DSREG: PROVABLY RECOVERING INDIVIDUAL WORLD LATENTS WITHOUT RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Methods that recover individual latent variables of the world, from nonlinear ICA to dictionary learning and causal representation learning, anchor the latents to observations through reconstruction, auxiliary supervision, or distributional asymmetries such as non-Gaussianity. Methods without these anchors, including joint-embedding predictive architectures (JEPAs), identify the latent state only up to a linear transformation, so individual latents remain mixed. We close this gap: individual world latents can be provably recovered with no reconstruction, no decoder, and no labels. The key condition is *Structural Diversity*: different latents leave distinct dependency footprints on observations, just as *no two snowflakes are alike*. Building on the linear identifiability that LeJEPA provides, we prove that under Structural Diversity and a standard faithfulness condition, ***DSReg*** (***D***ependency-***S***parsity ***Reg***ularization) recovers individual world latents up to signed permutation. It applies post hoc to any linearly identified representation, reusing trained checkpoints at no loss over joint training, and establishes the first fully identifiable JEPA that recovers every world latent. Moreover, as a condition on dependency footprints, Structural Diversity is strictly weaker than all structural conditions of prior identifiable latent variable models. Across synthetic regimes, world model probes, learned visual encoders, and external renderers, DSReg preserves dense prediction while improving individual-latent recovery and downstream use with scales.

1 INTRODUCTION

World models are useful when their learned variables can be treated as variables of the world. A representation may contain the full latent state and still hide every physical factor inside a dense mixture of learned variables. Such mixing can leave predictive performance untouched while taking away the variables needed for intervention, precise control, visual editing, and mechanistic inspection. Many modules built on top of a world model act through a few variables at a time, such as a controller moving one object or a monitor watching one factor, and for such uses a well-spanned mixture may not be enough. A model that represents a scene through position, velocity, contact, and object attributes should isolate those individual latents rather than fold them into one dense state vector.

Joint-embedding predictive architectures (JEPAs) learn world models by predicting in representation space instead of reconstructing observations (LeCun et al., 2022). The choice, however, removes the anchor that grounds latents elsewhere in the identifiability literature. No decoder, observation likelihood, or auxiliary signal ties each learned variable to a variable of the world. Collapse, the

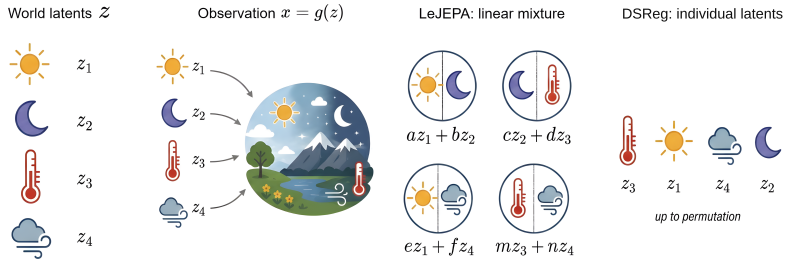


Figure 1: DSReg turns a mixed latent state into individual latents. World latents z generate observations $x = g(z)$. LeJEPA identifies the latent state up to a linear transformation, $\hat{z} = Az$, so each learned variable can remain a mixture of ground-truth latents. DSReg recovers individual latents up to signed permutation.

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.

best-known symptom of this missing anchor, is handled by an established toolkit (Grill et al., 2020; Chen & He, 2021; Zbontar et al., 2021; Bardes et al., 2022). At the same time, avoiding collapse does not recover the latents of the world: a non-collapsed representation may still encode them as an invertible mixture. What is needed is identifiability, the guarantee that every representation consistent with the training criterion matches the true latents up to a known ambiguity, such as a rotation.

Theory has begun to catch up. Under Gaussian latent-world assumptions, LeJEPA identifies the latent state up to a linear transformation (Klindt et al., 2026), separating the representations that merely predict well from those that align with the hidden generative process of the world. Yet each individual latent can still be a mixture of several ground-truth factors (Figure 1), and the mixture does harm where world models are expected to help. Consider a household robot asked to slide a cup to the left while keeping a nearby knife fixed: if one learned latent mixes the cup position with the knife position, changing that latent is no longer a targeted cup movement, and a correction meant for the cup can also shift the knife. Thus, a fundamental question about learning world models remains open:

Can we learn individual world latents without reconstruction?

The question has deep roots in the identifiability literature. In nonlinear ICA, one line of work uses auxiliary variables as weak supervision (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019; Yao et al., 2021; Hälvä et al., 2021; Lachapelle et al., 2022), while another constrains the mixing function itself (Taleb & Jutten, 1999; Moran et al., 2022; Kivva et al., 2022; Zheng et al., 2022; Buchholz et al., 2022). A reconstruction-free line reaches component-wise recovery through contrastive objectives, but leans on nonstationarity, non-Gaussian temporal dependence, or a non-Gaussian conditional (Hyvarinen & Morioka, 2016; Hyvärinen & Morioka, 2017; Zimmermann et al., 2021), none of which a stationary Gaussian world state offers (Appendix B). Meanwhile, in causal representation learning, identifiability typically rests on interventional data (von Kügelgen et al., 2023; Jiang & Aragam, 2023; Jin & Syrgkanis, 2023; Zhang et al., 2024) or counterfactual views (Von Kügelgen et al., 2021; Brehmer et al., 2022). Each of these routes anchors the latents in reconstruction, an explicit likelihood, or symmetry-breaking side information. In contrast, methods that discard every anchor stop at the linear guarantee above, and recovering individual latents without any of these anchors has remained an open problem.

Contributions. We close this gap. We prove, to the best of our knowledge, the first component-wise identifiability result that requires no reconstruction, labels, or distributional asymmetry. Under *Structural Diversity*, which asks only that different latents leave distinct dependency footprints on observations, together with a standard faithfulness condition, each latent variable of the world is identified up to signed permutation. Therefore, each learned variable is provably a variable of the hidden world. Moreover, we prove that Structural Diversity is strictly weaker than prior structural conditions across the identifiability literature. Our theory further prescribes an actionable objective, named ***DSReg*** (***D***ependency-***S***parsity ***Reg***ularization): it applies post hoc to any linearly identified representation, reusing trained checkpoints at no loss relative to joint training. Instantiated on a JEPA whose predictive objective supplies the linear-identifiability premise, DSReg resolves the Gaussian rotational invariance and pins down individual latents. We validate the theory and the method across synthetic regimes, world model probes, learned visual encoders, and external renderers: DSReg preserves dense prediction while improving individual-latent recovery and its downstream use, from sparse control and visual editing to few-shot latent readout of world state.

2 PRELIMINARIES

Identifiability asks when the latent variables that generated the observations can be recovered, up to relabeling and sign. Throughout, $z = (z_1, \ldots, z_d)$ is the latent state of the world, $x = g(z)$ an observation, and $h(z) = f_\theta(g(z))$ the estimated latents of a JEPA encoder f_θ. The recovery we target is component-wise. Each estimated variable should be one variable of the world rather than an arbitrary unknown mixture of several of them, as the next definition makes precise.

Definition 1 (Identifiability of individual world latents)**.** *A representation $\hat{z} = h(z)$* recovers the individual world latents*, and is called* signed-permutation identifiable*, if $\hat{z}_i = s_i z_{\pi(i)}$ almost surely for every $i \in [d]$, for some permutation π of $[d]$ and signs $s_1, \ldots, s_d \in \{\pm 1\}$.*

Identifiability of latent variable models. Nonlinear Independent Component Analysis obtains component-wise guarantees by adding structure beyond the marginal distribution of x: auxiliary

variables (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019; Yao et al., 2021; Hälvä et al., 2021; Lachapelle et al., 2022; Hyvärinen et al., 2024) or constraints on the mixing function (Taleb & Jutten, 1999; Moran et al., 2022; Kivva et al., 2022; Zheng et al., 2022; Buchholz et al., 2022). Causal representation learning extends such guarantees to dependent latents through interventions (von Kügelgen et al., 2023; Jiang & Aragam, 2023; Jin & Syrgkanis, 2023; Varici et al., 2025), counterfactual or multi-view observations (Von Kügelgen et al., 2021; Brehmer et al., 2022; Yao et al., 2024; Morioka & Hyvärinen, 2023), distributional shifts (Zhang et al., 2024; Ng et al., 2026), or combinations of complementary constraints (Li et al., 2025; Reizinger et al., 2025; Yao et al., 2025). Meanwhile, parallel results exist for factor analysis (Hu, 2008) and dictionary learning (Zheng et al., 2026). Nearly all of these strategies, however, anchor the latents through observation reconstruction or an explicit likelihood. That anchor is exactly what the JEPA objective removes, so the existing techniques do not transfer to settings without reconstruction.

JEPAs and linear identifiability. A JEPA learns by predicting the embedding of a target observation from the embedding of a related context observation, rather than reconstructing the observation itself. In a world model the pairs are consecutive states of a temporal process, $x_t = g(z_t)$ and $x_{t+1} = g(z_{t+1})$; as in Klindt et al. (2026) we assume the process is stationary, so all time slices share one law and we drop the subscript, writing (x, x^+). Given such a pair, an encoder f_θ is trained so that the representation of x aligns with that of x^+, while constraints prevent collapse. The objective is therefore deliberately non-reconstructive: no decoder maps representations back to observations, and no reconstruction error or observation likelihood enters training. What it can identify on its own is the weaker notion of linear identifiability.

Definition 2 (Linear identifiability)**.** *A representation $\hat{z} = h(z)$ is* linearly identifiable *if $\hat{z} = Az$ almost surely for some invertible $A \in \mathbb{R}^{d\times d}$.*

LeJEPA (Balestriero & LeCun, 2025) instantiates this design with SIGReg, a scalable regularizer that drives the embedding toward an isotropic Gaussian. In a Gaussian latent dynamical system, Klindt et al. (2026) proves the following characterization of the ambiguity DSReg must resolve.

Theorem 1 (LeJEPA linear identifiability (Klindt et al., 2026))**.** *Let $z \sim \mathcal{N}(0, I_d)$, and let the target latent follow the stationary Gaussian transition $z^+ = \rho z + \sqrt{1-\rho^2}\,\eta$ with $\eta \sim \mathcal{N}(0, I_d)$ independent of z, and $\rho \in (0, 1)$. Let $x = g(z)$, $x^+ = g(z^+)$, and let $h = f_\theta \circ g : \mathbb{R}^d \to \mathbb{R}^d$ be measurable with $h(z) \sim \mathcal{N}(0, I_d)$. Then the population alignment loss satisfies*

$$\mathcal{L}_{\text{align}}(h) = \mathbb{E}\big[\|h(z^+) - h(z)\|_2^2\big] \;\geq\; 2(1-\rho)d, \tag{1}$$

with equality if and only if $h(z) = Qz$ almost surely for some orthogonal $Q \in O(d)$. At the alignment optimum, h is linearly identifiable in the sense of Definition 2, with orthogonal $A = Q$.

Under LeJEPA's Gaussian-world conditions, alignment plus the Gaussian constraint identifies the latent process up to a rotation. Each estimated variable may still be a mixture of true latents, and the next section explains why *no distributional criterion can ever do better*.

3 Identifiability from Structural Diversity

The theory develops in three steps: Section 3.1 isolates the obstacle, a distributionally undetectable rotation that can mix every latent; Section 3.2 builds the dependency structure that this rotation cannot hide from and proves the main identifiability theorem; Section 3.3 turns DSReg into a practical estimator whose recovery guarantee survives estimation error in the Jacobians.

3.1 From linear to signed-permutation identifiability

Linear is not component-wise. Linear identifiability does not imply recovery of individual latents. The counterexample is any rotation. If $\hat{z} = Az$ is linearly identifiable and $U \in O(d)$ is not a signed permutation, then $U\hat{z} = (UA)z$ satisfies Definition 2 equally well, yet each of its variables is a mixture of several world latents. The two notions are separated by exactly this freedom. In the Gaussian JEPA setting no per-latent rescaling enters the comparison, because both z and the representation are standardized and the remaining ambiguity from Theorem 1 is orthogonal.

The gap matters. A representation whose variables mix object position with illumination predicts the scene as well as one that separates them, but a controller that is allowed to change only one latent

cannot move the object without also changing the lighting. Under the premise $h(z) = Qz$, a generic row of Qz mixes all d latents at once, so a linearly correct representation may still be ill suited to the downstream tasks that motivate learning a world model in the first place.

The obstacle: rotation invariance. The linear indeterminacy is not unique to LeJEPA. It limits any objective that evaluates representations only through their distribution, because the Gaussian distribution is rotation invariant. An isotropic Gaussian is like a ball. Its density $p(z) \propto \exp(-\|z\|^2/2)$ depends on z only through its norm, so rotating changes nothing and $Uz \sim \mathcal{N}(0, I_d)$ for every $U \in O(d)$. A non-Gaussian law such as Laplace would break this symmetry, which is how contrastive nonlinear ICA separates components. The Gaussian state gives it nothing to work with. Thus, h and Uh are distributionally indistinguishable, even though the rotation U may arbitrarily mix the latent variables. An unknown rotation breaks any correspondence h already achieves, and downstream use can then silently destroy the alignment without any warning signal that it happened.

From obstacle to signal. The same observation says where a fix must come from. If no function of the latent distribution can detect the rotation, the identifying signal must live outside that distribution, in something observable that does change when the latents are rotated. Fortunately, we find such a signal in how observations depend on the estimated latents, as we elaborate in the next section.

Insight. Rotations of a Gaussian latent state are undetectable to every distributional criterion, yet they mix the variables. Recovering individual latents needs a rotation-sensitive signal. DSReg takes that signal from how the observations depend on the estimated latents.

3.2 Dependency supports and Structural Diversity

DSReg takes the rotation-sensitive signal from how observed variables depend on latents. Let $x = g(z) \in \mathbb{R}^p$ and define the dependency Jacobian $D(z) = \partial x/\partial z$. Throughout, μ is the law of z (standard Gaussian under Theorem 1). The map g is differentiable μ-almost everywhere with measurable Jacobian, so $D(z)$ is defined μ-a.e. The *dependency support* declares an entry active when its partial derivative is nonzero on a set of positive μ measure. For matrix-valued $M(z)$ we count active entries by

$$\|M(\cdot)\|_{0,\mu} = \left|\{(r,i) : M_{ri}(z) \neq 0 \text{ on a set of positive } \mu \text{ measure}\}\right|,$$

a count with values in $\{0, 1, \ldots, pd\}$. Only finitely many values occur, so some rotation attains the smallest value and minimizers over $O(d)$ exist. Neither g nor D is assumed known. Section 3.3 estimates the dependency structure from observations and estimated latents alone.

What rotation disturbs. The support is what rotation changes while the latent distribution stays fixed. If one latent controls position and affects one set of observed variables while another controls color and affects a different set, the rotation $(z_i \pm z_j)/\sqrt{2}$ leaves the Gaussian law unchanged but makes each rotated latent act on the union of the two sets. The Jacobian support records this mixing.

The same example shows what structure must supply. Define the *dependency footprint* of latent z_i as

$$\mathcal{S}_i = \{\, r : D_{ri}(z) \neq 0 \text{ on a set of positive } \mu \text{ measure} \,\},$$

the set of observed variables that z_i affects. Figure 2 shows the same object as a support matrix. If two latents affected exactly the same observed variables, structure alone could not distinguish them, since a rotation confined to the pair changes nothing dependency structure can see. Identifiability from structure therefore requires different latents to leave different dependency footprints on the observed variables.

Assumption 1 (Structural Diversity)**.** *The dependency footprints are pairwise distinct: $\mathcal{S}_i \neq \mathcal{S}_j$ for all $i \neq j$.*

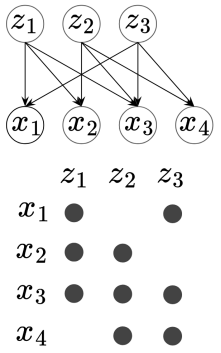


Figure 2: Dependency footprints. Each column of $\partial x/\partial z$ marks which observed variables one latent affects.

A structural condition. Structural Diversity constrains the dependency structure alone and places no requirement on the latent distribution. It asks only that no two latents affect exactly the same set of observed variables. Footprints may overlap, differ in a single variable, or even be nested.

To connect the dependency structure to the residual ambiguity, recall that training leaves one unknown rotation: $h(z) = Qz$ with $Q \in O(d)$. DSReg searches over a further

rotation $R \in O(d)$, producing candidate latents $\tilde{z} = Rh(z)$. The composite map from true to candidate latents is RQ, and the goal is to choose R so that RQ is a signed permutation. The next lemma computes how the dependency Jacobian transforms under this search.

Lemma 1 (Dependency bridge)**.** *Suppose a representation is linearly identifiable with orthogonal ambiguity, so that $h(z) = Qz$ for some $Q \in O(d)$, as Theorem 1 provides for JEPA representations, and let $x = g(z)$ have dependency Jacobian $D(z) = \partial x / \partial z$ almost everywhere. Then for candidate latents $\tilde{z} = Rh(z)$ with $R \in O(d)$,*

$$\frac{\partial x}{\partial \tilde{z}} = D(z) Q^\top R^\top \qquad \textit{almost everywhere}. \tag{2}$$

The identity is nothing but the chain rule, since the premise makes the map between candidate and true latents the linear bijection $z = Q^\top R^\top \tilde{z}$. It is worth noting that no invertibility of the encoder on observations is needed, in keeping with an encoder that discards low-level detail (Appendix A). The ambiguity thus enters the dependency structure as right multiplication by the orthogonal $U = Q^\top R^\top$, and DSReg selects the remaining rotation by minimizing the dependency support. Among the representations in the identified orthogonal class, it prefers the one whose latents have the sparsest dependency on the observations,

$$\min_{R \in O(d)} \left\| \frac{\partial x}{\partial \tilde{z}} \right\|_{0,\mu}, \qquad \tilde{z} = Rh(z). \tag{3}$$

Why the support sees the rotation. For a Givens rotation on a fixed generic matrix, the possible effects on the support form a short taxonomy (Ghassami et al., 2020). Apart from column swaps, any rotation that mixes two columns deactivates the targeted entry and activates entries wherever the two column supports differ, so it always leaves a visible trace (Proposition 2 in Appendix A restates the taxonomy in our setting). For the function-valued dependency Jacobian, the support of a mixed candidate latent follows the union-support formula of Lemma 2 (Appendix A), under Assumption 2 below. Under Structural Diversity, a rotation mixing latents must change the support.

Assumption 2 (Functional no-cancellation)**.** *For each observation row r, write*

$$I_r = \{i : D_{ri}(z) \neq 0 \textit{ on a set of positive } \mu \textit{ measure}\}.$$

Then any constant-coefficient relation among the active derivative functions is trivial:

$$\sum_{i \in I_r} c_i D_{ri}(z) = 0 \quad \textit{for } \mu\textit{-a.e. } z \quad \Longrightarrow \quad c_i = 0 \textit{ for all } i \in I_r.$$

Theorem 2 (Component-wise identifiability without reconstruction)**.** *In the setting of Lemma 1, assume Structural Diversity (Assumption 1) and Functional no-cancellation (Assumption 2). Then for every minimizer R of the support-sparsity criterion* (3)*, the matrix RQ is a signed permutation: there exist a permutation π of $[d]$ and signs $s_1, \ldots, s_d \in \{\pm 1\}$ such that the candidate latents $\tilde{z} = Rh(z)$ satisfy $\tilde{z}_i = s_i z_{\pi(i)}$ almost surely for every $i \in [d]$.*

After the DSReg rotation, each estimated latent is therefore one individual world latent up to sign and relabeling, the only ambiguity left, and one that keeps each variable individually meaningful.

Insight. Theorem 2 takes reconstruction-free world models beyond the linear indeterminacy. A representation that spans the world state becomes one whose variables are the world's individual latents up to signed permutation, without decoder, likelihood, or labels. This is the difference between holding all the latent information and being able to reach any one piece of it.

Functional no-cancellation is a faithfulness-type genericity requirement. It rules out constant-coefficient cancellations that would hide genuine Jacobian edges. It plays the same role as sufficient nonlinearity and sufficient variability in nonlinear identifiability, where Jacobian variation across samples must span each active support (Hyvarinen & Morioka, 2016; Khemakhem et al., 2020; Sorrenson et al., 2020; Lachapelle et al., 2022; Zheng et al., 2022; Kong et al., 2023; Yan et al., 2023; Zhang et al., 2024; Lachapelle et al., 2023). Moreover, within smooth observation families with fixed supports and analytic Gram determinants, and provided each row's determinant is nonzero somewhere in the family, the condition holds for Lebesgue-almost every parameter value (Proposition 3).

Strictly weaker than prior conditions. Structure-based identifiability rests on sparsity conditions on the mixing Jacobian in nonlinear ICA (Zheng et al., 2022) and on column supports in linear ICA (Ng et al., 2023). Proposition 1 states these conditions directly on the footprints and orders them. Every one forces pairwise-distinct footprints, and no converse implication holds.

Proposition 1 (Structural Diversity is strictly weaker)**.** *Stated on the dependency footprints $\mathcal{S}_1, \ldots, \mathcal{S}_d$, each of the following conditions implies Structural Diversity (Assumption 1), and none of the converse implications holds:*

1. Structural Sparsity, intersection form *(Zheng et al., 2022): for every $k \in [d]$ there is a nonempty set $\mathcal{C}_k$ of observed variables with $\bigcap_{r \in \mathcal{C}_k} \{\, i \in [d] : r \in \mathcal{S}_i \,\} = \{k\}$;*
2. Structural Sparsity, overlap-rank form *(Zheng et al., 2022): for every $\mathcal{C} \subseteq [d]$ with $|\mathcal{C}| \geq 2$ and every $k \in \mathcal{C}$, $\left|\bigcup_{j \in \mathcal{C}} \mathcal{S}_j\right| - \operatorname{rank}(M^{\mathcal{C}}) > |\mathcal{S}_k|$, where $M^{\mathcal{C}}$ is the binary matrix with $M^{\mathcal{C}}_{rj} = 1$ exactly when $j \in \mathcal{C}$ and r lies in $\mathcal{S}_j$ and in $\mathcal{S}_{j'}$ for some other $j' \in \mathcal{C}$;*
3. Non-Inclusion*, a common weakening of both forms above: $\mathcal{S}_i \not\subseteq \mathcal{S}_j$ for all $i \neq j$;*
4. Structural Variability *(Ng et al., 2023): $|\mathcal{S}_i \triangle \mathcal{S}_j| \geq 2$ for all $i \neq j$.*

The difference matters in practice. A global latent that touches every observed variable (illumination, camera gain, a global style factor) has a footprint that strictly contains those of localized latents. Nested footprints of this kind violate Non-Inclusion, and hence both forms of Structural Sparsity, yet satisfy Structural Diversity, so Theorem 2 still applies. The boundary is equally sharp in the other direction. If $\mathcal{S}_i = \mathcal{S}_j$, every rotation in the (i, j) plane leaves both rotated columns active on the same footprint, the support count acquires non-permutation minimizers, and distinguishing the pair would then require information beyond the dependency support pattern itself.

Insight. Prior structural conditions were built for reconstruction-based models, where a decoder anchors the latents. Proposition 1 brings structure-based identifiability to the reconstruction-free setting while even weakening the condition, suggesting that Structural Diversity is the weakest among all structural conditions, possibly even a necessary one for recovering individual latents.

3.3 The DSReg estimator

The full method, LeJEPA + DSReg, has two objectives: predictive Gaussian learning identifies the orthogonal solution class, and DSReg selects the sparsest representative within that class. This decomposition is exact at the population level. For every $R \in O(d)$, alignment obeys $\mathcal{L}_{\text{align}}(Rh) = \mathcal{L}_{\text{align}}(h)$ and the Gaussian constraint is preserved. The orthogonal DSReg head may therefore be optimized alongside predictive learning or after it. We recommend the decoupled schedule, which is simpler, cheaper, and reuses existing checkpoints, and an ablation confirms the convenience costs nothing. The schedules recover equally well, with the joint head marginally lower until the decoupled fit completes it (Appendix Table 2). Specifically, on the whitened representation, DSReg estimates local maps $B_a \approx \partial x / \partial h$ by ridge regression at m anchors and minimizes the anchor-averaged ℓ_1 relaxation of the population support criterion (3):

$$\mathcal{L}_{\text{DSReg}}(R) = \frac{1}{m} \sum_{a=1}^{m} \|B_a R^\top\|_1, \qquad R \in O(d). \tag{4}$$

Here $B_a R^\top$ estimates how observations depend on the candidate latents Rh. The local regressions read the observations at the anchors, as any identifying signal ultimately must, but they provide no global observation model, train no decoder, and never optimize reconstruction. Theorems 2 and 3 analyze the support criterion and its thresholded perturbation, while Equation (4) is the continuous finite-sample relaxation used in every experiment. Because the ℓ_1 norm weighs magnitude as well as count, the relaxation and the support criterion can in principle select different rotations. On the benchmarks of Figure 8 an exact support count recovers at least as well as the ℓ_1 default, a refinement of the ℓ_1 solution by the exact count closes the remaining gap, and differentiability is what lets the objective scale to large d and p (Appendix C.2). Ground-truth latents are never accessible during estimation and enter only through benchmark construction and evaluation.

Scaling. Equation (4) appears to need m matrices of size $p \times d$, which would explode as soon as the observed dimension grows. Fortunately, it does not. Each B_a is a ridge fit over the k neighbors of its anchor, so it factors as $B_a = \Delta X_a^\top P_a$ with $P_a \in \mathbb{R}^{k \times d}$, and the observations ΔX_a are already in memory as the dataset. The criterion is therefore determined by factors costing $O(mkd)$ rather than $O(mpd)$, and can be evaluated without the large array ever existing. Appendix C.2 turns this one fact into three interchangeable ways to evaluate the same objective, materialized, exactly factored, and an unbiased row-sampled estimator whose per-step cost is independent of p and m, and says which

Algorithm 1 DSReg: recovering individual latents without reconstruction

1: Train the predictive Gaussian objective to identify the orthogonal solution class; whiten and freeze the representation h.
2: Choose the observed vector used for dependency fitting; sample m anchor points $h_1, \ldots, h_m$ from the data.
3: For each anchor, estimate $B_a \approx \partial x/\partial h$ by ridge regression of observations on h over the k nearest neighbors of h_a in representation space.
4: Select $R = \arg\min_{R \in O(d)} \mathcal{L}_{\mathrm{DSReg}}(R)$.
5: **return** the DSReg representation $\tilde{z} = Rh$.

Joint schedule: the rotation may equally be trained alongside step 1, as a head updated during training that never feeds back into the encoder, with steps 2–4 as a final refit; recovery matches the decoupled schedule.

to use when. The full estimated procedure reaches $d = 8192$ on one 48 GB GPU, with samples to 10^6 and observed dimensions to 2^{17} leaving the per-step cost essentially flat.

Robustness to estimation error. Whitening reduces any invertible linear ambiguity to an orthogonal one: $\mathrm{Cov}(h)^{-1/2}h(z) = \tilde{Q}z$ with $\tilde{Q} \in O(d)$ (Lemma 4, Appendix A.4). One may then wonder how estimated-Jacobian error affects the support target that DSReg relaxes.

Recovery degrades gracefully, governed by two constants with mechanical meanings. The first, $\sigma > 0$, measures how strongly an active dependency registers in the data: the smallest $\lambda_{\min}(G_r)^{1/2}$ over the row Gram matrices $G_r = [\langle D_{ri}, D_{ri'} \rangle_{L^2(\mu)}]_{i,i' \in I_r}$, positive under Assumption 2. The second, the pattern margin $\rho^*(\delta) > 0$, measures how visibly a rotation δ-far from every signed permutation must disturb the support pattern: the (s^*+1)-th largest of the sub-vector norms $\|U_{I_r,j}\|_2$ over such rotations, with $s^* = \sum_i |\mathcal{S}_i|$ the population minimum (Lemma 6). A strong signal cannot be hidden by weak noise. Whenever the sum of the threshold and the noise level stays below $\sigma\rho^*(\delta)$, every minimizer lands within δ of a signed permutation.

Theorem 3 (Approximate recovery under Jacobian perturbation)**.** *Let the assumptions of Theorem 2 hold with $d \geq 2$, together with*

$$\operatorname{ess\,sup}_z \max_r \|D_{r,:}(z)\|_2 \leq M < \infty.$$

Let $\widehat{D}(z) = D(z) + E(z)$ with E measurable and $\operatorname{ess\,sup}_z \max_r \|E_{r,:}(z)\|_2 \leq \varepsilon$, and for a threshold $\tau > \varepsilon$ define

$$\widehat{N}_\tau(R) = \Big| \big\{ (r,j) : \mu\big(|(\widehat{D}(z)\, Q^\top R^\top)_{rj}| > \tau \big) > 0 \big\} \Big|, \qquad R \in O(d).$$

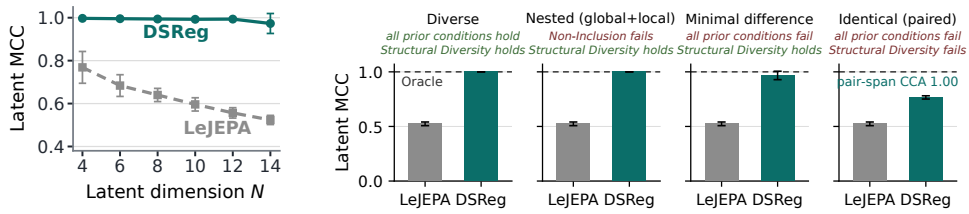
If $\delta > 0$ satisfies $\tau + \varepsilon < \sigma\, \rho^*(\delta)$, then minimizers of $\widehat{N}_\tau$ over $O(d)$ exist and every minimizer $\widehat{R}$ satisfies $\min_P \|\widehat{R}Q - P\|_F < \delta$ over signed permutations P.

Exact recovery is impossible under noise, since thresholding at any $\tau > 0$ ties the count on rotations $O(\tau)$-close to a signed permutation, so δ-closeness is the right notion. The statement concerns the thresholded population criterion. Appendix Figure 9 stress-tests the ℓ_1 estimator under noise.

Insight. DSReg needs no retraining or decoder. Predictive learning fixes the orthogonal class and DSReg, fit jointly with it or afterwards, selects individual latents. Theorem 3 shows the thresholded population target is stable to Jacobian estimation error, and the ℓ_1 implementation is exercised directly in ablations. Dense performance is untouched because a rotation loses no information.

4 Experiments

The experiments validate the theory and support downstream applications. Encoders trained from scratch recover individual latents wherever Structural Diversity holds (Section 4.1). The regime map, failure included, is exactly the theory's (Section 4.2). The recovered latents pay off in sparse control, editing, prediction, and monitoring (Section 4.3). The gains survive encoders trained from pixels and renderers we did not design, reaching a supervised oracle (Section 4.4). Dense $R^2(h \to z)$, the best linear readout from estimated to ground-truth latents, measures whether the state is retained as a span; the mean correlation coefficient (MCC) after maximum-weight matching, the standard identifiability metric in nonlinear ICA (Hyvarinen & Morioka, 2016; Khemakhem et al., 2020), measures one-to-one recovery; and sparse-use probes ask whether a module can act through a few estimated latents without a dense unmixing. Ground-truth latents define benchmark pairs and evaluation scores. DSReg is fit from estimated latents with raw observations or a fixed preprocessing chosen before fitting.



(a) Learned encoders **(b)** Footprint regimes ($N = 16$; full sweep in Figure 7)

Figure 3: DSReg recovers individual latents wherever Structural Diversity holds, and only there. (a) DSReg (solid) is near ceiling at every N, twenty runs. LeJEPA (dashed) stays mixed. (b) On analytic orbits $h = Qz$, recovery succeeds wherever the condition holds and falls back to the shared pair subspace where it fails.

Throughout, LeJEPA labels the frozen representation before the DSReg rotation, so each pair isolates what the rotation does to one representation. Below, N denotes the benchmark's latent dimension, and results report mean $\pm$ std over all runs with random seeds, five to fifteen for the learned visual encoders, and ten to twenty for the remaining families, with the exact count listed in every figure and table caption. Full setups and the supporting suite, led by the scaling study, are in Appendix C, which every claim in the subsections that follow points into.

4.1 The full method recovers individual latents end to end

The first question is whether the whole method works end to end. We train a generic encoder from scratch, fit the rotation from observations alone, and check whether individual latents come out. The synthetic benchmark makes Structural Diversity hold by construction, in the hardest form we consider. An MLP observation map $x_j = z_j + f_j(z_{<j})$ gives z_k the footprint $\{x_k, \ldots, x_N\}$, so all footprints are distinct yet nested, the regime that prior conditions exclude. Figure 3a traces recovery across latent dimension. LeJEPA retains the predictive span but leaves its variables mixed, while DSReg lifts them to near-ceiling recovery at every dimension tested. Recovery tracks the linear-identifiability premise of Theorem 1, which DSReg inherits rather than establishes. Once the encoder reaches the premise, the rotation does the rest. Latent-only rotations (PCA, Varimax (Kaiser, 1958; Rohe & Zeng, 2023), FastICA) fail on the same estimated latents, so the Jacobian dependency signal rather than the optimizer is what does the work, on exactly the same inputs (Appendix C.8, Table 5).

4.2 Every regime lands where the theory puts it

The theory makes a sharp and testable prediction. Recovery must succeed whenever footprints differ, however slightly, and may fail only when they coincide. Each panel of Figure 3b is an analytic orbit ($h = Qz$) realizing one regime of Proposition 1: *diverse* footprints satisfy the prior conditions, *nested* footprints add a global latent containing all local ones, *minimal-difference* footprints differ in one observed feature, and *identical* footprints violate Structural Diversity outright.

Wherever Structural Diversity is satisfied, recovery is near-perfect at every scale while unrotated baselines decay with dimension. The nested regime, which violates Non-Inclusion and both sparsity certificates, is recovered as cleanly as the easy one. Even the failure is the one the theory dictates. With identical footprints, individual recovery is impossible from support alone, yet DSReg still recovers the pair's span almost perfectly. The map is the theory's (Appendix Figures 7 and 8).

4.3 Sparse modules can act through recovered latents

One may wonder why individual latents should be preferred to a well-spanned mixture. The modules built on a world model are sparse, and a rotation breaks their interface, so one estimated latent moves several physical factors at once. We therefore evaluate four sparse-use settings on states $h = Qz$ for a random orthogonal Q: visual edits in three environments, sparse model predictive control (MPC) fit from few transitions, short horizon rollout prediction, and transition-surprise detection, six probes in total, with environments and probe procedures detailed in Appendix C.4.

DSReg improves every probe and wins in nearly every run (Figure 4), while dense-span checks stay matched (Appendix Table 6). The same signature persists on encoders trained from pixels, where DSReg is ahead in every single run of all three probes (Appendix C.5).

4.4 THE GAINS SURVIVE LEARNED ENCODERS AND EXTERNAL RENDERERS

Nothing in the method depends on the analytic setting, so we move to encoders trained from pixels and renderers we did not design (Appendix C.5). Figure 5a shows the outcome on the visual-factor and object-scene benchmarks. DSReg lifts convolutional encoders trained from scratch to the supervised Procrustes oracle that sees labels, while PCA, Varimax, and FastICA leave the latents nearly as mixed as no rotation, the rotational symmetry at work (Appendix C.8). Figure 5b widens the representation. The top-N right-singular subspace of the estimated Jacobians selects the directions the observations depend on, and the match holds at every width, so one can start wide and let the criterion find the subspace it needs.

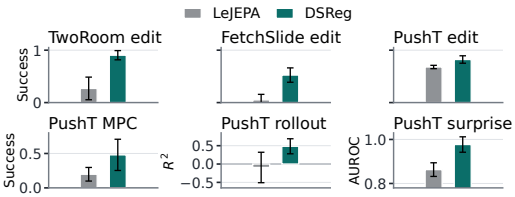
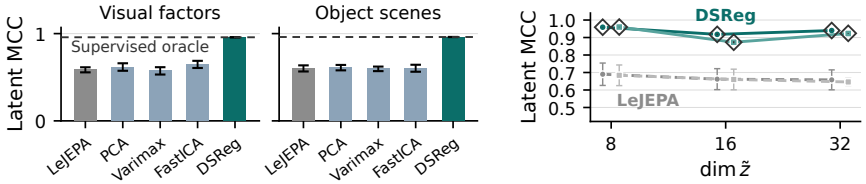


Figure 4: DSReg improves all six sparse-use probes. DSReg in teal, LeJEPA in gray. Tasks are (1) editing (top) and (2) control, prediction, and monitoring (bottom).

On external renderers the same signature holds. Gaussian 3DShapes (Kim & Mnih, 2018) rises from mixed to near-ceiling recovery at matched dense R^2 (Appendix Table 4), and on quarter-orientation dSprites (Matthey et al., 2017), where the premise is only partially attained, DSReg still beats swept β-VAE and β-TCVAE baselines and the premise sets the ceiling (Appendix Fig. 18).



(a) Rotation baselines on learned encoders **(b)** Overparameterized: $\dim \tilde{z}$ grows

Figure 5: DSReg matches the supervised Procrustes oracle on encoders trained from pixels. (a) Latent-only rotations leave the variables mixed. DSReg closes the gap to the label-fitted oracle (fifteen seeds). (b) With $\dim z = 8$ fixed, the match holds within 0.001 as the estimate widens to $\dim \tilde{z} = 32$ (five seeds).

5 DISCUSSION

Individual world latents can be recovered without reconstruction. To the best of our knowledge, prior guarantees at this resolution ran through a decoder or auxiliary access, and objectives free of any such anchor stopped at a linear class. This paper closes that gap: under the LeJEPA premise and Structural Diversity, ***DSReg*** provably recovers each individual world latent through a sparsity regularization on the dependency Jacobian. The condition is strictly weaker than prior structural conditions, and the guarantee is stable to estimation error. The experiments confirm the theory where it is testable and show why it matters. After the DSReg rotation, a controller, a dynamics model, or a monitor acts through individual variables, while dense readouts remain unchanged on every benchmark.

A world model whose variables are the variables of the world is not a predictive black box. Its latents can be intervened on and inspected one at a time, and all of it comes from structure the observations already carry. Structural Diversity is a mild requirement: factors with identical footprints cannot be distinguished without extra information, and the real world rarely produces them. Other conditions may also be leveraged to conduct alternative strategies to recover individual world latents, such as the orthogonality of the Jacobian in independent mechanism analysis (Gresele et al., 2021).

The clearest limitation is scope: our benchmarks are simulated or rendered, and DSReg has not yet been tested on a physical robot. That gap is the future work that excites us most. How does a robot behave when each factor of its world model is its own variable, and how far does that carry planning? Identical footprints, the one boundary that support alone cannot cross, invite signals such as temporal structure or cheap interventions. Moreover, the Gaussian latent-world premise is a foothold rather than a ceiling. Every extension of linear identifiability to richer worlds widens the guarantee.

AI DISCLOSURE

In this paper, we used generative AI tools to assist polishing and plotting. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nilin Abrahamsen and Philippe Rigollet. Sparse Gaussian ICA. *arXiv preprint arXiv:1804.00408*, 2018.

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. IEEE, 2023.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. In *Advances in Neural Information Processing Systems*, 2022.

Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15745–15753. IEEE, 2021.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018.

AmirEmad Ghassami, Alan Yang, Negar Kiyavash, and Kun Zhang. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *International conference on machine learning*, pp. 3494–3504. PMLR, 2020.

Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34:28233–28248, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284, 2020.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvärinen. Disentangling identifiable features from noisy data with structured nonlinear ICA. *Advances in Neural Information Processing Systems*, 34, 2021.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Yingyao Hu. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61, 2008.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in neural information processing systems*, 29, 2016.

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

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868. Pmlr, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear: A. hyvärinen et al. *Annals of the Institute of Statistical Mathematics*, 76(1): 1–33, 2024.

Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. *Advances in Neural Information Processing Systems*, 36:60468–60513, 2023.

Jikai Jin and Vasilis Syrgkanis. Learning causal representations from general environments: Identifiability and intrinsic ambiguity. *arXiv preprint arXiv:2311.12267*, 2023.

Henry F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Bobak Kiani, Randall Balestriero, Yann LeCun, and Seth Lloyd. projUNN: efficient method for training deep networks with unitary matrices. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14448–14463, 2022.

Hyunjik Kim and Andriy Mnih. Disentangling by Factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model? *arXiv preprint arXiv:2605.26379*, 2026.

Lingjing Kong, Biwei Huang, Feng Xie, Eric Xing, Yuejie Chi, and Kun Zhang. Identification of nonlinear latent hierarchical models. *Advances in Neural Information Processing Systems*, 36:2010–2032, 2023.

Yilun Kuang, Yash Dagade, Tim GJ Rudner, Randall Balestriero, and Yann LeCun. Rectified LpJEPA: Joint-embedding predictive architectures with sparse and maximum-entropy representations. *arXiv preprint arXiv:2602.01456*, 2026.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, 2022.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *Advances in Neural Information Processing Systems*, 36:25112–25150, 2023.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Zijian Li, Shunxing Fan, Yujia Zheng, Ignavier Ng, Shaoan Xie, Guangyi Chen, Xinshuai Dong, Ruichu Cai, and Kun Zhang. Synergy between sufficient changes and sparse mixing procedure for disentangled representation learning. In *International Conference on Learning Representations*, volume 2025, pp. 44065–44089, 2025.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

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.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Gemma Elyse Moran, Dhanya Sridhar, Yixin Wang, and David M. Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022.

Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. *arXiv preprint arXiv:2310.15709*, 2023.

Ignavier Ng, Yujia Zheng, Xinshuai Dong, and Kun Zhang. On the identifiability of sparse ICA without assuming non-gaussianity. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47960–47990, 2023.

Ignavier Ng, Shaoan Xie, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Causal representation learning from general environments under nonparametric mixing. *arXiv preprint arXiv:2604.23800*, 2026.

Frosti Palsson, Magnus O Ulfarsson, and Johannes R Sveinsson. Sparse Gaussian noisy independent component analysis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4224–4228. IEEE, 2014.

William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *European conference on computer vision*, pp. 581–597. Springer, 2020.

Patrik Reizinger, Siyuan Guo, Ferenc Huszár, Bernhard Schölkopf, and Wieland Brendel. Identifiable exchangeable mechanisms for causal structure and representation learning. In *International Conference on Learning Representations*, volume 2025, pp. 62196–62223, 2025.

Karl Rohe and Muzhe Zeng. Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1037–1060, 2023.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). *arXiv preprint arXiv:2001.04872*, 2020.

Anisse Taleb and Christian Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on signal Processing*, 47(10):2807–2820, 1999.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112): 1–90, 2025.

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. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36:48603–48638, 2023.

Hanqi Yan, Lingjing Kong, Lin Gui, Yuejie Chi, Eric Xing, Yulan He, and Kun Zhang. Counterfactual generation with identifiability guarantees. *Advances in neural information processing systems*, 36:56256–56277, 2023.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, volume 2024, pp. 34817–34848, 2024.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, volume 2025, pp. 53847–53890, 2025.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Kun Zhang, Heng Peng, Laiwan Chan, and Aapo Hyvärinen. ICA with sparse connections: Revisited. In *International Conference on Independent Component Analysis and Signal Separation*, pp. 195–202. Springer, 2009.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *Forty-first International Conference on Machine Learning*, 2024.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ICA beyond structural sparsity. In *Advances in Neural Information Processing Systems*, 2023.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yujia Zheng, Zijian Li, Shunxing Fan, Andrew Gordon Wilson, and Kun Zhang. Diverse dictionary learning. *International Conference on Learning Representations*, 2026.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2025.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, 2021.

CONTENTS OF THE APPENDIX

A Proofs . . . 13

A.1 Dependency supports under the LeJEPA result . . . 14

A.2 Identifiability under Structural Diversity . . . 16

A.3 Comparison of structural conditions . . . 20

A.4 Robustness to Jacobian perturbation . . . 21

B Additional discussions . . . 24

C Additional experiments . . . 25

C.1 Metrics and protocol . . . 25

C.2 Scaling the dependency criterion . . . 27

C.3 Verification and robustness . . . 30

C.4 Sparse-use probes: environments and procedures . . . 33

C.5 Learned visual encoders . . . 34

C.6 Gaussian 3DShapes . . . 36

C.7 Quarter-orientation dSprites . . . 37

C.8 Rotation baselines and dense-use sanity checks . . . 38

A PROOFS

The linear-identifiability premise. The proofs take the linear-identifiability theorem of Klindt et al. (2026) as their starting point. Its assumptions, in our notation, are: (i) the world latents are independent Gaussians, $z \sim \mathcal{N}(0, I_d)$; (ii) positive pairs are generated by a stationary additive-noise transition that preserves this marginal, $z' = \rho z + \sqrt{1-\rho^2}\,\eta$ with $\eta \sim \mathcal{N}(0, I_d)$ independent of z and $\rho \in (0, 1)$; (iii) observations $x = g(z)$ come from an unknown generating process, and the estimated latents $h(z) = f_\theta(g(z))$ are given by a measurable map $h : \mathbb{R}^d \to \mathbb{R}^d$; (iv) the learned representation satisfies the Gaussian constraint $h(z) \sim \mathcal{N}(0, I_d)$, as enforced by SIGReg (Balestriero & LeCun, 2025), and attains the alignment lower bound $2(1-\rho)d$ of Theorem 1. Under (i)–(iv), the equality case of Theorem 1 gives $h(z) = Qz$ for some $Q \in O(d)$. Every statement below operates inside this Gaussian latent-world premise: it fixes the orthogonal equivalence class, and DSReg selects the signed permutation within it, from the observed dependency structure alone.

Proof sketch for linear identifiability in Klindt et al. (2026). So that the paper is self-contained on its premise, we sketch why the alignment bound holds with equality exactly at linear maps; the full proof is in Klindt et al. (2026). Since $h(z) \sim \mathcal{N}(0, I_d)$ and the pair (z, z^+) is stationary,

$$\mathcal{L}_{\text{align}}(h) = \mathbb{E}\|h(z^+)\|_2^2 + \mathbb{E}\|h(z)\|_2^2 - 2\sum_{i=1}^{d} \mathbb{E}\big[h_i(z^+)\, h_i(z)\big] = 2d - 2\sum_{i=1}^{d} \mathbb{E}\big[h_i(z^+)\, h_i(z)\big],$$

so minimizing alignment maximizes the summed cross-correlations. Expand each component in the multivariate Hermite basis $\{H_\alpha\}$, orthonormal for the standard Gaussian: $h_i = \sum_{|\alpha|\geq 1} c_{i\alpha} H_\alpha$ with $\sum_\alpha c_{i\alpha}^2 = 1$, where the Gaussian constraint supplies the zero mean and unit variance. The additive-noise transition acts diagonally on this basis (the Mehler kernel): $\mathbb{E}[H_\alpha(z^+) H_\beta(z)] = \rho^{|\alpha|}\, \mathbf{1}\{\alpha = \beta\}$. Therefore

$$\mathbb{E}\big[h_i(z^+)\, h_i(z)\big] = \sum_{|\alpha|\geq 1} \rho^{|\alpha|}\, c_{i\alpha}^2 \;\leq\; \rho,$$

with equality exactly when all coefficient mass sits at degree one, since $\rho^k < \rho$ for every $k \geq 2$ and $\rho \in (0, 1)$. Summing over i gives $\mathcal{L}_{\text{align}}(h) \geq 2(1-\rho)d$, and equality forces each h_i to be a degree-one polynomial, so $h(z) = Az$ almost surely for some matrix A; the constraint $h(z) \sim \mathcal{N}(0, I_d)$ then

gives $AA^\top = I_d$, hence $A \in O(d)$. To summarize, among unit-variance functions of a Gaussian, linear functions are the most predictable across the stationary transition, and the Gaussian constraint turns the optimal linear map into a rotation, exactly the premise our proofs consume.

We now give proof details for the identifiability statements of the main text. The JEPA-specific step is a bridge from linear identifiability to a concrete matrix problem: once a JEPA representation is identified up to an orthogonal transformation, DSReg must resolve the remaining orthogonal ambiguity in the dependency Jacobian. The structural steps below are stated directly in that matrix language, so the proofs are self-contained and do not revisit the JEPA objective itself.

Proof notation. For an orthogonal matrix $U \in O(d)$, write

$$\mathcal{I}_j(U) = \{i : U_{ij} \neq 0\}$$

for the row support of column j. When the matrix is clear, we write $\mathcal{I}_j$. A column is a *singleton* if $|\mathcal{I}_j| = 1$ and *mixed* if $|\mathcal{I}_j| \geq 2$. In the main minimality proof, the set of mixed columns is

$$\mathcal{J} = \{j : |\mathcal{I}_j| \geq 2\}, \qquad \mathcal{I} = \bigcup_{j \in \mathcal{J}} \mathcal{I}_j,$$

so $\mathcal{I}$ is the set of true-latent rows that participate in at least one mixed candidate latent.

A.1 Dependency supports under the LeJEPA result

Lemma 1 (Dependency bridge)**.** *Suppose a representation is linearly identifiable with orthogonal ambiguity, so that $h(z) = Qz$ for some $Q \in O(d)$, as Theorem 1 provides for JEPA representations, and let $x = g(z)$ have dependency Jacobian $D(z) = \partial x / \partial z$ almost everywhere. Then for candidate latents $\tilde{z} = Rh(z)$ with $R \in O(d)$,*

$$\frac{\partial x}{\partial \tilde{z}} = D(z) Q^\top R^\top \qquad \textit{almost everywhere}. \tag{2}$$

Proof. We prove the identity by writing the change of variables explicitly.

Step 1: replace the almost-sure identity by an a.e. representative. The LeJEPA premise gives

$$h(z) = Qz \qquad \mu\text{-a.s.}$$

for some $Q \in O(d)$. Changing h on a μ-null set does not change any a.e. dependency support, because the support criterion only asks whether an entry is nonzero on a set of positive μ measure. We may therefore take $h(z) = Qz$ on the differentiability points used below.

Step 2: compute the candidate latent map. For a candidate DSReg rotation $R \in O(d)$,

$$\tilde{z} = Rh(z) = RQz.$$

Since R and Q are orthogonal,

$$(RQ)^{-1} = Q^\top R^\top, \qquad z = Q^\top R^\top \tilde{z}.$$

Step 3: apply the chain rule. Let $D(z) = \partial x / \partial z$. At every point where g is differentiable and the linear change of variables is defined,

$$\frac{\partial x}{\partial \tilde{z}} = \frac{\partial x}{\partial z} \frac{\partial z}{\partial \tilde{z}}.$$

The first factor is $D(z)$, and the second factor is the constant matrix $Q^\top R^\top$. Thus

$$\frac{\partial x}{\partial \tilde{z}} = D(z) Q^\top R^\top.$$

The orthogonal change of variables preserves Gaussian null sets, so this identity holds at the corresponding almost-everywhere differentiability points of the map g as well. □

Proposition 2 (Support rotations and dependency supports)**.** *Let A be a fixed matrix, and apply a Givens rotation in the (j, k) column plane with the angle chosen to zero an active entry in row i of column j. Assume A is* generic for the triple (i, j, k)*: for every row $r \neq i$ with $A_{rj} A_{rk} \neq 0$,*

$$A_{rj} A_{ik} \neq A_{ij} A_{rk} \qquad \textit{and} \qquad A_{rj} A_{ij} \neq -A_{rk} A_{ik}.$$

These are finitely many polynomial equalities, so the matrices excluded for any triple form a finite union of measure-zero algebraic sets. Then the support of the rotated matrix changes only in the two rotated columns, and exactly one of the following four cases occurs:

1. Support Reduction*: the entry (i, k) is active and the two column supports agree outside row i; the entry (i, j) is deactivated and no other entry changes;*

2. Reversible Acute Rotation*: the entry (i, k) is active and the two column supports differ in exactly one row other than i; the entry (i, j) is deactivated and exactly one previously inactive entry in the rotated columns becomes active;*

3. Irreversible Acute Rotation*: the entry (i, k) is active and the two column supports differ in at least two rows other than i; the entry (i, j) is deactivated and all previously inactive entries in the rows where the supports differ become active;*

4. Column Swap*: the entry (i, k) is inactive; the supports of columns j and k are exchanged.*

Here a Givens rotation with $\sin\theta\cos\theta \neq 0$ *is called* acute*, and an acute rotation applied to A is called* reversible *if some Givens rotation applied to the rotated matrix restores the support of A, and* irreversible *otherwise; the terminology is from Ghassami et al. (2020). In particular, except for signed Column Swaps, a nontrivial Givens rotation changes the dependency support by deleting or adding active entries in the rotated columns.*

Proof. Only the two rotated columns can change, so it is enough to analyze those columns row by row. Let b_j and b_k be columns j and k before the rotation. A Givens rotation in the (j, k) column plane gives

$$b'_j = b_j \cos\theta + b_k \sin\theta, \qquad b'_k = -b_j \sin\theta + b_k \cos\theta.$$

The angle is chosen so that the target entry (i, j) vanishes:

$$b'_{ij} = b_{ij} \cos\theta + b_{ik} \sin\theta = 0, \qquad b_{ij} \neq 0.$$

Case 1: the paired target entry is inactive. If $b_{ik} = 0$, then the equation above becomes $b_{ij} \cos\theta = 0$. Since $b_{ij} \neq 0$, we must have $\cos\theta = 0$ and hence $\sin\theta = \pm 1$. Therefore

$$b'_j = \pm b_k, \qquad b'_k = \mp b_j.$$

The two column supports are exchanged. This is the Column Swap case.

Case 2: the paired target entry is active. If $b_{ik} \neq 0$, then

$$\tan\theta = -\frac{b_{ij}}{b_{ik}},$$

so both $\sin\theta$ and $\cos\theta$ are nonzero. The target row after rotation is

$$(b'_{ij}, b'_{ik}) = (0, b'_{ik}).$$

Because a Givens rotation preserves the Euclidean norm of the row pair,

$$|b'_{ik}|^2 = |b_{ij}|^2 + |b_{ik}|^2 > 0.$$

Thus the target row loses the entry in column j and remains active in column k.

Now consider any other row $r \neq i$. The pair (b_{rj}, b_{rk}) is multiplied by the same invertible two-dimensional rotation. If both entries are zero, they remain zero. If exactly one entry is nonzero, both entries of the image are nonzero, because $\sin\theta\cos\theta \neq 0$. If both entries are nonzero, both entries of the image remain nonzero outside the excluded coefficient ratios. Indeed, an additional cancellation in column j would require

$$b_{rj} \cos\theta + b_{rk} \sin\theta = 0 \quad \Longleftrightarrow \quad \frac{b_{rj}}{b_{rk}} = \frac{b_{ij}}{b_{ik}},$$

and an additional cancellation in column k would require

$$-b_{rj} \sin\theta + b_{rk} \cos\theta = 0 \quad \Longleftrightarrow \quad \frac{b_{rj}}{b_{rk}} = -\frac{b_{ik}}{b_{ij}}.$$

Both conditions are exactly the polynomial equalities excluded by genericity for the triple (i, j, k), so neither cancellation occurs.

Outside the exceptional set, the support outcome is determined only by how the two original column supports agree away from the target row:

1. If the supports agree on every row $r \neq i$, then no missing entry is created away from the target row. The only support change is deletion of (i, j), giving Support Reduction.

2. If the supports differ on exactly one row $r \neq i$, then that row had exactly one active entry before the rotation and two active entries after it. The target deletion is balanced by one activation, giving a Reversible Acute Rotation.

3. If the supports differ on at least two rows $r \neq i$, then each such row gains the missing entry. The target deletion is accompanied by at least two activations, giving an Irreversible Acute Rotation.

Together with the inactive-paired-entry case, these are the four cases in the proposition. □

A.2 Identifiability under Structural Diversity

We first record the auxiliary genericity facts and then prove Theorem 2. The support-minimality proof works directly with the data-generating dependency Jacobian. Meanwhile, Functional no-cancellation (Assumption 2) rules out a.e. cancellations by a constant orthogonal mixing matrix, so a mixed candidate latent has the union of the source dependency footprints as its own.

Proposition 3 (Functional no-cancellation within analytic families)**.** *Let $\{F_\theta : \theta \in \Theta\}$ be a smooth observation family with $\Theta \subset \mathbb{R}^q$ open and connected. For each row and parameter let*

$$I_r(\theta) = \{\, i : D_{ri,\theta} \neq 0 \text{ on a set of positive } \mu \text{ measure} \,\}.$$

Assume there are fixed sets I_r such that $I_r(\theta) = I_r$ for all $\theta \in \Theta$. Suppose each active derivative $D_{ri,\theta} = \partial(F_\theta)_r/\partial z_i$, $i \in I_r$, is square integrable and the Gram determinant

$$G_r(\theta) = \det\big[\langle D_{ri,\theta}, D_{ri',\theta}\rangle_{L^2(\mu)}\big]_{i,i' \in I_r}$$

is real analytic in θ. If, for every row r, there exists $\theta_r \in \Theta$ with $G_r(\theta_r) \neq 0$, then Assumption 2 holds for Lebesgue-almost every $\theta \in \Theta$.

Proof of Proposition 3. Fix an observation row r. If $|I_r| \leq 1$, then any relation

$$\sum_{i \in I_r} c_i D_{ri,\theta} = 0 \quad \text{in } L^2(\mu)$$

forces all coefficients to be zero, because there is at most one active nonzero function.

Now suppose $|I_r| > 1$. Let

$$H_r(\theta) = \big[\langle D_{ri,\theta}, D_{ri',\theta}\rangle_{L^2(\mu)}\big]_{i,i' \in I_r}$$

be the row-wise Gram matrix. For any coefficient vector $c = (c_i)_{i \in I_r}$,

$$c^\top H_r(\theta) c = \left\| \sum_{i \in I_r} c_i D_{ri,\theta} \right\|^2_{L^2(\mu)}.$$

Therefore the active row functions are linearly dependent in $L^2(\mu)$ if and only if $H_r(\theta)$ is singular, equivalently

$$G_r(\theta) = \det H_r(\theta) = 0.$$

By assumption, G_r is real analytic and is not identically zero on the connected open set Θ. A nonzero real analytic function has a Lebesgue-null zero set. Thus the no-cancellation condition fails for this row only on a measure-zero subset of Θ. Taking the finite union over the observation rows gives the result. The argument is the same one used for faithfulness in Ng et al. (2023, Proposition 3). □

An empirical diagnostic for no-cancellation. The no-cancellation condition also admits a rotation-invariant empirical diagnostic. Specifically, for a fixed observation row r, the estimated derivative family is $D_{r,:}(\cdot)U$, an invertible recombination of the true row functions, so the functional row rank across anchors is invariant to the unknown rotation. The local Jacobians in Equation (4) estimate this rank by a row-wise Gram matrix; a row whose estimated rank falls below its recovered row-support size signals a spurious cancellation and flags that row for closer inspection.

Lemma 2 (Functional support of a mixed column)**.** *For a vector-valued function $v(\cdot)$, write $\mathrm{supp}_\mu(v) = \{r : v_r(z) \neq 0$ on a set of positive μ measure$\}$. Assume Functional no-cancellation (Assumption 2), let $U \in O(d)$, and define*

$$\mathcal{I}_j = \{i : U_{ij} \neq 0\}.$$

Then the a.e. support of the j-th column of $D(\cdot)U$ is

$$\mathrm{supp}_\mu\left((D(\cdot)U)_{:,j}\right) = \bigcup_{i \in \mathcal{I}_j} \mathcal{S}_i.$$

Proof. Fix a column j and an observation row r. The row entry of the mixed column is

$$(D(z)U)_{rj} = \sum_{i=1}^{d} D_{ri}(z)U_{ij} = \sum_{i \in \mathcal{I}_j} D_{ri}(z)U_{ij}.$$

We prove both inclusions.

First inclusion. If

$$r \notin \bigcup_{i \in \mathcal{I}_j} \mathcal{S}_i,$$

then $r \notin \mathcal{S}_i$ for every $i \in \mathcal{I}_j$. By definition of $\mathcal{S}_i$, each $D_{ri}(\cdot)$ is zero μ-a.e. Hence $(D(\cdot)U)_{rj}$ is zero μ-a.e., so r is not in the support of the mixed column, which proves the first inclusion.

Second inclusion. Now suppose

$$r \in \bigcup_{i \in \mathcal{I}_j} \mathcal{S}_i.$$

Let

$$K_{rj} = \{i \in \mathcal{I}_j : r \in \mathcal{S}_i\}.$$

This set is nonempty. Terms with $i \notin K_{rj}$ are zero μ-a.e., so

$$(D(z)U)_{rj} = \sum_{i \in K_{rj}} D_{ri}(z)U_{ij} \qquad \mu\text{-a.e.}$$

For every $i \in K_{rj}$, the coefficient U_{ij} is nonzero by the definition of $\mathcal{I}_j$. Thus the last display is a nontrivial constant-coefficient combination of the active row functions $\{D_{ri}(\cdot) : i \in K_{rj}\}$. This family is a subfamily of the row-r active family in Assumption 2. Functional no-cancellation says that no such nontrivial combination can be zero μ-a.e. Therefore $(D(\cdot)U)_{rj}$ is nonzero on a set of positive μ measure, so r belongs to the support of the mixed column, as required. □

Lemma 3 (Functional support minimality)**.** *Under the assumptions of Theorem 2, every solution of*

$$\min_{U \in O(d)} \|D(\cdot)U\|_{0,\mu}$$

is a signed permutation matrix.

Proof. Let $\hat{U}$ be a minimizer. We show that $\hat{U}$ cannot contain any mixed column.

Step 1: if $\hat{U}$ is not a signed permutation, it has a nonempty mixed block. For each column j, let

$$\mathcal{I}_j = \{i : \hat{U}_{ij} \neq 0\}.$$

Every column of an orthogonal matrix has unit norm, so $|\mathcal{I}_j| \geq 1$. If $|\mathcal{I}_j| = 1$, say $\mathcal{I}_j = \{r\}$, then

$$1 = \|\hat{U}_{:,j}\|_2^2 = \hat{U}_{rj}^2,$$

so $\hat{U}_{rj} = \pm 1$. Since row r also has unit norm,

$$1 = \|\hat{U}_{r,:}\|_2^2 = \hat{U}_{rj}^2 + \sum_{\ell \neq j} \hat{U}_{r\ell}^2 = 1 + \sum_{\ell \neq j} \hat{U}_{r\ell}^2.$$

Hence $\hat{U}_{r\ell} = 0$ for every $\ell \neq j$. A singleton column therefore uses its row completely.

This observation has two consequences. First, two singleton columns cannot use the same row. Second, a row used by a singleton column cannot appear in any mixed column. Indeed, if a singleton

column uses row r, then the calculation above gives $\hat{U}_{r\ell} = 0$ for every other column ℓ. Therefore $r \notin \mathcal{I}_\ell$ for every mixed column ℓ, which is precisely the second consequence.

If every column were a singleton, the d singleton columns would occupy d distinct rows, each with a single ± 1 entry. Then $\hat{U}$ would be a signed permutation matrix. Arguing by contradiction, suppose $\hat{U}$ is not a signed permutation. Then the set of mixed columns

$$\mathcal{J} = \{j : |\mathcal{I}_j| \geq 2\}$$

is nonempty. Let

$$\mathcal{I} = \bigcup_{j \in \mathcal{J}} \mathcal{I}_j.$$

The row-completion argument above also shows

$$\hat{U}_{i\ell} = 0 \quad \text{for every } i \in \mathcal{I} \text{ and every singleton column } \ell \notin \mathcal{J}. \tag{i}$$

In words, rows touched by mixed columns have support only inside mixed columns.

Step 2: the mixed rows and mixed columns form a square orthogonal block. Consider the submatrix

$$B = \hat{U}_{\mathcal{I},\mathcal{J}}.$$

By definition of $\mathcal{I}$, every nonzero entry of every mixed column $j \in \mathcal{J}$ lies in a row from $\mathcal{I}$. Restricting such a column to rows $\mathcal{I}$ therefore removes only zeros. Hence for $j, j' \in \mathcal{J}$,

$$\langle B_{:,j}, B_{:,j'} \rangle = \langle \hat{U}_{:,j}, \hat{U}_{:,j'} \rangle = \mathbf{1}\{j = j'\}.$$

The columns of B are orthonormal. Since B has $|\mathcal{J}|$ orthonormal columns in $\mathbb{R}^{|\mathcal{I}|}$, we must have

$$|\mathcal{J}| \leq |\mathcal{I}|. \tag{ii}$$

Now use (i). If $i \in \mathcal{I}$, then row i has zero entries in every singleton column, and every column is either singleton or mixed. Thus restricting row i to columns $\mathcal{J}$ removes only zeros. For $i, i' \in \mathcal{I}$,

$$\langle B_{i,:}, B_{i',:} \rangle = \langle \hat{U}_{i,:}, \hat{U}_{i',:} \rangle = \mathbf{1}\{i = i'\}.$$

The rows of B are orthonormal. Since B has $|\mathcal{I}|$ orthonormal rows in $\mathbb{R}^{|\mathcal{J}|}$, we must have

$$|\mathcal{I}| \leq |\mathcal{J}|. \tag{iii}$$

Combining (ii) and (iii) gives $|\mathcal{I}| = |\mathcal{J}|$. Therefore B is square. Since its columns are orthonormal,

$$B^\top B = I_{|\mathcal{J}|},$$

so B is invertible. This is the precise meaning of saying that the mixed block is square orthogonal.

Step 3: invertibility gives a distinct representative for each mixed column. Expand the determinant of B:

$$\det(B) = \sum_{\tau : \mathcal{J} \to \mathcal{I} \text{ bijection}} \operatorname{sgn}(\tau) \prod_{j \in \mathcal{J}} B_{\tau(j),j}.$$

Since B is invertible, $\det(B) \neq 0$. Thus at least one product in this sum is nonzero. For that bijection, call it σ, every factor $B_{\sigma(j),j}$ is nonzero. Equivalently,

$$\sigma(j) \in \mathcal{I}_j \qquad \text{for every } j \in \mathcal{J}.$$

So the bijection $\sigma : \mathcal{J} \to \mathcal{I}$ selects, for each mixed column, one true latent row that the column actually uses, and it selects distinct rows for distinct mixed columns of the block.

Step 4: construct a signed-permutation competitor. Define $\tilde{U}$ as follows. For each mixed column $j \in \mathcal{J}$, put a single nonzero entry ± 1 at row $\sigma(j)$. For each singleton column, keep the same singleton row and sign as in $\hat{U}$. The rows $\sigma(\mathcal{J}) = \mathcal{I}$ are disjoint from the singleton rows by Step 1, and the singleton rows are distinct. Therefore every row and every column of $\tilde{U}$ has exactly one nonzero entry of magnitude one. Hence $\tilde{U}$ is a signed permutation matrix.

Step 5: the signed-permutation competitor is never less sparse. For a mixed column $j \in \mathcal{J}$, Lemma 2 gives

$$\operatorname{supp}_\mu\left((D(\cdot)\hat{U})_{:,j}\right) = \bigcup_{i \in \mathcal{I}_j} \mathcal{S}_i.$$

The competitor $\tilde{U}$ makes column j pure row $\sigma(j)$, so

$$\operatorname{supp}_\mu\left((D(\cdot)\tilde{U})_{:,j}\right) = \mathcal{S}_{\sigma(j)}.$$

Because $\sigma(j) \in \mathcal{I}_j$, one set in the union is $\mathcal{S}_{\sigma(j)}$. Therefore

$$\left|\bigcup_{i\in\mathcal{I}_j} \mathcal{S}_i\right| \geq |\mathcal{S}_{\sigma(j)}| = \|(D(\cdot)\tilde{U})_{:,j}\|_{0,\mu}. \tag{iv}$$

Singleton columns are unchanged, so their support counts are equal under $\hat{U}$ and $\tilde{U}$. Summing (iv) over mixed columns and adding the singleton columns gives

$$\|D(\cdot)\hat{U}\|_{0,\mu} \geq \|D(\cdot)\tilde{U}\|_{0,\mu}. \tag{v}$$

Step 6: Structural Diversity makes the inequality strict. Suppose equality held in (v). Then equality must hold in (iv) for every mixed column:

$$\bigcup_{i\in\mathcal{I}_j} \mathcal{S}_i = \mathcal{S}_{\sigma(j)} \qquad \text{for every } j \in \mathcal{J}. \tag{vi}$$

Since every set in a union is contained in the union, (vi) implies

$$\mathcal{S}_k \subseteq \mathcal{S}_{\sigma(j)} \qquad \text{for every } k \in \mathcal{I}_j. \tag{vii}$$

Choose an inclusion-minimal footprint among $\{\mathcal{S}_i : i \in \mathcal{I}\}$, and call it $\mathcal{S}_{i_*}$. Inclusion-minimal means that there is no $i \in \mathcal{I}$ such that

$$\mathcal{S}_i \subsetneq \mathcal{S}_{i_*}. \tag{viii}$$

The bijection $\sigma : \mathcal{J} \to \mathcal{I}$ is onto, so there is a mixed column $j_* \in \mathcal{J}$ with

$$\sigma(j_*) = i_*.$$

Because j_* is mixed, $\mathcal{I}_{j_*}$ has at least two elements. Also $i_* = \sigma(j_*) \in \mathcal{I}_{j_*}$. Therefore we can choose

$$k \in \mathcal{I}_{j_*}, \qquad k \neq i_*.$$

Applying (vii) to column j_* gives

$$\mathcal{S}_k \subseteq \mathcal{S}_{\sigma(j_*)} = \mathcal{S}_{i_*}. \tag{ix}$$

By (ix) we have $\mathcal{S}_k \subseteq \mathcal{S}_{i_*}$, so exactly one of two cases holds. If $\mathcal{S}_k = \mathcal{S}_{i_*}$ (in particular if $\mathcal{S}_{i_*} = \emptyset$, which forces $\mathcal{S}_k = \emptyset$), then two different latents have the same footprint, contradicting Structural Diversity. If $\mathcal{S}_k \subsetneq \mathcal{S}_{i_*}$, then (viii) is violated, contradicting the inclusion-minimal choice of $\mathcal{S}_{i_*}$. Hence equality in (v) is impossible in either one of the two cases above.

Thus

$$\|D(\cdot)\hat{U}\|_{0,\mu} > \|D(\cdot)\tilde{U}\|_{0,\mu},$$

which contradicts the optimality of $\hat{U}$. Therefore a minimizer cannot have a mixed column. Every column is singleton, and orthogonality then forces the matrix to be a signed permutation. □

Remark 1 (Global minimum of the support criterion)**.** *Steps 1–5 above do not use Structural Diversity: for every $U \in O(d)$ they construct a signed-permutation competitor and yield*

$$\|D(\cdot)U\|_{0,\mu} \;\geq\; \sum_{i=1}^{d} |\mathcal{S}_i|,$$

with equality attained by every signed permutation. Hence $\sum_{i=1}^{d} |\mathcal{S}_i|$ is the global minimum of the support criterion and signed permutations are always among its minimizers; Structural Diversity, invoked only in Step 6, is exactly what removes the non-permutation minimizers. This is the formal basis for the sharp-boundary discussion after Proposition 1: when $\mathcal{S}_i = \mathcal{S}_j$ for some $i \neq j$, a Givens rotation in that plane produces a non-permutation matrix that still attains $\sum_i |\mathcal{S}_i|$.

Theorem 2 (Component-wise identifiability without reconstruction)**.** *In the setting of Lemma 1, assume Structural Diversity (Assumption 1) and Functional no-cancellation (Assumption 2). Then for every minimizer R of the support-sparsity criterion* (3)*, the matrix RQ is a signed permutation: there exist a permutation π of $[d]$ and signs $s_1, \ldots, s_d \in \{\pm 1\}$ such that the candidate latents $\tilde{z} = Rh(z)$ satisfy $\tilde{z}_i = s_i z_{\pi(i)}$ almost surely for every $i \in [d]$.*

Proof. The proof reduces the DSReg criterion over R to the matrix problem solved by Lemma 3.

Step 1: express the candidate Jacobian through one orthogonal matrix. By the LeJEPA premise, $h(z) = Qz$ μ-a.e. for some $Q \in O(d)$. As in Lemma 1, changing h on a null set does not change the a.e. support, so we use the representative $h(z) = Qz$. For candidate latents $\tilde{z} = Rh(z)$, the dependency bridge gives

$$\frac{\partial x}{\partial \tilde{z}} = D(z)Q^\top R^\top.$$

Since $Q, R \in O(d)$, the following matrix is again orthogonal:

$$U = Q^\top R^\top.$$

Step 2: show that optimizing over R is the same as optimizing over U. The map

$$R \mapsto U = Q^\top R^\top$$

is a bijection from $O(d)$ to $O(d)$. Indeed, for any $U \in O(d)$, choosing

$$R = (QU)^\top$$

gives

$$Q^\top R^\top = Q^\top (QU) = U.$$

Therefore

$$\min_{R \in O(d)} \left\| \frac{\partial x}{\partial \tilde{z}} \right\|_{0,\mu} = \min_{U \in O(d)} \|D(\cdot)U\|_{0,\mu}.$$

A rotation R minimizes the left-hand side exactly when its induced U minimizes the right-hand side.

Step 3: apply support minimality. Lemma 3 says every minimizer of the right-hand side is a signed permutation. Write such a minimizer as

$$U = SP,$$

where S is diagonal with entries in $\{\pm 1\}$ and P is a permutation matrix. Since $U = Q^\top R^\top$, we have

$$RQ = U^\top = P^\top S.$$

Thus RQ is a signed permutation.

Step 4: translate the matrix statement into latent recovery. The candidate latents are

$$\tilde{z} = Rh(z) = RQz = P^\top Sz.$$

Hence each component of $\tilde{z}$ is one component of z, up to a sign and relabeling:

$$\tilde{z}_i = s_i z_{\pi(i)} \qquad \mu\text{-a.e.}$$

This is the claimed signed-permutation identifiability of Definition 1. □

A.3 Comparison of structural conditions

Proposition 1 (Structural Diversity is strictly weaker)**.** *Stated on the dependency footprints $\mathcal{S}_1, \ldots, \mathcal{S}_d$, each of the following conditions implies Structural Diversity (Assumption 1), and none of the converse implications holds:*

1. Structural Sparsity, intersection form *(Zheng et al., 2022): for every $k \in [d]$ there is a nonempty set $\mathcal{C}_k$ of observed variables with $\bigcap_{r \in \mathcal{C}_k} \{\, i \in [d] : r \in \mathcal{S}_i \,\} = \{k\}$;*
2. Structural Sparsity, overlap-rank form *(Zheng et al., 2022): for every $\mathcal{C} \subseteq [d]$ with $|\mathcal{C}| \geq 2$ and every $k \in \mathcal{C}$, $\left|\bigcup_{j \in \mathcal{C}} \mathcal{S}_j\right| - \operatorname{rank}(M^{\mathcal{C}}) > |\mathcal{S}_k|$, where $M^{\mathcal{C}}$ is the binary matrix with $M^{\mathcal{C}}_{rj} = 1$ exactly when $j \in \mathcal{C}$ and r lies in $\mathcal{S}_j$ and in $\mathcal{S}_{j'}$ for some other $j' \in \mathcal{C}$;*
3. Non-Inclusion*, a common weakening of both forms above: $\mathcal{S}_i \not\subseteq \mathcal{S}_j$ for all $i \neq j$;*
4. Structural Variability *(Ng et al., 2023): $|\mathcal{S}_i \triangle \mathcal{S}_j| \geq 2$ for all $i \neq j$.*

Proof. The proposition compares conditions only through column-support sets, so we can reason directly at the level of the footprints $\mathcal{S}_1, \ldots, \mathcal{S}_d$ throughout.

Step 1: prior structural conditions imply Structural Diversity. Ng et al. (2023, Theorem 2) establish that the overlap-rank form of Structural Sparsity implies Non-Inclusion (termed Column Subset there), and Non-Inclusion implies Structural Variability; Ng et al. (2023, Theorem 3) establish the same implication chain starting from the intersection form of Structural Sparsity. Hence

$$\text{Structural Sparsity (either form)} \implies \text{Non-Inclusion} \implies \text{Structural Variability}.$$

It remains to check that Structural Variability implies Structural Diversity.

Structural Variability says

$$|\mathcal{S}_i \triangle \mathcal{S}_j| \geq 2 \qquad \text{for all } i \neq j.$$

If two footprints were equal, $\mathcal{S}_i = \mathcal{S}_j$, then their symmetric difference would be empty:

$$\mathcal{S}_i \triangle \mathcal{S}_j = \emptyset, \qquad |\mathcal{S}_i \triangle \mathcal{S}_j| = 0,$$

contradicting Structural Variability. Hence all footprints are distinct: Structural Diversity holds.

Step 2: Structural Diversity does not imply the prior conditions. It suffices to give one support pattern satisfying Structural Diversity but violating the stronger conditions. Let $d = 2$, $p = 2$, and

$$\mathcal{S}_1 = \{1\}, \qquad \mathcal{S}_2 = \{1, 2\}.$$

The footprints are distinct, so Structural Diversity holds. They are realized, for example, by

$$F_1(z) = z_1 + \sin z_2, \qquad F_2(z) = z_2 + \sin z_2.$$

Indeed,

$$\frac{\partial F_1}{\partial z_1} = 1, \quad \frac{\partial F_1}{\partial z_2} = \cos z_2, \quad \frac{\partial F_2}{\partial z_1} = 0, \quad \frac{\partial F_2}{\partial z_2} = 1 + \cos z_2,$$

so the first, second, and fourth derivatives are nonzero on sets of positive measure while the third vanishes identically, giving exactly the two footprints above. Moreover, the active partial derivatives in observation row 1, namely $\partial F_1/\partial z_1 = 1$ and $\partial F_1/\partial z_2 = \cos z_2$, are linearly independent in $L^2(\mu)$, so this witness also satisfies Functional no-cancellation (Assumption 2) and Theorem 2 therefore applies to it without modification.

However,

$$\mathcal{S}_1 \subsetneq \mathcal{S}_2,$$

so Non-Inclusion fails. Also

$$\mathcal{S}_1 \triangle \mathcal{S}_2 = \{2\}, \qquad |\mathcal{S}_1 \triangle \mathcal{S}_2| = 1,$$

so Structural Variability fails. Since both structural-sparsity forms imply Non-Inclusion through the cited implication chains, they fail as well. Hence Structural Diversity is strictly weaker. □

A.4 Robustness to Jacobian perturbation

This section proves the stability results of Section 3.3. The first lemma removes general invertible-linear ambiguity at the population level; the later lemmas bound the constants the theorem uses.

Lemma 4 (Whitening reduction)**.** *Let $z \sim \mathcal{N}(0, I_d)$ and $h(z) = Az$ for an invertible $A \in \mathbb{R}^{d\times d}$. Then $W = \mathrm{Cov}(h)^{-1/2} = (AA^\top)^{-1/2}$ satisfies $Wh(z) = \tilde{Q}z$ with $\tilde{Q} = (AA^\top)^{-1/2}A \in O(d)$.*

Proof. The covariance of the mixed representation is

$$\mathrm{Cov}(h) = A\,\mathrm{Cov}(z)A^\top = AA^\top.$$

This matrix is symmetric positive definite, so its inverse square root exists and is symmetric. For $\tilde{Q} = (AA^\top)^{-1/2}A$,

$$\tilde{Q}\tilde{Q}^\top = (AA^\top)^{-1/2}AA^\top(AA^\top)^{-1/2} = I_d,$$

and likewise

$$\tilde{Q}^\top\tilde{Q} = A^\top(AA^\top)^{-1}A = I_d,$$

so $\tilde{Q}$ is orthogonal. Finally,

$$Wh(z) = (AA^\top)^{-1/2}Az = \tilde{Q}z.$$

□

Constants of the support pattern. Since μ is a probability measure, the bounded-rows condition of Theorem 3 places every D_{ri} in $L^2(\mu)$, and under Assumption 2 the active row functions $\{D_{ri}\}_{i\in I_r}$ are linearly independent as elements of $L^2(\mu)$, so each row Gram matrix $G_r = [\langle D_{ri}, D_{ri'}\rangle_{L^2(\mu)}]_{i,i'\in I_r}$ is positive definite. Write $\sigma_r = \lambda_{\min}(G_r)^{1/2} > 0$ and $\sigma = \min\{\sigma_r : I_r \neq \emptyset\}$; under Structural Diversity with $d \geq 2$ at most one footprint is empty, so some I_r is nonempty and σ is well defined. For $U \in O(d)$ let $v_{rj}(U) = \|U_{I_r,j}\|_2$ (zero when $I_r = \emptyset$); each v_{rj} is continuous on $O(d)$, and by Lemma 2 the population count satisfies $N(U) = \|D(z)U\|_{0,\mu} = |\{(r,j) : v_{rj}(U) > 0\}|$. Let $SP(d)$ denote the signed permutation matrices, and for $\delta > 0$ let $K_\delta = \{U \in O(d) : \min_{P\in SP(d)} \|U - P\|_F \geq \delta\}$. Define $\rho^*(U)$ as the (s^*+1)-th largest of the pd values $v_{rj}(U)$, where $s^* = \sum_i |\mathcal{S}_i|$, and $\rho^*(\delta) = \min_{U\in K_\delta} \rho^*(U)$, with $\rho^*(\delta) = +\infty$ when $K_\delta = \emptyset$ (making Theorem 3 trivial).

Lemma 5 (Quantitative activity)**.** *Let Assumption 2 and the bounded-rows condition of Theorem 3 hold, let r be a row with $I_r \neq \emptyset$, let $c \in \mathbb{R}^{I_r}$ with $\|c\|_2 \leq 1$, and set $f_c = \sum_{i\in I_r} c_i D_{ri}$. Then for every $0 \leq t < \sigma_r\|c\|_2$,*

$$\mu\big(|f_c| > t\big) \;\geq\; \frac{\sigma_r^2\|c\|_2^2 - t^2}{M^2} \;>\; 0.$$

Proof. Step 1: pointwise upper bound. By Cauchy–Schwarz, for μ-a.e. z,

$$|f_c(z)| \leq \|D_{r,:}(z)\|_2\|c\|_2 \leq M.$$

Step 2: second-moment lower bound. The Gram bound gives

$$\|f_c\|^2_{L^2(\mu)} = c^\top G_r c \geq \sigma_r^2\|c\|_2^2.$$

Step 3: split the integral. Combining the two bounds,

$$\begin{aligned}\sigma_r^2\|c\|_2^2 &\leq \int |f_c|^2\, d\mu \\ &\leq t^2\,\mu(|f_c| \leq t) + M^2\,\mu(|f_c| > t) \\ &\leq t^2 + M^2\,\mu(|f_c| > t).\end{aligned}$$

Rearranging the resulting inequality proves the claimed lower bound. □

Lemma 6 (Uniform pattern margin)**.** *Let Assumptions 1 and 2 hold with $d \geq 2$, and let $\delta > 0$ be such that $K_\delta \neq \emptyset$. Then $\rho^*(\delta) > 0$.*

Proof. Step 1: reduce to pointwise positivity. Each v_{rj} is continuous on $O(d)$, so the order statistic $\rho^*(U)$ (the (s^*+1)-th largest of finitely many continuous functions) is continuous, and K_δ is a closed subset of the compact group $O(d)$, hence compact. A function that is positive and continuous on a compact set attains a strictly positive minimum over that whole set, so it suffices to show $\rho^*(U) > 0$ for every $U \in K_\delta$.

Step 2: the population count exceeds the minimum. Fix $U \in K_\delta$; it is not a signed permutation. By Remark 1, $\min_{U'} N(U') = s^*$ and every signed permutation attains it, and by Lemma 3 any U with $N(U) = s^*$ is a signed permutation; both statements hold under Assumptions 1 and 2 alone. Since the count is integer-valued,

$$N(U) \geq s^* + 1.$$

Step 3: convert the count into a margin. By the union-support identity,

$$N(U) = |\{(r,j) : v_{rj}(U) > 0\}|,$$

at least $s^* + 1$ of the values $v_{rj}(U)$ are strictly positive, i.e. $\rho^*(U) > 0$. Under Structural Diversity with $d \geq 2$ not every footprint equals $[p]$, so $s^* + 1 \leq pd$ and $\rho^*(U)$ is well defined. By Step 1, $\rho^*(\delta) > 0$ as claimed. □

Remark 2 (The constants on the witness)**.** *It is worth noting that the constants are not vacuous. For the two-dimensional witness of Proposition 1 ($F_1 = z_1 + \sin z_2$, $F_2 = z_2 + \sin z_2$), a direct computation gives $\sigma \approx 0.374$ and $\rho^*(0.4) \approx 0.28$, so the recovery condition $\tau + \varepsilon < \sigma\rho^*(\delta)$ of Theorem 3 is satisfiable at nontrivial noise levels on an example that meets every assumption.*

Theorem 3 (Approximate recovery under Jacobian perturbation)**.** *Let the assumptions of Theorem 2 hold with $d \geq 2$, together with*

$$\operatorname{ess\,sup}_z \max_r \|D_{r,:}(z)\|_2 \leq M < \infty.$$

Let $\widehat{D}(z) = D(z) + E(z)$ with E measurable and $\operatorname{ess\,sup}_z \max_r \|E_{r,:}(z)\|_2 \leq \varepsilon$, and for a threshold $\tau > \varepsilon$ define

$$\widehat{N}_\tau(R) = \Big|\big\{(r,j) : \mu\big(|(\widehat{D}(z)\, Q^\top R^\top)_{rj}| > \tau\big) > 0\big\}\Big|, \qquad R \in O(d).$$

If $\delta > 0$ satisfies $\tau + \varepsilon < \sigma\, \rho^*(\delta)$, then minimizers of $\widehat{N}_\tau$ over $O(d)$ exist and every minimizer $\widehat{R}$ satisfies $\min_P \|\widehat{R}Q - P\|_F < \delta$ over signed permutations P.

Proof. Step 1: change of variables. Throughout write $U = (RQ)^\top$. By Lemma 1,

$$\frac{\partial x}{\partial \tilde{z}} = D(z)U.$$

The map $R \mapsto U$ is a bijection of $O(d)$, and

$$\min_P \|RQ - P\|_F = \min_P \|U - P\|_F$$

because transposition is an isometry and $SP(d)$ is closed under transposition. Moreover, $\widehat{N}_\tau$ is integer-valued with finitely many attainable values, so its minimum over $O(d)$ is attained.

Step 2: uniform bound on the perturbation. For any fixed U and μ-a.e. z, Cauchy–Schwarz with unit columns gives

$$|(E(z)U)_{rj}| \leq \|E_{r,:}(z)\|_2 \leq \varepsilon \qquad \text{for all } (r,j).$$

The exceptional null set may depend on U, and each step below fixes one U, so the U-dependence of the null set costs nothing.

Step 3: value at signed permutations. Choose $R_0 = P_0 Q^\top$ for any $P_0 \in SP(d)$, so $U_0 = (R_0 Q)^\top \in SP(d)$. If entry (r,j) is counted by $\widehat{N}_\tau(R_0)$, then

$$|(\widehat{D}(z)U_0)_{rj}| > \tau$$

on a set of positive measure. Since $|(E(z)U_0)_{rj}| \leq \varepsilon$ a.e., on that set also

$$|(D(z)U_0)_{rj}| > \tau - \varepsilon > 0,$$

hence (r,j) is active for $D(z)U_0$. Therefore

$$\widehat{N}_\tau(R_0) \leq N(U_0) = s^*,$$

using Remark 1 for the last equality.

Step 4: value far from signed permutations. Let R satisfy $\min_P \|RQ - P\|_F \geq \delta$, so $U = (RQ)^\top \in K_\delta$. By the definition of $\rho^*(\delta)$, and Lemma 6 for its positivity, at least $s^* + 1$ entries (r,j) have

$$v_{rj}(U) \geq \rho^*(\delta).$$

Fix such an entry and let $c = U_{I_r,j}$. The inactive terms $\sum_{i \notin I_r} U_{ij} D_{ri}(z)$ vanish μ-a.e., so for μ-a.e. z,

$$(D(z)U)_{rj} = f_c(z),$$

with $\|c\|_2 = v_{rj}(U) \geq \rho^*(\delta)$ and $\|c\|_2 \leq 1$ as a sub-vector of a unit column. Applying Lemma 5 with $t = \tau + \varepsilon < \sigma\rho^*(\delta) \leq \sigma_r \|c\|_2$ gives

$$\mu(|(D(z)U)_{rj}| > \tau + \varepsilon) > 0.$$

On that set,

$$|(\widehat{D}(z)U)_{rj}| \geq |(D(z)U)_{rj}| - \varepsilon > \tau,$$

so (r,j) is counted by $\widehat{N}_\tau(R)$. Hence $\widehat{N}_\tau(R) \geq s^* + 1$ holds for every rotation R in this set.

Step 5: combine. For every such R,

$$\widehat{N}_\tau(R) \geq s^* + 1 > s^* \geq \widehat{N}_\tau(R_0),$$

so no such R can minimize $\widehat{N}_\tau$. □

B ADDITIONAL DISCUSSIONS

Two literatures meet for our goal. Methods that recover individual latents anchor them to observations through a decoder or a likelihood, through auxiliary variables or interventions, or through an objective built around the true positive-pair conditional. In contrast, methods that dispense with all of these identify the latent state at most up to an orthogonal transformation, leaving the individual latents mixed. The three groups below trace each side and the premise that connects them.

Nonlinear ICA and self-supervised identifiability. Identifiability has been studied most extensively in nonlinear ICA, where latent sources cannot be recovered from i.i.d. observations alone. One family of results restores identifiability through auxiliary structure: nonstationary segments (Hyvarinen & Morioka, 2016), temporal dependence (Hyvärinen & Morioka, 2017; Klindt et al., 2020), conditioning on a general auxiliary variable (Hyvarinen et al., 2019), and conditional priors in deep generative models (Khemakhem et al., 2020). A second family analyzes self-supervised objectives directly. Contrastive learning provably inverts the data-generating process when the positive-pair conditional matches the contrastive similarity (Zimmermann et al., 2021), and augmentation-based learning block-identifies the invariant content variables (Von Kügelgen et al., 2021). These contrastive results separate components only when the assumed positive-pair conditional breaks rotational symmetry. With Gaussian latents the distribution carries no component-wise signal, and separation must come from outside both the conditional and the distribution. A third family keeps i.i.d. Gaussian sources and instead restricts the mixing to structured function classes such as conformal maps (Buchholz et al., 2022) or volume-preserving, orthogonal-column mixings (Gresele et al., 2021). Thus, in every case identifiability rests on structure from outside the representation objective, whether an auxiliary variable, a latent conditional built into the likelihood or the loss, or a restriction on the mixing function. Prediction in representation space alone, the setting of the present paper, supplies none of the three, and that is the gap the present theory sets out to fill.

Identifiability from sparse structure. A separate line obtains identifiability from sparse latent-to-observation structure. Structural sparsity of the mixing Jacobian yields permutation identifiability in nonlinear ICA (Zheng et al., 2022), later extended to undercomplete, partially sparse, and grouped settings (Zheng & Zhang, 2023); in the linear Gaussian case, sparsity replaces non-Gaussianity (Zhang et al., 2009; Palsson et al., 2014; Abrahamsen & Rigollet, 2018) under the Structural Variability condition (Ng et al., 2023), rooted in support-equivalence characterizations (Ghassami et al., 2020). Sparse mechanisms (Lachapelle et al., 2022), sparse decoders (Moran et al., 2022), and additive decoders (Lachapelle et al., 2023) carry the principle to latent dynamics and deep generative models. Each of these anchors recovery through a decoder or an observation likelihood. DSReg needs neither: Structural Diversity plus a faithfulness/no-cancellation condition yields signed-permutation identifiability with no decoder and no likelihood, and the support-pattern requirement itself is strictly weaker than the structural conditions above (Proposition 1). Besides, the disentanglement literature pursues individual latent recovery with regularized generative models (Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018; Peebles et al., 2020) and quantitative evaluation protocols (Eastwood & Williams, 2018). The impossibility result of Locatello et al. (2019) shows that some signal beyond the marginal distribution is necessary; here that signal is the dependency structure itself rather than any further generative assumption about how the observations were produced.

JEPA world models and the linear premise. DSReg takes its linear-identifiability premise from the JEPA line of world-model learning. I-JEPA established prediction in representation space as a scalable alternative to pixel reconstruction (Assran et al., 2023), LeJEPA grounded it in an isotropic-Gaussian embedding objective, SIGReg, that is provably optimal for downstream prediction (Balestriero & LeCun, 2025), and JEPA-style world models support planning and control directly in latent space (Zhou et al., 2025; Maes et al., 2026). Klindt et al. (2026) proved that in a Gaussian latent world this recipe identifies the latent state up to an orthogonal transformation. Within that world, the orthogonal class is where identifiability without reconstruction previously stopped. The state is recovered while the individual latents stay mixed by an unknown rotation. DSReg starts from the theorem and resolves the rotation, with dependency sparsity selecting the individual latents and with no reconstruction or likelihood entering at any point of the procedure.

Sparse codes in JEPAs. A parallel JEPA line pursues sparsity in the code itself. Rectified LpJEPA replaces the isotropic Gaussian target with a rectified generalized Gaussian, so that representations become sparse and non-negative, with most coordinates zero for a given sample (Kuang et al., 2026).

The two designs answer the same rotational symmetry in complementary ways, and they sparsify different objects. Changing the target law breaks the symmetry distributionally and sparsifies the code; DSReg keeps the Gaussian target and breaks the symmetry structurally, sparsifying the support of the dependency Jacobian $\partial x/\partial \tilde{z}$, and a dense code can carry a very sparse dependency structure. Sparsity enters DSReg only through what it selects for, never as an assumption about the world: the guarantee asks only that the true footprints be pairwise distinct, which allows them to be almost fully dense in the observed variables while remaining pairwise distinguishable.

Independent mechanism analysis. Independent mechanism analysis (IMA) (Gresele et al., 2021) also constrains a latent representation through the Jacobian of the observation map. IMA asks the columns of the observation Jacobian $\partial x/\partial z$ to be orthogonal, so that its Gram matrix is diagonal, whereas DSReg asks the columns to have distinct functional supports (Assumption 1) that do not cancel (Assumption 2). Neither condition implies the other, since orthogonal columns may share every row and columns with distinct supports need not be orthogonal. Under the premise $h = Qz$, IMA also suggests a post hoc baseline, jointly diagonalizing the estimated Gram matrices $B_a^\top B_a$ across anchors, which pins down the rotation when the diagonal profiles satisfy the uniqueness condition of joint diagonalization, namely that no two of them coincide across the anchors. The distinction between the two approaches is therefore support structure against orthogonality, under a shared linear identifiability premise and a shared estimate of local Jacobians that trains no decoder. The same premise also admits regularizers built on IMA rather than on supports. The IMA contrast of Gresele et al. (2021) measures how far the Jacobian columns are from orthogonal and is differentiable, so it could take the place of Equation (4) on the same local Jacobians. A term penalizing both non-orthogonality and shared support would let the two conditions cover for each other where either holds only approximately. Which of these recovers individual latents, and under what condition on the observation map, is an open question, which we leave to future work, along with the joint diagonalization baseline sketched earlier in this paragraph.

Estimation in practice. The identifiability statements use exact a.e. dependency supports, with Theorem 3 covering their thresholded perturbation. Equation (4) replaces the exact support count by local Jacobian estimates and an anchor-averaged ℓ_1 relaxation; the surrogate and noise experiments of Appendix C.3 (Figures 8 and 9) evaluate how faithfully this practical objective tracks its population target.

C Additional experiments

This appendix expands the empirical part of the paper. Section C.1 fixes the metrics and the protocol shared by every benchmark, and Section C.2 then opens the experiments with the study readers of Section 3.3 most often ask for, scaling DSReg far beyond the paper's benchmarks, to $d = 8192$ on a single GPU with samples to 10^6 and observed dimensions to 2^{17}. Section C.3 verifies the regime map and stress-tests the criterion from every side, Section C.4 specifies the sparse-use environments and probes, the next three sections cover the learned visual encoders and the two external renderers, and Section C.8 closes by asking whether any latent-only rotation, or any baseline reading the same signal, could have matched the dependency criterion from the same inputs.

C.1 Metrics and protocol

Every benchmark in the paper shares one measurement vocabulary and one experimental protocol. This section fixes both once, so the sections that follow only state what changes.

Research questions. Every experiment in the paper asks one of three questions: is the latent state present in the representation at all, has it been resolved into individual variables, and can we act through those variables for downstream tasks? One metric family answers each, defined below. Keeping the three separate lets the experiments show that DSReg changes the answers to the second and third while leaving the first answer exactly where LeJEPA had already put it.

Recovery metrics. Let $Z, \widehat{Z} \in \mathbb{R}^{n \times d}$ collect the ground-truth and the estimated latents over n evaluation samples, write $Z_{:,i}$ for a column, and let $\rho(u, v)$ denote the Pearson correlation. The first

question is answered by the dense readout

$$R^2(h \to z) \;=\; 1 - \frac{\min_{W,b} \left\| Z - (\widehat{Z}W + \mathbf{1}b^\top) \right\|_F^2}{\left\| Z - \mathbf{1}\bar{z}^\top \right\|_F^2}, \tag{5}$$

the coefficient of determination of the best linear map from the whole estimated vector onto all ground-truth latents. Any invertible A acting as $\widehat{Z} \mapsto \widehat{Z}A$ leaves it unchanged. A high value therefore settles the first question and is silent on the second.

The second question is answered by the latent mean correlation coefficient, which scores the best one-to-one matching of estimated to true coordinates,

$$\mathrm{MCC} \;=\; \max_{\pi \in S_d} \frac{1}{d} \sum_{i=1}^{d} \left| \rho\big(Z_{:,i}, \widehat{Z}_{:,\pi(i)}\big) \right|, \tag{6}$$

maximized over permutations by the Hungarian algorithm for $d \leq 128$ and by a greedy assignment above it. Besides being sensitive to mixing, which (5) is not, this quantity reaches one precisely when the representation recovers the individual world latents in the sense of Definition 1, which is why we report it as the main individual-latent recovery score on every benchmark below.

Disentanglement scores on the external renderers. The renderer benchmarks also report the standard scores, so that the comparison with β-VAE and β-TCVAE is made in their own vocabulary. Write $v_1, \ldots, v_K$ for the ground-truth factors and $c_1, \ldots, c_d$ for the estimated coordinates. DCI (Eastwood & Williams, 2018) fits gradient-boosted trees predicting each factor from all coordinates and collects the absolute feature importances into $R \in \mathbb{R}^{d \times K}$. It then measures how far each coordinate concentrates on a single factor,

$$\mathrm{DCI} = \sum_{i=1}^{d} \rho_i \Big(1 - H_K\big(P_{i,:}\big)\Big), \qquad P_{ij} = \frac{R_{ij}}{\sum_k R_{ik}}, \qquad \rho_i = \frac{\sum_j R_{ij}}{\sum_{i',j} R_{i'j}}, \tag{7}$$

with H_K the entropy in base K. The informativeness column of Table 4 is the held-out R^2 of those same regressions. MIG (Chen et al., 2018) discretizes coordinates and factors into twenty uniform-width bins, then averages the normalized gap between the two most informative coordinates for each factor. SAP (Kumar et al., 2017) replaces mutual information by the squared correlation of the univariate fit,

$$\mathrm{MIG} = \frac{1}{K} \sum_{k=1}^{K} \frac{I(v_k; c_{(1)}) - I(v_k; c_{(2)})}{H(v_k)}, \qquad \mathrm{SAP} = \frac{1}{K} \sum_{k=1}^{K} \Big(\rho^2\big(v_k, c_{(1)}\big) - \rho^2\big(v_k, c_{(2)}\big) \Big), \tag{8}$$

where $c_{(1)}$ and $c_{(2)}$ are the two coordinates ranked highest for v_k by the score each metric uses.

Sparse-use probes. The third question has no closed form. What matters is whether a downstream module restricted to a few coordinates can do its job, so each probe is a task with its own success criterion. The four are visual editing, sparse model predictive control, five-step rollout prediction, and the detection of physically implausible next states. Section C.4 gives the environments and the exact construction of each, and the readout probe that recurs across benchmarks is described next.

Design principle. One principle governs every experiment. Ground-truth latents define benchmarks and evaluation scores, never the method. The dependency rotation is always fit from estimated latents and observations, or a fixed observation preprocessing chosen before fitting; labels enter only through benchmark construction, small-label evaluation probes, and metrics. The benchmarks differ only in where the representation comes from: exact orbits $h = Qz$ isolate the rotation step under known structure, encoders trained from pixels test whether the gain survives learning, and external renderers remove the observation map from our control. Each step takes one more piece of our own design out of the loop, and the gain survives each removal, as the sections below elaborate.

Few-shot single-latent readout. This recurring probe appears on several benchmarks and deserves a procedural description. For each ground-truth factor, a handful of labeled examples select the single estimated coordinate that correlates best with the factor and fit a scalar affine map from that coordinate alone; the probe is forbidden to mix coordinates, so its score is the price of naming the variables. On a representation whose coordinates are individual factors, a few labels suffice and the readout approaches the dense ceiling; on a mixture no scalar map works at any label budget, which is why this probe separates the two representations more sharply than any dense score.

Optimization. The dependency step regresses local Jacobians over nearest neighbors (ridge 10^{-3}; $k = 32$ on the small estimated-Jacobian benchmarks, raised where the latent dimension requires it, since the neighborhood must exceed d) and optimizes the rotation by Adam on the matrix-exponential parameterization of $O(N)$; at larger dimension the exponential gives way to the Cayley chart of Appendix C.2. Anchor counts, step budgets, and restarts grow with the dimension of the benchmark. These choices carry no load. Figure 8 varies the sparsity surrogate and the anchor budget directly and finds recovery insensitive to both, so we report defaults rather than tuned values.

C.2 Scaling the dependency criterion

Section 3.3 claims that nothing in DSReg is bound to the small worlds of the benchmarks. This section backs the claim: the factorized algebra first, then one axis at a time.

The Jacobians are never materialized. The DSReg criterion of Section 3.3 averages the rotated local Jacobians over m anchors, and written as Equation (4) it appears to require m matrices of size $p \times d$, which is the binding cost once the observed dimension p or the latent dimension d grows. That reading is pessimistic. The ridge estimate at anchor a is built from the k neighbors of h_a, and if $\Delta H_a \in \mathbb{R}^{k\times d}$ and $\Delta X_a \in \mathbb{R}^{k\times p}$ stack the neighbor differences in representation and observation space, then

$$\begin{aligned} B_a &= \Delta X_a^\top \Delta H_a \left(\Delta H_a^\top \Delta H_a + \lambda I\right)^{-1} = \Delta X_a^\top P_a, \\ P_a &= \Delta H_a \left(\Delta H_a^\top \Delta H_a + \lambda I\right)^{-1} \in \mathbb{R}^{k\times d}. \end{aligned} \tag{9}$$

Two consequences follow immediately. First, B_a factors through the k dimensional neighborhood, so $\operatorname{rank}(B_a) \le \min(k, d)$, and since the estimator needs $k > d$ the rank is at most d: whenever p exceeds the latent dimension, the $p \times d$ array encodes an object with far fewer degrees of freedom than its size suggests. Second, B_a is determined by the raw observations, which are already in memory as the dataset, together with the neighbor indices and the factor P_a. Storing $\{(a, \mathcal{N}_a, P_a)\}_{a=1}^m$, which is Algorithm 2(a), costs $O(mkd)$ and replaces the $O(mpd)$ of the materialized form, a reduction by p/k, which is close to two orders of magnitude at the largest observed dimension we run.

The exact objective streams in two nested blocks. Nothing about the factorization is an approximation, and the criterion can be evaluated from it exactly, as Algorithm 2(c) does. The only requirement is to associate the products in the right order: forming $W_a = P_a R^\top$ first, at cost $O(kd^2)$, and then $\Delta X_a^\top W_a$, at cost $O(kpd)$, avoids ever building B_a, whereas the natural left-to-right order pays $O(pd^2)$ per anchor and allocates the $p \times d$ result. The subgradient comes out of the same quantities. Writing $M_a = B_a R^\top$ for the rotated Jacobian,

$$\nabla_R \mathcal{L}_{\mathrm{DSReg}}(R) = \frac{1}{m}\sum_{a=1}^m \operatorname{sign}(M_a)^\top B_a = \frac{1}{m}\sum_{a=1}^m \Big(\operatorname{sign}(M_a)^\top \Delta X_a^\top\Big) P_a, \tag{10}$$

where the parenthesized factor is $d \times k$ and the product with P_a is $d \times d$, so the gradient is assembled at the same cost as the value and again never touches a $p \times d$ array. In the implementation the anchors are traversed in chunks and, inside each chunk, the observed coordinates in blocks, so peak memory is set by the two block sizes rather than by m or p, and both can be tuned freely to the device at hand without changing the value that the pass computes in exact arithmetic.

An unbiased estimator makes the cost independent of p and m. Streaming keeps memory bounded but still reads every anchor and every observed coordinate at each step. Because the criterion is an average, it also admits an unbiased stochastic estimate, which is Algorithm 2(d). Normalizing Equation (4) by the number of entries, the objective is the expectation

$$\mathcal{L}_{\mathrm{DSReg}}(R) \propto \mathbb{E}_{(a,r)}\left[\tfrac{1}{d}\left\|R\, b_{a,r}\right\|_1\right], \qquad b_{a,r} = P_a^\top \Delta X_a[:, r] \in \mathbb{R}^d, \tag{11}$$

over a uniformly drawn anchor-coordinate pair (a, r), since $b_{a,r}$ is exactly the r-th row of B_a and $(B_a R^\top)_{r,:} = R\, b_{a,r}$. Averaging Equation (11) over s independent draws is therefore unbiased for the full criterion, and its subgradient $\operatorname{sign}(R b_{a,r})\, b_{a,r}^\top$ is rank one in R. A step touches $O(sd)$ memory for the sampled rows and reads one anchor's factor at a time, so per-step cost no longer depends on p or m at all. Drawing the pairs grouped by anchor matters in practice: a flat gather of the factor tensor would materialize an $s \times k \times d$ copy, which at large k and d is tens of gigabytes per step.

Algorithm 2 DSReg at scale. All three evaluation modes consume the same factors from part (a) and return the same objective, differing only in what they hold in memory. Annotations on the right give the shape of the array formed on that line, and its cost where that is the point. Here n is the sample count, p the observed dimension, d the latent dimension, m the anchor count and $k > d$ the neighborhood size, and all three modes return the value together with its subgradient.

(a) Neighborhood factors. Run once, shared by (b), (c) and (d).

Require: $h \in \mathbb{R}^{n\times d}$, $x \in \mathbb{R}^{n\times p}$, anchors m, neighbors $k > d$, ridge λ

1: sample anchor indices $\mathcal{A} \subset [n]$ with $|\mathcal{A}| = m$
2: **for** $a \in \mathcal{A}$ **do**
3: $\quad \mathcal{N}_a \leftarrow$ the k nearest neighbors of h_a in h; $\quad \Delta H_a \leftarrow h[\mathcal{N}_a] - h_a$ — shape $k \times d$
4: $\quad P_a \leftarrow \Delta H_a(\Delta H_a^\top \Delta H_a + \lambda I)^{-1}$ — shape $k \times d$
5: **end for**
6: **return** $\{(a, \mathcal{N}_a, P_a)\}_{a\in\mathcal{A}}$ — storage $O(mkd)$, not $O(mpd)$

(b) Materialized. Fastest while it fits; forms the $p \times d$ array explicitly.

Require: factors, x, rotation R; $\quad L \leftarrow 0, G \leftarrow 0_{d\times d}$

1: **for** each anchor a **do**
2: $\quad \Delta X_a \leftarrow x[\mathcal{N}_a] - x_a$ — shape $k \times p$
3: $\quad B_a \leftarrow \Delta X_a^\top P_a$ — shape $p \times d$; (c) and (d) avoid this
4: $\quad M_a \leftarrow B_a R^\top$; $\quad L \mathrel{+}= \sum_{r,j} |(M_a)_{rj}|$; $\quad G \mathrel{+}= \operatorname{sign}(M_a)^\top B_a$ — cost $O(pd^2)$
5: **end for**
6: **return** $L/(mpd)$, $\quad G/(mpd)$

(c) Factored. Same value as (b), about half the memory, more time.

Require: factors, x, rotation R, anchor chunk size, feature block size; $\quad L \leftarrow 0, G \leftarrow 0_{d\times d}$

1: **for** each chunk of anchors **do**
2: $\quad W_a \leftarrow P_a R^\top$ for a in the chunk — shape $k \times d$, cost $O(kd^2)$
3: $\quad$ **for** each block $F \subset [p]$ of observed coordinates **do**
4: $\quad\quad \Delta X_a^F \leftarrow x[\mathcal{N}_a, F] - x_a[F]$ — shape $k \times |F|$, recomputed each pass
5: $\quad\quad M \leftarrow (\Delta X_a^F)^\top W_a$ — shape $|F| \times d$, cost $O(k|F|d)$
6: $\quad\quad L \mathrel{+}= \sum |M|$; $\quad G \mathrel{+}= \big(\operatorname{sign}(M)^\top (\Delta X_a^F)^\top\big) P_a$ — inner factor $d \times k$
7: $\quad$ **end for**
8: **end for**
9: **return** $L/(mpd)$, $\quad G/(mpd)$

(d) Row-sampled. Unbiased; per-step cost independent of p and m.

Require: factors, x, rotation R, rows per step s

1: draw $(a_1, r_1), \ldots, (a_s, r_s)$ independently and uniformly from $[m] \times [p]$
2: group the draws by anchor — a flat gather would allocate $s \times k \times d$
3: **for** each distinct anchor a in the draw, with its sampled columns F_a **do**
4: $\quad \Delta X_a^{F_a} \leftarrow x[\mathcal{N}_a, F_a] - x_a[F_a]$; $\quad b_{a,r} \leftarrow P_a^\top \Delta X_a[:, r]$ for $r \in F_a$ — rows of B_a, shape d
5: **end for**
6: **return** $\widehat{L} = \frac{1}{sd} \sum_t \|R b_{a_t,r_t}\|_1$, $\quad \widehat{G} = \frac{1}{s} \sum_t \operatorname{sign}(R b_{a_t,r_t})\, b_{a_t,r_t}^\top$ — $\mathbb{E}[\widehat{L}] \propto \mathcal{L}_{\mathrm{DSReg}}(R)$

At large d the rotation chart matters more than the criterion. Once the criterion is cheap, the remaining cost is the orthogonality constraint. Optimizing over $O(d)$ with a gradient method requires writing R in terms of unconstrained parameters. We use the matrix exponential $R = \exp(A - A^\top)R_0$ throughout the paper, which is the standard choice and is well behaved up to a few thousand dimensions. Its autograd workspace is a $2d \times 2d$ block exponential, and beyond $d \approx 16000$ that workspace, rather than the Jacobians, becomes the binding memory cost. The Cayley transform $R = (I+S)^{-1}(I-S)R_0$ with $S = A - A^\top$ is an alternative chart on the same manifold, orthogonal for every A because S is skew-symmetric, and it needs one linear solve in place of the exponential. It is what we use at the largest dimension; low-rank orthogonal updates (Kiani et al., 2022) are the natural next step for anyone who needs to go beyond the dimensions we report.

Which mode to use. Parts (b), (c) and (d) of Algorithm 2 compute the same objective from the same factors and differ only in what they hold in memory, so the choice among them is purely one of cost. Materialize while the Jacobians still fit on the device, since below $d \approx 512$ the dense product is also the fastest of the three modes (Table 1). Use the factored form when they stop fitting but the objective must stay exact. It is the memory-frugal mode rather than the fast one: it roughly halves peak memory at $d = 2048$ and pays for that in step time, because the neighbor differences

Table 1: Scaling across latent dimension. The full estimated procedure, three seeds each. At $d \geq 1024$ the sampled columns use a compute-matched budget whose wall-clock stays at or below the materialized one; at smaller d the standard budget already suffices, and sampling holds no advantage there since its per-step cost is nearly d-independent while the materialized cost is what grows. All three modes optimize the same objective, and the materialized and factored columns agree on MCC to within 0.002 everywhere, which is a live check that the factorization is exact. The factored mode trades time for memory, roughly halving peak memory at $d = 2048$ at about $2.8\times$ the step time, since it recomputes the neighbor differences on each pass. Axes without a panel of their own, each measured on its own sweep with its own baseline rather than on the rows above: raising the sample count to 10^6 leaves recovery and step time flat to within seed noise ($0.826 \to 0.833$ at $d = 1024$, 49 ms throughout; 0.822 at $d = 2048$), and growing the observed dimension to 2^{17} holds the sampled step at 47 ms while MCC improves from 0.806 to 0.884 across the full doubling sweep.

		MCC			step time (ms)			peak memory (GB)		
d	n	mat.	fact.	samp.	mat.	fact.	samp.	mat.	fact.	samp.
128	20k	0.853	0.853	0.850	7.5	4.9	42.6	0.14	0.11	0.11
256	20k	0.843	0.844	0.840	7.4	6.8	43.9	0.47	0.31	0.30
512	24.6k	0.833	0.831	0.829	13.1	34.0	43.8	1.76	1.07	1.02
1024	49.2k	0.827	0.828	0.823	101.6	231.6	53.5	6.97	4.20	3.26
2048	98.3k	0.821	0.820	0.809	699.8	1933.2	92.5	27.76	13.73	11.42

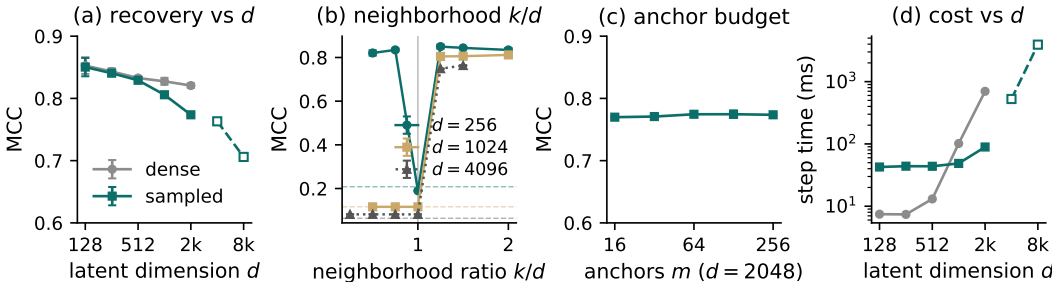


Figure 6: Scaling the full estimated procedure. (a) Recovery declines gently with d; open markers are the single-GPU frontier, where anchor count shrinks to fit memory. (b) Recovery against the neighborhood ratio k/d at three scales; dashed lines mark the analytic score of an unrotated mixing, $\sqrt{2 \ln d / d}$. (c) Sixteen anchors already match 256 at $d = 2048$. (d) The sampled step cost is nearly flat in d while the materialized cost grows two orders of magnitude, with the curves crossing between $d = 512$ and $d = 1024$.

are recomputed on every pass instead of being held as a $p \times d$ array. Use row sampling when p or the anchor set is large, since it is the only mode whose per-step cost is flat in both and the only one that still runs once the materialized route no longer fits the card, and accept that the objective is then estimated rather than computed exactly at every step. One requirement survives every mode and is not negotiable: the neighborhood must satisfy $k > d$, since $k \leq d$ makes ΔH_a rank deficient and the recovered rotation degenerates, with $k = 1.5d$ the default we recommend and $k = 1.25d$ the smallest ratio that still worked at every scale we tested (Figure 6b), a requirement no mode can rescue.

Setup. The testbed is a banded synthetic world with known footprints: each observed coordinate depends on two adjacent latents through distinct nonlinear components, footprints are length-four circular windows at distinct offsets with $p = 2d$, and the representation is a random orthogonal mixing of the true latents, so recovery is scored by MCC at any scale. Unless stated otherwise the full estimated procedure runs with $k = 1.5d$ neighbors, 256 anchors, and three seeds on one 48 GB GPU. Table 1 scales the latent dimension, and Figure 6 summarizes recovery and cost across every axis of the study.

One GPU carries the full method to $d = 8192$. Recovery declines gently with d while the sampled step cost stays nearly flat against a steep rise for the materialized evaluation, at consistently lower memory (Table 1). Past the point where the materialized route stops fitting on the card, row sampling is the only remaining option, and it carries recovery to $d = 8192$ within the card's memory.

No other axis is the bottleneck. The remaining cost axes are flat in the way the algebra predicts (Figure 6). Raising the sample count to 10^6 leaves recovery and per-step cost unchanged, and doubling the observed dimension up to 2^{17} leaves the sampled step cost flat while recovery improves, since each latent leaves its footprint on proportionally more observed coordinates. Moreover, the criterion is anchor-efficient: a small fraction of the default anchor budget already recovers as well as the full budget, which is what makes the frontier's reduced anchor counts benign, and a protocol-matched control confirms the leaner protocol costs almost nothing. Meanwhile, the decline

with d is independent of the footprint family, with nested and random-overlap constructions tracking the banded family at every dimension tested, with no family-specific tuning.

The neighborhood must span the latent dimension. Scaling the estimator exposes one requirement of its own, with a margin (Figure 6b). Neighborhoods with $k < d$ make B_a rank-deficient by construction; at small d this only dents recovery, but the surviving anisotropy washes out with dimension, and at larger d the criterion loses the rotation signal entirely, with the optimizer returning its starting point and the measured score matching an unrotated random orthogonal mixing. Meanwhile, square neighborhoods $k = d$ fail at every scale for a different reason, since the neighborhood Gram matrix then sits at the hard edge of its spectrum and the ridge solve amplifies noise. Recovery is restored from $k = 1.25d$ at every scale tested, which motivates the $k = 1.5d$ default and its margin. The practical frontier on one 48 GB card is therefore $d = 8192$, set by the memory a well-posed local regression needs rather than by the optimizer. The next rung has a measured price: at $d = 16384$ the well-posed construction needs an 80 GB device (with a Cayley parameterization replacing the matrix exponential, whose autograd workspace dominates at this scale), and there the binding constraint shifts from memory to the compute budget available to the rotation search.

Shortcuts that discard structure fail. Two natural approximations fail, for reasons the theory predicts. Orthogonally invariant sketches of the Jacobian, such as Hutchinson-style estimates of $\|B_a R^\top\|_F$, are blind to R by construction, so nothing remains to optimize. Dense random projections of the observed coordinates destroy the footprints the criterion reads, since each projected observed coordinate mixes rows across different footprints; only locality-preserving reductions such as the row subsampling used above keep the footprints that the criterion needs intact.

C.3 Verification and robustness

This subsection verifies the synthetic evidence behind Section 4.1 and then stress-tests it from every side: the regime construction itself, the end-to-end procedure, the optimizer and the training schedule, corrupted dependency signals, the stability guarantee of Theorem 3, and the boundary of Assumption 2. Each block below states its conclusion in its own heading.

Recovery follows Structural Diversity at every dimension. The regime map of Figure 3b carries weight only if each panel realizes the regime the theory names, so the sweep begins by verifying the conditions on the support masks themselves: the diverse pattern satisfies the prior certificates, the nested pattern violates Non-Inclusion and both structural-sparsity certificates, the minimal-difference pattern also violates Structural Variability, and the identical pattern violates Structural Diversity itself. Figure 7 then reports the full sweep from $N = 8$ to $N = 128$, five runs per point, on analytic orbits with nonlinear observed variables and exact anchor-wise dependency Jacobians, so estimation error plays no role in this comparison. Recovery stays at ceiling in every regime that satisfies Structural Diversity while the unrotated baseline decays steadily with dimension. Meanwhile, the identical regime sits at its shared-subspace level, which is the best that support information allows, and the pair span is still recovered. Each orbit draws a generator whose Jacobian support equals the prescribed footprint mask, with random nonlinear components on the active entries, so each panel's regime is enforced by construction and verified on the support masks before any run is scored.

The end-to-end procedure uses no oracle information. The learned-encoder benchmark of Figure 3a runs the whole method end to end on the same family. An encoder is trained from scratch on consecutive-state pairs of the MLP world, its representation is whitened and frozen, and the rotation is fit from observations and estimated latents with the default settings, with recovery scored on held-out data. Nothing about the world, the footprints, or the latents enters training, so this benchmark measures the procedure a practitioner would run on new data.

Optimizer choices carry no load. Figure 8 swaps the sparsity surrogate and varies the anchor budget on the $N = 16$ orbits, five runs each; each panel changes one choice while the other stays at the default (ℓ_1, 128 anchors), and every surrogate starts from the same six rotations. Recovery is insensitive to the surrogate on diverse footprints, and larger anchor budgets help exactly where the theory says the evidence is scarcest. In the minimal-difference regime a single observation row separates two footprints, and finding it takes more anchors. The two rightmost bars test the relaxation against the criterion it relaxes. A hard support count, thresholded at $\tau = 0.05$ as in Theorem 3 and minimized by coordinate search over Givens rotations, recovers at least as well as the ℓ_1 default

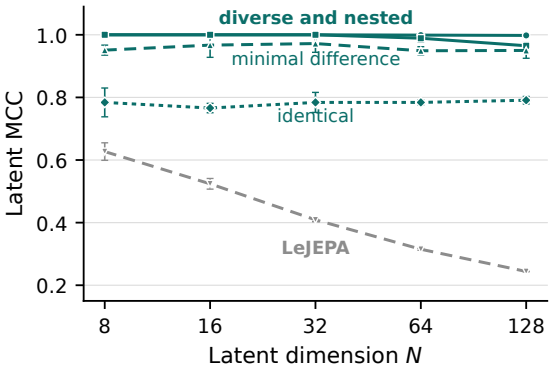


Figure 7: Footprint regimes across dimension. Recovery follows Structural Diversity at every dimension: diverse, nested, and minimal-difference footprints stay at ceiling while identical footprints sit at the shared-subspace level, far above the decaying baseline of the unrotated representation; the identical pair's span is still recovered, with CCA at least 0.998 at every dimension of the full regime sweep.

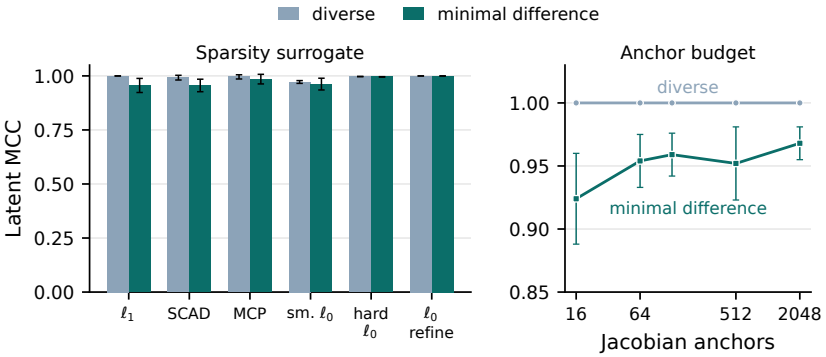


Figure 8: Recovery is insensitive to the optimizer's choices. Recovery is insensitive to the sparsity surrogate and to the anchor budget; only the minimal-difference regime rewards more anchors, exactly where the evidence is scarcest. The two rightmost bars replace the relaxation by a hard support count at $\tau = 0.05$, minimized by coordinate search from the same six starts, and by that count used to refine the ℓ_1 solution afterwards. Both series are DSReg selections, with color denoting the underlying footprint regime.

at roughly twice the cost, and where a gap to exact recovery remains, as in the minimal-difference regime, refining the ℓ_1 solution by the same count closes it within seconds.

The training schedule is immaterial. Table 2 compares fitting the rotation after training (the paper's default), jointly with it, and jointly followed by the decoupled fit, on Gaussian 3DShapes with twenty runs per arm on a shared encoder trajectory. The joint head tracks the decoupled solution and the final fit closes the remaining gap, so the schedule is purely a matter of convenience.

Table 2: Training-schedule ablation on Gaussian 3DShapes. Latent MCC (mean $\pm$ std), twenty runs per arm on a shared encoder trajectory; the rotation head never feeds back into the encoder.

Schedule	Latent MCC
LeJEPA (no rotation)	0.640 ± 0.041
Joint	0.906 ± 0.014
Joint, then decoupled fit	0.922 ± 0.005
Decoupled (default)	0.921 ± 0.005

Recovery degrades gracefully under corrupted Jacobians. Figure 9 corrupts the dependency Jacobian consumed by the optimizer with Gaussian noise, on fresh analytic orbits at $N = 16$ with $h = Qz$ and five runs per level. The observation rows carry the same nonlinear channels as the rest of the appendix, so Assumption 2 holds and the noiseless cell is a genuine signed permutation rather than a point on the excluded stratum of Figure 12. Recovery stays at ceiling across the whole noise range, and the identical-footprint control stays at its shared-subspace level throughout, so the corruption degrades the two regimes by the same small amount without moving either off its own ceiling. The behavior is what one expects from an estimator of a stable population criterion.

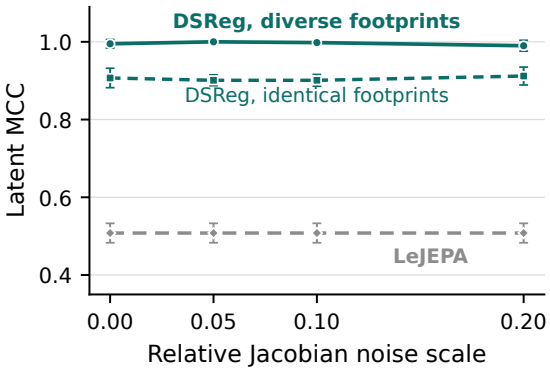


Figure 9: Corrupted Jacobians leave recovery at ceiling. Diverse footprints stay above 0.99 through noise 0.20, and identical footprints stay at their predicted shared-subspace level. Nonlinear observation rows, so Assumption 2 holds; $N = 16$, five seeds. LeJEPA applies no rotation, so its curve is flat by construction.

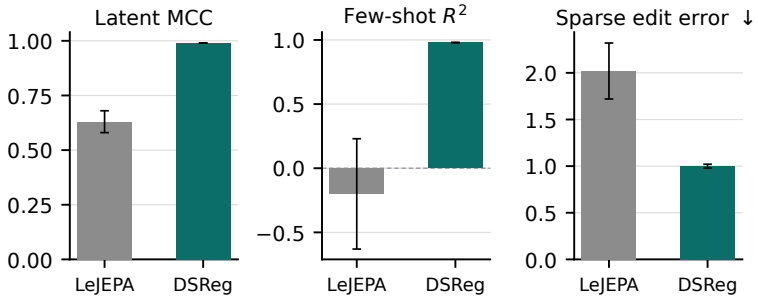


Figure 10: Sparse access improves while dense content is unchanged. At identical dense recovery ($R^2 = 0.988$ for both methods), the rotation transforms individual-latent recovery, few-shot readout from eight labels, and sparse editing, each moving decisively in DSReg's favor on the MLP benchmark at $N = 8$.

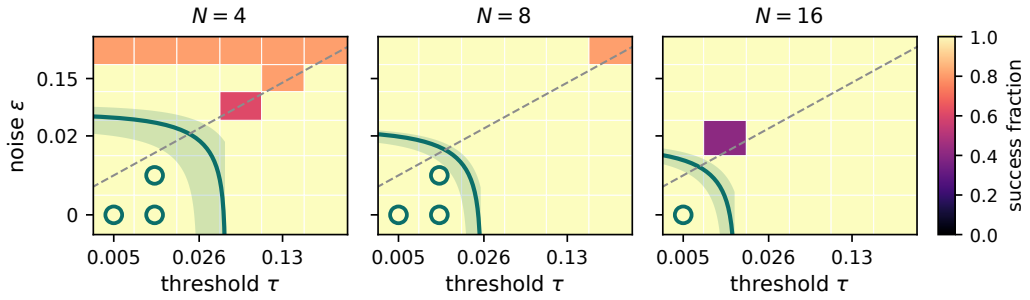


Figure 11: Success strictly contains the guarantee. Success fraction of the thresholded criterion over the (τ, ε) grid, five seeds. Solid curve: the guarantee boundary $\tau + \varepsilon = \sigma\rho^*(\delta^*)$ of Theorem 3, with seed-range shading; circled cells satisfy the guarantee for every seed, and all of them succeed. The ℓ_1 rotation used throughout the paper reproduces this map on all but four of the 108 cells, again succeeding on every guaranteed cell.

Dense readouts are unchanged by the rotation. A further check returns to the MLP benchmark of Section 4.1. Figure 10 compares the two methods at $N = 8$. Dense recovery is identical for the two methods, while individual-latent recovery, few-shot readout from eight labels, and sparse edit error move decisively in DSReg's favor. Notably, the few-shot readout swings from below zero, worse than predicting the mean, to within a hair of the dense readout ceiling.

The stability guarantee holds with room to spare. Theorem 3 guarantees recovery whenever $\tau + \varepsilon < \sigma\rho^*(\delta)$, and a dedicated sweep instantiates every quantity in that inequality on analytic orbits with nonlinear observation rows and verified distinct footprints ($N \in \{4, 8, 16\}$, five seeds), measuring σ and $\rho^*(\delta)$ directly from the data, corrupting the Jacobians at magnitude ε, and running both the thresholded criterion and the ℓ_1 implementation over a grid of (τ, ε) cells. Every cell satisfying the guarantee succeeds, for both estimators, at every dimension, and success extends well beyond the guaranteed region, so the bound is sufficient rather than tight (Figure 11). A control with linear observation rows trips the Assumption 2 check, confirming that the verification is not vacuous.

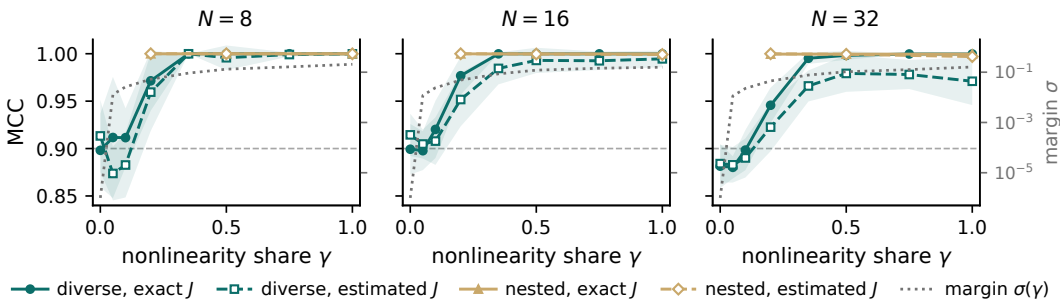


Figure 12: The excluded linear stratum is thin. MCC against the nonlinearity share γ with the measured no-cancellation margin σ overlaid (dotted, right axis). Where Assumption 2 fails ($\gamma = 0, \sigma = 0$) the criterion returns a mixed rotation; recovery crosses the 0.9 level (dashed) at $\gamma^* \approx 0.1$ and reaches the ceiling by $\gamma \approx 0.35$ at every dimension, in both regimes, with exact and estimated Jacobians in close agreement.

A small nonlinear component already restores recovery. Assumption 2 excludes, by design, worlds whose observation rows are all exactly linear in their active latents, and a second sweep measures how thin that excluded boundary is. Each row interpolates between a linear and a nonlinear function, $x_r = (1-\gamma)\,\ell_r + \gamma f_r$, so $\gamma = 0$ places every row on the excluded linear stratum; the measured margin σ grows from exactly 0 with γ, and recovery follows it. On the stratum the criterion returns a mixed rotation rather than a signed permutation, and recovery is restored from $\gamma^* \approx 0.1$ at every dimension tested ($N \in \{8, 16, 32\}$, ten seeds, exact and estimated Jacobians in close agreement), identically in the diverse and nested regimes (Figure 12).

C.4 Sparse-use probes: environments and procedures

This section specifies the environments and probe constructions behind Figure 4, in enough detail to reproduce them and to check that no probe smuggles in oracle information.

Every probe sees the full state through a random rotation. The probes of Section 4.3 run in three environments, TwoRoom (navigation between two rooms), PushT (block pushing), and FetchSlide (robot-arm puck sliding); a scripted waypoint policy with small action noise collects trajectories rendered at 64×64. The representation is the analytic orbit $h = Qz$ of Figure 4, the standardized environment state mixed by a random orthogonal Q drawn per run. The dependency rotation is fit as everywhere else in the paper: a fixed pooled-pixel feature map chosen before fitting, nearest-neighbor ridge Jacobians at anchor points, and the anchor-averaged ℓ_1 criterion of Section 3.3 under the default dependency settings of Section C.1, with no benchmark-specific tuning.

Editing succeeds only when one coordinate moves one factor. Start and goal frames are chosen so that the goal moves one ground-truth factor substantially while leaving the others nearly fixed (two factors in the compositional PushT variant). The planner selects the estimated coordinate with the largest start-to-goal change, interpolates only that coordinate (the two largest in the compositional variant), and retrieves the nearest dataset frame at each step. An edit succeeds when the retrieved endpoint reaches the goal in ground-truth state space while the off-target factors stay close to their start values. A mixed representation fails structurally: one coordinate drags several factors.

A capped sparse model is well specified only in the physical basis. The control, rollout, and surprise probes share one protocol. Train and test episodes are disjoint, and from 64 logged training transitions the module fits a linear model of the representation change: the candidate inputs are the estimated latents and the actions, each output coordinate keeps only its top few inputs by absolute correlation (two for control and monitoring, one for rollout) for a ridge regression on those alone, and a separate linear readout maps representations to states for scoring goals and endpoints in state terms. In the physical basis the dynamics are approximately sparse and the capped model is close to well specified; in a rotated basis every input and output mixes all physical variables and the same capped model becomes systematically misspecified rather than merely noisy under the same cap.

Control skill lies entirely in ranking real action sequences. Each query pairs a start frame with the state actually reached eight steps later, far enough to make planning nontrivial. The planner, given no policy and no gradients, ranks eight-step action subsequences replayed from the training episodes, each rolled through the fitted model from the start representation; the one whose predicted endpoint

lands closest to the goal is executed in the simulator from the true start state, and success means reaching the goal within a fixed relative error. Ranking is what a misspecified model cannot do, and the simulator scores the executed plan honestly either way, so no model error can hide.

Open-loop iteration compounds model error. With one input feature per output coordinate, the model is iterated open loop for five steps from held-out starts, feeding the logged actions so that only dynamics prediction is tested, and the final representation is decoded to a state and scored by R^2 against the state truly reached. On the unrotated representation the probe does worse than predicting the dataset mean, while the identical procedure on the rotated representation recovers a usable five-step model from the same budget of 64 logged transitions.

A misspecified monitor hides violations inside its own error. A held-out transition is corrupted by replacing one factor of the true next state with that factor's value from a different transition, chosen so that the jump is large and physically implausible. Clean and corrupted transitions are scored by the sparse model's prediction error, and the probe reports the separation as AUROC. Orthogonal maps preserve norms, so the corruption is exactly as large in the mixed representation as in the recovered one; what differs is the monitor's baseline, since a well-specified sparse model predicts clean transitions tightly while a misspecified one carries a large and variable clean error inside which the same violation can hide and slip past the detector without raising surprise.

No probe uses oracle information. None of the probes needs the permutation or the signs that Theorem 2 leaves free: the editing probes select their coordinate from the start-to-goal difference and the model-based probes select their features by correlation, both name-free procedures, so every probe is invariant to exactly the ambiguity the theory does not resolve. All thresholds and budgets are fixed once and shared by both representations, with the same data, model class, and planner; the fitted rotation is the only difference between the two series in each panel of Figure 4.

C.5 Learned visual encoders

The experiments of Section 4.4 replace the analytic orbit with encoders trained from pixels. This section details those encoders, their probes, and the protocol behind them.

Setup. Both visual suites render $N = 8$ independent Gaussian factors at 64×64: in the visual-factor suite each factor sets the intensity of one localized colored part of the image, and in the object-scene suite each factor moves one colored object along its own lane. Consecutive frames follow the stationary transition of Section 2. The encoder is a small strided convolutional network with a linear head, trained with the alignment objective and LeJEPA's Gaussianity regularizer; nothing in the architecture or the objective is specific to DSReg. The supervised Procrustes oracle reported alongside solves the orthogonal Procrustes problem from the whitened representation onto the ground-truth latents, so it uses labels and marks the ceiling available to any method that can only rotate. The dependency rotation itself is fit exactly as on the analytic benchmarks, from pooled-pixel features and local ridge Jacobians. Labels enter only through benchmark construction and the evaluation metrics.

The gain survives learned encoders. Learning replaces the exact orbit with an imperfect one, and these experiments ask whether the rotation's gain survives that replacement. The protocol guards the answer against overfitting. The rotation is fit on one split of estimated latents and fixed pooled-pixel features, and every recovery and sparse-use score in Table 3 is computed on the held-out split. The ten-run aggregate shows the same signature as the analytic benchmarks, with dense R^2 unchanged while latent MCC, few-shot single-latent readout, and sparse-use scores all improve (Figure 13). Moreover, the probe families of Section 4.3 carry over to the same pixel-trained encoders, with control error falling, rollout R^2 rising, surprise detection improving, and DSReg ahead in every run of all three probes from pixels (Figure 13). The adaptation to pixels is minimal: transitions come from actuated episodes whose actions are recorded, the budgets and feature caps of Section C.4 are kept, control is scored by relative endpoint error under the benchmark's mean transition, and the surprise violations are rendered and re-encoded rather than transformed analytically, so the monitor sees the corruption only through the encoder, exactly as it would in deployment.

Table 3: Split-audited learned visual encoders. The analytic signature survives learning: dense recovery is unchanged while individual-latent recovery, few-shot readout, and sparse use improve, with every score computed on a held-out split (the rotation is fit on a separate split of estimated latents and fixed pooled-pixel features). Arrows are ten-run means for LeJEPA $\rightarrow$ DSReg; sparse error is lower better.

Benchmark	N	Dense R^2	MCC	Few-shot R^2	Sparse error $\downarrow$
Visual factors	8	0.918 ± 0.002	$0.625 \rightarrow 0.958$	$-0.081 \rightarrow 0.882$	$2.13 \rightarrow 1.35$
Object scenes	8	0.924 ± 0.002	$0.591 \rightarrow 0.961$	$-0.143 \rightarrow 0.899$	$2.29 \rightarrow 1.41$

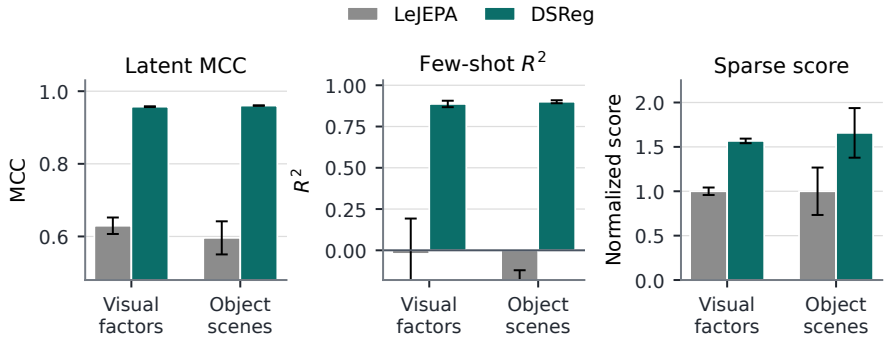


Figure 13: The analytic signature survives pixel training. The gain concentrates where individual latents matter: component-wise recovery and single-latent readout improve sharply, and sparse use follows. A negative few-shot value underperforms the held-out mean; the sparse-use score is normalized so that LeJEPA sits at one. On the pixel probes, sparse control error drops from 0.63 to 0.52, five-step sparse rollout R^2 rises from 0.74 to 0.81, and transition-surprise AUROC from 0.995 to 0.999 (LeJEPA / DSReg, ten runs).

Width mismatch costs nothing. Figure 5b in the main text removes the one remaining piece of oracle knowledge in the encoder design, the assumption that representation width matches the latent dimension. With width $M \in \{16, 32\}$ at true $N = 8$, the top-N right-singular subspace of the estimated dependency Jacobians selects the directions the observed features depend on. It is worth noting that variance could not make this selection, since every direction has unit variance under the Gaussian constraint. Within that subspace the dense span is recovered and DSReg again matches the supervised oracle to within 0.001 at every width: the dependency structure finds the latent subspace inside a larger representation on its own, with no other use of the true dimension.

Training imbalance is tolerated. Real training data is rarely balanced, so Figure 14 stress-tests the same protocol under imbalanced training distributions: long-tail settings make some factors rare during training, correlated settings make pairs of factors co-vary, and evaluation returns to a balanced held-out distribution in both cases, with the rotation fit on an unlabeled dependency split from the same training distribution (five runs). Dense R^2 stays between 0.80 and 0.93, unchanged by the rotation, and DSReg continues to improve individual-latent recovery in every setting, so the rotation does not depend on the training distribution matching the isotropic ideal of the theory, only on the dependency structure itself, which imbalance leaves intact by construction.

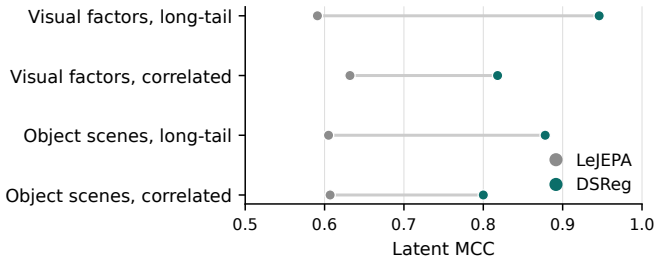


Figure 14: Training imbalance does not break the rotation. Long-tail and correlated training distributions leave the recovery gain intact, with the dense span fully preserved throughout.

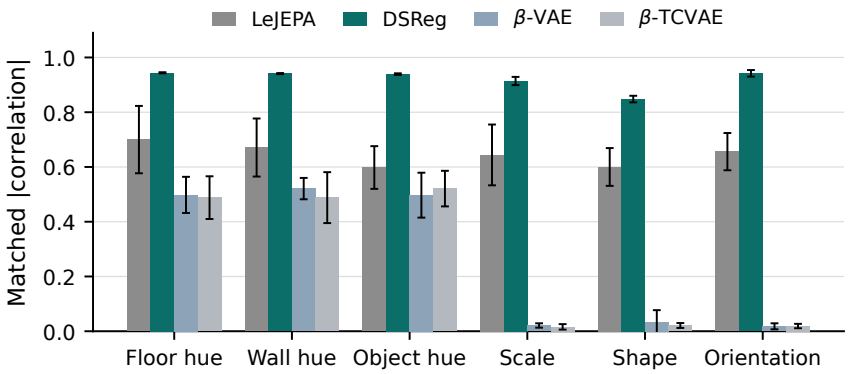


Figure 15: Per-factor recovery on Gaussian 3DShapes. DSReg improves every factor and alone recovers scale, shape, and orientation, where the VAE baselines collapse almost entirely.

C.6 Gaussian 3DShapes

The first external-renderer benchmark tests DSReg on images whose generator we did not build, so the footprint structure the criterion reads is entirely the renderer's own.

The recovery carries over to an external renderer. The 3DShapes benchmark uses the DeepMind renderer (Kim & Mnih, 2018): $z \sim \mathcal{N}(0, I_6)$ with OU pairing ($\rho = 0.95$), each latent CDF-mapped onto one factor grid (floor hue, wall hue, object hue, scale, shape, orientation), and the pre-rendered RGB image as the observation, so the latent-to-pixel footprint structure is not of our design. Table 4 reports the aggregate scores; its caption explains why the low dense R^2 of the external β-VAE and β-TCVAE references measures axis-alignment rather than lost factor information.

The dependency criterion recovers the factors that move few pixels. Figure 15 splits the scores per factor. The three hue factors paint large pixel regions and every method captures them to some degree, while scale, shape, and orientation move few pixels, and there the VAE baselines collapse outright. Reconstruction-driven objectives weight factors by the pixels they explain; the dependency criterion asks only which features a factor touches, so pixel-poor and pixel-rich factors fare alike.

Table 4: DSReg improves latent MCC on every run of Gaussian 3DShapes while matching LeJEPA's dense recovery. Scores are latent MCC, DCI, MIG, SAP, and informativeness (test R^2 of a boosted readout), twenty runs, on an external renderer. The LeJEPA row is the same frozen representation DSReg rotates, which is why the two share dense R^2; the β-VAE/β-TCVAE rows are external references with the matched trunk, trained on i.i.d. images at an untuned β=4; their low dense R^2 reflects a linear readout only, with nonlinear informativeness staying high, so the gap measures axis alignment rather than lost information.

Method	Latent MCC	DCI	MIG	SAP	Info. R^2	Dense R^2
LeJEPA	0.640 ± 0.041	0.298 ± 0.044	0.069 ± 0.026	0.212 ± 0.074	0.854 ± 0.003	0.861 ± 0.002
β-VAE	0.272 ± 0.025	0.318 ± 0.086	0.080 ± 0.038	0.102 ± 0.030	0.738 ± 0.076	0.191 ± 0.010
β-TCVAE	0.277 ± 0.043	0.345 ± 0.117	0.102 ± 0.067	0.109 ± 0.052	0.751 ± 0.087	0.199 ± 0.027
DSReg	0.921 ± 0.005	0.901 ± 0.017	0.419 ± 0.013	0.842 ± 0.016	0.868 ± 0.003	0.861 ± 0.002

The recovery is visible without a decoder. Sweeping one factor with the others held fixed selects real rendered images from the exhaustive factor grid, and the learned latents are evaluated on exactly those frames; factor labels enter only in this evaluation and never during learning. In Figure 16 each sweep moves a single rotated latent while the others stay flat and the unrotated LeJEPA coordinates respond weakly and jointly. The matched correlation matrices of Figure 17 carry the same contrast in aggregate over the evaluation set, with no decoder anywhere in the loop.

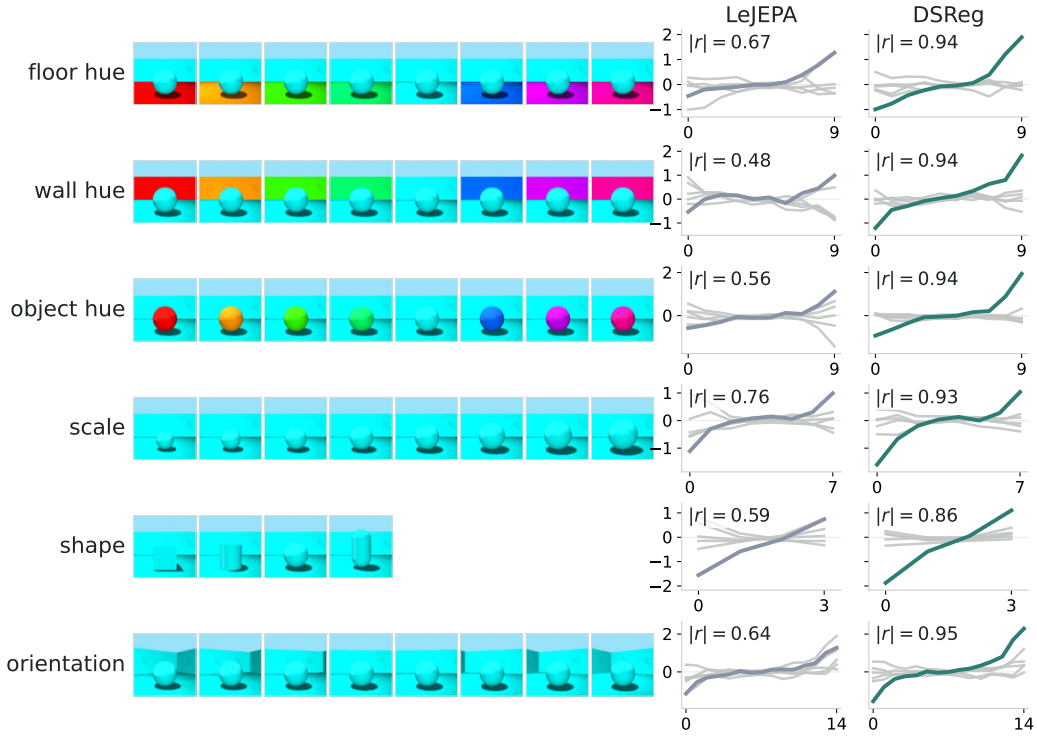


Figure 16: The rotation turns entangled responses into one selective response per factor. One run of Gaussian 3DShapes. Each row sweeps one factor with all others fixed (rendered frames on the left); the right panels show every learned latent's standardized response in LeJEPA and in DSReg coordinates, on a shared scale per row. The flat gray curves carry the no-mixing claim across all six factors.

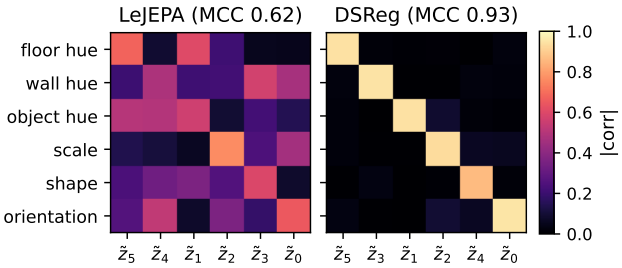


Figure 17: Matched correlation heatmaps. For the run of Figure 16, columns are permuted by each method's own best assignment. The same encoder yields a smeared matrix in the unrotated coordinates and a near-diagonal one after rotation, in agreement with the per-factor responses of the figure above.

C.7 Quarter-orientation dSprites

The second external renderer is harder by construction: a symmetry of the sprites breaks the linear premise unless the factor grid is restricted, making it a test under an imperfect linear stage.

Setup. The dSprites benchmark (Matthey et al., 2017) follows the same protocol as 3DShapes: $z \sim \mathcal{N}(0, I_5)$ with OU pairing ($\rho = 0.95$), each latent CDF-mapped onto one factor grid (shape, scale, orientation, position x, position y), and the rendered image used as the observation. Orientation is restricted to a quarter turn because the sprites are rotationally symmetric, so orientation over a full turn is not a function of the image; without the restriction the linear-identifiability premise fails for every method. What remains is a benchmark on which the premise is only partially attained, a stress test of the rotation under an imperfect linear stage rather than another clean win. The dense R^2 is also the binding ceiling: it is invariant to any rotation, so no selection criterion can recover more of the factors than the encoder linearly contains, and the scores should be read against that ceiling.

The premise, once attained, is the only bottleneck. Figure 18 shows the outcome over twenty runs, with each VAE at its best β from the sweep $\{1, 2, 4, 8, 16\}$ and twenty runs per β. DSReg improves over LeJEPA and exceeds both VAE baselines (one-sided Mann–Whitney $p = 8 \times 10^{-5}$ and $p = 4 \times 10^{-5}$), which fits the pattern seen throughout the paper, where the premise rather than the rotation step is what binds recovery on every benchmark in the paper.

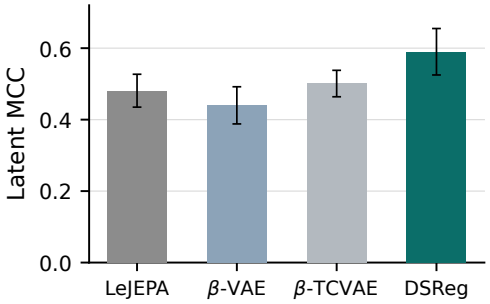


Figure 18: DSReg beats both VAE baselines under a partially attained premise. Quarter-orientation dSprites, twenty runs, dense $R^2 \approx 0.53$ as the rotation-invariant ceiling; each VAE at its best β. One-sided Mann–Whitney tests are against β-VAE and β-TCVAE at their own best β; higher MCC is better.

C.8 Rotation baselines and dense-use sanity checks

The last question is whether anything simpler could have done the work. The baselines get two chances: rotations that never see observations, and a family reading the same signal.

Latent-only rotations cannot break the rotational symmetry. The latents of the exact analytic orbit are isotropic Gaussian, so PCA, Varimax, and FastICA face the very symmetry that motivates DSReg and stay at the unrotated level alongside LeJEPA and a random orbit, while DSReg from exact Jacobians approaches the oracle and drives the dependency objective to the oracle level (Table 5). The residual gap reflects the ℓ_1 relaxation rather than estimation error: on a fraction of random worlds the relaxation's global minimizer is a slightly mixed rotation even though the support criterion of Theorem 2 still identifies the permutation. Each baseline, with Varimax and FastICA applied after whitening, returns an orthogonal transformation of the same estimated latents DSReg receives.

Freed from the symmetry, the baselines still leave the latents mixed. On the learned, non-isotropic representations the baselines have a real opening, yet Figure 5a in the main text shows the latents still mixed while DSReg nearly matches the supervised Procrustes oracle of Section C.5. Whatever asymmetry learning leaves in the representation does not point toward the world's variables; the observation-side dependency signal does, on both learned visual suites.

Dense readouts and retrieval do not move. Table 6 checks the other side: dense readouts and retrieval are unchanged to two decimals on every task, so everything DSReg gains over these baselines, it gains in sparse access rather than in the content the readouts measure.

Access to the dependency signal alone does not close the gap. A stronger baseline consumes the signal itself. On ten fresh runs of each renderer benchmark we optimize the scale-invariant column-support sparse-ICA criterion of Ng et al. (2023) on the same estimated anchor Jacobians that

Table 5: Latent-only rotation baselines. The observation-side signal rather than the span is what does the work: all methods start from the same linearly identifiable representation, but PCA, Varimax, and FastICA see only estimated-latent samples and cannot break the rotational symmetry, while the dependency criterion $\partial x/\partial(Rh)$ can, because it also sees how the observed variables respond to the applied rotation.

Method	Latent MCC	$R^2(h \to z)$	Dependency sparsity ↓
LeJEPA	0.65 ± 0.03	1	0.32 ± 0.02
Random orbit	0.64 ± 0.03	1	–
PCA	0.62 ± 0.04	1	–
Varimax	0.64 ± 0.03	1	–
FastICA	0.64 ± 0.05	1	–
DSReg	0.90 ± 0.05	1	0.18 ± 0.02
Oracle	1	1	0.18 ± 0.02

Table 6: Dense versus sparse-use sanity check. Everything the rotation gains, it gains in sparse access: dense readout and ordinary retrieval are unchanged, with dense R^2 at $1/1$ (LeJEPA / DSReg) for every task, while sparse-control error drops whenever only a few estimated latents may be used.

Task	Retrieval cost LeJEPA / DSReg	Sparse err. LeJEPA / DSReg
TwoRoom	$1.31/1.31$	$0.81/0.58$
PushT	$1.45/1.45$	$1.45/1.10$
FetchSlide	$1.43/1.43$	$1.24/0.85$

DSReg consumes, with the same orthogonal parameterization and restarts, and it recovers part of the structure while DSReg stays ahead on both benchmarks (Table 7). The advantage lies in the criterion itself rather than in mere access to the signal that both criteria consume.

Table 7: Sparse ICA on the same dependency signal recovers less. Latent MCC, mean $\pm$ std over ten fresh runs per renderer benchmark; both criteria consume the same estimated anchor Jacobians, orthogonal parameterization, and restarts. Both rows share every setting except the criterion.

Criterion	Gaussian 3DShapes	Quarter-orientation dSprites
Column-support sparse ICA	0.76 ± 0.06	0.57 ± 0.05
DSReg	0.92 ± 0.01	0.60 ± 0.08