a fixed camera while an Ornstein–Uhlenbeck exploration policy drives the actions; on Reacher the policy has $\rho = 0.9$, $\sigma = 0.6$ and a burn-in of 100 steps, and the target is pinned. A positive pair is two frames of one trajectory δ steps apart, with $\delta = 56$ on Reacher and $\delta = 16$ on Cartpole. Reacher uses 850 training episodes (90,100 pairs) and 32 evaluation episodes, and Cartpole $240/30/30$ training, development and test trajectories; test episodes never appear in training. Reacher is scored on the shoulder angle and the wrist angle, Cartpole on the cart position and the pole angle; velocities are not visible in a single frame and are not scored. Finger-spin, used only in Appendix D.8, follows the same protocol at stride 8.

Image datasets. MPI3D (Gondal et al., 2019), Falcor3D and Isaac3D have no time axis, so a pair consists of two images whose scored factors differ by a small random step. For each of the real, toy and realistic versions of MPI3D we step the two 40-cell arm axes independently by a uniform draw in $\{-3, \ldots, +3\}$, reflect at the boundaries, and redraw zero net motion; the five categorical scene factors are held fixed within a pair. The two arm axes are scored, and each version is split by scene into 458 training, 88 development and 102 test scenes (50,380 training pairs). MPI3D-real is photographed and the other two are rendered, so the body averages them as one dataset family rather than three independent sources. Falcor3D and Isaac3D were both released by Nie et al. (2020); we use the 128×128 archives of Zenodo record 3669344, downsampled to 64×64. On Falcor3D each pair steps each camera-position axis by ± 1 cell with reflection and redraws the target's lighting tuple uniformly from the training tuples; the three camera coordinates are scored, and the data are split by lighting tuple into $756/162/162$ groups. On Isaac3D each scored factor (robot x and y, camera height, object scale, light intensity and light y-direction) steps independently in $\{-1, 0, +1\}$, with the all-zero step redrawn and reflection at the boundaries, while the target's categorical scene (shape, object colour, wall colour) is redrawn from the training tuples; the data are split by categorical tuple into $34/7/7$ groups. Both use 100,000 training pairs.

C.2 Models and Training

C.2.1 Pixel-based benchmarks

The pixel models trained from scratch share the CNN encoder of Klindt et al. (2026) (≈1.1M parameters, BatchNorm) on raw frames and are trained without labels; simulator state is used only to score the frozen representation. The four internal arms differ only in the predictor (identity, dense, or diagonal) and the regularizer (SIGReg or PairSIGReg). The identity+SIGReg arm is the alignment-plus-SIGReg core of LeJEPA applied to our pairs without a projection head, and we label it LeJEPA throughout. The dense predictor is the closed-form ridge least-squares matrix from y to y' on each batch, and the anisotropic predictor uses the closed-form coefficients $r_j = \overline{y_j y'_j} / \overline{y_j^2}$; neither has learned parameters. The ablations of Tables 5 and 10 add a shared-scalar predictor $P = aI$, whose single closed-form coefficient $a = \sum_j \overline{y_j y'_j} / \sum_j \overline{y_j^2}$ is shared by all coordinates, and pair every predictor with either regularizer. All pixel models use a 32-dimensional representation and are trained with AdamW (learning rate 3×10^{-3}, weight decay 10^{-4}, cosine schedule) at batch size 256 for 100 epochs, with 256 SIGReg slices.

External baselines. VICReg, BYOL and Barlow Twins are trained on the same pairs with the same encoder, optimizer and schedule, and are scored on the same encoder output; their method-specific heads are discarded before scoring. VICReg uses a three-layer projector of width 256 with

invariance, variance and covariance weights $(25, 25, 1)$. Barlow Twins uses the same projector shape with off-diagonal weight 0.0051. BYOL uses projector and predictor MLPs of width 256 and an EMA target network whose momentum rises from 0.99 to 1 on a cosine schedule. SimDINO (Wu et al., 2025) uses its published coding-rate objective (two views, $\varepsilon = 0.5$, no local crops), a DINO head $32 \to 2048 \to 2048 \to 256$ with ℓ_2-normalized output, and an EMA teacher whose momentum rises from 0.996 to 1 on a cosine schedule; as in its reference implementation, the scored representation is the teacher encoder's output before the head. On OU-Spiral the same four losses replace the JEPA objective on the MLP encoder of that benchmark, with head input width 2.

Within a setting, all internal arms share one regularization weight: $\lambda = 0.03$ on the $(0.95, 0.60)$ and $(0.97, 0.50)$ synthetic settings and Cartpole, $\lambda = 0.01$ on the $(0.70, 0.60)$ synthetic setting, Reacher, MPI3D and Falcor3D, and $\lambda = 0.1$ on Isaac3D. Reacher uses $\mathcal{L}_{\text{pred}} + \lambda R$ and the synthetic settings use $(1 - \lambda)\mathcal{L}_{\text{pred}} + \lambda R$. The pair weight is $\eta = 1$ on every pixel benchmark, the default, which was not searched there, and $\eta = 0.1$ on the $(0.95, 0.60)$ synthetic setting. The SIGReg weights and the synthetic η were chosen on development seeds (Appendix D.6). The reported runs use five fresh confirmation seeds (for example 41–45 on Reacher-hard and 25–29 on MPI3D-real) and a held-out test split.

C.2.2 OU–Spiral confirmation runs

OU–Spiral uses an MLP encoder $2 \to 128 \to 128 \to 2$ with GELU, trained for 16k steps at batch size 1024, with SIGReg weight 0.1, joint PairSIGReg weight 0.2, and no explicit covariance penalty. Its confirmation seeds are 41–45; development seeds 0–2 are not pooled with them.

C.2.3 Controlled mechanism studies

The controlled studies in Appendices D.1 and D.2 use a lighter setup. Observations are $x = g(z)$ for a frozen random MLP $g : \mathbb{R}^n \to \mathbb{R}^{d_{\text{obs}}}$, and the encoder is an MLP $\mathbb{R}^{d_{\text{obs}}} \to 64 \to 64 \to n$ with ReLU. Unless stated otherwise the objective is the rotation-invariant whitening loss $\text{MSE}(\hat{y}, y') + 25\|\text{Cov}(y) - I\|_F^2/n$, and the anisotropic predictor uses the closed-form coefficient $r_j = \overline{y_j y'_j}/\overline{y_j^2}$, with gradients flowing through y and y'. We train with Adam at learning rate 10^{-3} and batch size 512, drawing fresh OU pairs at every step.

These controlled studies use three development seeds unless stated otherwise; they are kept separate from the five-seed benchmark confirmation results.

C.3 Metrics and Probe Evaluation

Linear probes are fitted on training frames and scored on the held-out split. MCC is computed on held-out frames without any fitted map: we form the matrix of absolute Pearson correlations between the d coordinates and the scored factor columns, match each column to a distinct coordinate with the Hungarian algorithm, and average the matched entries, so a representation scores well only when each factor sits on its own coordinate. Full R^2 regresses each factor on all coordinates with a ridge probe (penalty 10 on coordinates standardized with training statistics). The optimal population linear readout is invariant to invertible linear reparameterizations; fitted probes also depend on the fitting procedure. We average it over physical factors, first averaging the $(\cos, \sin)$ columns of each angle so that a circular factor is not counted twice, and also report the worst factor. Worst-factor R^2 is the lowest per-factor full R^2 and exposes a lost factor. Single-coordinate R^2 fits each factor affinely on its MCC-matched coordinate only. On OU-Spiral, the axis deviation of Figure 3(b) is the angle between the learned coordinate axes and the factor axes, from $0°$ (aligned) to $45°$ (fully mixed in two dimensions). In the controlled studies, the He_2 content is the squared correlation between the coordinate matched to the fast factor z_2 and the harmonic $\text{He}_2(z_1) \propto z_1^2 - 1$ of the slow factor. Dispersion is the sample standard deviation over seeds ($\text{ddof} = 1$). Both AP regularizers are reported; neither is selected on test performance.

C.4 Frozen-Representation Transformation

Appendix B.7 derives the post-hoc mode used in Table 4: for a fixed representation, the optimum of the criterion over whitening linear maps is the eigenbasis of the symmetrized lag-one covariance

(Proposition 7). Our implementation whitens the stable subspace exactly and rescales the low-variance complement. On the released backbones the pooled covariance has full rank (1280 for I-JEPA, 1024 for V-JEPA), and the implementation agrees with the closed form to six decimals on every column of Table 4. The controlled readout comparisons are in Appendix D.4.

MPI3D pairs for Table 4. Step (b) needs distinct lag correlations, and the MPI3D pairs used to *train* the encoders of Table 3 do not provide them: one uniform step in $\{-3, \dots, 3\}$ on each axis, under the same law, gives $\rho = .985$ for both arm axes, and the scene is frozen (Table 18). The transformation of Table 4 is therefore fitted on a different unlabelled pair construction, a frozen walk manifest shared by both backbones. Each pair consists of the two ends of a 32-step walk on the arm grid, frames 15 and 31 for I-JEPA and the clips 0:16 and 16:32 for V-JEPA; features are the spatial mean of the EMA target encoder's output before its final layer normalization. The two axes move independently under different laws. The horizontal axis moves with probability .8 by $\pm$1–2 cells, the vertical axis with probability .85 by $\pm$1–4, and proposals that leave the grid are not taken. Between the two halves the scene is redrawn from the same pool with probability .5. The transformation uses 8,784 pairs from 366 training scenes. The factor labels of 458 training scenes serve only for the MCC matching, and scores are computed on the 102 test scenes, which were encoded only after the map was frozen. On these pairs the arm axes have $\rho = .896$ and .716.

Action recognition on K400 and SSv2. The top-1 columns of Table 4 check that the transformation removes no information useful beyond MPI3D. We use the official I-JEPA ViT-H/14 (ImageNet-1K, 300 epochs) and V-JEPA ViT-L/16 checkpoints with no further training, and the same pooled pre-normalization EMA-target features as above. Each video contributes one deterministic center-cropped clip of 16 frames (frame step 4 on K400 and 1 on SSv2); I-JEPA encodes four of its frames separately and averages their spatial means, and V-JEPA takes the spatial mean of its last temporal tubelet. For the native features and for their AP transform, obtained with the MPI3D map above applied unchanged, we fit separate but identically specified linear softmax heads with bias on features standardized with training statistics, minimizing the mean cross-entropy plus $\frac{\lambda}{2}\|W\|_F^2$ with full-batch L-BFGS. The weight $\lambda \in \{10^{-6}, 10^{-5}, \dots, 10^{-1}\}$ is selected by top-1 on a class-stratified $80/20$ split of the training videos, after which the head is refitted on all 59,629 K400 or 168,913 SSv2 training videos. Top-1 is reported on the official validation sets (19,877 and 24,777 videos), which are used only for scoring. These linear heads on pooled features measure linear readability, so their absolute accuracies are not comparable with the attentive-probe results published for these models.

C.5 Calibration and Overlay Evaluation

For the few-label readout, K counts the labelled calibration frames used after the encoder is frozen. The optional post-hoc rotation uses training pairs, and test labels are used only for scoring. The overlay evaluation uses development data to fix coordinate identities (Appendix D.5).

Coordinate selection in the K-label experiment. All rules are those of a frozen readout script, fixed before the independent test trajectories were generated and applied unchanged to every arm. The shoulder is Reacher's circular joint, so it is read from a *pair* of coordinates, as the angle of the whitened pair. The pair is selected *without labels*: among all $\binom{32}{2} = 496$ pairs we take the one whose whitened radius has the smallest coefficient of variation on unlabelled training frames, since a $(\cos, \sin)$ embedding of one angle has constant radius. The K labelled training frames then fit two parameters, a sign and an angular offset, which are exactly the $\mathrm{O}(2)$ freedom that the tied $(\cos, \sin)$ correlations leave (Proposition 1). The wrist is a bounded scalar, so its coordinate is selected with the same K labels, by $\arg\max_j |\mathrm{corr}(h_j, y)|$; this is undefined at $K = 2$.

The rule, step by step. (1) Take the 4096 training frames that serve as the label pool, without their labels, and embed them with the frozen encoder. (2) For every coordinate pair (j, k), estimate the 2×2 mean and covariance of (h_j, h_k) on these frames, whiten, and compute the radius $r = \|u\|$ of every frame. (3) Select the pair with the smallest $\mathrm{std}(r)/\mathrm{mean}(r)$. (4) Read the angle $\phi = \mathrm{atan2}(u_2, u_1)$. (5) For a draw, take the first K frames of a permutation of the label pool seeded by $1000 +$ seed; the same frames serve every K, rule and arm. For both signs, fit the offset as the circular mean of $\theta - \pm\phi$, and keep the sign with the smaller calibration residual. (6) Score the

median absolute angular error on the 2736 frames of the 48 held-out test episodes; report the median over 100 draws and the mean over five seeds. Steps (1)–(4) use no label, so the selected pair is a deterministic function of the encoder and the unlabelled frames; Appendix D.5.1 summarizes the use of labels in the readout.

The overlay in Figure 2 uses a separate readout evaluation with coordinate identities fixed on development data and 4096 training labels. Calibration-budget results and overlay errors are reported separately in Appendix D.5.

D Additional Experimental Results

We first examine basis selection (Appendix D.1) and nonlinear feature competition (Appendix D.2) on controlled OU data. Full benchmark and matched-ablation results follow in Appendix D.3. Appendices D.4 and D.5 connect frozen coordinate transformations to calibration, and Appendix D.6 collects sensitivity results. Finally, Appendix D.7 measures the timescales behind the benchmarks, and Appendix D.8 relates the timescale gap to recovery on real dynamics.

D.1 Basis Selection and Temporal Separation

These diagnostics examine the coordinate preference induced by the predictor. Training settings and the distinction between development and confirmation seeds are given in Appendix C.2.3.

D.1.1 Predictor classes and rotational ambiguity

Table 6: **Predictor class and linear identifiability.** Optimal predictor within each admissible class for $h = Qz$, with $S(Q) := Q\,\mathrm{diag}(\rho)\,Q^\top$, its invariance group, and what can be recovered. The anisotropic row is AP-JEPA.

Predictor	Optimal P for $h = Qz$	Invariance group	Recovers
Identity (LeJEPA)	I	$\mathrm{O}(n)$	subspace
Dense linear	$S(Q)$	$\mathrm{O}(n)$	subspace
Unconstrained nonlinear	arbitrary	larger than $\mathrm{O}(n)$	nothing
Anisotropic (ours)	$\mathrm{diag}\big(S(Q)_{11}, \ldots, S(Q)_{nn}\big)$	signed permutations	each coordinate

The predictor ladder of Table 5 follows Table 6. Every predictor class shares the same unconstrained target $S(Q)$; what differs is how much of it the class can represent. A dense predictor realizes $S(Q)$ exactly for every Q, so it attains the same loss on every rotation and cannot prefer a frame, even though its optimum is the richest. The diagonal predictor has to discard the off-diagonal part of $S(Q)$, which equation 4 prices at $\frac{1}{2}(\rho_1 - \rho_2)^2 \sin^2(2\theta)$ in two dimensions. That price is zero only at signed permutations when the correlations are distinct, and it vanishes when $\rho_1 = \rho_2$, so the mechanism is driven by the spread of the spectrum rather than by diagonality as such.

D.1.2 Timescale gap and coordinate access

Table 7: **Anisotropy-gap sweep under whitening.** Five seeds at $\rho_1 = 0.95$ (the sweep in Figure 4(a)). Aggregate $R^2(h \to z) \approx 0.97$ for both predictors throughout.

ρ	$\Delta\rho$	Identity MCC	Anisotropic MCC	Learned r
$(0.95, 0.50)$	0.45	0.79	**0.983**	$(0.94, 0.50)$
$(0.95, 0.55)$	0.40	0.80	**0.982**	$(0.94, 0.55)$
$(0.95, 0.65)$	0.30	0.83	**0.974**	$(0.93, 0.65)$
$(0.95, 0.75)$	0.20	0.83	**0.877**	$(0.90, 0.78)$
$(0.95, 0.85)$	0.10	0.84	0.839	$(0.91, 0.88)$
$(0.95, 0.95)$	0.00	0.82	0.818	$(0.95, 0.94)$

Table 7 tests the dependence on the factor gap with controlled OU data. Full readout remains high for both predictors, while the anisotropic advantage in MCC disappears as the gap closes. The coefficients approximate the latent correlations where factor alignment is strong. This agrees with

the within-factor-span penalty in equation 4. The MPI3D pair construction treats the two arm axes symmetrically (Appendix C.1); both have lag correlation .985 (Table 18), so this temporal criterion is not expected to separate them, consistent with Table 3.

Restricting the readout to one coordinate per factor hurts dense+SIGReg much more than AP+PairSIGReg, despite their similar full readout. On the five confirmation seeds of Figure 3(b), their full probe scores are 0.949 and 0.962, while their single-coordinate scores are 0.834 and 0.950. The alignment gap, defined as full minus single-coordinate R^2, is therefore 0.115 for dense+SIGReg and 0.012 for AP+PairSIGReg. AP+SIGReg has a gap of 0.008 and identity+SIGReg 0.177. The external baselines, trained on the same pairs with the same encoder and budget, lie in between: VICReg has $10.1 \pm 5.1^\circ$ axis deviation and a gap of 0.036, Barlow Twins $11.4 \pm 7.6^\circ$ and 0.054, and BYOL $17.9 \pm 15.2^\circ$ and 0.117.

D.2 Nonlinear Feature Competition and Regularization

The following studies examine whether improved predictability is accompanied by substitution of a nonlinear harmonic for a factor. They use the controlled training setup in Appendix C.2.3.

D.2.1 Harmonic content across the regularizers

Figure 3(a) sweeps ρ_1 at $\rho_2 = 0.70$ with five seeds per point, and the three arms share everything except the regularizer. Alongside recovery we track the He_2 content (Appendix C.3) to measure competition with the degree-two harmonic. PairSIGReg has lower harmonic content than SIGReg throughout this sweep, while whitening develops substantial harmonic content as persistence increases. In the safe regime it is near zero for every arm. Whitening stays below 0.006 up to $\rho_1 = 0.85$, then jumps to 0.43 at 0.92 and to 0.52–0.57 beyond. SIGReg stays below 0.136 across the whole sweep, and PairSIGReg stays below 0.043 and under SIGReg at all fourteen spectra. Past $\rho_1 \approx 0.944$, where the SIGReg certificate no longer applies, PairSIGReg leads at all five spectra, with mean MCC 0.906–0.957 against 0.816–0.909 for SIGReg and 0.381–0.421 for whitening. Over the same range SIGReg's He_2 rises from 0.059 to 0.136, and PairSIGReg beats it on $3/5$ seeds at $\rho_1 \in \{0.95, 0.955, 0.965\}$, $4/5$ at 0.975, and $5/5$ at 0.98.

D.2.2 Response of the implemented joint penalty

We checked that the implemented PairSIGReg responds to the direction the theory singles out. We moved a fixed representation along the obstruction direction with strength s and measured both terms, using a detection criterion of $|\Delta|/\mathrm{sem} \geq 3$ fixed in advance. The marginal term does not change at any s (ratio 0.97–1.00, $|\Delta|/\mathrm{sem} \leq 1.2$), whereas the joint term rises 31.9-fold at $s = 0.7$ ($\Delta/\mathrm{sem} = 26.2$). At training batch size $B = 1024$ the rise is essentially the same (2.27×10^{-4} against 2.07×10^{-4} at population scale), with a noise floor 47 times larger. No encoder was trained for this check.

D.3 Benchmark Results and Matched Ablations

D.3.1 Full benchmark comparisons

Table 8 adds three synthetic OU pixel settings, one inside and two outside the whitening regime, to the benchmarks of the body. An AP variant has the highest MCC in every setting, and the anisotropic predictor improves on the dense one under SIGReg in five of the six; the exception is MPI3D-real, whose arm axes have tied timescales. There AP+SIGReg loses a factor (worst-factor R^2 of 7.0%), and PairSIGReg restores it (89.3%), as the regularizer analysis of Section 4.3 anticipates. On the synthetic settings the two AP variants are close, and full readout stays within about one point of the dense control.

Table 9 lists the MPI3D versions separately. AP+PairSIGReg is first on both metrics for every version, so the averaged column of Table 3 is not driven by a single rendering.

Figure 5 shows the MPI3D-real result at the level of individual coordinates. Under AP+PairSIGReg the matched coordinate stands out from the rest of its row. Under dense+SIGReg the same factors are spread over many coordinates of similar magnitude, so no single coordinate reads out the factor well, even though a full linear probe can.

Table 8: **Four-arm comparison on six settings.** Five-seed mean ± sample SD (%) on the held-out split; AP rows are shaded and bold marks the best arm in each column of each block.

	Synthetic OU pixels, ρ			Control suite		Images
Method	$(0.70, 0.60)$	$(0.95, 0.60)$	$(0.97, 0.50)$	Reacher	Cartpole	MPI3D-real
MCC						
LeJEPA	$47.0_{\pm 4.4}$	$39.5_{\pm 5.4}$	$36.8_{\pm 3.1}$	$34.7_{\pm 3.0}$	$35.6_{\pm 3.1}$	$36.0_{\pm 7.1}$
dense+SIGReg	$90.2_{\pm 0.7}$	$71.3_{\pm 5.6}$	$60.6_{\pm 1.3}$	$56.2_{\pm 5.4}$	$56.3_{\pm 4.6}$	$33.1_{\pm 7.6}$
AP+SIGReg	$\mathbf{93.3}_{\pm 0.2}$	$\mathbf{85.6}_{\pm 2.8}$	$\mathbf{79.8}_{\pm 2.2}$	$88.8_{\pm 8.3}$	$66.8_{\pm 5.7}$	$19.0_{\pm 4.7}$
AP+PairSIGReg	$93.2_{\pm 0.1}$	$85.5_{\pm 3.8}$	$76.0_{\pm 3.9}$	$\mathbf{94.7}_{\pm 1.9}$	$\mathbf{69.9}_{\pm 3.4}$	$\mathbf{38.3}_{\pm 3.4}$
Full R^2						
LeJEPA	$67.7_{\pm 0.5}$	$77.2_{\pm 0.6}$	$74.2_{\pm 1.2}$	$58.3_{\pm 0.7}$	$77.2_{\pm 1.3}$	$80.0_{\pm 3.3}$
dense+SIGReg	$87.0_{\pm 0.5}$	$\mathbf{86.9}_{\pm 0.5}$	$81.2_{\pm 1.2}$	$\mathbf{95.9}_{\pm 0.1}$	$\mathbf{86.7}_{\pm 0.4}$	$71.2_{\pm 23.0}$
AP+SIGReg	$\mathbf{87.1}_{\pm 0.3}$	$86.3_{\pm 0.5}$	$\mathbf{81.3}_{\pm 0.8}$	$92.8_{\pm 4.1}$	$86.2_{\pm 0.7}$	$31.9_{\pm 13.0}$
AP+PairSIGReg	$87.0_{\pm 0.3}$	$86.3_{\pm 0.6}$	$79.9_{\pm 0.6}$	$95.7_{\pm 0.3}$	$85.4_{\pm 0.2}$	$\mathbf{89.8}_{\pm 0.8}$
Worst-factor R^2						
LeJEPA	$62.1_{\pm 0.7}$	$67.7_{\pm 0.8}$	$59.9_{\pm 2.1}$	$31.4_{\pm 1.2}$	$60.9_{\pm 2.0}$	$74.7_{\pm 5.1}$
dense+SIGReg	$\mathbf{84.8}_{\pm 0.5}$	$\mathbf{82.6}_{\pm 0.9}$	$\mathbf{71.5}_{\pm 2.2}$	$\mathbf{93.4}_{\pm 0.1}$	$\mathbf{77.2}_{\pm 0.4}$	$53.8_{\pm 44.7}$
AP+SIGReg	$84.5_{\pm 0.5}$	$82.0_{\pm 0.9}$	$\mathbf{71.5}_{\pm 1.6}$	$87.2_{\pm 8.1}$	$76.0_{\pm 1.0}$	$7.0_{\pm 4.3}$
AP+PairSIGReg	$84.7_{\pm 0.4}$	$81.9_{\pm 0.7}$	$69.0_{\pm 1.3}$	$92.8_{\pm 0.5}$	$74.7_{\pm 0.6}$	$\mathbf{89.3}_{\pm 0.9}$

Table 9: **MPI3D versions.** The three versions averaged in Table 3. Five-seed mean ± sample SD (%); bold and underline mark first and second.

	MPI3D-real		MPI3D-toy		MPI3D-realistic	
Method	MCC	R^2	MCC	R^2	MCC	R^2
BYOL	$33.7_{\pm 3.7}$	$59.2_{\pm 7.2}$	$37.9_{\pm 3.0}$	$73.3_{\pm 3.5}$	$33.0_{\pm 2.7}$	$62.6_{\pm 8.6}$
Barlow Twins	$24.9_{\pm 5.0}$	$37.3_{\pm 12.0}$	$36.7_{\pm 4.0}$	$66.9_{\pm 2.8}$	$26.5_{\pm 4.6}$	$49.0_{\pm 5.0}$
VICReg	$4.6_{\pm 0.9}$	$0.7_{\pm 1.3}$	$8.5_{\pm 0.5}$	$2.0_{\pm 1.0}$	$6.1_{\pm 0.5}$	$1.7_{\pm 1.0}$
SimDINO	$24.0_{\pm 4.0}$	$23.5_{\pm 4.5}$	$31.7_{\pm 5.3}$	$47.9_{\pm 2.0}$	$26.6_{\pm 6.8}$	$26.8_{\pm 10.1}$
LeJEPA	$\underline{36.0}_{\pm 7.1}$	$\underline{80.0}_{\pm 3.3}$	$\underline{38.4}_{\pm 3.6}$	$\underline{84.5}_{\pm 5.0}$	$\underline{34.7}_{\pm 2.5}$	$\underline{83.4}_{\pm 3.7}$
dense+SIGReg	$33.1_{\pm 7.6}$	$71.2_{\pm 23.0}$	$36.3_{\pm 3.7}$	$82.9_{\pm 7.8}$	$24.7_{\pm 1.4}$	$43.7_{\pm 2.8}$
AP+SIGReg	$19.0_{\pm 4.7}$	$31.9_{\pm 13.0}$	$29.7_{\pm 6.8}$	$58.0_{\pm 16.1}$	$18.8_{\pm 7.6}$	$32.5_{\pm 17.9}$
AP+PairSIGReg	$\mathbf{38.3}_{\pm 3.4}$	$\mathbf{89.8}_{\pm 0.8}$	$\mathbf{38.8}_{\pm 3.3}$	$\mathbf{89.9}_{\pm 1.3}$	$\mathbf{44.0}_{\pm 5.6}$	$\mathbf{90.1}_{\pm 0.8}$

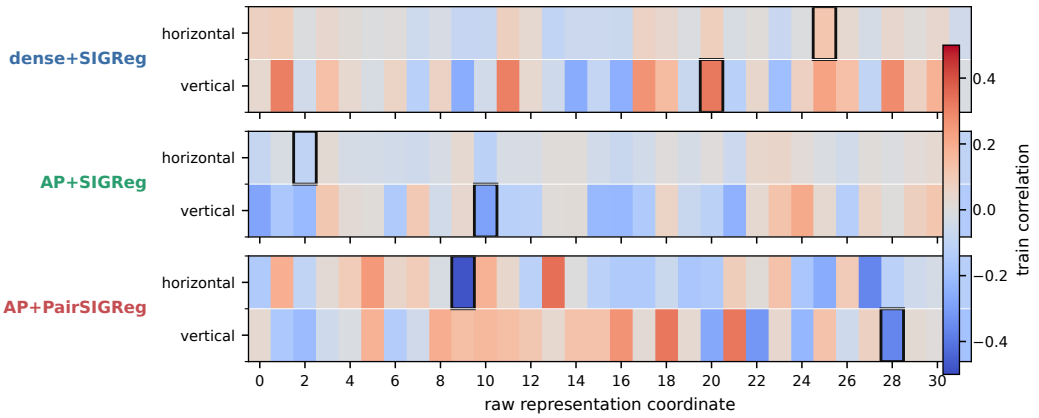


Figure 5: **MPI3D-real coordinate correlations.** Training-set correlations between the 32 representation coordinates and the two arm factors for the seed-25 runs of Table 3. Boxes mark the one-to-one assignment by absolute correlation; no evaluation labels are used.

D.3.2 MATCHED PREDICTOR AND REGULARIZER COMPARISONS

Table 10 extends the predictor ablation of Table 5 to MPI3D-real, the setting with tied timescales, and reports worst-factor R^2, the lowest per-factor full R^2, which exposes a lost factor. All rows use the same data splits, encoder and training protocol.

Table 10: **Predictor and regularizer ablation on Reacher-hard and MPI3D-real.** Mean±SD (%); small numbers give the change from identity + SIGReg (the LeJEPA baseline). Bold and underline mark first and second; ⋆: full method.

		Reacher-hard		MPI3D-real	
Predictor	Regularizer	MCC	worst R^2	MCC	worst R^2
identity	SIGReg	34.7±3.0	31.4±1.2	36.0±7.1	74.7±5.1
shared scalar	SIGReg	49.3±4.8 +14.6	90.1±1.7 +58.7	15.3±8.5 −20.7	8.0±8.8 −66.7
dense	SIGReg	56.2±5.4 +21.5	**93.4**±0.1 +62.0	33.1±7.6 −2.9	53.8±44.7 −20.9
anisotropic	SIGReg	<u>88.8</u>±8.3 +54.1	87.2±8.1 +55.8	19.0±4.7 −17.0	7.0±4.3 −67.7
identity	PairSIGReg	31.2±3.4 −3.5	26.9±0.8 −4.5	33.8±2.6 −2.2	66.3±9.7 −8.4
shared scalar	PairSIGReg	50.3±4.3 +15.6	89.0±0.8 +57.6	**41.3**±1.5 +5.3	88.0±1.7 +13.3
dense	PairSIGReg	55.5±3.0 +20.8	<u>93.1</u>±0.5 +61.7	38.2±3.2 +2.2	**90.5**±1.2 +15.8
anisotropic ⋆	PairSIGReg	**94.7**±1.9 +60.0	92.8±0.5 +61.4	<u>38.3</u>±3.4 +2.3	<u>89.3</u>±0.9 +14.6

On Reacher-hard, whose factors have distinct timescales, changing the predictor under fixed PairSIGReg raises MCC from 55.5% to 94.7% at essentially the same worst-factor readout. On MPI3D-real, whose arm axes share one timescale, adding PairSIGReg restores the worst factor under every fitted predictor. The two settings thus isolate the two components of Theorem 1: the predictor selects the basis, and the regularizer secures the subspace.

The seed-level scores explain the large uncertainty of dense+SIGReg on MPI3D-real. Three runs have high worst-factor readout and two have very low readout, whereas all five dense+PairSIGReg runs have high readout, so PairSIGReg makes the factor readout consistent across seeds. The individual worst-factor scores are:

Representation	Run 1	Run 2	Run 3	Run 4	Run 5
dense+SIGReg	0.071	0.856	0.885	0.027	0.850
dense+PairSIGReg	0.918	0.898	0.904	0.889	0.916

D.4 POST-HOC ALIGNMENT OF A FIXED REPRESENTATION

We apply the transformation described in Appendix C.4 to fixed representations. Table 11 compares the resulting coordinate readout with the full-linear ceiling in Proposition 7(iii).

Table 11: **Post-hoc coordinate recovery and full linear readout.** Reacher-hard, five seeds per arm (mean ± SD, %); the arms differ only in the training predictor and regularizer. Full-dimensional R^2 is reported once because the map leaves it unchanged to within 0.1 points.

		Single-coordinate R^2		MCC	
Training predictor	Full R^2	native	post-hoc	native	post-hoc
identity	$48.7_{\pm 0.8}$	$13.5_{\pm 2.7}$	$32.1_{\pm 0.9}$	$35.4_{\pm 3.4}$	$47.6_{\pm 1.7}$
dense	$95.2_{\pm 0.1}$	$33.5_{\pm 7.2}$	$90.5_{\pm 2.5}$	$55.8_{\pm 6.2}$	$95.1_{\pm 1.3}$
AP+SIGReg	$90.5_{\pm 6.1}$	$80.6_{\pm 11.3}$	$87.3_{\pm 6.5}$	$88.6_{\pm 8.5}$	$93.2_{\pm 3.8}$
AP+PairSIGReg	$94.7_{\pm 0.3}$	$89.9_{\pm 3.6}$	$91.7_{\pm 1.5}$	$94.7_{\pm 1.9}$	$95.7_{\pm 0.8}$

Post-hoc alignment changes coordinates within the fixed feature span. Under an invertible linear transformation, the population full-linear readout is unchanged and bounds the single-coordinate readout (Proposition 7(iii)). In Table 11, dense-trained features benefit substantially from alignment, whereas the identity-trained features remain limited by lower full readout. A high full readout is necessary for high coordinate recovery, but the temporal criterion also requires suitable lagged structure to find those coordinates.

Training with AP and post-hoc alignment are complementary, as Theorem 1 suggests. Post-hoc alignment fixes the basis of a given span and can be applied to an existing encoder; training with AP also shapes the span, which matters when a non-factor direction competes with a factor (Table 19). The released-backbone results in Table 4 evaluate the latter use. Table 11 scores the same five encoders per arm as Table 2 (identical checkpoints) on the $1{,}824$-frame development split.

Table 12: **Whitened lag spectrum behind Table 4.** Leading eigenvalues of C_s (Proposition 7) on the $8{,}784$ walk pairs of Appendix C.4. $\lambda_{(1)}$ and $\lambda_{(2)}$ are the coordinates matched to the horizontal and vertical axes; the next eigen-directions have $|\mathrm{corr}| \leq .12$ with either axis. The last two columns report a 20-fold scene bootstrap of the transformation: the absolute correlation of rank 1 with the horizontal axis and of rank 2 with the vertical axis (mean $\pm$ SD, minimum).

	$\lambda_{(1)}$	$\lambda_{(2)}$	$\lambda_{(3)}$	gaps $\lambda_{(1)}-\lambda_{(2)}$, $\lambda_{(2)}-\lambda_{(3)}$	boot. horizontal	boot. vertical
I-JEPA	.855	.765	.723	.090, .042	$.889 \pm .004$ (.881)	$.780 \pm .063$ (.554)
V-JEPA	.903	.828	.821	.075, .007	$.907 \pm .003$ (.900)	$.755 \pm .077$ (.529)

Table 12 reports the eigenvalues behind Table 4. Recovery follows the gaps, as Proposition 8 predicts. The horizontal axis is separated from the rest of the spectrum by .075–.09; it lands on rank 1 in all 20 bootstrap refits, and its correlation varies by less than .02. The vertical axis lies close to a cluster of scene-related eigenvalues, with a gap of .042 for I-JEPA and .007 for V-JEPA; it stays on rank 2 in 19 of the 20 refits, with a wider spread of correlations (.53–.86). Resampling scenes with replacement leaves about 63% of them distinct, so bootstrap values are lower than the full-data fit on the same frames (.918 horizontal for both backbones; .902 and .871 vertical). Using the Table 3 training pairs instead gives smaller measured eigengaps and poorer recovery (Table 14).

Test protocol and whitening-only control. The transformation of Table 4 was fitted and its coordinates matched before the 102 test scenes of MPI3D-real were encoded. We then encoded the 4096 test windows once, with the same representation, and applied the frozen maps without refitting; coordinates stay matched on the 458 factor-fit scenes. Table 13 adds a whitening-only control, the same map without the temporal rotation: it raises MCC only to 38.3% and 43.3%, so the recovery comes from the rotation step of post-hoc AP.

Table 13: **Whitening-only control for Table 4** (MPI3D test scenes, %). Maps are fitted on the unlabelled walk pairs of 366 training scenes and coordinates are matched on 458 factor-fit scenes. Whitening only is post-hoc AP without the temporal rotation.

Backbone	map	MCC	single R^2	full R^2
I-JEPA	native	28.3	8.4	91.5
	whitening only	38.3	14.7	91.3
	post-hoc AP	**90.9**	**82.6**	91.3
V-JEPA	native	32.3	10.7	89.2
	whitening only	43.3	18.9	89.0
	post-hoc AP	**89.3**	**79.8**	89.0

Pair construction and the lag spectrum. The two pair constructions differ in three ways. The Table 3 training pairs take one step of the same law on both axes (uniform in $\{-3, \ldots, 3\}$, reflected, joint zero step redrawn), are one frame apart, and keep the scene fixed. The Table 4 pairs are the ends of a 16-step walk with a different law per axis, and redraw the scene with probability .5. To test the effect directly we built the Table 3 construction in the same manifest format, on the same scenes and seed, and refitted the map on the same frozen backbones (first row of Table 14). Every slow direction then has a lag eigenvalue above .97: the fixed scene contributes directions with correlation one, and one step of the grid barely moves the arm. The two arm directions sit at ranks 16–17 (I-JEPA) and 5 and 14 (V-JEPA), with gaps of at most .004 to their neighbours, and test MCC falls to .633 and .410. This is the tied case of Table 18; Table 4 uses a pair construction that has the gap.

Table 14: **Pair law, eigengap and post-hoc recovery on frozen backbones.** Each row refits the map of Table 4 on a new unlabelled manifest (366 training scenes, 24 pairs each, same seed); only the vertical walk law changes (move probability q, step range in cells), while the horizontal law stays at q=.8, ±1–2; the scene is redrawn with probability .5 except in the first row. ρ_h, ρ_v: lag correlation of the grid positions (reporting only). g: distance of the lag eigenvalue of the direction matched to each axis from the rest of the whitened spectrum (no labels needed once the direction is identified). $|\mathrm{corr}|$: matched test correlation of that axis on the 102 sealed test scenes.

	label lag corr.		I-JEPA: eigengap g / test $\|\mathrm{corr}\|$				V-JEPA: eigengap g / test $\|\mathrm{corr}\|$			
Vertical walk law	ρ_h	ρ_v	g_h	h	g_v	v	g_h	h	g_v	v
Table 3 pairs	0.984	0.985	0.003	0.50	0.004	0.77	0.001	0.44	0.001	0.38
q=.5, ±1	0.900	0.972	0.008	0.78	0.062	0.96	0.010	0.80	0.023	0.86
q=.8, ±1–2 (tied)	0.896	0.900	0.036	0.89	0.036	0.95	0.005	0.82	0.016	0.80
q=.9, ±1–2	0.899	0.886	0.025	0.87	0.025	0.93	0.012	0.75	0.012	0.70
q=.7, ±1–3	0.897	0.846	0.009	0.84	0.009	0.86	0.014	0.84	0.014	0.78
q=.9, ±1–3	0.899	0.803	0.040	0.89	0.040	0.91	0.032	0.91	0.018	0.86
q=.85, ±1–4 (Table 4)	0.896	0.716	0.090	0.91	0.042	0.91	0.075	0.92	0.007	0.87
q=.9, ±1–5	0.899	0.601	0.127	0.91	0.016	0.72	0.082	0.92	0.009	0.87
q=.9, ±1–6	0.899	0.511	0.124	0.91	0.002	0.30	0.082	0.92	0.004	0.74
q=.9, ±1–8	0.899	0.311	0.125	0.92	0.000	0.10	0.088	0.92	0.011	0.67

D.5 Few-Label Calibration and Overlay Results

The calibration and overlay protocols are specified in Appendix C.5. Table 15 compares native and post-hoc coordinates under the same reported calibration budgets.

Table 15: **Reacher-hard shoulder readout.** Test error in degrees (lower is better) across calibration budgets K: median over 100 calibration draws per seed, then mean over five training seeds. The lower block applies the label-free post-hoc rotation of Appendix D.4 before reading out.

Representation	K=2	K=8	K=64	K=1024
LeJEPA	87.9	85.8	82.1	79.6
dense+SIGReg	85.9	83.4	78.7	76.4
AP+SIGReg	12.6	11.3	10.8	10.7
AP+PairSIGReg	**8.9**	**8.4**	**8.2**	**8.2**
with post-hoc rotation				
LeJEPA	87.3	87.5	84.8	83.8
dense+SIGReg	10.9	9.8	9.6	9.6
AP+SIGReg	18.4	16.6	16.0	15.8
AP+PairSIGReg	9.3	8.5	8.3	8.2

Table 15 shows that additional calibration labels have limited benefit when the native coordinates are poorly suited to the prescribed readout. At $K = 2$, AP+PairSIGReg reaches $8.9°$ shoulder error, compared with $85.9°$ for dense+SIGReg. At $K = 1024$, the dense control still has $76.4°$ error; Appendix D.5.1 compares with a full-dimensional decoder. For the wrist, from $K = 8$ upward AP+SIGReg and AP+PairSIGReg reach $R^2 = 0.981$ and 0.982, against 0.259 for dense+SIGReg. We omit the wrist at $K = 2$, because with two samples every nonconstant coordinate has absolute sample correlation one and this criterion cannot select a coordinate.

The lower block holds each encoder fixed and applies post-hoc alignment before calibration. Dense+SIGReg then reaches $10.9°$ at $K = 2$, closing most of the native-coordinate gap. AP+PairSIGReg is already aligned and changes little after this transformation.

The overlay evaluation in Figure 2 has mean shoulder error $16.6°$ for AP+PairSIGReg and $59.1°$ for dense+SIGReg; both AP variants have wrist error of about $11°$, compared with $57.3°$ for the dense control. These measurements describe the overlay evaluation, rather than the $K = 2$ calibration setting.

D.5.1 SUPERVISION LEDGER OF THE TWO-LABEL READOUT

The K in Table 15 counts only the labels spent after the encoder is frozen. The readout behind the $K = 2$ entry uses labels and prior knowledge as follows, identically for every arm.

1. *Encoder training*: 90,100 unlabelled frame pairs; no labels.
2. *Choice of the shoulder pair*: the rule of Appendix C.5, on the 4096 unlabelled training frames; no labels. It uses one piece of prior knowledge, that the shoulder is a circular joint read from two coordinates. The prior is part of the readout definition and is given to every arm.
3. *Calibration*: K labelled training frames fit a sign and an offset.
4. *Scoring*: test labels, only to compute the error.

The overlay of Figure 2 is a different protocol (4096 labels, identities fixed on development data) and is not part of this ledger.

Is the label-free pair the right one? For each run we also compute the *oracle* pair: the pair with the smallest test error when sign and offset are fitted on all 4096 training labels. This bounds what any pair rule could achieve with more supervision. Table 16 applies the same code, frames and draws to all eight Reacher-hard arms, including VICReg, BYOL, Barlow Twins and dense+PairSIGReg. For both AP arms the label-free pair is the oracle pair in $5/5$ seeds (Table 17), with a clear CV margin to the runner-up in 9 of the 10 runs ($\approx$ .19 against $\approx$ .33). For every control it is not, since no pair stands out: the selected CV is within .017 of the runner-up, and the selected pair ranks 32–495 of 496. The $K = 2$ gap is not a property of the selection rule, however. Even with all 4096 labels, the best native pair of any control has an error of 35–79° (arm means). The shoulder is simply not carried by any coordinate pair of those representations. A full-dimensional ridge readout recovers it given enough labels (Figure 3(c)): dense+SIGReg needs about 120 labels (log-log interpolation between $K = 64$ and 256) to match the two-label AP error and reaches 7.4° at $K = 1024$ (dense+PairSIGReg: 7.6°), while LeJEPA, VICReg, BYOL and Barlow Twins remain at 18–48° even then; with two labels every arm is at 81–90°. Only the post-hoc rotation, which also uses no labels, brings the dense controls to the AP regime (10.7–10.9°).

Table 16: **Two-label readout under one budget for all Reacher-hard arms.** Sealed test trajectories; five confirmation seeds; unmodified readout code. Oracle: the selected pair equals the pair with the smallest test error given all 4096 training labels. Rank: position of the selected pair in that ordering.

	label-free pair, native			best pair	ridge, all 32	label-free pair,
Representation	= oracle	rank / 496	K=2	K=4096	K=2 / K=1024	post-hoc, K=2
LeJEPA	0/5	138–484	$87.9_{\pm 1.0}$	$61.5_{\pm 2.2}$	88.0 / 38.5	$87.3_{\pm 2.3}$
VICReg	0/5	350–495	$87.0_{\pm 2.2}$	$47.1_{\pm 4.2}$	86.9 / 23.1	$88.4_{\pm 1.5}$
BYOL	0/5	143–466	$75.5_{\pm 9.6}$	$34.9_{\pm 2.4}$	81.4 / 18.4	$77.0_{\pm 8.6}$
Barlow Twins	0/5	124–477	$89.4_{\pm 0.7}$	$78.7_{\pm 3.3}$	90.0 / 47.8	$85.9_{\pm 6.4}$
dense+SIGReg	0/5	227–472	$85.9_{\pm 1.8}$	$40.6_{\pm 3.4}$	84.7 / 7.4	$10.9_{\pm 0.4}$
dense+PairSIGReg	0/5	32–475	$82.6_{\pm 7.2}$	$44.4_{\pm 3.5}$	84.3 / 7.6	$10.7_{\pm 0.5}$
AP+SIGReg	5/5	1	$12.6_{\pm 4.9}$	$10.7_{\pm 3.9}$	86.0 / 10.4	$18.4_{\pm 15.6}$
AP+PairSIGReg	5/5	1	$8.9_{\pm 0.4}$	$8.2_{\pm 0.4}$	85.7 / 7.7	$9.3_{\pm 0.5}$

D.6 REGULARIZATION WEIGHTS

D.6.1 MARGINAL REGULARIZATION WEIGHT

The λ sweep in Figure 4(b) covers $\lambda \in \{0.003, 0.01, 0.03, 0.1\}$ on two synthetic settings: the safe $\rho = (0.70, 0.60)$ and the dangerous $\rho = (0.97, 0.50)$. It includes all four arms with three development seeds each, $d_{\text{latent}} = 32$, and 100 epochs, for 96 runs in total. Each job differs from its $\lambda = 0.01$ counterpart only in λ. The panel shows the dangerous setting, where the regularizer matters most; the safe setting is flatter.

So that no arm is favoured through its hyperparameter, each arm also selects its own λ under one rule: average the worst-factor probe R^2 over seeds, take the best λ, and break ties by MCC. On the

Table 17: **Per-seed audit of the label-free pair rule.** Coordinate indices refer to the 32 native coordinates of each run. †: selected pair differs from the oracle pair. The other five controls behave like dense+SIGReg (Table 16).

		radius CV (unlabelled)				test error (°)	
Arm	seed	selected	runner-up	selected pair	oracle pair	K=2	oracle, K=4096
dense+SIGReg	41	0.415	0.426	$(7, 26)$	$(11, 15)^\dagger$	86.2	40.6
	42	0.438	0.448	$(1, 20)$	$(8, 30)^\dagger$	83.2	37.4
	43	0.449	0.449	$(1, 9)$	$(6, 23)^\dagger$	87.6	44.0
	44	0.437	0.442	$(10, 23)$	$(8, 19)^\dagger$	87.5	37.1
	45	0.433	0.440	$(8, 15)$	$(7, 20)^\dagger$	85.1	44.1
AP+SIGReg	41	0.224	0.329	$(12, 31)$	$(12, 31)$	10.9	9.6
	42	0.193	0.337	$(4, 31)$	$(4, 31)$	9.1	8.1
	43	0.186	0.304	$(14, 16)$	$(14, 16)$	8.6	7.4
	44	0.341	0.345	$(14, 19)$	$(14, 19)$	20.5	17.0
	45	0.273	0.350	$(11, 28)$	$(11, 28)$	13.8	11.5
AP+PairSIGReg	41	0.183	0.313	$(3, 12)$	$(3, 12)$	8.8	8.0
	42	0.188	0.333	$(24, 31)$	$(24, 31)$	8.5	7.9
	43	0.201	0.321	$(3, 30)$	$(3, 30)$	9.4	8.8
	44	0.198	0.327	$(10, 15)$	$(10, 15)$	8.6	7.7
	45	0.197	0.336	$(2, 26)$	$(2, 26)$	9.4	8.4

dangerous setting the rule returns $\lambda = 0.1$ for identity and 0.03 for the other three arms, and on the safe setting 0.1 and 0.01. The ordering of the arms is the same under per-arm selection as under a shared λ.

D.6.2 Pair-penalty weight

The pair weight η scales only the additive joint term. Figure 4(c) sweeps $\eta \in \{0.03, 0.1, 0.3, 1, 3\}$ on the $(0.95, 0.60)$ setting at the shared $\lambda = 0.03$, with five development seeds per point. MCC stays within 82.1–85.0% across this 100-fold range, against 83.9% for AP+SIGReg, 72.0% for dense+SIGReg and 32.2% for LeJEPA, and full R^2 stays within 83–84% up to $\eta = 1$, falling to 79.7% only at $\eta = 3$. The choice of η is therefore not critical. The confirmation runs use $\eta = 0.1$ on this setting; on its five confirmation seeds the two AP variants are tied (85.6 and 85.5 MCC in Table 8), as in the sweep.

D.7 Timescales and the Approximate-Identifiability Bound

We measure the two quantities the theory depends on, the timescale spectrum of the factors and the readout deficit δ of Proposition 8, on the pairs used for training. No encoder is retrained, and labels are used only to score.

Table 18: **Lag-one correlations of the scored factors** on the training pairs, at the stride used for training. A $(\cos, \sin)$ pair of one angle forms a tied block by construction. $\Delta\rho$ is the smallest gap between distinct blocks, and it prices the basis through equation 4.

Benchmark (pairs)	Lag-one correlation per factor	$\Delta\rho$
Reacher-hard (stride 56)	wrist .746; shoulder $(\cos, \sin)$ .263/.251	.49
Cartpole (stride 16)	slider .967; hinge $(\cos, \sin)$ .661/.658	.31
Finger (stride 8)	$(\cos, \sin)$: prox. .591/.887; distal .815/.941; hinge .883/.872	.004
MPI3D, all versions (Table 3)	horizontal .985, vertical .985; scene 1	$< 10^{-3}$
MPI3D walk (Table 4 rotation)	horizontal .896, vertical .716; scene $\approx .50^a$	.18
Falcor3D	camera $x/y/z$ .875/.874/.874	$< 10^{-3}$
Isaac3D	robot x/y .935/.819; light dir. .879; three others .699–.701	$.002^b$

[a] Fraction of pairs that keep the scene. [b] Between camera height, object scale and light intensity; the other gaps are at least .056.

Timescales. Table 18 separates the benchmarks into two groups. On Reacher and Cartpole the joints move on clearly different timescales, and the predictor edit is where the MCC gain comes from (Tables 5 and 8). On our MPI3D and Falcor3D pairs the scored factors share one correlation, so equation 4 gives the predictor no basis preference. The MPI3D result is consistent with this: the gain over dense+SIGReg comes from PairSIGReg, and the predictor edit under PairSIGReg is neutral. On Isaac3D the predictor can separate at most the three factors whose correlations are distinct.

Table 19: **Bound of Proposition 8 on Reacher-hard.** Five seeds per arm (mean, SD in the text). Whitened 32-dimensional representations on 20,000 training pairs; the factor block uses the measured factor lag matrix. "Bound" is $2\|K - K_0\|/g$, capped at 1 (vacuous); "meas." is the measured sine of the largest principal angle between the K-eigen-directions and the factor directions. The shoulder is a tied $(\cos, \sin)$ block.

			Wrist			Shoulder block		
Training arm	δ	$\|K - K_0\|$	gap g	bound	meas.	gap g	bound	meas.
identity+SIGReg	1.44	.436	.020	n/a*	.100	.171	n/a*	.998
dense+SIGReg	.146	.050	.399	.25	.049	.094	1	.222
AP+SIGReg	.266	.056	.473	.24	.037	.167	.70	.164
AP+PairSIGReg	.155	.047	.493	.19	.029	.191	.49	.133

*$\delta > 1$, outside the scope of the proposition; the identity arm has lost the shoulder (block $R^2 = .003$).

Approximation bound. For each run we compute δ, $\|K - K_0\|$, the bound of Proposition 8(ii) and the actual angle between the eigen-directions of K and the factor directions (Table 19). Summed over the three factor columns, δ averages .15–.27 for the dense and AP arms, while the identity arm loses the shoulder ($\delta = 1.44$). The measured $\|K - K_0\|$ ($.047 \pm .005$ for AP+PairSIGReg) is well below the worst case $\varepsilon(\delta)$, and the bound is informative: for AP+PairSIGReg it guarantees the wrist to $\sin\theta \leq .19$ (actual .029) and the shoulder block to .49 (actual .133). The bound holds in all 15 dense and AP runs, supporting Proposition 8 empirically. It also shows how the regularizer and the predictor interact. The dense arm retains a non-factor direction whose lag eigenvalue (.361) lies .094 above the shoulder block, so its shoulder is not certified. In the AP arms every non-factor eigenvalue is below .06, and AP+PairSIGReg gives the tightest bounds of all arms.

D.8 Timescale Gap and Coordinate Recovery on Real Dynamics

Proposition 8 makes a specific prediction. What matters for a factor is its gap g_i to the rest of the lag spectrum, including the non-factor directions, whose lag correlation is near zero once a factor decorrelates. The gap between two factors is only one part of g_i. We test this prediction in three settings with real, non-Gaussian dynamics, choosing them so that small gaps occur (Figure 6). All scores are on held-out data.

Reacher-hard stride sweep. The stored training set is a 113-frame render grid per episode (hop 8). We rebuilt it exactly and formed pairs at strides $\Delta t \in \{8, \ldots, 480\}$ from the same trajectories. Each stride has the same number of pairs (90,100 where available, otherwise all of them), the same number of gradient steps, and the same development and test frames. Only the factor lag correlations change. At $\Delta t = 56$ the pairs are exactly those of Table 2. We train AP+PairSIGReg, dense+PairSIGReg (the predictor-only control) and dense+SIGReg with the settings of the confirmation runs (Table 20). While both factors are predictable ($\Delta t \leq 56$), the predictor gain is positive and larger at larger gaps: $+16.8$ points at $g = .04$ and $+16.6$ at .13, $+21.0$ at .24, and $+24.9$ and $+36.7$ at .40 and .26 ($\Delta t = 40, 56$). Two-label shoulder readout reaches the regime of Table 15 only at the two longest strides (9.4–9.7° at $\Delta t \in \{40, 56\}$, against 39–59° for $\Delta t \leq 24$). Even at the smallest tested gap (.04), the predictor still adds 16.8 points.

The long strides confirm the second half of the prediction. For $\Delta t \geq 160$ the shoulder decorrelates ($\rho_s \leq .005$) and joins the near-zero lag spectrum of the non-factor directions, so its gap g vanishes even though $\rho_w - \rho_s = .39$ at $\Delta t = 160$. The shoulder is then unpredictable from one frame to the next, no method recovers it (MCC below 40% for every arm), and the predictor gain is negative, as

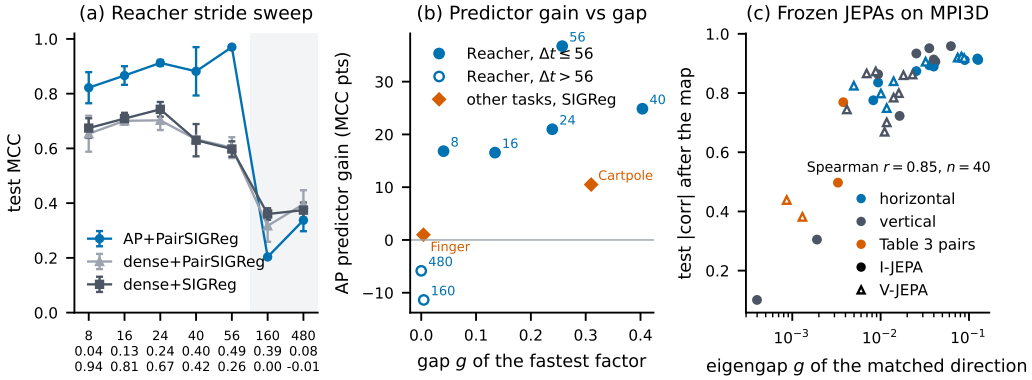


Figure 6: **Timescale gap and coordinate recovery. (a)** Reacher-hard trained from scratch at seven strides (three seeds, sealed test trajectories). Tick labels give the stride, $\rho_w - \rho_s$ and ρ_s of the training pairs. Grey band: the shoulder is unpredictable ($\rho_s \approx 0$). **(b)** Predictor-only MCC gain against the gap of the fastest factor, $g = \min(\rho_w - \rho_s, \rho_s)$. Reacher: AP+PairSIGReg minus dense+PairSIGReg. Other control tasks: AP+SIGReg minus dense+SIGReg, five confirmation seeds. **(c)** Frozen I-JEPA/V-JEPA with the map of Table 4, refitted on ten unlabelled MPI3D pair laws (Table 14). The plot shows each axis's sealed-test correlation against the eigengap of its matched direction.

expected when $g = 0$. Recovery thus requires every factor to stand apart from the whole spectrum, not only from the other factors, exactly as Proposition 8 states.

Table 20: **Reacher-hard stride sweep.** Three seeds (41–43), sealed test trajectories. ρ_w, ρ_s: lag correlations of the wrist and the shoulder $(\cos, \sin)$ block on the training pairs. $g = \min(\rho_w - \rho_s, \rho_s)$. AP gain: AP+PairSIGReg minus dense+PairSIGReg. Full R^2: mean over the three factor columns. K=2: label-free two-label shoulder readout, 20 draws. P and S denote PairSIGReg and SIGReg.

	training pairs			test MCC (%)			AP gain	full R^2 (%)			K=2 (°)
Δt	ρ_w	ρ_s	g	AP+P	dense+P	dense+S	(pts)	AP+P	dense+P	dense+S	AP+P
8	0.98	0.94	0.04	$82.2_{\pm 5.7}$	$65.4_{\pm 6.6}$	$67.4_{\pm 3.6}$	+16.8	99.8	99.8	99.8	59.1
16	0.95	0.81	0.13	$86.6_{\pm 3.4}$	$70.1_{\pm 1.0}$	$70.9_{\pm 2.1}$	+16.6	94.5	97.3	96.9	39.9
24	0.91	0.67	0.24	$91.2_{\pm 1.1}$	$70.3_{\pm 3.6}$	$74.3_{\pm 2.7}$	+21.0	91.9	96.2	95.8	38.9
40	0.82	0.42	0.40	$88.2_{\pm 8.8}$	$63.3_{\pm 0.9}$	$63.0_{\pm 5.9}$	+24.9	96.1	95.3	96.0	9.7
56	0.75	0.26	0.26	$97.0_{\pm 0.3}$	$60.3_{\pm 3.8}$	$59.8_{\pm 2.8}$	+36.7	94.9	94.9	95.0	9.4
160	0.39	0.00	0.00	$20.3_{\pm 0.2}$	$31.7_{\pm 5.8}$	$36.0_{\pm 2.1}$	−11.4	25.9	44.6	53.8	88.5
480	0.06	−0.01	0.00	$33.8_{\pm 4.0}$	$39.6_{\pm 5.1}$	$37.4_{\pm 2.8}$	−5.8	40.4	47.4	46.5	63.6

Other control tasks. Finger-spin is the natural small-gap case. At its stride the lag correlations of proximal sin (.887) and hinge cos (.883) differ by .004. The whitened factor-lag spectrum (.941, .892, .879, .870, .814, .588) has three eigenvalues within .022. The predictor gain there is +1.0 point under SIGReg, against +10.5 on Cartpole ($g = .31$) and +32.6 on Reacher. On Finger every arm reaches a full R^2 of about .40, so δ lies outside the range of Proposition 8.

Frozen JEPAs on MPI3D. The post-hoc map needs no training, so we can vary the pair law freely (Table 14). Across ten laws, two axes and two backbones, the eigengap of the matched direction predicts the sealed-test recovery of that axis (Spearman $r = .85$, $n = 40$), and low recovery occurs only at small eigengaps: .38–.77 in the Table 3 construction ($g \leq .004$), and .30 and .10 for I-JEPA when the vertical axis falls into the cluster of scene directions ($\rho \approx .5$), although $\rho_h - \rho_v$ is then largest. Conversely, equal position correlations ($\rho_h = .896$, $\rho_v = .900$) still give I-JEPA an eigengap of .036 and recovery of .89/.95, because the two axes change the image differently. The quantity that decides recovery is therefore the gap in the representation's own lag spectrum, which is measurable without labels.