RANK IS NOT QUALITY: LABEL-FREE MODEL SELECTION INVERTS UNDER INSTANCE-PRIVATE NUISANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised encoders are increasingly selected without labels, most often by the effective rank of their embeddings (RankMe), on the premise that a more spread-out embedding is a more useful one; anti-collapse objectives such as VICReg and LeJEPA push the same quantity up during training. Because such criteria are used where labels are absent, a failure goes unnoticed unless labels are brought back. We show that effective rank can rank models backwards. When a label-irrelevant, instance-specific panel is appended beside each CIFAR-10 image, effective rank rises $1.49\times$ (9/9 paired runs) while linear-probe accuracy falls from 0.782 to 0.115. Across that population RankMe's correlation with accuracy is $\rho = -0.686$ ($p = 0.002$), it inverts within each of SimCLR, VICReg and BYOL, and a preregistered 30-run replication reproduces the inversion before convergence ($\rho = -0.684$ at epoch 5). A first proposition shows why the objectives cannot rule this out: an encoder that whitens (and, where required, Gaussianises) instance-private content attains the alignment optimum and the optimum of the marginal health terms in use, so no such objective can prefer semantic content. Nor is the failure specific to rank. On a preregistered dial from shared to instance-private nuisance, every criterion we measured prefers a nuisance-bearing model to the clean one somewhere: spread criteria from mid entropy upward, cluster criteria such as CLID's cluster learnability wherever the nuisance is shared; a second proposition shows that no criterion reading only the embedding distribution can be guaranteed safe. Three of the dial's five preregistered tests failed, as did an anti-collapse term designed to fix the problem; we report them all, claim no graded inversion, and restricted our claim to instance-private nuisance only after seeing them. The accuracy collapse itself is Chen et al.'s RandBit result; its consequence for label-free selection is, to our knowledge, new. Where candidate models may differ in what they capture, rank should not select among them, and no distribution-only criterion replaces a labelled check.

1 INTRODUCTION

Choosing among self-supervised models without labels is attractive exactly when labels are scarce, and several criteria computed on the embeddings alone have been proposed for it. RankMe (Garrido et al., 2023) proposes the effective rank of the embedding as a proxy for downstream performance; LiDAR (Thilak et al., 2024) refines it with augmentation-induced surrogate classes; α-ReQ (Agrawal et al., 2022) reads the eigenspectrum's decay; CLID (Lu et al., 2023) combines cluster learnability with intrinsic dimension. The same quantity appears inside objectives: VICReg's variance and covariance terms (Bardes et al., 2022) and Barlow Twins' redundancy reduction (Zbontar et al., 2021) are motivated as maximising information content, and LeJEPA (Balestriero & LeCun, 2025) proves the isotropic Gaussian is the optimal embedding distribution and regularises toward it with SIGReg. Most of these criteria, and these objectives, rest on one premise, that a more spread-out embedding is a more useful one, and the criteria have been validated on populations whose quality varies with training length, architecture and hyperparameters (Section 9); because they are used where labels are absent, a population on which the premise fails would go unnoticed.

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

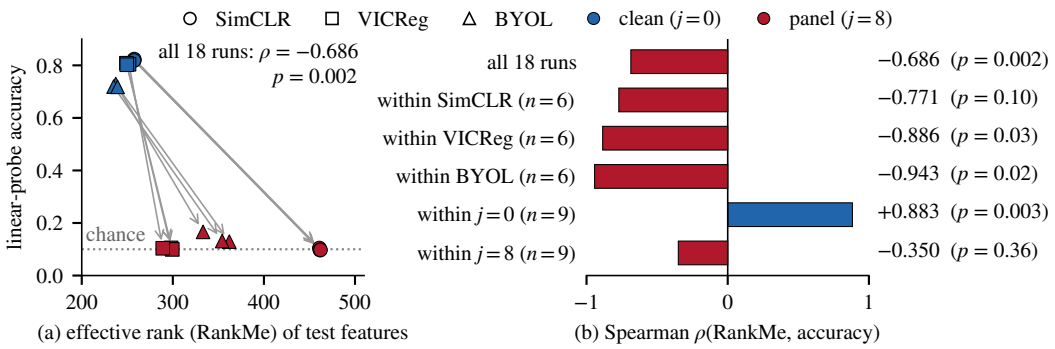

Figure 1: Effective rank ranks models backwards across a nuisance axis. (a) The 18 runs of Section 3; each arrow joins the clean run (j=0) and the panel run (j=8) of one method and seed: rank rises in all nine units while linear-probe accuracy falls to near chance. (b) RankMe's Spearman correlation with accuracy, pooled and within each population (two-sided permutation tests, 20 000 permutations; exact for $n = 6$). Within the clean population RankMe is correct; that agreement does not hold in our other clean populations (Section 4).

We construct such a population. On it, the rank-based members of that family report *improvement* on a representation that has learned nothing, and the most widely used of them ranks models in the wrong order, also among runs of a single method that differ in their input (Figure 1).

Contributions.

1. A controlled, model-free-verified nuisance axis on CIFAR-10, on which effective rank rises $1.49\times$ while linear-probe accuracy falls to near chance (Section 3).
2. The consequence for label-free *selection*: RankMe's Spearman correlation with downstream accuracy is -0.686 across the axis, and it inverts within each objective as well (-0.77, -0.89, -0.94 for SimCLR, VICReg, BYOL), so the failure follows the nuisance rather than the mixing of methods (Section 4); it replicates before convergence across three objectives (Section 5).
3. A worked account of the inversion (Section 7) and a first proposition explaining why objectives with an explicit marginal health term (VICReg, Barlow Twins, SIGReg; SimCLR at its optimum) cannot prefer semantic content, which applies to LeJEPA's isotropic target; BYOL, which fails empirically too, lies outside its scope (Section 8).
4. A preregistered entropy dial from shared to private nuisance, three of whose five preregistered tests failed and are reported as such. Descriptively, every criterion we measured on it prefers a nuisance-bearing model to the clean one in some unit: spread criteria (RankMe, α-ReQ) fail from mid entropy upward and cluster criteria (CLID's cluster learnability, a structural-entropy criterion) wherever the nuisance is shared; a proposition shows that no distribution-only criterion can be guaranteed safe (Sections 6, 8).
5. Evidence that the obvious confound is not the explanation: the training loss is *not* a substitute predictor across methods on the instance-private axis ($\rho = +0.420$, $p = 0.08$; $+0.36$ on the dial), because loss scales differ by orders of magnitude between objectives (Appendix C).
6. A negative result reported in full: a neighbour-anchored anti-collapse term, designed specifically to fix this, fails in four preregistered verdicts (Appendix D).

What is and is not ours. The accuracy collapse under instance-private content is *not* our finding. Chen et al. (2021) constructed the RandBit datasets — random bits concatenated in the channel dimension, "not altered by augmentation, so they are identical for both augmented views" — and reported linear-evaluation accuracy falling from $\sim$99% to $\sim$10% on MNIST, persisting on CIFAR-10 and ImageNet, and unfixable by changing the loss, the batch size, or the momentum memory;

preliminary experiments showed BYOL suffering too. They proposed no remedy beyond augmentation and called it an open challenge. We reproduce their phenomenon under a different injection geometry, confirm it for BYOL and extend it to VICReg. Our claim is confined to what happens to label-free *selection criteria* on such axes, which they did not measure and which largely postdate them. This work does not argue that these criteria are poor on the populations they were validated on, nor does it disavow them; it identifies a population on which their premise fails and on which, without labels, the failure would go unnoticed.

2 THE AXIS, AND WHY IT IS AN AXIS

The axis adds content that identifies each image but says nothing about its class, without touching the scene. Each CIFAR-10 image is widened from 32×32 to 32×48 by appending a 32×16 panel holding j coloured squares whose positions and colours are a deterministic function of the *image index*. The panel is therefore instance-specific (the same panel accompanies both augmented views, so a contrastive objective *can* use it for instance discrimination), independent of the image, and independent of the label. Augmentation is applied to the image only; the panel is appended afterwards. Test images are indexed from 10^6, so their panels never occur in training (all settings: Appendix G). Two gates were run before any training: (i) a linear probe on the *panel alone* scores 0.0895 against a chance level of 0.10, so the panel carries no class information; (ii) the scene region at $j = 8$ is *bitwise identical* to the same batch at $j = 0$, so any loss of class information is the encoder's doing, not the input's.

3 RANK RISES WHILE USEFULNESS COLLAPSES

	effective rank		linear-probe accuracy	
method	j=0	j=8	j=0	j=8
SimCLR	257.5	461.5	0.820	0.100
VICReg	250.1	295.5	0.805	0.103
BYOL	237.2	349.8	0.722	0.142
pooled	248.3	**368.9**	0.782	**0.115**

Table 1: Effective rank rises $1.49\times$ while accuracy falls to near chance. Paired over the nine (method, seed) units: rank $+120.6$, exact sign-flip $p = 0.0039$, 9/9 same sign; accuracy -0.668, $p = 0.0039$.

Every method fails, including the two non-contrastive ones: with the panel, linear-probe accuracy falls to near chance (Table 1). We train SimCLR (Chen et al., 2020), VICReg (Bardes et al., 2022) and BYOL (Grill et al., 2020) with and without it (ResNet-18 with a CIFAR stem, 512-d backbone features, 60 epochs, three seeds per cell; every run in Appendix Table 5). VICReg — whose variance term is *justified* as maintaining information content — reaches chance accuracy with that term essentially satisfied: its final loss of 0.59 bounds the mean variance hinge by 0.024.

4 THE CONSEQUENCE FOR LABEL-FREE SELECTION

RankMe, whose purpose is to rank models without labels, ranks the eighteen runs above backwards, pooled ($\rho = -0.686$, $p = 0.002$) and within each objective (Figure 1b). It inverts when the population spans a nuisance axis, which is exactly what a practitioner selecting without labels cannot check. The inversion is not a by-product of pooling methods: within each objective the model captured by the nuisance has the higher rank in every seed. Within the clean population above RankMe is correct ($+0.883$), but this cross-method agreement is fragile: in the clean levels of our 30-epoch populations ($n = 6$ each) it is $+0.37$ at epoch 5 and -0.20 at the end (Section 5) and $+0.14$ on the dial (Section 6, $b = 0$).

5 CONFIRMATION ACROSS THREE OBJECTIVES

Section 4 rests on a two-point axis and 18 runs, all at 60 epochs. We repeat it, preregistered before the runs existed, on $j \in \{0, 1, 2, 4, 8\}$ across *all three* objectives (30 runs, 2 seeds, 30 epochs), and read every criterion at **epoch 5 of 30** — before convergence, which is when a practitioner would want it. A separability gate was evaluated before anything else: a criterion that merely restates the training loss would reproduce every correlation below for an uninteresting reason, so we fixed in advance that no headline may be reported unless (S1) the pooled $\rho(\text{loss}, \text{accuracy})$ is at most 0.50 in absolute value and not significant, and (S2) the candidate — an anchored variance ratio we introduced, the variance of neighbour-averaged embeddings relative to that of the embeddings (Appendix D) — correlates with the loss at $|\rho| \leq 0.90$. Observed: **S1** $\rho = -0.152$ ($p = 0.423$), **S2** $\rho = -0.376$. Both pass, so the loss is genuinely separable here — unlike within a single objective, where it can be collinear with accuracy: per method the loss reaches $\rho = +0.479$ (SimCLR), $+0.122$ (VICReg) and $+0.964$ (BYOL), which is why the study spans three objectives.

criterion (epoch 5 of 30)	ρ with final accuracy	p	ρ with loss
α-ReQ (directed)	-0.846	$< 10^{-4}$	$+0.233$
RankMe (effective rank)	-0.684	$< 10^{-4}$	$+0.313$
training loss	-0.152	0.4225	—
LiDAR-style	$+0.102$	0.5919	$+0.740$
cluster learnability (CL of CLID)	$+0.769$	$< 10^{-4}$	-0.524
feature variance	$+0.850$	$< 10^{-4}$	-0.419
anchored variance ratio	$+0.869$	$< 10^{-4}$	-0.376

Table 2: Label-free criteria against downstream accuracy, 30 runs spanning three objectives, read before convergence. Two-sided permutation test, 20 000 permutations. α-ReQ (directed) is $-|\alpha - 1|$ with α fitted to the top 50 eigenvalues. Every criterion's reliability is high (ICC 0.91–0.99), so none of these correlations is limited by measurement noise.

What replicates. RankMe inverts again ($\rho = -0.684$), now across a wider population and read at epoch 5 rather than at the end. α-ReQ inverts more strongly still (-0.846). A single selection decision shows it starkly: in each of the six (method, seed) pools of five checkpoints, RankMe and α-ReQ read at epoch 5 keep a captured checkpoint ($j > 0$; probe accuracy 0.10 for SimCLR and VICReg), although the clean one is the most accurate in every pool. The preregistration also asked for held-out selection, reported but not gated: choose the (method, j) cell by each criterion on one seed and score it on the other. RankMe and α-ReQ select chance-level models (0.10 in both directions); feature variance, cluster learnability and the anchored ratio select BYOL at $j = 0$ (0.64); LiDAR selects the best model (SimCLR at $j = 0$, 0.78).

The boundary holds on this axis — and only on this axis. Cluster Learnability does not invert ($+0.769$), as expected from what it reads: the nuisance here is private, and private content has no shared structure to reward. This is no safe harbour (**Section 6**): when the nuisance is shared, cluster learnability inverts in turn.

Two refuted expectations. First, we predicted (confidence 0.7, in the preregistration of the earlier CPU study of Appendix E) that LiDAR would invert *hardest*, because the nuisance is identical across views and its within-class scatter is therefore exactly zero. Pooled at epoch 5 it does not invert: $\rho = +0.102$, $p = 0.59$ (-0.210 at the final checkpoint). **That prediction is refuted**. LiDAR is instead the criterion most entangled with the loss here ($+0.740$), and restricted to $j > 0$ it *does* go strongly negative (-0.752) — so the mechanism may be real but is not what governs the pooled ranking. Second, **plain feature variance ($+0.850$) matches the anchored ratio we introduced ($+0.869$)**: the bootstrap intervals of their difference contain zero at both checkpoints (Appendix D), so the contribution is the inversion and its boundary, not a new criterion.

A caveat about the axis itself. The nuisance axis is steep, not gradual: mean accuracy runs $0.729, 0.234, 0.202, 0.179, 0.162$ for $j = 0, 1, 2, 4, 8$. Almost all of the drop happens between $j = 0$ and $j = 1$, so the population is closer to two groups than to five levels. We therefore report every

correlation restricted to $j > 0$ as well, where the ordering is unchanged (ours $+0.831$, RankMe -0.648, CLID $+0.778$), and within a nuisance level, where all correlations weaken substantially (ours $+0.436$, RankMe -0.333). Is the inversion a two-group difference dressed as a correlation? Section 6 runs the preregistered check; three of its five tests failed.

6 THE ENTROPY DIAL: NO CRITERION IS SAFE ACROSS IT

Section 5 leaves one objection open: its population is closer to two groups than to five levels. The check we preregistered for it changes the dial from the *amount* of nuisance to its *entropy*. The panel ($j = 8$ squares) is seeded by the image index modulo 2^b, so 2^b distinct panels are shared across the dataset, $b \in \{0, 2, 4, 6, 8, 16\}$. At $b = 0$ every image carries the same panel; at $b = 16$ the raw index is used and every image carries its own, the instance-private case of Section 2. Test images are indexed from 10^6, so for $b \leq 8$ their panels come from the training alphabet, while at $b = 16$ they never occur in training. The scene and the panel generator are identical at every level (gates: panel-only probe ≤ 0.114 for every b against a threshold of 0.15; scene bitwise identical; $\min(2^b, 10^4)$ distinct panels among 10^4 images). Three objectives, two seeds, 30 epochs: 36 runs. Two deviations from the preregistration: the training script logs epochs 10, 20 and 30, so the verdicts are read at epoch 30 rather than at epoch 5; and it logs effective rank, α-ReQ and feature variance only, so cluster learnability was computed afterwards from saved features, and LiDAR and the anchored ratio were not computed.

The preregistered verdicts: three of five tests failed. Accuracy falls with b (Spearman -0.855, $p < 10^{-4}$), but the test that the fall is *graded* (S3-A) fails: no intermediate level lies more than two pooled standard deviations from both ends. The headline test (S3-B), that RankMe's correlation with accuracy is at most -0.4, fails: it is -0.208 ($p = 0.11$). Feature variance correlates at $+0.735$, as predicted (S3-C). The structural-entropy test (S4-A, below) fails too. As written, the decision rule then reads: the two-group caveat stands. That reading rests on an operationalisation the preregistration did not fix. We implemented "two pooled standard deviations" as the root-mean within-level variance over all six runs of a level (0.143), which mixes three objectives whose accuracies differ by up to 0.5 at the same b; this choice exists only in our analysis script, whose earliest record postdates the S3 data. With a seed-only deviation (0.017), levels $b = 2, 4, 6$ *are* graded, and the preregistered rule then declares the inversion a two-group artifact and requires retracting the headline as it then stood, an inversion along a graded axis; removing method means does not rescue S3-B (-0.22, $p = 0.10$). We therefore claim no graded inversion. Confining the headline to instance-private versus clean models, as this paper now does, is a post hoc re-scoping, and the rest of this section is descriptive.

Where rank inverts. Effective rank is U-shaped in b (Table 3; to read every criterion on the same embeddings we re-ran the 36 runs unchanged, saving the test-set features, and accuracy replicates to within 0.003 per level). From $b = 0$ to $b = 4$ it falls with accuracy in all six (method, seed) units. It turns while the nuisance is still shared: from $b = 6$ to $b = 8$ (64 and 256 patterns, each carried by about 780 and 195 training images) it rises by 102 in 6/6 units while accuracy falls in 6/6, and from $b = 8$ to $b = 16$ it rises by a further 121 (6/6) while accuracy does not move ($+0.007$). Over the five shared levels its correlation with accuracy is $+0.17$ ($p = 0.38$). Asked to choose among the six models of a unit, RankMe picks $b = 16$ in 6/6 units (mean accuracy of the pick 0.168; best available 0.729). What the dial leaves of the headline is the contrast between its ends: in every unit the instance-private model has higher effective rank and lower accuracy than the clean one.

A preregistered cluster criterion, and why it failed. Before any structural-entropy number was computed we preregistered the criterion we expected to survive: the two-level structural entropy of the 10-NN graph against degree-preserving nulls, $g = (\mathrm{SI} - \mathrm{SI}_{\mathrm{null}})/H_1$, label-free and blind to a fingerprint that spreads points evenly. It separates the two ends as predicted (S4-B: $g(b{=}0) > g(b{=}16)$ in 6/6 units, $p = 0.016$), but its primary test (S4-A) fails: its correlation with accuracy is -0.224 against a preregistered $+0.4$ ($p = 0.905$). The reason is the bottom row of Table 3. For $2 \leq b \leq 8$ at least 99.5% of neighbourhood edges join items with the same panel in every run, and from $b = 4$ on the graph's communities are the panels (normalised mutual information with panel identity $0.74, 1.00, 1.00, 1.00$ for $b = 2, 4, 6, 8$; exactly 64 and 256 modules at $b = 6, 8$). A shared nuisance manufactures exactly the community structure a cluster criterion rewards (Figure 2). CLID's cluster

criterion	b=0	b=2	b=4	b=6	b=8	b=16	ρ with acc.
probe accuracy	0.729	0.691	0.476	0.396	0.161	0.168	—
effective rank (RankMe)	220.2	177.4	109.4	119.8	221.8	343.1	−0.213
α-ReQ (directed)	−0.644	−1.050	−1.880	−0.253	−0.362	−0.306	−0.349
CLID's cluster learnability	0.846	0.924	0.996	0.952	1.000	0.582	−0.136
structural-entropy gain g	0.129	0.173	0.255	0.424	0.585	0.061	−0.224
feature variance	0.1138	0.1475	0.1287	0.0844	0.0376	0.0215	+0.737
k-NN edges joining same-panel items	1.000	0.998	1.000	1.000	1.000	0.000	

Table 3: The entropy dial, 36 re-run models at epoch 30; entries are means over the six (method, seed) units per level, ρ is Spearman over all 36 runs. Accuracy falls to 0.161 at $b = 8$ (SimCLR and VICReg at chance, BYOL 0.27). Rank is U-shaped; the cluster criteria rise while the panel is shared and collapse once it is private. Effective rank and feature variance use 10 000 test features, g 3 000, α-ReQ and cluster learnability 2 000. Bottom row (post hoc): fraction of 10-NN edges that join two items carrying the same panel (trivially 1 at $b = 0$, where all items share one panel).

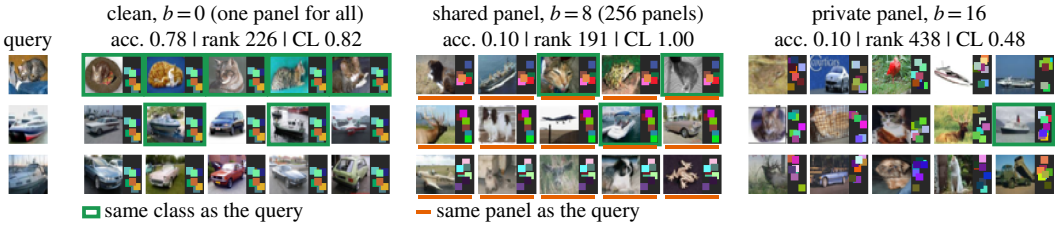

Figure 2: Nearest neighbours show what each criterion rewards. Queries: the first three CIFAR-10 test images (rule fixed in advance; shown without their panel, which differs by level). Neighbours: the five nearest by cosine similarity in the 10-NN graph of Table 3, for the seed-0 SimCLR models at $b = 0, 8, 16$; each block is headed by that model's accuracy, effective rank and cluster learnability (CL). The clean model groups by class, the $b = 8$ model by panel (CL 1.00), the private-panel model by neither (highest rank). Over the graph, 54% of edges join same-class items at $b = 0$ and 10% at $b = 8$ and 16; at $b = 8$ all join same-panel items.

learnability, computed as in Section 5 (the preregistration listed it but made no prediction), behaves the same way: it reaches 1.000 at $b = 8$, where accuracy is 0.161, and collapses (0.582) only when the panel becomes private. Over the five shared levels its correlation with accuracy is −0.74, and that of g is −0.85. By selection, g picks $b = 8$ in 6/6 units (accuracy 0.161), and cluster learnability picks $b \in \{4, 8\}$ (0.444).

What the dial shows. No criterion we measured is safe across the dial. Spread criteria track accuracy while the nuisance alphabet is small and fail from mid entropy upward: effective rank as described above, and α-ReQ (directed, $-|\alpha - 1|$), which prefers a $b = 6$ or $b = 8$ model to the clean one in 6/6 and 4/6 units (pooled $\rho = -0.349$). Cluster criteria fail wherever the nuisance is shared and recover only when it is private. Feature variance comes closest ($\rho = +0.737$; its picks average 0.704 against the best available 0.729), but it too prefers a $b = 2$ or $b = 4$ model to the clean one in 4/6 units, and it inverted outright in our voided CPU study (Appendix E). The boundary of Section 5 is therefore a property of the nuisance, not a safe region; Proposition 2 (Section 8) shows why no distribution-only criterion can be guaranteed safe.

7 Why the rank rises: a worked account

What effective rank measures. Effective rank measures how *evenly* energy is spread over directions and says nothing about what they encode. For singular values $\sigma_1 \geq \cdots \geq \sigma_d$ of the embedding matrix, RankMe reports the effective rank of Roy & Vetterli (2007), $\mathrm{erank} = \exp\left(-\sum_i p_i \log p_i\right)$, $p_i = \sigma_i / \sum_k \sigma_k$: d when the spectrum is flat, 1 when one direction carries everything. α-ReQ reads the decay exponent of the same spectrum, which is why the two move together in Section 5.

What the panel offers an encoder. The panel is a function of the image index: a high-entropy code, identical across the two views of an image, distinct across images, independent of the label. To the alignment term it is a perfect positive pair. To the health term it is content that can be spread as evenly as the encoder likes: read through a whitening map its spectrum is flat, and a flat spectrum maximises effective rank, VICReg's variance and covariance terms, and, once Gaussianised, the isotropy target (Section 8). Semantic content offers none of this: class structure is shared across thousands of images and, at $j = 0$, occupies a decaying spectrum — effective rank 248 of 512 in Table 1.

The arithmetic of composition. Whitened instance-private content spread over all directions cannot lower the effective rank of whatever else the encoder keeps; it can only raise it, most when it takes the largest share. A schematic built from synthetic spectra, with no trained model behind it (Figure 3a, Appendix A), shows the arithmetic: a decaying spectrum of effective rank 29.1 becomes 56.1 once a flat component is mixed into every direction, and a purely flat spectrum reaches the ceiling of 64. (A nuisance code confined to a few dominant directions can lower the rank, as at $j = 1$ for two methods, Appendix A, and at low b in Section 6.) Effective rank therefore counts *the share of the spectrum the nuisance occupies* — arithmetic, not training dynamics.

The real runs. Pooled over methods (Figure 3b–c), epoch-5 effective rank rises with the number of squares ($172 \to 311$ for $j = 0 \to 8$) while accuracy falls ($0.729 \to 0.162$). *The criterion counts what was captured, the probe counts what was kept*. The per-method picture, two readings of it, and what this account does not explain (why BYOL degrades gradually, why feature variance falls under capture) are in Appendix A.

8 Why whitening-optimal objectives cannot prefer signal

Proposition 1. *Let $x = (s, u)$, where u is instance-private content that is (i) identical across augmented views, (ii) high-entropy across instances, with at least as many degrees of freedom as the embedding has dimensions, and (iii) independent of the label. Consider any objective*

$$\mathcal{L}(f) \;=\; \mathcal{L}_{\text{align}}\big(f(x), f(x')\big) \;+\; \lambda\, \mathcal{L}_{\text{health}}\big(\textit{marginal of } f(x)\big).$$

Then $f(x) = Wu$ with W whitening u attains $\mathcal{L}_{\text{align}} = 0$ exactly, and, because its covariance is exactly the identity, the optimum of $\mathcal{L}_{\text{health}}$ for variance and covariance (VICReg), effective rank (RankMe as an objective) and redundancy reduction (Barlow Twins). If the encoder can further map u to $\mathcal{N}(0, I)$, the same holds for isotropic Gaussianity (SIGReg) and, after normalisation, for uniformity on the sphere (Wang & Isola, 2020). In each case f is a global minimiser carrying zero label information.

The failure is thus not a defect of any one method: **no objective whose second term is optimised by a whitened (or Gaussianised) code can prefer semantic content over instance-private nuisance**. In particular, reaching LeJEPA's isotropic-Gaussian target says nothing about what the embedding encodes.

But the proposition concerns objectives, not every criterion. It shows that a marginal objective *can* be satisfied by nuisance; it does not predict that every marginal criterion *is* defeated on every population, and empirically they are not. On our 30-run population (Table 2) the marginal family splits: feature variance ($+0.850$ at epoch 5, $+0.921$ at the end) and cluster learnability ($+0.769$, $+0.813$) do not invert, while effective rank (-0.684, -0.558) and α-ReQ (-0.846, -0.647) do, although all four read the same marginal law. We therefore state the practical claim narrowly: **do not select models by rank.** Scale-based criteria do not invert on this axis, but that is no guarantee: feature variance inverted outright in our voided CPU study ($\rho = -0.85$, Appendix E), and within each of our clean populations it orders models worse than RankMe (at epoch 5 of Section 5, -0.89 against $+0.37$). When the nuisance is shared instead, the cluster criteria fail (Section 6), and no criterion that reads only the embedding distribution can be immune:

Proposition 2. *Let the label c and a nuisance v be independent and identically distributed on $\{1, \dots, K\}$, and let ϵ be noise independent of both. For any map ϕ, the encoders $f_c(x) = \phi(c) + \epsilon$ and $f_v(x) = \phi(v) + \epsilon$ induce the same embedding distribution. Hence every criterion that is a*

function of the embedding distribution alone — spectral, cluster-based, graph-based or scale-based — gives them the same score, while a probe for c reaches, on f_c, the accuracy that $\phi(c) + \epsilon$ allows and, on f_v, only the majority-class rate.

The proof is one line: c and v have the same law, so $\phi(c) + \epsilon$ and $\phi(v) + \epsilon$ do too. The statement is elementary. It does not explain why the cluster criteria of Section 6 *prefer* nuisance clusters; it shows that no distribution-only criterion can be guaranteed to avoid them. Proposition 1 lets private nuisance attain the marginal optima; Proposition 2 lets shared nuisance impersonate class structure. Feature variance, the criterion that comes closest to being right across our dial, is no exception: in the constructed pair its value is identical, so what it tracks empirically is how trained encoders happen to scale nuisance features.

The SIGReg clause, and its limit. On our reimplementation of SIGReg's statistic, *any* exactly whitened code, including an instance-private one, reaches the finite-sample optimum (a whitened Gaussian sample beats an unwhitened one in 20/20 seeds); our preregistered test of whether *trained* encoders move toward it **failed** ($\rho = -0.427$ against -0.5; Appendix B).

Three caveats, stated rather than buried. (i) The claim is about global optima, not about zero loss for every method: for SimCLR the correct statement is that perfect alignment plus uniform directions on the sphere is NT-Xent's global optimum. (ii) It assumes the encoder class can whiten (and, for SIGReg and uniformity, Gaussianise) u. (iii) **BYOL is outside the stated form** — it has no explicit marginal health term, its anti-collapse coming from the predictor and EMA target (Tian et al., 2021). BYOL's failure in Section 3 is therefore *empirical evidence that the phenomenon is broader than the proposition's scope*, not an instance of it.

9 Related work

Feature suppression and shortcuts. That networks prefer whichever feature is easiest to learn, whether or not it is the one the task needs, is the general shortcut-learning observation of Geirhos et al. (2020). In self-supervised learning the specific form we study — content shared across the two views suppresses competing content — was established by Chen et al. (2021) with the RandBit and DigitOnImageNet constructions, and left open; Li et al. (2023) is an early treatment of the same effect. The remedies since proposed all modify the *alignment or pair-selection* side of the objective: Robinson et al. (2021) perturb features adversarially; Xue et al. (2023) analyse simplicity bias as the cause; Zhang et al. (2024) learn suppressed features in stages through feature-aware negative sampling. A withdrawn ICLR 2025 submission (Wu et al., 2024) counters suppression with a method that maintains the embedding dimensionality; on our axis dimensionality *rises* $1.49\times$ precisely when suppression is total, so it is not the quantity that fails. All are contrastive-only, and none touches the health term or the label-free criteria derived from it.

Collapse, and the metrics built to detect it. The field's diagnostic for a failed joint-embedding run is a *low* rank: complete collapse to a point, or the dimensional collapse in which the spectrum decays to a low-dimensional subspace (Jing et al., 2022). Anti-collapse terms push the other way — VICReg's variance and covariance terms (Bardes et al., 2022), Barlow Twins' cross-correlation identity target (Zbontar et al., 2021), explicit feature decorrelation (Hua et al., 2021), LDReg's local intrinsic-dimension regulariser (Huang et al., 2024), and LeJEPA's isotropic-Gaussian target (Balestriero & LeCun, 2025) with its kernel (Zimmermann et al., 2025) and sketching (Wu et al., 2026) successors. The shared assumption is that more spread is healthier; on our axis spread rises while usefulness falls, because anti-collapse regularisation and instance-private nuisance push the spectrum the same way.

Label-free evaluation, and which members our result touches. Effective rank (Roy & Vetterli, 2007) became a selection criterion in RankMe (Garrido et al., 2023); α-ReQ (Agrawal et al., 2022) reads the power-law exponent of the same eigenspectrum; LiDAR (Thilak et al., 2024) a Fisher ratio over augmentation-induced surrogate classes; CLID (Lu et al., 2023) combines cluster learnability with intrinsic dimension; IdEst (Mordacq et al., 2026) estimates intrinsic dimension directly; a persistent-homology score (Shestov et al., 2026) reads multi-scale topology; Tsitsulin et al. (2023) compare such unsupervised quality measures; Xu et al. (2022) monitor training dynamics without labels; we did not measure IdEst, persistence or monitoring scores. All are validated on popula-

tions whose quality varies through training length, architecture and hyperparameters. LiDAR was itself motivated by the observation that covariance-rank criteria cannot tell informative from uninformative directions, and Arputharaj et al. (2026), comparing such metrics across many trained models, find α-ReQ anti-correlated with accuracy on natural data and hypothesise that RankMe cannot separate task-relevant variance from noise. Closest to us, DSE (Dai et al., 2025) pairs an effective-dimensionality measure with a class-relevance measure to select checkpoints for dense prediction without labels; on our axis the dimensionality term alone moves the wrong way. What we add is that on a controlled population this blindness reverses the sign of model selection, pooled and within each objective, and that cluster criteria fail on the opposite, shared side. We could find none of these criteria evaluated on a population captured by instance-private nuisance. Appendix F tabulates which of them it touches.

Is this just CLID? CLID's cluster learnability asks whether a k-NN classifier can recover k-means cluster labels, i.e. whether the embedding carries structure *shared across instances*. Instance-private content has none, so CLID does not invert on our first axis ($+0.769$) or on constructed embeddings (1.000 for clustered versus 0.620 for isotropic, where effective rank scores the isotropic case $2.2\times$ higher). But a nuisance shared by many images *has* shared structure, and CLID rewards it (Section 6). If a selection population might span a nuisance axis, do not select on rank, do not assume that any label-free criterion is safe, and check the selection with labels. JEPA-family objectives and information-theoretic limits are discussed in Appendix F.

10 LIMITATIONS

One nuisance construction, one dataset, one architecture. The inversion is measured on index-derived squares in an appended panel, on CIFAR-10, with a ResNet-18 CIFAR stem. The panel is appended rather than embedded, so augmentation cannot reach it — a real shortcut lives inside the scene, where it can. We have not shown the effect on ImageNet, on ViTs, or on a naturally occurring shortcut. Models train for 30 or 60 epochs with two or three seeds per cell, and every criterion is read on the 512-d backbone features rather than on projector embeddings.

We do not know how often this regime occurs. We show that rank-based selection inverts *when* a population spans a nuisance axis; we do not establish how often real model-selection populations do. RankMe is recommended only for comparing runs of one method (Garrido et al., 2023): our pooled correlations lie outside that scope, and even our within-objective contrasts change the input (the panel), not a hyperparameter on fixed data, the setting in which RankMe was validated. The claim is conditional and should be read that way.

Our permutation tests treat runs as exchangeable. Runs share methods and nuisance levels, so the effective sample size is well below 18 or 30 and those p-values overstate the evidence; the exact paired sign-flip test of Table 1 is the more conservative one.

The inversion is not a graded function of nuisance entropy, and our scope is post hoc. Three of the five preregistered tests on the entropy dial (Section 6) failed: gradedness (S3-A), RankMe's correlation on the dial (S3-B, $\rho = -0.21$ against -0.4) and the structural-entropy criterion (S4-A). The level-pooled reading of the gradedness rule, under which the rule leaves our earlier caveat standing, was fixed in analysis code we cannot show predates the data; under a seed-only reading the preregistered rule requires retracting the headline as it then stood. We therefore claim no graded inversion, and confining the claim to instance-private versus clean models is a re-scoping made after the data. Effective rank tracks accuracy only while few nuisance patterns exist and already inverts from $b = 6$ to $b = 8$, well before the nuisance identifies the instance. The cluster-learnability, selection and mechanism analyses of Section 6 were run after the verdicts and are descriptive. At $b = 16$ the test panels are also unseen in training, unlike at $b \leq 8$, so the last step of the dial changes the novelty of the nuisance as well as its entropy.

Proposition 1 is bounded by its three caveats (Section 8); BYOL lies outside it, and Section 7 is an account, not a mechanism. **A confound voided one of our experiments** (Appendix E), and **we do not fix the problem** (Appendix D); the open challenge of Chen et al. (2021) remains open.

REPRODUCIBILITY STATEMENT

Every number in the paper is computed from per-run data files in the supplementary material, and every table and figure is regenerated from them by its scripts; an index in the supplement maps each section, table and figure to its data and command. Table 1 and the per-seed tables of Appendix I come from one table script, which also prints the pooled statistics quoted in the text; Figure 1 is drawn from the same per-run file with the same statistics code, and Table 3 comes from the dial summary script. Panel (a) of Figure 3 is synthetic and labelled so. Correlations are Spearman. P-values are two-sided permutation tests (20 000 permutations, seed 0), except the preregistered dial tests, which are one-sided in the preregistered direction; paired tests are exact sign-flip tests over (method, seed) units, and S4-B is an exact sign test. The training code, the patches that produced each population, the gate checks and all analysis scripts are included; Appendix G lists every setting. The saved test-set features and fp16 backbones of the 36 dial models (1.7 GB) are too large for a review supplement and will be released with the code upon publication; Figure 2 is drawn from nine of these feature files, and the neighbour lists it shows are in the supplement with the script that computed them. The supplement contains the preregistrations, with dated amendments and verdicts, of the anti-collapse term (Appendix D), the voided single-objective study (Appendix E), the 30-run replication (Section 5), the SIGReg test (Appendix B), the entropy dial and the structural-entropy criterion (Section 6). They are dated documents kept, with the analysis code, which is not public during review and will be released at publication. We can document that the entropy-dial preregistration is dated two days before its runs and that the analysis code carrying the structural-entropy thresholds predates its data; the dial's operationalisation of "two pooled standard deviations" does not (Section 6). The SIGReg statistic is reimplemented from LeJEPA's description, not its code.

AI USE STATEMENT

This work made extensive use of large-language-model agents, primarily Anthropic's Claude models running as coding agents, under the direction of the authors, who set the research direction and authorised each preregistered study and its budget. In the categories of the ICLR 2027 policy: *hypotheses, methodology and experiments* — the agents proposed and refined the hypotheses, designed the experiments, drafted the preregistration documents (the first-person voice in the supplement's preregistrations is the agents') and recorded their verdicts; *synthetic data and implementation* — they designed and implemented the appended nuisance panel and the synthetic populations of the CPU studies, wrote the training, analysis and plotting code, rented and operated the GPU machines, and launched and monitored the runs; *mathematical claims* — they formulated both propositions and wrote their proofs (Proposition 1 was corrected after an LLM-based review); *analysis and writing* — they analysed and interpreted the results, identified and summarised related work, formatted the references, and drafted and revised the text, tables and figures. No AI system generated images or data values: the images are CIFAR-10, the panels come from a deterministic procedure in the training script, every reported number is computed by the released scripts from logged runs, and no language model is part of any method or evaluation in the paper. Checks: independent LLM-agent audits recomputed the numbers of the abstract, contributions, tables and figures from the per-run files and checked the claims against their thresholds; the references were checked against Crossref, arXiv or Semantic Scholar records where such records exist; the code was run from the supplement layout and reproduces the shipped outputs. The authors reviewed the results and the paper and take full responsibility for all content of this submission, including AI-generated text, code, analyses and references.

REFERENCES

Kumar Krishna Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-ReQ: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Daniel Richards Arputharaj, Daniel Jönsson, and Gabriel Eilertsen. A comparative study of label-free representation quality metrics in deep learning. *Transactions on Machine Learning Research*, 2026. arXiv:2608.23182.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *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 (ICLR)*, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Siran Dai, Qianqian Xu, Peisong Wen, Yang Liu, and Qingming Huang. Exploring structural degradation in dense representations for self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.17299.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning (ICML)*, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Hanxun Huang, Ricardo J. G. B. Campello, Sarah Monazam Erfani, Xingjun Ma, Michael E. Houle, and James Bailey. LDReg: Local dimensionality regularized self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2110.09348.

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

Tianhong Li, Lijie Fan, Yuan Yuan, Hao He, Yonglong Tian, Rogerio Feris, Piotr Indyk, and Dina Katabi. Addressing feature suppression in unsupervised visual representations. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1411–1420, 2023. doi: 10.1109/wacv56688.2023.00146.

Yuchen Lu, Zhen Liu, Aristide Baratin, Romain Laroche, Aaron Courville, and Alessandro Sordoni. Using representation expressiveness and learnability to evaluate self-supervised learning methods. *Transactions on Machine Learning Research (TMLR)*, 2023.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884. PMLR, 2020.

Julie Mordacq, Vicky Kalogeiton, and Steve Oudot. IdEst: Assessing self-supervised learning representations via intrinsic dimension. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2606.03338.

Gollam Rabby and Sören Auer. When Graph-JEPA learns the wrong thing: Diagnosing and repairing category-conditional collapse. *arXiv:2608.20516*, 2026.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, 2007.

Aleksei Shestov, Anton Klenitskiy, Daria Denisova, Amurkhan Dzagkoev, Daniil Petrovich, Andrey Savchenko, and Maksim Makarenko. Topological metric for unsupervised embedding quality evaluation. In *Advances in Information Retrieval (ECIR 2026)*, Lecture Notes in Computer Science, pp. 596–605. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-21300-6_51.

Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua M Susskind, and Etai Littwin. LiDAR: Sensing linear probing performance in joint embedding SSL architectures. In *International Conference on Learning Representations (ICLR)*, 2024.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning (ICML)*, 2021.

Anton Tsitsulin, Marina Munkhoeva, and Bryan Perozzi. Unsupervised embedding quality evaluation. In *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, volume 221 of *Proceedings of Machine Learning Research*, pp. 169–188. PMLR, 2023. arXiv:2305.16562.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

Haiyu Wu, Randall Balestriero, and Morgan Levine. VISReg: Variance-invariance-sketching regularization for JEPA training. *arXiv:2606.02572*, 2026.

Shushan Wu, Xilin Gong, Yongkai Chen, Yufang Liu, Wenxuan Zhong, and Ping Ma. Fisher contrastive learning: A robust solution to the feature suppression effect. OpenReview, ICLR 2025 withdrawn submission, `https://openreview.net/forum?id=Tl6hStJNYX`, 2024.

Isaac Xu, Scott Lowe, and Thomas Trappenberg. Label-free monitoring of self-supervised learning progress. In *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 78–84, 2022. doi: 10.1109/CCECE49351.2022.9918377. arXiv:2409.06612.

Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. In *International Conference on Machine Learning (ICML)*, 2023.

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 (ICML)*, 2021.

Jihai Zhang, Xiang Lan, Xiaoye Qu, Yu Cheng, Mengling Feng, and Bryan Hooi. Learning the unlearned: Mitigating feature suppression in contrastive learning. In *European Conference on Computer Vision (ECCV)*, 2024.

Eric Zimmermann, Harley Wiltzer, Justin Szeto, David Alvarez-Melis, and Lester Mackey. KerJEPA: Kernel discrepancies for euclidean self-supervised learning. *arXiv:2512.19605*, 2025.

A THE WORKED ACCOUNT ON THE REAL RUNS

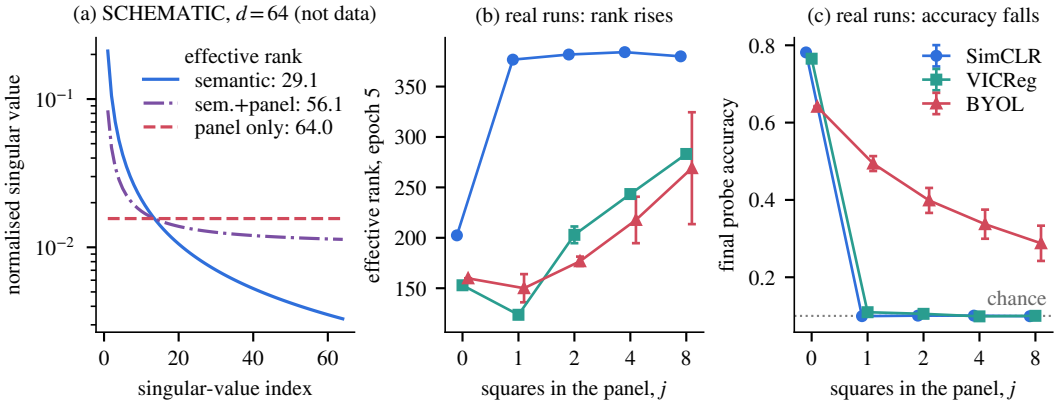

Figure 3: **(a) Schematic, not data:** three synthetic 64-d spectra and their effective ranks; mixing a flat (whitened) component into every direction of a decaying spectrum can only raise the effective rank. **(b), (c) Real runs** from the 30-run population of Section 5: effective rank at epoch 5 and final probe accuracy per method and nuisance level, mean over two seeds with min–max bars.

What the real runs show, and what they do not. Figure 3b–c are the 30 real runs. Pooled over methods, epoch-5 effective rank rises monotonically with the number of squares ($172, 217, 254, 282, 311$ for $j = 0, 1, 2, 4, 8$; $216 \rightarrow 337$ at the final checkpoint) while mean accuracy falls ($0.729, 0.234, 0.202, 0.179, 0.162$). Per method the picture is less tidy, and we report it as it is. SimCLR jumps at $j = 1$ ($202 \rightarrow 377$) and saturates; VICReg and BYOL *fall* at $j = 1$ ($153 \rightarrow 124$, $160 \rightarrow 150$) before rising past their clean value and continuing to rise with j. SimCLR and VICReg are at chance from $j = 1$; BYOL degrades gradually ($0.641, 0.494, 0.399, 0.337, 0.288$). Two readings are consistent with this; they are readings, not demonstrated mechanisms. (i) Rank tracks the *entropy of what was captured*: one square has few distinct configurations, so an encoder organised around it holds a lower-entropy code than the semantic content it displaced, and its rank can fall before rising with j, while the probe shows the semantic directions already gone at $j = 1$ for two of three methods. (ii) Rank is graded in j where accuracy is, for SimCLR and VICReg, a cliff: RankMe measures how much of the spectrum the panel occupies, which grows with j; the probe measures whether the class directions survive, which was decided at $j = 1$. That is the inversion in one sentence — *the criterion counts what was captured, the probe counts what was kept* — and here the two move in opposite directions.

What this account does not explain. Three things, so that they are not mistaken for claims. Why BYOL degrades gradually where the other two collapse: it has no marginal health term, and Section 8 keeps it outside the proposition for that reason. Why the scale-based criteria survive on this axis: feature variance *falls* under capture, by an order of magnitude for SimCLR ($0.042 \rightarrow 0.003$) and 3–4$\times$ for VICReg and BYOL (Table 5), so the captured content is not spread at the scale of the clean features; we have no mechanism for that and offer none, and it does not hold everywhere (Appendix E; Section 6). (Cluster learnability survives *this* axis because the panel is private; Section 6 shows it does not survive a shared one.) And whether an encoder is *driven* to the whitened optimum in general: our preregistered SIGReg test on trained encoders asked exactly that, and it failed (Appendix B).

B THE SIGREG CLAUSE: DIRECT EVIDENCE AND ITS LIMIT

The proposition's clause about isotropic Gaussianity can be probed on its own. Reimplementing SIGReg's statistic from LeJEPA's description — centre, scale globally, project onto 256 random directions, Epps–Pulley goodness-of-fit against $\mathcal{N}(0, 1)$, averaged — an exactly whitened Gaussian sample scores better than an unwhitened Gaussian sample in $20/20$ seeds, and the relative gap does not shrink with n ($0.446 \rightarrow 0.416$ across a $16\times$ increase). This is not the target being *exceeded* by

content carrying no information: whitening fixes the first two sample moments, and the Epps–Pulley statistic is n-scaled, so matching those moments lowers it by a constant fraction at every n: whitening the *same* Gaussian sample lowers it by 41% at $n = 400$ and 40% at $n = 1600$ (supplementary material). Both samples come from the target distribution and neither carries information. What the check does show is that the statistic's finite-sample optimum is reached by *any* exactly whitened code, including an instance-private one.

We are careful about what this does and does not show. It is a statement about the *objective*. Whether a *trained* encoder walks to that optimum when nuisance is available is a separate question, and our preregistered test of it on pixel-JEPA embeddings **failed**: the statistic improves sharply once any nuisance is present ($26.5 \to 11.8$) but not monotonically in the amount ($\rho = -0.427$, $p = 0.064$, against a preregistered threshold of $\rho \leq -0.5$, $p \leq 0.05$). We report the failure and make no claim about trained encoders. A control is worth recording: on an axis where information provably *rises*, the same statistic is flat ($\rho = -0.098$, $p = 0.64$) — it does not track information in either direction. Every embedding scored is listed in Table 8.

C THE TRAINING LOSS IS NOT A SUBSTITUTE

The obvious objection is that the training loss already tells you: capture drives the loss down. It does — *within* a method. Pooled across the three objectives it does not, because loss scales differ by orders of magnitude (mean final loss from $j = 0$ to $j = 8$: VICReg $8.99 \to 0.59$, BYOL $0.219 \to 0.007$, SimCLR $2.97 \to 2.10$):

ρ(final loss, accuracy)	value
pooled over all methods	$+0.420$ ($p = 0.081$, n.s.)
within SimCLR / VICReg / BYOL	$+0.657$ / $+0.600$ / $+0.771$

This matters methodologically as well as substantively: a single-objective study cannot separate "measures representation quality" from "measures how low the loss went", because on one objective the two are collinear. Appendix E records a study voided for this reason.

D A FIX THAT DOES NOT WORK, REPORTED IN FULL

If the marginal is the vulnerability, the natural repair (NAVR, a neighbour-anchored variant of the health term) is to read the health term on the part of the representation that is *shared across instances*. We define a row-stochastic affinity over *other* instances (stop-gradient, fixed temperature, ℓ_2-normalised rows), $\hat{Z} = AZ$, and evaluate the health term on $\hat{Z}$. Instance-private content gives diffuse affinities, so $\hat{Z}$ collapses to the batch mean and the term is maximally violated; shared content clusters and survives. On constructed embeddings the statistic separates cleanly (0.987 clustered vs 0.028 isotropic).

It does not work. Preregistered controls, four verdicts, all negative: reading the health term on $\hat{Z}$ alone has a degenerate optimum and finds it (effective rank $\to 2.5$); keeping both terms cures the degeneracy but recovers no accuracy (-0.043 at one embedding width, $+0.000$ at another) while costing 0.074–0.083 accuracy with no nuisance present at all. The intervention demonstrably happened — the shared fraction of embedding spread rose 3–7$\times$ on average in each NAVR-2 condition — and bought nothing. Every control run is listed in Tables 9–11. We report this because the open problem Chen et al. (2021) named is still open, and because a paper that only reported the diagnosis would leave the reader to guess whether the obvious repair works. In the light of Section 6 the failure is less surprising: the term rewards shared structure, which a shared nuisance supplies directly and which an encoder can manufacture from private nuisance by quantising it — a counter-move the preregistration flagged in advance as unblocked.

Plain feature variance matches the anchored ratio. On the 30-run population of Section 5, feature variance, $\mathrm{mean}_d \mathrm{Var}(z_d)$, reaches $+0.850$ against the anchored ratio's $+0.869$ at epoch 5. A bootstrap over runs (20,000 resamples) puts the difference at $+0.018$, 95% CI $[-0.044, +0.081]$ at epoch 5 and -0.082, 95% CI $[-0.212, +0.010]$ at the final checkpoint: both intervals contain zero, so the two are statistically indistinguishable and we claim no margin in either direction.

E A VOIDED EXPERIMENT

Before the results reported here, we ran a CPU-scale study in which our diagnostic predicted downstream accuracy at $\rho = +0.988$. We discarded it: the diagnostic correlated $\rho = +0.967$ with the alignment loss, and on a single-objective population the loss alone predicts accuracy at $\rho = +0.962$, so nothing in that design could be distinguished from "how low the loss went". Appendix C is what remains after applying the same test to the results we do report. We flag this because the same confound is easy to reintroduce and, on a single objective, invisible. Three cells of that study were *not* voided (feature variance, LiDAR and cluster learnability); one bears on the main text: feature variance inverted, $\rho = -0.846$ at epoch 5 and -0.894 at the end ($|\rho| = 0.79$ with the alignment loss, below the voiding threshold of 0.9), and selecting by it picked the worst nuisance scale (supplementary material). Feature variance is therefore not a safe criterion either.

F FURTHER RELATED WORK

The JEPA family. Joint-embedding predictive architectures (LeCun, 2022; Assran et al., 2023) replace augmentation invariance with prediction in embedding space but keep an anti-collapse mechanism, which LeJEPA (Balestriero & LeCun, 2025) makes an explicit marginal target. The proposition of Section 8 applies to any objective of the stated form and so to LeJEPA's optimality argument as stated. The empirical population of this paper is SimCLR, VICReg and BYOL, not a JEPA, and the SIGReg test scores pixel-JEPA embeddings from a simulator rather than LeJEPA's own model; we do not claim more than that.

Information-theoretic limits. McAllester & Stratos (2020) show any distribution-free high-confidence lower bound on mutual information from N samples is $O(\ln N)$, and Goldfeld et al. (2019) that $I(X;T)$ for a deterministic network is constant or infinite. We therefore make no claim about a scalar information estimate; our quantities are proxies and we treat them as such.

Which label-free criteria the inversion touches (a summary of Sections 4–6):

criterion	reads	does the inversion apply?
RankMe	marginal spectrum	**Yes, measured twice**: $\rho = -0.686$ (§4), -0.684 (§5); on the dial from mid entropy upward (§6).
α-ReQ	marginal spectrum	**Yes, measured**: $\rho = -0.846$; on the dial from mid entropy upward.
LiDAR	Fisher ratio over augmentation-induced classes	**We predicted yes-and-hardest and were wrong**: pooled $\rho = +0.102$; it selects the best model held out; -0.752 within $j > 0$.
CLID (its CL term)	*shared cluster structure*	**Not under private nuisance** ($\rho = +0.769$); **yes under shared nuisance**: 1.000 at $b = 8$, where accuracy is 0.161 (§6).
structural entropy (ours, prereg.)	graph community structure	Same as CLID: preregistered test failed, $\rho = -0.224$ (§6).

Concurrent work. Arputharaj et al. (2026) compare label-free representation-quality metrics over 260 vision models and find intrinsic dimension the most reliable; they do not study nuisance-captured populations. Rabby & Auer (2026) report a Graph-JEPA whose effective rank reads healthy while its representation carries no usable instance information, a degenerate solution that is a global minimum of its objective — the same pattern as our Proposition 1 in another domain.

Criteria logged but not analysed, and an earlier study. The training script also logs uniformity and a cross-view rank (the entropy rank of the cross-covariance of two augmented views), a criterion from an earlier preregistered study of ours on the same construction, in which it failed. On the dial the cross-view rank inverts as well ($\rho = -0.39$; it picks $b \in \{8, 16\}$ in 6/6 units), so reading the joint distribution of two views does not escape the failure. That earlier study also found RankMe negative within every method and checkpoint under the instance-specific panel, and its fairness gate — that RankMe track accuracy on clean models pooled across methods — failed, as it does in our 30-epoch populations (Section 4).

G TRAINING AND EVALUATION DETAILS

All runs use a ResNet-18 with a CIFAR stem (3×3 stride-1 first convolution, no max-pool; 512-d backbone output) on 32×48 inputs. Augmentation acts on the 32×32 image only (random resized crop with scale 0.2–1, horizontal flip, colour jitter $(0.4, 0.4, 0.4, 0.1)$ with probability 0.8, grayscale with probability 0.2); the panel (grey background, j squares of 6×6 pixels whose positions and colours come from a generator seeded by the panel id) is appended afterwards and is identical in both views. The panel id is the image index (Sections 3–5, and $b = 16$ in Section 6) or the index modulo 2^b; test images use the index plus 10^6. Projector 512–1024–1024–256 with batch normalisation and ReLU. SimCLR: NT-Xent with temperature 0.2. VICReg: coefficients $(25, 25, 1)$. BYOL: predictor 256–512–256, target EMA from 0.996 to 1 on a cosine schedule. Optimiser: SGD (momentum 0.9, weight decay 10^{-4}), batch 512, one-cycle schedule to learning rate 0.3 with 10% warm-up, mixed precision, gradient clipping at 1.0; 60 epochs (Section 3) or 30 (Sections 5 and 6). Linear probe: frozen, standardised backbone features of the 50 000 training images, one linear layer trained for 100 epochs with AdamW (learning rate 10^{-3}, weight decay 10^{-5}, batch 1024, cosine schedule), top-1 accuracy on the 10 000 test images.

Criteria, all on test-set backbone features. Effective rank: exponentiated entropy of the normalised singular values of the centred features (10 000 features in Sections 3 and 6; the first 2 000 in Section 5). α-ReQ (directed): $-|\alpha - 1|$, where α is the negative log–log slope of the top 50 covariance eigenvalues (2 000 features). Cluster learnability (CLID's CL term): accuracy of a 5-NN classifier trained on k-means ($k = 10$) labels of 1 200 features and scored against the nearest-centroid labels of the next 800. LiDAR-style: effective rank of $\Sigma_w^{-1/2}\Sigma_b\Sigma_w^{-1/2}$ with instances as classes and four augmented views of 300 test images ($\delta = 10^{-3}$). Anchored variance ratio: total variance of AZ over that of Z, with A a row softmax of cosine similarities at temperature 0.1, self excluded. Feature variance: $\mathrm{mean}_d \,\mathrm{Var}(z_d)$. Structural-entropy gain g: as preregistered (first 3 000 test features, cosine 10-NN union graph, Louvain, five degree-preserving nulls with $10|E|$ swaps).

H PREREGISTERED TESTS AT A GLANCE

Table 4 collects every preregistered test of the studies in this paper, with its threshold as preregistered, the observed value and the verdict; each is discussed where the table points. The preregistrations, with dated amendments and verdicts, are in the supplementary material.

I PER-SEED TABLES

Every table in this appendix, and Table 1, is written by the supplement's table script from the named per-run data file; no value is hand-entered. The script also prints, for cross-checking, the pooled statistics that the main text quotes: the method and pooled means of Table 1, the paired differences and sign-flip p-values, the Spearman correlations of Sections 4 and 5 with their permutation p-values (Monte Carlo; computed with the rows in file order, which reproduces every printed value), the SIGReg level means and correlations, and the NAVR arm means and differences. Effective rank is reported on test-set features throughout.

study	test (preregistered threshold)	observed	verdict
30-run replication (Sec. 5)	S1: $\lvert\rho(\text{loss}, \text{accuracy})\rvert \leq 0.50$ and $p > 0.05$	-0.152 ($p = 0.4225$)	pass
	S2: $\lvert\rho(\text{anchored ratio}, \text{loss})\rvert \leq 0.90$	-0.376	pass
	G1: $\rho(\text{anchored ratio}, \text{accuracy}) \geq +0.60$, $p \leq 0.05$	$+0.869$ ($p < 10^{-4}$)	pass
	G2: RankMe $\rho \leq 0$	-0.684	pass
	G3: anchored ratio above every named rival	$+0.869$ vs. $+0.850$ (feature variance)	pass on point estimates; difference not significant
	held-out selection (reported, not gated)	RankMe and α-ReQ pick 0.10 models	reported
earlier CPU study (App. E)	LiDAR inverts hardest (confidence 0.7)	$+0.539$ ($p = 0.005$) there; $+0.102$ in Sec. 5 (epoch 5)	refuted
anti-collapse term (App. D)	four verdicts	all negative	fail
SIGReg (App. B)	trained encoders move toward the optimum ($\rho \leq -0.5$)	-0.427	fail
entropy dial (Sec. 6)	S3-A: accuracy graded in b	not graded under the level-pooled sd; graded under the seed-only sd	fail as operationalised
	S3-B: $\rho(\text{RankMe}, \text{accuracy}) \leq -0.4$	-0.208 ($p = 0.11$)	fail
	S3-C: $\rho(\text{feature variance}, \text{accuracy}) \geq +0.4$	$+0.735$	pass
structural entropy (Sec. 6)	S4-A: $\rho(g, \text{accuracy}) \geq +0.4$	-0.224 ($p = 0.905$)	fail
	S4-B: $g(b{=}0) > g(b{=}16)$ in 6/6 units	6/6 ($p = 0.016$)	pass
	S4-C: void if $\lvert\rho(g, \text{loss})\rvert > 0.9$	-0.137	not voided

Table 4: Preregistered tests. Dial and structural-entropy tests are read at epoch 30 (a deviation stated in Section 6) with one-sided permutation tests in the preregistered direction; the others are two-sided. S3-A depends on an operationalisation that the preregistration did not fix (Section 6).

method	j	seed	eff. rank	feat. var.	final loss	probe acc.
SimCLR	0	0	257.6	0.0423	2.965	0.8238
SimCLR	0	1	257.4	0.0408	2.967	0.8183
SimCLR	0	2	257.6	0.0425	2.964	0.8193
SimCLR	8	0	460.9	0.0035	2.100	0.1047
SimCLR	8	1	462.1	0.0034	2.100	0.0975
SimCLR	8	2	461.5	0.0032	2.100	0.0971
VICReg	0	0	251.9	0.0227	9.006	0.8043
VICReg	0	1	249.1	0.0229	8.959	0.8081
VICReg	0	2	249.5	0.0232	9.014	0.8024
VICReg	8	0	288.9	0.0077	0.593	0.1038
VICReg	8	1	299.4	0.0070	0.594	0.0995
VICReg	8	2	298.1	0.0062	0.593	0.1052
BYOL	0	0	237.3	0.0599	0.220	0.7288
BYOL	0	1	239.3	0.0613	0.217	0.7183
BYOL	0	2	235.0	0.0630	0.219	0.7190
BYOL	8	0	333.2	0.0160	0.007	0.1662
BYOL	8	1	362.0	0.0139	0.007	0.1291
BYOL	8	2	354.2	0.0140	0.007	0.1312

Table 5: The 18 runs behind Section 3 and Section 4, one row per (method, nuisance level j, seed). ResNet-18 CIFAR stem, 512-d backbone features, 60 epochs. Effective rank is RankMe's exponentiated spectral entropy on the test-set features; feature variance is $\text{mean}_d \, \text{Var}(z_d)$.

method	j	seed	RankMe	α-ReQ	feat. var.	CL	LiDAR	anch. ratio	final loss	probe acc.
SimCLR	0	0	204.4	-0.802	0.271	0.766	144.4	0.048	3.142	0.7780
SimCLR	0	1	200.4	-0.825	0.304	0.794	142.9	0.041	3.137	0.7851
SimCLR	1	0	378.0	-0.309	0.016	0.550	97.7	0.000	2.102	0.1062
SimCLR	1	1	375.3	-0.338	0.015	0.554	106.6	0.000	2.102	0.0925
SimCLR	2	0	384.6	-0.239	0.018	0.511	100.2	0.000	2.100	0.1052
SimCLR	2	1	378.9	-0.248	0.018	0.534	98.3	0.000	2.100	0.0962
SimCLR	4	0	385.2	-0.231	0.020	0.529	84.3	0.000	2.100	0.1039
SimCLR	4	1	383.1	-0.214	0.014	0.537	94.7	0.000	2.100	0.0975
SimCLR	8	0	379.3	-0.252	0.021	0.584	66.2	0.000	2.100	0.0941
SimCLR	8	1	380.6	-0.215	0.018	0.549	78.0	0.001	2.100	0.1043
VICReg	0	0	152.5	-1.050	0.423	0.851	137.9	0.136	10.963	0.7663
VICReg	0	1	153.4	-1.043	0.404	0.841	124.6	0.135	10.998	0.7643
VICReg	1	0	119.7	-0.403	0.091	0.839	29.9	0.009	0.533	0.1073
VICReg	1	1	127.6	-0.315	0.071	0.833	32.9	0.007	0.499	0.1118
VICReg	2	0	194.7	-0.334	0.055	0.731	36.1	0.002	0.563	0.1012
VICReg	2	1	211.2	-0.299	0.045	0.708	38.9	0.002	0.571	0.1095
VICReg	4	0	242.7	-0.174	0.038	0.627	45.4	0.001	0.582	0.0988
VICReg	4	1	244.1	-0.153	0.040	0.629	46.8	0.001	0.583	0.0988
VICReg	8	0	280.6	-0.159	0.035	0.559	59.8	0.000	0.594	0.0999
VICReg	8	1	285.7	-0.122	0.028	0.584	61.2	0.000	0.596	0.1005
BYOL	0	0	159.7	-1.075	0.561	0.911	139.8	0.275	0.302	0.6425
BYOL	0	1	159.9	-1.043	0.562	0.915	115.0	0.317	0.316	0.6392
BYOL	1	0	164.0	-1.035	0.373	0.886	16.3	0.094	0.082	0.5133
BYOL	1	1	136.0	-1.028	0.485	0.899	17.6	0.102	0.058	0.4748
BYOL	2	0	172.1	-0.982	0.371	0.874	25.1	0.059	0.058	0.4310
BYOL	2	1	181.3	-0.886	0.313	0.834	28.0	0.056	0.036	0.3664
BYOL	4	0	194.7	-0.892	0.310	0.877	33.7	0.046	0.044	0.3749
BYOL	4	1	240.7	-0.717	0.186	0.818	38.8	0.036	0.025	0.2996
BYOL	8	0	213.6	-0.776	0.313	0.848	45.6	0.030	0.037	0.3334
BYOL	8	1	324.5	-0.338	0.069	0.766	59.0	0.014	0.022	0.2422

Table 6: The 30 runs of Section 5, criteria read at the epoch 5 of 30, one row per (method, nuisance level j, seed). 30 epochs; probe accuracy and loss are final. CL is the cluster-learnability term of CLID. One cell was logged twice; its last row is kept.

method	j	seed	RankMe	α-ReQ	feat. var.	CL	LiDAR	anch. ratio	final loss	probe acc.
SimCLR	0	0	218.8	-0.518	0.115	0.819	136.1	0.026	3.142	0.7780
SimCLR	0	1	220.7	-0.507	0.113	0.811	134.8	0.025	3.137	0.7851
SimCLR	1	0	430.8	-0.492	0.009	0.542	134.4	0.000	2.102	0.1062
SimCLR	1	1	426.6	-0.454	0.008	0.480	134.4	0.000	2.102	0.0925
SimCLR	2	0	430.3	-0.451	0.009	0.401	149.0	0.000	2.100	0.1052
SimCLR	2	1	428.1	-0.468	0.009	0.405	145.2	0.000	2.100	0.0962
SimCLR	4	0	429.8	-0.460	0.009	0.458	143.0	0.000	2.100	0.1039
SimCLR	4	1	429.3	-0.462	0.008	0.341	145.1	0.000	2.100	0.0975
SimCLR	8	0	426.0	-0.467	0.008	0.374	129.7	0.000	2.100	0.0941
SimCLR	8	1	425.0	-0.485	0.008	0.347	132.9	0.000	2.100	0.1043
VICReg	0	0	207.0	-0.583	0.083	0.882	117.0	0.034	10.963	0.7663
VICReg	0	1	204.1	-0.616	0.092	0.860	112.8	0.038	10.998	0.7643
VICReg	1	0	156.3	-0.245	0.044	0.877	46.2	0.004	0.533	0.1073
VICReg	1	1	161.0	-0.285	0.041	0.834	48.2	0.004	0.499	0.1118
VICReg	2	0	209.4	-0.005	0.027	0.642	46.6	0.001	0.563	0.1012
VICReg	2	1	220.6	-0.077	0.025	0.635	45.6	0.001	0.571	0.1095
VICReg	4	0	236.4	-0.057	0.022	0.584	46.2	0.001	0.582	0.0988
VICReg	4	1	236.2	-0.078	0.021	0.581	46.9	0.001	0.583	0.0988
VICReg	8	0	260.7	-0.265	0.017	0.571	58.3	0.000	0.594	0.0999
VICReg	8	1	268.5	-0.173	0.016	0.530	57.8	0.000	0.596	0.1005
BYOL	0	0	220.1	-0.786	0.150	0.825	138.7	0.103	0.302	0.6425
BYOL	0	1	222.6	-0.761	0.155	0.838	139.1	0.121	0.316	0.6392
BYOL	1	0	209.5	-0.779	0.073	0.841	10.3	0.063	0.082	0.5133
BYOL	1	1	197.6	-0.746	0.069	0.865	13.3	0.053	0.058	0.4748
BYOL	2	0	249.3	-0.590	0.054	0.765	17.0	0.042	0.058	0.4310
BYOL	2	1	255.6	-0.496	0.050	0.806	23.6	0.038	0.036	0.3664
BYOL	4	0	279.6	-0.484	0.048	0.801	25.9	0.030	0.044	0.3749
BYOL	4	1	293.9	-0.376	0.042	0.780	31.4	0.026	0.025	0.2996
BYOL	8	0	308.4	-0.313	0.040	0.739	35.9	0.016	0.037	0.3334
BYOL	8	1	332.6	-0.145	0.034	0.677	47.3	0.011	0.022	0.2422

Table 7: The 30 runs of Section 5, criteria read at the final checkpoint (epoch 30), one row per (method, nuisance level j, seed). 30 epochs; probe accuracy and loss are final. CL is the cluster-learnability term of CLID. One cell was logged twice; its last row is kept.

axis	level	seed	n	SIGReg	eff. rank	axis	level	seed	n	SIGReg	eff. rank
control	2	0	3000	38.54	10.3	control	32	3	3000	26.45	9.0
control	2	1	3000	16.92	21.5	control	32	4	3000	12.65	23.5
control	2	2	3000	40.51	11.8	private	0	0	3000	21.63	17.7
control	2	3	3000	23.50	8.7	private	0	1	3000	32.10	13.7
control	2	4	3000	17.03	20.9	private	0	2	3000	23.14	7.5
control	4	0	3000	26.87	9.8	private	0	3	3000	38.54	14.4
control	4	1	3000	13.90	19.2	private	0	4	3000	17.09	20.9
control	4	2	3000	19.59	16.7	private	2	0	3000	9.15	26.6
control	4	3	3000	27.31	8.2	private	2	1	3000	9.97	27.5
control	4	4	3000	10.92	23.0	private	2	2	3000	12.97	21.3
control	8	0	3000	23.38	12.4	private	2	3	3000	9.99	26.5
control	8	1	3000	20.67	19.8	private	2	4	3000	7.93	26.8
control	8	2	3000	24.59	14.9	private	4	0	3000	8.39	26.9
control	8	3	3000	32.16	7.0	private	4	1	3000	8.22	26.1
control	8	4	3000	11.59	23.9	private	4	2	3000	26.16	19.9
control	16	0	3000	22.04	11.0	private	4	3	3000	18.21	23.8
control	16	1	3000	26.25	16.5	private	4	4	3000	8.11	26.6
control	16	2	3000	25.80	13.3	private	8	0	3000	9.96	27.5
control	16	3	3000	20.86	9.5	private	8	1	3000	15.99	26.6
control	16	4	3000	11.34	24.8	private	8	2	3000	11.05	28.0
control	32	0	3000	23.08	10.0	private	8	3	3000	10.33	26.6
control	32	1	3000	19.24	18.0	private	8	4	3000	9.92	26.9
control	32	2	3000	26.01	14.9						

Table 8: The 45 archived pixel-JEPA embeddings scored by the reimplemented SIGReg statistic (lower = closer to isotropic Gaussian). Axis *private*: instance-private nuisance (information verified constant model-free); axis *control*: the positive control on which information rises as $2\log_2 g$ with level g.

nuisance	arm	seed	probe acc.	anchored ratio	eff. rank	align loss
absent	baseline	0	1.000	0.999	10.2	0.0055
absent	baseline	1	1.000	0.999	10.2	0.0035
absent	baseline	2	1.000	0.999	10.2	0.0047
absent	NAVR	0	0.764	1.000	4.0	0.0000
absent	NAVR	1	0.672	1.000	2.2	0.0000
absent	NAVR	2	0.768	1.000	2.2	0.0000
present	baseline	0	0.999	0.034	59.8	0.0089
present	baseline	1	1.000	0.035	59.5	0.0085
present	baseline	2	1.000	0.027	59.3	0.0077
present	NAVR	0	0.487	1.000	2.1	0.0000
present	NAVR	1	0.482	1.000	2.1	0.0000
present	NAVR	2	0.520	1.000	3.2	0.0000

Table 9: NAVR v1 control (health term on $\hat{Z}$ only), three seeds, synthetic two-factor data with a linearly separable semantic factor. Baseline accuracy is 1.000 with the nuisance present, so this control could not test suppression (Appendix D); it is reported because it exposed the degenerate optimum (effective rank $\to$ 2–3).

nuisance	arm	seed	probe acc.	anchored ratio	eff. rank	align loss
absent	baseline	0	0.748	0.134	14.9	0.0492
absent	baseline	1	0.650	0.159	14.3	0.0436
absent	baseline	2	0.728	0.147	15.3	0.0445
absent	baseline	3	0.710	0.129	15.0	0.0471
absent	baseline	4	0.712	0.132	15.3	0.0446
absent	baseline	5	0.692	0.143	14.7	0.0433
absent	NAVR	0	0.649	0.570	12.1	0.0393
absent	NAVR	1	0.594	0.563	11.6	0.0381
absent	NAVR	2	0.632	0.608	11.1	0.0331
absent	NAVR	3	0.629	0.578	12.2	0.0395
absent	NAVR	4	0.613	0.568	12.0	0.0367
absent	NAVR	5	0.627	0.589	11.2	0.0335
present	baseline	0	0.138	0.073	16.0	0.0036
present	baseline	1	0.147	0.051	16.0	0.0027
present	baseline	2	0.157	0.055	15.9	0.0058
present	baseline	3	0.157	0.057	16.0	0.0045
present	baseline	4	0.152	0.045	15.9	0.0042
present	baseline	5	0.160	0.035	16.0	0.0024
present	NAVR	0	0.105	0.396	13.3	0.0035
present	NAVR	1	0.125	0.404	13.2	0.0062
present	NAVR	2	0.090	0.395	13.3	0.0036
present	NAVR	3	0.126	0.403	13.5	0.0044
present	NAVR	4	0.097	0.373	13.9	0.0054
present	NAVR	5	0.111	0.375	14.1	0.0051

Table 10: NAVR-2 primary control (both health terms), embedding width 16, six seeds, 40 epochs, nonlinear semantic factor.

nuisance	arm	seed	probe acc.	anchored ratio	eff. rank	align loss
absent	baseline	0	0.790	0.105	24.9	0.0424
absent	baseline	1	0.752	0.117	25.8	0.0421
absent	baseline	2	0.784	0.117	25.4	0.0423
absent	baseline	3	0.770	0.106	25.0	0.0432
absent	baseline	4	0.772	0.110	25.1	0.0410
absent	baseline	5	0.771	0.127	25.9	0.0407
absent	NAVR	0	0.697	0.354	26.5	0.0375
absent	NAVR	1	0.673	0.365	26.5	0.0360
absent	NAVR	2	0.692	0.379	26.1	0.0357
absent	NAVR	3	0.698	0.372	26.5	0.0381
absent	NAVR	4	0.728	0.372	26.3	0.0358
absent	NAVR	5	0.709	0.370	26.8	0.0332
present	baseline	0	0.228	0.021	38.6	0.0177
present	baseline	1	0.223	0.021	39.1	0.0194
present	baseline	2	0.221	0.020	38.5	0.0198
present	baseline	3	0.230	0.021	38.7	0.0190
present	baseline	4	0.215	0.020	39.6	0.0203
present	baseline	5	0.207	0.020	38.8	0.0181
present	NAVR	0	0.223	0.136	37.1	0.0173
present	NAVR	1	0.218	0.132	37.9	0.0190
present	NAVR	2	0.228	0.135	38.0	0.0189
present	NAVR	3	0.219	0.131	38.0	0.0186
present	NAVR	4	0.223	0.133	37.7	0.0187
present	NAVR	5	0.218	0.136	37.4	0.0179

Table 11: NAVR-2 secondary control, embedding width 64, six seeds, 40 epochs.