PREDICTING THE PREDICTABLE: FEATURE SUPPRESSION IN JOINT-EMBEDDING PREDICTIVE ARCHITECTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised methods that train an encoder to map different augmented views of an image to similar embeddings must prevent the trivial solution in which all embeddings become identical. Recent work does so by constraining the embedding distribution itself: LeJEPA requires the embeddings to follow an isotropic Gaussian, and proves this distribution minimizes worst-case downstream risk. However, we observe that fixing the distribution turns the prediction objective into a competition for a bounded variance budget. We find that this competition orders directions by the fraction of their variance that is stable across views, an intraclass correlation, not by how easily a feature is learned. Whichever feature ranks higher occupies the budget and suppresses the other; because the constraint fixes every direction's own variance rather than which feature fills it, numerical rank never drops even as effective rank, how evenly that variance is used, does. In contrast, I-JEPA, which predicts masked representations rather than comparing views, is unaffected by a competing feature that substantially degrades LeJEPA. To address this, we propose a penalty that removes variance which is stable across views of an image but not shared with other images, leaving the isotropy constraint intact. Empirically, we show that it reduces feature suppression and improves downstream performance.

1 INTRODUCTION

Contrastive self-supervised objectives train an encoder to discriminate each image from a large set of negatives, producing representations competitive with supervised training across a wide range of vision benchmarks (Wang & Isola, 2020). But discriminating an image from everything else does not tell the encoder which of its many features to rely on: a contrastive encoder can settle on whichever feature is easiest to discriminate with and let a harder, more useful one go unrepresented (Chen et al., 2021; Robinson et al., 2021; Shah et al., 2020); the preference grows sharper as embedding capacity runs out (Xue et al., 2023); and a feature learned quickly can suppress a slower one before it is ever given a chance (Tamkin et al., 2023), a pattern already visible in which feature is linearly decodable before training even starts (Hermann & Lampinen, 2020). Recovering a suppressed feature carries its own cost, tending to suppress whichever feature was well represented before (Robinson et al., 2021; Wang et al., 2022). None of this work tests whether the same failure reaches an objective built around no discrimination task at all.

Instance discrimination is one way to make two views of an image agree, by pushing each apart from every other image in view. Self-distillation reaches the same agreement without a discrimination step, matching a student network's output to a slowly-updated teacher's instead (Caron et al., 2021), and I-JEPA groups both under one family, invariance-based methods (Assran et al., 2023). I-JEPA introduces a third design, distinct from every invariance-based method and from generative methods that reconstruct pixels directly: from a single visible context block, it predicts the representation of masked target blocks in the same image, with no augmentation-built view pairs and no negatives anywhere in the objective (Assran et al., 2023). LeJEPA follows the same template from full augmented views rather than a masked context, and adds one requirement: the embedding distribution must stay isotropic Gaussian, enforced by a regularizer called SIGReg in place of the stop-gradient and teacher-student schedule self-distillation relies on (Caron et al., 2021; Balestriero

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.

& LeCun, 2025). Neither paper frames this departure as a fix for suppression, or studies it: whether removing instance discrimination also removes the pressure that causes it has never been tested.

There is reason to expect that pressure survives regardless. A contrastive criterion built on negatives and a covariance-based criterion built on none, VICReg, are algebraically equivalent under mild conditions on the embeddings (Garrido et al., 2023b); LeJEPA's own SIGReg reduces exactly to VICReg once its distribution test is restricted to matching only two moments (Balestriero & LeCun, 2025) (Section 2 details both results). Under the conditions each requires, SIGReg and VICReg rank a candidate feature by the same criterion a negatives-based objective does, how reliably it survives across views, not how useful it is downstream; neither paper claims this holds unrestricted, and we do not either. Plain masked prediction carries no distributional constraint at all, so it sits outside that equivalence, and nothing in this literature licenses expecting it to inherit the same preference.

This licenses a precise question: does suppression appear only where an isotropy-style constraint sits alongside a predictive objective, or does prediction across views alone already carry the same preference? We answer this by building three systems from the same encoder and the same planted competing feature: masked prediction alone, as in I-JEPA (Assran et al., 2023); masked prediction with VICReg's variance and covariance terms added (Bardes et al., 2022), as also done in Mo & Tong (2024); and LeJEPA, predicting each view's embedding toward the mean of the other views under an isotropic Gaussian constraint (Balestriero & LeCun, 2025). The first pair is architecture-matched, isolating the isotropy-style constraint as the only difference between them; LeJEPA differs in more than the objective, no masking, no predictor, no momentum target, so the three-way comparison bears on which part of the objective matters rather than isolating it outright.

Answering that second question means asking what the invariance term already does alone, and what isotropy adds: left alone it already prefers whichever feature is most stable across an image's own views, but isotropy hardens that soft preference into a competition for a literal budget, where a reliable but task-irrelevant feature can win a slot from a less reliable but useful one. This predicts that the competition should turn on whether a competing feature is confirmed by other, similar images, not merely on whether it survives one image's own augmentations.

We summarize our contributions as follows:

- Under an identical planted competing feature, suppression is not universal across joint-embedding objectives: it appears under LeJEPA, which pushes full views to agree, but not under plain masked prediction, which predicts one region from another instead. Adding VICReg's terms to masked prediction does not reproduce LeJEPA's pattern either, predicting across disjoint regions changes what a competing feature needs to survive, an empirical wrinkle we report separately from the isotropy mechanism.
- We trace suppression, where it occurs, to the invariance term itself, which already ranks candidate features by how reliably they survive across views, not to whether the objective uses negatives; isotropy only hardens this ranking into a fixed budget in which a merely reliable feature can take the place of a useful one.
- We test this diagnosis with a penalty on image-private variance, taxing exactly the variance an image's neighbours do not confirm; recovering suppressed accuracy under it is evidence for the mechanism we diagnose, not a proposal we ask readers to adopt on its own terms.

2 Related Work

Feature suppression and competing features. Contrastive objectives can drop features a downstream task requires (Chen et al., 2021; Robinson et al., 2021). Chen et al. (2021) demonstrate this with datasets carrying explicit and controllable competing features, reporting that "a few bits of easy-to-learn shared features can suppress, and even fully prevent, the learning of other competing features". Robinson et al. (2021) find the same effect in instance discrimination, and observe that recovering a suppressed feature costs whichever feature was previously well represented. Li et al. (2023) prove the contrastive objective is vulnerable to suppression, and Wang et al. (2022) show it retains only what two augmented views share, discarding task-relevant information they do not. None of this work says what decides which feature is kept.

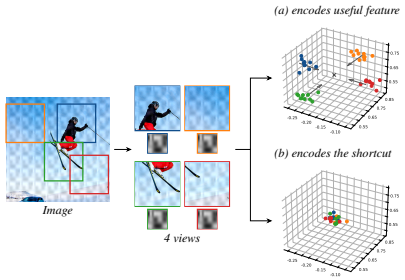

Figure 1: Why the invariance term alone cannot tell a useful feature from a shortcut. One watermarked image (real crops, watermark shown at the training opacity) is split into four views: an encoder representing the useful feature disagrees across its own crops and pays $\mathcal{L}_{\text{pred}}$, while one representing the shortcut agrees on every crop and pays nothing. Embedding positions are illustrative, not measured. Section 4 shows this preference is real; Section 5 proposes a penalty against it.

Feature selection and simplicity bias. Which of several available features a network comes to represent has been studied from several directions. Shah et al. (2020) show a network can rely exclusively on the simplest available feature and remain invariant to more complex ones that are equally predictive, and Xue et al. (2023) attribute suppression in contrastive learning to the same simplicity bias of stochastic gradient descent. Tamkin et al. (2023) give a complementary account in which learning one feature quickly suppresses the learning of others, so augmentations that destroy the easiest feature act as a form of feature dropout. Hermann & Lampinen (2020) find that when two features predict the label redundantly a network represents only one, whichever was already linearly decodable from the untrained network. In each account the competition is settled by the learning process rather than by anything the objective enforces, and the clearest evidence for that reading is Xue et al. (2023)'s remedy, which is simply to enlarge the embedding.

Objectives that constrain the embedding distribution. A separate account locates the competition in the objective itself. Wang & Isola (2020) decompose the contrastive loss into alignment of augmented views and uniformity over the hypersphere, making stability under augmentation a property the loss optimizes directly rather than a side effect of training. Non-contrastive methods replace the uniformity term with an explicit constraint on the embedding distribution: VICReg (Bardes et al., 2022) regularizes the covariance of the embeddings, and LeJEPA (Balestriero & LeCun, 2025) constrains their marginal distribution to an isotropic Gaussian. Garrido et al. (2023b) show dimension-contrastive criteria like these are equivalent to sample-contrastive ones under assumptions on the embeddings' normalization, and Balestriero & LeCun (2025) prove a second, tighter link: restricted to matching only the first two moments, their own SIGReg reduces exactly to VICReg, a case they flag as prone to the same shortcut solutions VICReg already exhibits. Both results place LeJEPA and VICReg in the family the suppression results above were shown for, but whether suppression carries over to it has not been tested.

Collapse and its measurement. A further body of work measures how much of the embedding space a representation uses. Jing et al. (2022) show contrastive methods suffer dimensional collapse, spanning a lower-dimensional subspace than the one available, and Zhang et al. (2022) identify the same effect in masked autoencoders and add a uniformity term to remove it. Li et al. (2022) find the collapse partial rather than total, and build a rank-based measure of it after observing that a low training loss does not by itself predict downstream accuracy. Garrido et al. (2023a) make the effective rank of the embeddings a label-free predictor of that accuracy. These measures report how much of the space is occupied, not what occupies it, so a representation that gives a direction to a competing feature can satisfy them while losing the feature the task needs. Neither line has been brought to joint-embedding predictive architectures (Assran et al., 2023; Balestriero & LeCun, 2025), which is where we test it.

3 FEATURE SUPPRESSION IN JEPAS

3.1 PROBLEM SETUP AND DEFINITIONS

A joint-embedding predictive objective trains an encoder so that a predictor's output for one part of an image matches an encoding of another part of the same image (Assran et al., 2023). Two instantiations of this template appear in this work. LeJEPA (Balestriero & LeCun, 2025) augments an image into V views and trains every view to predict the mean of the global views, subject to SIGReg, a term constraining the marginal distribution of each view's embedding to an isotropic Gaussian. I-JEPA (Assran et al., 2023) instead masks an image into a visible context and several target blocks, and trains a predictor to output a target encoder's representation of each masked block from a context encoder's representation of the visible one; the target encoder is an exponential moving average of the context encoder, and no explicit distributional constraint is placed on either encoder's output. Where we add VICReg's variance and covariance terms (Bardes et al., 2022) to I-JEPA, that addition sits on the same masked context-to-target prediction: masking is unchanged, and VICReg supplies the explicit constraint I-JEPA otherwise lacks.

Both instantiations must avoid collapse, the trivial solution where the encoder outputs a constant regardless of input. LeJEPA rules it out by constraining the embeddings' marginal distribution; I-JEPA rules it out through the asymmetry between a context encoder trained by gradient descent, a target encoder trained only by momentum, and a predictor with no access to the target encoder's parameters (Assran et al., 2023; Grill et al., 2020). I-JEPA with VICReg carries both: the architectural asymmetry and an explicit constraint on the context encoder's pooled output. Section 4 varies this axis, whether the objective constrains its own embeddings' marginal distribution, while holding the injected competing feature fixed.

A competing feature is a feature added to an image, uncorrelated with the class label the task requires, present across every view or crop the objective sees (Chen et al., 2021). Feature suppression is the resulting drop in downstream accuracy on the class label: linear-probe accuracy for an encoder trained with a competing feature present, minus the same probe on an encoder trained without it (Chen et al., 2021; Robinson et al., 2021).

We also report feature obsession: how much of the competing feature a frozen encoder's representation carries, measured against what a clean-trained encoder already decodes of it. This floor is not zero, a clean encoder never sees the competing feature but still decodes it above chance, so every obsession number is the excess over this baseline, not raw decodability. Suppression and obsession need not move together, and we report both separately throughout.

Every experiment here rests on a triple of training conditions, differing only in the competing feature: a clean condition never sees it; a test condition sees a feature that is a deterministic function of image identity, present across every view; a control condition sees a feature matched in magnitude and coverage to the test condition's, resampled so it carries no information the encoder could exploit. Test minus control isolates obsession, holding the raw perturbation fixed; control minus clean isolates the perturbation's own cost, independent of whether anything in it is learnable. Appendix A's Figure 3 shows both constructions.

3.2 DATASETS WITH CONTROLLABLE COMPETING FEATURES

We construct the competing feature two ways, both following the strategy of Chen et al. (2021): a feature is overlaid on the image before the augmentations that create the views, so it is present in every one. They differ in what they make controllable: DigitOnImageNet (Chen et al., 2021) lets us vary how many images share the same digit, and whether those images share a class, while TileOnImageNet lets us vary whether the pattern is shared between an image's own augmented views, giving a control condition where the pattern is still added to every image but no single pattern is shared across its views.

DigitOnImageNet. Following Chen et al. (2021), we overlay one MNIST digit (LeCun et al., 1998) on each ImageNet (Deng et al., 2009; Tian et al., 2020) image and replicate it at nine fixed locations before the augmentations that create the views, which at our crop scale guarantees that every view carries at least one whole copy. We blend the digit into the image in proportion to its

intensity, rather than by channel addition as the paper describes, which would retain the digit only where it is brighter than the image beneath it and so make its strength depend on the content behind it. Digits are drawn from the whole MNIST training split, so each of the MNIST digit images is shared by two or three ImageNet images, the largest number of distinct digits they consider and where they report the largest drop in ImageNet accuracy (Chen et al., 2021, Figure 6b). Appendix A gives the full construction.

TileOnImageNet. We overlay a faint achromatic pattern on each ImageNet image before the augmentations that create the views. The pattern is generated from a 12-bit key that depends on the image but not on its content, and is rendered as a small square tile. In the test condition one tile is repeated across the frame, so any crop spanning a tile carries the whole key. In the control condition the pattern is resampled at each tile position, which leaves the per-tile magnitude and the frame coverage unchanged and removes only the periodicity. Appendix A gives the full construction.

4 Suppression Depends on the Objective

4.1 LeJEPA and Masked I-JEPA

We inject the same watermark into LeJEPA and into masked I-JEPA, holding it fixed, and train each system under a test and a control condition. The two systems diverge: LeJEPA's test-condition invariance loss falls below its control condition over training while the online probe on the same checkpoints degrades, but masked I-JEPA's two conditions track one another for the whole run. Appendix E's Figure 7 shows both; Section 6 reports the accuracy and key-tracking numbers this crossover predicts.

LeJEPA augments an image into V views and trains every view's embedding toward the mean of the global ones,

$$\mathcal{L}_{\text{pred}} = \frac{1}{V}\sum_{v'=1}^{V} \|z_{n,v'} - \mu_n\|_2^2, \qquad \mu_n = \frac{1}{V_g}\sum_{v=1}^{V_g} z_{n,v},$$

weighed against SIGReg at a fixed ratio, 0.95 to 0.05, the authors' own recommended default (Balestriero & LeCun, 2025). Restricted to the global views, this term is, up to a constant, their sample variance around μ_n: a direction costs nothing whenever it does not change from one global view of an image to the next, whether it carries the image's content or an incidental shortcut, as Appendix B's Lemma 1 states precisely. I-JEPA instead predicts patch-level representations from a visible context block rather than toward a shared mean: given M masked target blocks $B_1, \ldots, B_M$, a predictor conditioned on the context outputs $\hat{s}_y(i)$ for each (Assran et al., 2023):

$$\mathcal{L}_{\text{pred}} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j\in B_i} \|\hat{s}_{y_j} - s_{y_j}\|_2^2.$$

Both losses average a squared distance between a prediction and a target, satisfied whenever the predicted direction agrees with the target's; neither says which direction that should be.

The two losses differ in what accompanies them. LeJEPA pairs its loss with SIGReg, pushing every view's embedding covariance toward the identity; we analyze the idealization where that push is exact, the surrogate Appendix B makes precise. I-JEPA pairs its loss with no such constraint: collapse is instead ruled out architecturally, by the asymmetry between a context encoder trained end to end and a target encoder updated only by momentum (Assran et al., 2023; Grill et al., 2020), not by any limit on how many directions the representation may use. Appendix B's Lemma 2 makes the difference precise: minimizing a prediction loss under an exact covariance constraint is a linear program over a fixed number of unit-variance slots, filled by the same criterion the invariance term already prefers, how reliably a direction survives across views; with no such constraint there are no slots to fill, and a direction that costs nothing to predict, such as a tile repeated over every crop, can sit alongside genuine content instead of displacing it. Figure 1 shows what happens once that choice is forced on LeJEPA: the shortcut wins the invariance term outright, costing nothing against the covariance SIGReg enforces either, though not necessarily against its full distributional target (Appendix B). Masked prediction poses no equivalent trap: a masked target block and its visible context are drawn from disjoint parts of the same image, so agreeing with a signal is not the same task as predicting it across a gap the context may know nothing about.

4.2 MASKED I-JEPA WITH VICREG

Masked prediction with VICReg's variance and covariance terms added is the crossed cell the other two systems leave confounded: LeJEPA constrains isotropy on top of view agreement, masked I-JEPA constrains nothing on top of context-to-target prediction, and this system adds the same explicit covariance constraint (Bardes et al., 2022) to masked prediction instead, the combination Mo & Tong (2024) use for a different purpose, fixing masked prediction's own collapse modes rather than studying feature suppression. If suppression tracks the isotropy constraint rather than the choice between view agreement and masked prediction, this system should suppress the competing feature much as LeJEPA does, even though it never compares two views of the same image directly.

On TileOnImageNet, the two conditions do not resolve this cleanly: the control-trained encoder scores lower than the test-trained one, the reverse of what suppression predicts. The reason is geometric, not the isotropy constraint: under the control condition the competing feature's payload changes from one tile position to the next rather than repeating, so most masked target blocks land somewhere their visible context carries no information about. TileOnImageNet's control condition is therefore not an informationally inert counterpart to its test condition for a masked-prediction objective, and we do not read a suppression result off this comparison either way.

DigitOnImageNet does not share this confound: its glyph sits at nine fixed positions rather than covering every tile, so a masked target block and its context are far more likely to share it, or share none of it, regardless of condition. Here, masked I-JEPA with VICReg's terms added shows the direction suppression predicts: clean-condition accuracy exceeds stamped-condition accuracy, and the digit remains linearly decodable well above chance from the same features (Section 6 reports both numbers). The two competing features thus say different things: TileOnImageNet's result is a property of how masked prediction samples its context relative to a periodic signal, not evidence for or against the isotropy account, while DigitOnImageNet's result is consistent with that account extending beyond LeJEPA to a masked-prediction objective carrying the same constraint.

4.3 WHY THE INVARIANCE TERM SELECTS OVERLAID FEATURES

Corollary 1 gives a precise answer to why the invariance term settles on a competing feature at all: among the directions available to the encoder, isotropy keeps only the ones whose variance survives across an image's own views rather than dissolving into within-image noise, and a competing feature that is deterministic given the image and unchanged by any crop survives this test as well as a direction possibly can. Making this precise: isotropy pins $\mathrm{tr}(W\Sigma_T W^\top) = D$, so by Lemma 1 minimizing the loss is the same as maximizing $\mathrm{tr}(W\Sigma_B W^\top)$, equivalently minimizing

$$L(W) = \frac{V-1}{V}\Big[D - \mathrm{tr}(W\Sigma_B W^\top)\Big].$$

Writing $M = W\Sigma_T^{1/2}$ turns the constraint into $MM^\top = I_D$ and the trace above into $\mathrm{tr}(MSM^\top)$ for the whitened matrix $S = \Sigma_T^{-1/2}\Sigma_B\Sigma_T^{-1/2}$: maximizing it places M's rows along S's D largest eigenvectors, the D candidate features with the highest $\rho(u)$, not the D most prominent in the raw signal or the D easiest to learn. Appendix B states the idealization this requires (a shared factor basis, no ties among the closest competitors, finite second moments) and proves it in full as Lemma 2 and Proposition 2. Section 6's Figure 2(b) reads this ranking directly off the trained embeddings, computing $\rho(u) = B(u)/T(u)$ for every retained direction and sorting from most to least view-stable, on the same LeJEPA checkpoints behind Figure 7. Under the control condition the spectrum stays high across nearly every direction. Under the test condition it falls away across most of them, not only the weakest one: the shortcut does not simply take the one slot a genuinely useful direction would otherwise have held, it depresses the reliability of the whole retained set.

This spectrum is graded rather than a single-slot swap, more than Corollary 2 strictly predicts, since a trained encoder does not hold every other factor fixed the way that idealized argument does; the whole-spectrum picture is an empirical observation the mechanism is consistent with, not one Appendix B's population result derives on its own.

5 PENALIZING IMAGE-PRIVATE VARIANCE

5.1 MOTIVATION

LeJEPA's predictive term, under SIGReg's isotropy constraint, is a per-direction reliability maximizer: the loss equals D minus the summed intraclass correlation of the D output directions, where the image is the object of measurement and the crop is the error facet. The constraint fixes the representation's total occupancy at exactly D unit-variance directions while leaving the allocation of that occupancy across candidate features continuous, a feature may in principle be half-represented. What rules out half-representation is not the constraint but the objective: the reliability total is linear in the allocation, and a linear functional over a polytope is maximized at a vertex. The optimum is therefore a hard selection of the D most view-stable features, with everything below the cut not attenuated but deleted, at zero weight in any linear readout.

Proposition 1 (Hard top-D selection under isotropy, informal)**.** *Under an idealized, exactly enforced isotropy constraint and no ties among the closest competitors, the predictive loss is minimized by representing exactly the D candidate features with the largest fraction of variance attributable to image identity rather than within-image nuisance, and no others. Introducing a feature more view-stable than the least stable currently represented one evicts that feature and no other.*

Full assumptions, proof, and the discreteness and projector-probe caveats are in Appendix B. The selection criterion is a ratio, invariant to rescaling, so it is blind to amplitude: whitening puts a faint but perfectly view-stable feature on equal footing with a loud one. This is the one respect in which the result departs from a dimension-count budget rather than restating it, and it is the reason a low-opacity watermark can displace object semantics at all.

5.2 MAIN IDEA

Proposition 1 selects directions by how stable they are across one image's own crops, and grants no credit for anything else. That selection alone cannot tell apart two directions that look identical by it. A real object feature is stable within an image and, because similar images share the same objects, should also be stable across that image's near neighbours. A private pattern, a per-image key rendered identically in every crop of that one image, is exactly as stable within the image, but no other image's crops confirm it. We add a second check the isotropy budget does not otherwise apply: whether an image's own neighbours confirm a direction, not only whether the image agrees with itself.

Recall LeJEPA's loss, $L = 0.05 \cdot \text{SIGReg} + 0.95 \cdot L_{\text{pred}}$. We add one term to it and change nothing else in the recipe.

We set up notation. For a batch of B images, let $p_{v,b} \in \mathbb{R}^D$ be the projector output of view $v \in \{1, \dots, V\}$ of image b, let $\bar{p} = \frac{1}{BV} \sum_{v,b} p_{v,b}$ be the batch mean, and write $\tilde{p}_{v,b} = p_{v,b} - \bar{p}$ for the centered crop. For C_{cn}, let $t_b \in \mathbb{R}^D$ be the mean embedding of image b's k nearest neighbours in a persistent queue of other images' embeddings, held under a stop-gradient, let $\bar{p}_b = \frac{1}{V} \sum_v \tilde{p}_{v,b}$ be image b's own average centered crop, and write $\tilde{t}_b = t_b - \frac{1}{B} \sum_{b'} t_{b'}$. The queue and the top-$k$ retrieval are the same mechanism NNCLR (Dwibedi et al., 2021) and MSF (Koohpayegani et al., 2021) use to find a positive pair; we reuse it to measure what a direction is private to versus shared with, not to pick a target to predict.

Definition 1 (Cross-crop and cross-neighbour covariance)**.**

$$C_{cc} = \frac{1}{BV(V-1)} \sum_{v \neq w} \sum_{b=1}^{B} \tilde{p}_{v,b} \, \tilde{p}_{w,b}^{\top}, \qquad C_{cn} = \frac{1}{2B} \sum_{b=1}^{B} \left(\bar{p}_b \, \tilde{t}_b^{\top} + \tilde{t}_b \, \bar{p}_b^{\top} \right).$$

C_{cc} sums over ordered pairs of an image's own crops, so it is symmetric by construction and carries gradient. C_{cn} pairs two different objects, an image's own average crop and its neighbours' mean, so we symmetrize it explicitly, since Δ below is only meaningful as a symmetric matrix; it is fully detached regardless, neither side receives gradient.

In other words, C_{cc} measures what an image reproduces across its own crops, and C_{cn} measures what its neighbours reproduce of the same thing. We form

$$\Delta = C_{cc} - \text{sg}(C_{cn}).$$

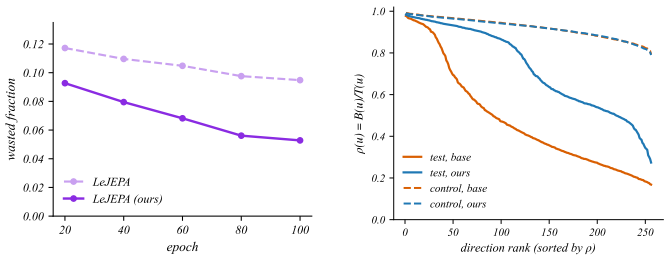

Figure 2: (a) Wasted fraction, $\sum_j (\lambda_j^+)/\mathrm{tr}(C_{cc})$, across training, DigitOnImageNet's private-glyph condition. Ours carries less private, uncorroborated capacity than LeJEPA at every checkpoint measured. (b) Sorted view-stability ρ of every retained direction, test versus control, before and after our penalty. The test condition's spectrum sits below the control condition's across most of its length, not only at the margin.

A direction where Δ is large and positive is one an image reproduces on its own but its neighbours do not: variance private to that image. A direction where Δ is at or below zero is at least as consistent among an image's neighbours as within the image itself, the signature of shared, class-like structure. We add

$$\mu \sum_j (\lambda_j^+)^2$$

to LeJEPA's loss, where λ_j^+ are the positive eigenvalues of Δ only. Δ is a proxy for the private variance Proposition 1 identifies, an image's own crops standing in for its views and a neighbour queue standing in for the population the theory's ICC is defined against, not a direct measurement of it. The penalty prices this proxy and leaves shared variance, whatever its stability, untaxed; it does not remove a direction architecturally, the way a smaller projector or a hand-picked subspace would.

6 Experiments

TileOnImageNet and DigitOnImageNet experiments (Table 2 and Appendix D's Table 4) train ViT-S/16 on ImageNet-100, plain views, no multi-crop; the clean-data control (Table 1) trains ViT-B/16 on the full ImageNet-1k, 2 global and 6 local multi-crop views. Appendix C gives the full recipe for each.

6.1 Does the Penalty Reduce Obsession?

Figure 2 compares LeJEPA against our penalty at matched checkpoints, tracking the quantity the penalty is built to move directly, rather than reading its effect off downstream accuracy alone: the wasted fraction, $\sum_j (\lambda_j^+)/\mathrm{tr}(C_{cc})$, the share of an image's own crop-variance that is private rather than confirmed by its neighbours. Ours carries less of it at every checkpoint measured – 0.053 against 0.095 by epoch 100, a 44% reduction – and the gap is already visible at epoch 20, before most of training has happened.

One limitation belongs here, beside the number it qualifies, rather than beside the recipe in Section 5. μ is not hand-tuned: it is set once, at the first training step the neighbour queue fills, to a fixed fraction of the invariance loss measured at that instant, and held fixed for the rest of the run. That instant is also the queue's noisiest, populated by the least-trained embeddings the run will ever produce, so every number the penalty earns in this section may be conservative; whether recalibrating μ later in training would recover more has not been tested, only argued from when it is set.

6.2 Performance on Downstream Tasks

Table 1 reports the control every fix owes before anything else: does the penalty cost accuracy when there is no competing feature to correct? On a full ImageNet-1k pretraining run, ViT-B/16, 100 epochs, with nothing injected, the penalty costs 0.20 points of linear-probe accuracy against the LeJEPA baseline, while raising RankMe's effective rank from 288 to 338 (Garrido et al., 2023a) and making the embedding more uniform on Wang & Isola's own measure (Wang & Isola, 2020).

Table 1: Linear-probe top-1 accuracy and representation health on clean ImageNet-1k (no injected competing feature), ViT-B/16, 100 epochs.

System	Top-1	RankMe	Uniformity
LeJEPA	71.07	288.01	−0.248
LeJEPA (ours)	70.87	338.19	−0.509

Table 2: Linear-probe top-1 accuracy (%) on clean ImageNet-100 validation images, for encoders trained with the TileOnImageNet pattern present. The test condition repeats one tile across the frame, so the same key survives every crop; the control condition resamples the tile at each position, matching its magnitude and coverage while removing the periodicity that makes it learnable. Appendix A gives the full construction.

System	Test	Control
I-JEPA	50.02	47.02
I-JEPA + VICReg	37.46	29.64
LeJEPA	**58.46**	**70.00**
LeJEPA (ours)	**65.78**	**68.70**

A small, real cost, paid for a representation that looks healthier by two independent, label-free measures.

Appendix D's Table 4 reads two signals off the DigitOnImageNet stimulus, which has no control condition (Section 3): how much accuracy the digit costs, and how well its features still decode the digit itself. LeJEPA's own gap is far smaller here than on TileOnImageNet, 0.32 points against 11.54. The two competing features differ in exactly the property the mechanism ranks on: TileOnImageNet's pattern covers the entire frame, so it survives any crop at any scale intact, while DigitOnImageNet's glyph sits at nine fixed, small positions a random crop can miss entirely. A feature a crop can miss is a feature whose view-stability, and so its claim on the isotropy budget, is itself weaker; we read the smaller gap as consistent with this, not as independent confirmation of it.

7 Discussion

Isotropy earns its keep by ruling out collapse, but the same move opens the door this paper studies: the invariance term already prefers whichever direction survives an image's own views most reliably, and isotropy turns that soft preference into a competition for a fixed budget, one a reliable but useless direction can win outright. Any joint-embedding objective that adds an isotropy-style constraint inherits this risk, negatives or not, since the constraint never asks what a direction is good for, only how stable it is. Masked prediction escapes here because context and target are drawn from disjoint parts of one image, a geometric fact about this stimulus, not a guarantee against every masked objective.

The competing features here are planted, but the mechanism is not: a capture artifact or a compression trace, reproduced identically across an image's own crops and uncorrelated with its neighbours, is exactly the direction Section 4 shows an isotropy-constrained objective rewards. We have not tested that claim on naturally occurring shortcuts. Whether more embedding dimensions or longer training relieve the competition, Section 6's open questions, is the natural next step toward finding out.

8 Conclusion

Feature suppression is not a property of instance discrimination alone. It appears under LeJEPA, which pushes full views to agree under an isotropic constraint, and disappears under plain masked prediction, which agrees with nothing and predicts across a gap instead; adding VICReg's isotropy to masked prediction suggests it returns, though on a single stimulus and a single seed, evidence to weigh rather than settle on. The common thread is the invariance term, not negatives: isotropy hardens its existing preference for view-stable directions into a fixed budget a merely reliable feature can win outright. A penalty taxing exactly the variance an image's neighbours do not confirm

recovers some of what this costs, at a small price on clean accuracy, evidence for the mechanism rather than a settled fix.

AI USE STATEMENT

We used generative AI to aid and polish writing, and for retrieval and discovery (e.g., finding related work). All AI-assisted text and citations were checked by the authors against the underlying sources and experimental logs before inclusion; the authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This work trains self-supervised encoders on ImageNet-100 and ImageNet-1k, both standard, publicly available benchmarks, and studies a planted, synthetic competing feature rather than any sensitive or personally identifying attribute. We do not foresee ethical concerns beyond those already attached to the base datasets and pretraining recipes this work builds on.

REPRODUCIBILITY STATEMENT

Appendix A gives the full construction of both competing features and the planted-feature protocol; Appendix C gives the training recipes and hyperparameters for every system compared. Appendix B states every assumption behind the theoretical results and proves them in full. Code is included in the supplementary material.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 (ICLR)*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Ky Fan. On a theorem of Weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

P. A. Fillmore and J. P. Williams. On operator ranges. *Advances in Mathematics*, 7(3):254–281, 1971.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023a.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2023b.

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.

Katherine L. Hermann and Andrew K. Lampinen. What shapes feature representations? exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alfred Horn. Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76(3):620–630, 1954.

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.

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Alexander C. Li, Alexei A. Efros, and Deepak Pathak. Understanding collapse in non-contrastive siamese representation learning. In *European Conference on Computer Vision (ECCV)*, 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1411–1420, 2023.

Shentong Mo and Shengbang Tong. Connecting joint-embedding predictive architecture with contrastive self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

Alex Tamkin, Margalit Glasgow, Xiluo He, and Noah Goodman. Feature dropout: Revisiting the role of augmentations in contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, 2020.

John von Neumann. Some matrix-inequalities and metrization of matric space. *Tomsk University Review*, 1:286–300, 1937.

Vincent Q. Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16041–16050, 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, 2020.

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 *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023.

Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A Dataset and Planted-Feature Construction

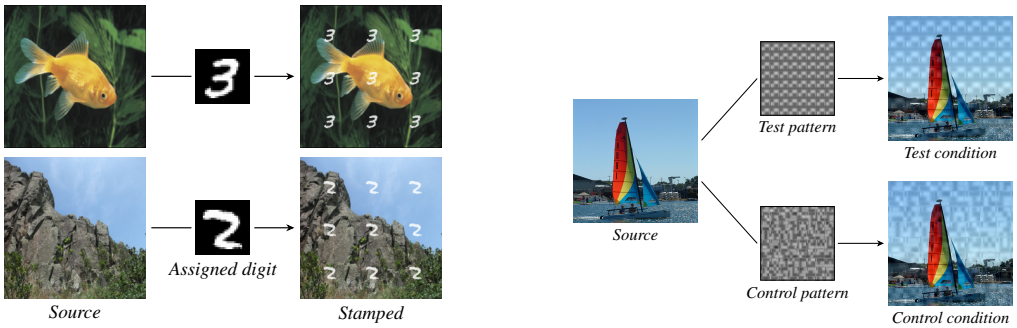

Figure 3: Construction of the two competing-feature stimuli. (a) DigitOnImageNet (Chen et al., 2021): one MNIST digit is assigned to each image and blended in at nine fixed locations before the augmentations that create the views, digits drawn from the whole MNIST training split. (b) TileOnImageNet: a 12-bit key computed from the image sets the pattern of a small tile added to the source before augmentation. The test condition repeats one tile across the frame, so every crop at least one tile wide carries the same 12 bits; the control condition resamples the pattern at each tile position, leaving the magnitude and the frame coverage unchanged and removing only the periodicity. Added fields are shown amplified $2\times$ for visibility.

A.1 DigitOnImageNet

Geometry. We follow the released code of Chen et al. (2021). Each image is resized to 224×224 before the digit is overlaid, and the digit is kept at its native 28×28, so it occupies 0.125 of the frame side for every image in the set. The nine copies are placed with their top-left corners at $\lfloor 224r \rfloor$ for $r \in \{0.15, 0.45, 0.75\}$ along each axis, giving offsets of 33, 100 and 168 px. Resizing before compositing fixes that ratio, since the shorter side of ImageNet-100 (Deng et al., 2009; Tian et al., 2020) ranges from 128 to 2304 px and a 28 px digit composited on the raw source would vary by an order of magnitude in extent once the image reached a 224 view. The digit is achromatic, a single plane composited into all three channels identically, and it is applied to the source before `RandomResizedCrop`.

Coverage under cropping. With a crop scale of $(0.3, 1.0)$ and aspect ratios in $(3/4, 4/3)$, the smallest crop has area 0.3×224^2 and, at the extreme aspect ratios, a shortest side of about 106 px. Such a crop begins no later than $x = 118$, while leaving all three digit bands, $[33, 61]$, $[100, 128]$ and $[168, 196]$, uncontained would require beginning past $x = 168$. Every crop therefore contains a whole band along each axis, and since the nine positions are the full product of the three bands on each axis, it contains at least one whole copy of the digit. The guarantee depends on the crop scale: under the $(0.08, 1.0)$ scale of Chen et al. (2021) the shortest side falls to about 55 px and 4.3% of crops carry no whole copy.

Compositing. Chen et al. (2021) describe channel addition and their released code uses an elementwise maximum. We use neither. A stroke of intensity g over a pixel of value I gives $(1 - g)I + g$, where g is the MNIST intensity rescaled to a peak of 0.85, so the digit replaces the image beneath it in proportion to its own intensity. Under the maximum a digit survives only where it is brighter than the image, which makes both its coverage and its amplitude depend on the content behind it; channel addition saturates on bright regions with the same consequence.

Assignment. The digit pool is the whole MNIST (LeCun et al., 1998) training split, $60{,}000$ digits, drawn under a fixed seed so that the evaluation probe can rebuild the identical pool in a separate process. The image-to-digit table is built once over the dataset and held fixed for the run. This follows the paper, which assigns each image a digit; the released code instead draws a fresh digit on every call, under which the digit would be shared by no two views and the size of the pool would control nothing. Digits are dealt evenly across images rather than by a hash of the image, so the number of images per digit is fixed by construction. With $60{,}000$ digits over $130{,}000$ images, each digit is carried by two or three images.

Evaluation. Validation images are stamped with digits drawn from the MNIST test split, so no digit seen during pretraining reappears at evaluation and the digit probe cannot be answered by recalling a bitmap. Assignment at evaluation is random with respect to class. This departs from Chen et al. (2021), who evaluate on the training digits.



Figure 4: DigitOnImageNet (Chen et al., 2021) at the amplitude used in training. Each image carries one MNIST digit, replicated at the nine positions described above, and a different digit is assigned to each image. The middle panel is the weakest case in the set: a digit with thin strokes over a bright background stays faint even at a peak stroke intensity of 0.85.

A.2 TileOnImageNet

Key. The key is a hash of the last two components of the image's file path, its class directory and its filename, which makes it independent of image content and stable across data roots so that it can be recomputed at evaluation from the path alone. We take 12 bits, giving a key space of 4096. The key is therefore image-dependent rather than unique: roughly 30 images share a key across the $130{,}000$ training images of ImageNet-100 (Deng et al., 2009; Tian et al., 2020), and roughly 310 across the $1{,}281{,}167$ training images of ImageNet-1k (Deng et al., 2009).

Tile. The tile is square, with side one eighth of the image's shorter dimension, matching the ratio of a 32 px tile on a 256 px source. Its pattern is a superposition of 12 wrapped Gaussian bumps on a 3×4 lattice, one per bit, each with standard deviation one eighth of the tile side and centred at

$(i + \frac{1}{2})/3$ and $(j + \frac{1}{2})/4$ of the tile. A bit set to one contributes its bump with positive sign and a bit set to zero with negative sign. The bump on the first lattice cell is replaced by an oriented Gabor, which fixes the origin: a crop meets the lattice at an arbitrary phase, and without an anchor the bit order would not be recoverable under the tile's translation symmetry.

Amplitude. The field is centred to zero mean and scaled so that its root mean square equals the opacity, which we set to 0.05. The root mean square is fixed rather than the peak, which leaves the two conditions below matched in energy by construction. The field is achromatic and added to all three channels identically.

Conditions. In the test condition one tile is generated from the image's key and repeated across the frame, so the field is seamless and spatially periodic. In the control condition a separate tile is generated at each position, from a key formed from the image together with the row and column of that position, so the per-tile magnitude and the frame coverage are unchanged and only the periodicity is removed.

Position in the pipeline. The field is applied once to the source image, before augmentation, in both conditions. Because the control field is aperiodic rather than redrawn at each view, the two conditions pass through an identical pipeline and differ only in the field itself.

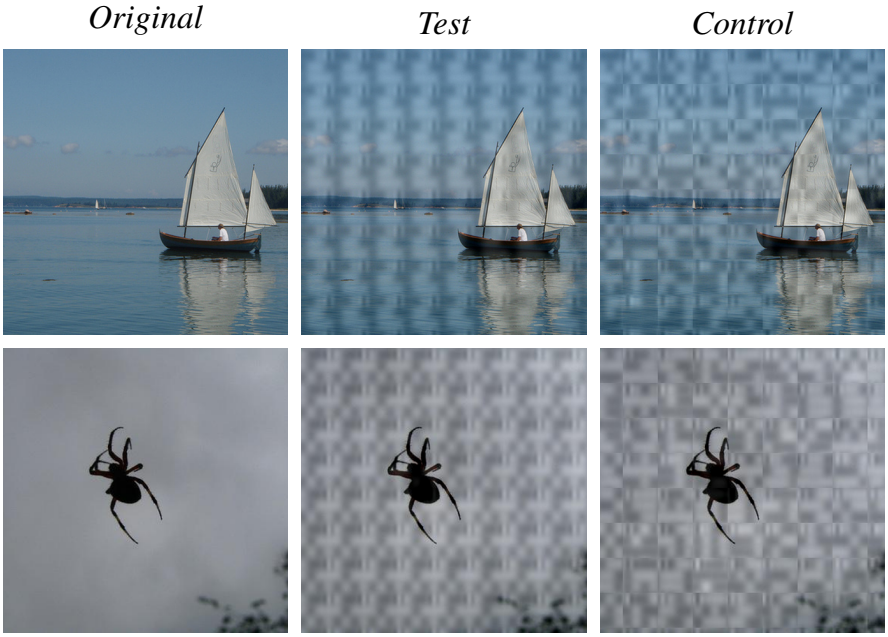

Figure 5: TileOnImageNet at the opacity used in training, on two images. The field is most visible over flat regions such as sky and uniform background, and is close to imperceptible over textured ones. In the test condition one tile is repeated across the frame, giving a regular grid; in the control condition a separate tile is generated at each position, so the field is matched to the test condition in per-tile magnitude and in frame coverage but is not periodic.

B Theoretical Grounding

This section formalizes and proves the allocation result stated informally as Proposition 1 in Section 5: what SIGReg's isotropy constraint does to the predictive loss's choice of which directions to represent, and what happens to that choice when a new competing feature is introduced. Xue et al. (2023, Theorem 5.1) prove an analogous displacement result under a dimension-count budget, where unused coordinates carry exactly zero variance at no cost; what follows isolates what changes when the budget is instead a hard constraint on the total variance of every retained direction.

Crops are nested in images. For a candidate direction $u \in \mathbb{R}^d$, let $\Sigma_B = \mathrm{Cov}_n(\mathbb{E}_v[x \mid n])$ and $\Sigma_W = \mathbb{E}_n \mathrm{Cov}_v(x \mid n)$ denote the between-image and within-image covariance of the pre-encoder

representation x, and $\Sigma_T = \Sigma_B + \Sigma_W$. Write $B(u) = u^\top \Sigma_B u$, $T(u) = u^\top \Sigma_T u$, and $\rho(u) = B(u)/T(u)$, the fraction of u's variance attributable to image identity rather than within-image nuisance across views – an intraclass-correlation quantity (Shrout & Fleiss, 1979), with an image's views standing in for a rater panel; under Lemma 1's $C = 0$ assumption below, it is also exactly the correlation between two views' own values of u, the simplest ICC form Shrout and Fleiss consider. This decomposition is definitional, the law of total variance, and since $\Sigma_T - \Sigma_B = \Sigma_W \succeq 0$, $\rho(u) \in [0, 1]$ wherever $T(u) > 0$; everything below is stated on the support where $T(u) > 0$, so this ratio is always well defined where it is used. All quantities below are population quantities under V views with finite second moments, conditionally exchangeable given image identity n (not merely exchangeable unconditionally, which alone would not pin down $\mathbb{E}[x_v \mid n]$), so Σ_W does not depend on the view index; as in Xue et al.'s own theorems, this is a statement about the population optimum, not about an SGD iterate, and it says nothing about whether training reaches it.

Lemma 1 (The prediction loss is within-image variance)**.** *Assume, in addition to the conditional exchangeability (given n) already in force, that the V views share a common conditional mean given image identity, $\mathbb{E}[x_v \mid n] =: m(n)$ not depending on v (a property conditional exchangeability itself already implies, by symmetry). The predictor is trained to minimize*

$$L(W) = \mathbb{E}\Big[\frac{1}{V}\sum_{v=1}^{V} \|Wx_v - W\bar{x}\|^2\Big], \qquad \bar{x} = \frac{1}{V}\sum_v x_v,$$

the average squared distance from each view's representation to their common mean. Write $C := \mathbb{E}_n \operatorname{Cov}(x_v, x_{v'} \mid n)$ for $v \neq v'$, the average conditional cross-covariance between two views of the same image. Then $L(W) = \frac{V-1}{V} \operatorname{tr}\big(W(\Sigma_W - C)W^\top\big)$. If views are additionally conditionally independent given image identity, so that $C = 0$, this is

$$L(W) = \frac{V-1}{V} \operatorname{tr}(W\Sigma_W W^\top).$$

Proof. Write $y_v = x_v - m(n)$ for the within-image residual, using the common conditional mean assumed above, and $\bar{y} = \frac{1}{V}\sum_v y_v = \bar{x} - m(n)$. Since $m(n)$ cancels from the difference, $x_v - \bar{x} = y_v - \bar{y}$ for every v; without a common conditional mean this step fails; two views deterministically fixed at different values, say, are trivially uncorrelated and would give $C = 0$ while still contributing a gap between per-view means that is not within-image variance in any symmetric sense, exactly the case the shared-mean hypothesis rules out. Since $\frac{1}{V}\sum_v \|z_v - \bar{z}\|^2 = \frac{1}{V}\sum_v \|z_v\|^2 - \|\bar{z}\|^2$ for any vectors z_v with mean $\bar{z}$, taking $z_v = Wy_v$ gives

$$L(W) = \mathbb{E}\Big[\tfrac{1}{V}\sum_v \|Wy_v\|^2\Big] - \mathbb{E}\|W\bar{y}\|^2 = \operatorname{tr}(W\Sigma_W W^\top) - \operatorname{tr}\big(W \operatorname{Cov}(\bar{y})\, W^\top\big).$$

By exchangeability, $\operatorname{Cov}(\bar{y}) = \frac{1}{V}\Sigma_W + \frac{V-1}{V}C$, and substituting gives the claimed identity; $C = 0$ simplifies it further. □

For $V = 2$ this is one quarter of Wang & Isola's alignment loss at their $\alpha = 2$, the expected squared distance between a positive pair's embeddings (Wang & Isola, 2020): $\bar{x} = (x_1 + x_2)/2$ puts each view a distance $\|Wx_1 - Wx_2\|/2$ from the mean, so $L(W) = \frac{1}{4}\mathbb{E}\|Wx_1 - Wx_2\|^2$. The lemma is that loss's generalization to $V > 2$ views sharing one mean, up to this fixed rescaling, and it is this identity, loss equals within-image variance, that lets Section 4 read $\rho(u)$ off the loss's own preference rather than introducing it separately.

Exchangeability alone does not make C vanish: two views built from one shared random augmentation (so that $x_1 = x_2$ exactly, given n) are exchangeable, yet $C = \Sigma_W \neq 0$ in general and $L(W) \equiv 0$ for every W, since nothing about such views is ever disagreed with. We assume $C = 0$ throughout what follows – views are conditionally independent given image identity, which is ensured whenever each view's augmentation is drawn independently, the setting this paper's stimuli are built in. The competing feature itself is, in every system studied here, a deterministic function of image identity (the watermark's payload is keyed off a hash of the image, never sampled on its own), so conditioning on n already determines it; there is no separate source of randomness left unconditioned on.

Lemma 2 (The constrained problem is a Fantope program)**.** *Assume there are $K > D$ generative factors whose directions $v_1, \dots, v_K$ form an orthonormal basis of $\mathbb{R}^d$ (so $K = d$), a linear readout $z = Wx$ with $W \in \mathbb{R}^{D\times d}$, and factor coordinates $c_k = v_k^\top x$ mutually uncorrelated at both levels: $\Sigma_B = \sum_k b_k v_k v_k^\top$, $\Sigma_T = \sum_k t_k v_k v_k^\top$ with every $t_k > 0$ (so Σ_T is invertible), and $\rho_k := b_k/t_k$. Assume further that SIGReg enforces $\mathrm{Cov}(z) = I_D$ exactly. Then, under Lemma 1's conditional-independence assumption, minimizing $L(W)$ subject to this constraint is equivalent to*

$$\max_{A\in\mathcal{F}_D} \langle A, S\rangle, \qquad S := \Sigma_T^{-1/2}\Sigma_B\Sigma_T^{-1/2}, \qquad \mathcal{F}_D := \{A = A^\top : 0 \preceq A \preceq I,\ \mathrm{tr}\, A = D\},$$

with $A = \Sigma_T^{1/2} W^\top W \Sigma_T^{1/2}$. Under the factor-basis assumption above, $S = \sum_k \rho_k v_k v_k^\top$, and writing $\pi_k := v_k^\top A v_k$ for the fraction of a unit-variance slot occupied by factor k, the problem is

$$\max_{\pi} \sum_{k=1}^{K} \rho_k \pi_k \qquad \textit{s.t.} \qquad \pi \in [0,1]^K, \quad \sum_k \pi_k = D.$$

Proof. By Lemma 1, $L(W) = \frac{V-1}{V}\mathrm{tr}(W\Sigma_W W^\top)$. Since $\Sigma_W = \Sigma_T - \Sigma_B$ and $W\Sigma_T W^\top = \mathrm{Cov}(z) = I_D$ fixes $\mathrm{tr}(W\Sigma_T W^\top) = D$, minimizing $L(W)$ is equivalent to maximizing $\mathrm{tr}(W\Sigma_B W^\top)$ subject to the same constraint. With $M = W\Sigma_T^{1/2}$, the constraint becomes $MM^\top = I_D$, and the objective becomes $\mathrm{tr}(MSM^\top) = \langle M^\top M, S\rangle$. As M ranges over matrices with orthonormal rows, $A = M^\top M$ ranges over the rank-D orthogonal projections of $\mathbb{R}^d$; these are exactly the extreme points of the Fantope $\mathcal{F}_D$ (Vu et al., 2013; Fillmore & Williams, 1971), whose convex hull they generate, so a linear functional's supremum over $\mathcal{F}_D$ is attained at one of them and equals its supremum over the projections alone – the relaxed and original problems share an optimal value, and an actual rank-D orthogonal readout attains it, even though a generic point of $\mathcal{F}_D$ (e.g. $\frac{1}{2}I_2$ for $D = 1$ in two dimensions) is not itself a projection and so does not correspond to any readout. This is where the factor-basis assumption first enters: the v_k diagonalize S with $Sv_k = \rho_k v_k$ only because Σ_B and Σ_T were assumed to share that eigenbasis, so $\langle A, S\rangle = \sum_k \rho_k \pi_k$; by Horn's theorem (Horn, 1954) the achievable diagonals of $A \in \mathcal{F}_D$ in the $\{v_k\}$ basis are exactly the hypersimplex $\{\pi \in [0,1]^K : \sum_k \pi_k = D\}$. □

The Fantope equivalence itself, same optimal value, an actual projector attains it, needs only that S is symmetric and needs no shared eigenbasis at all; it is solely the *named-factor* reading, that slot k is occupied by generative factor k specifically, that requires Σ_B and Σ_T to be simultaneously diagonalizable. Correlated factors that do not share a basis with Σ_T are not individually selected by a raw ratio of variances; S's own eigenvectors, not the factors, are what Ky Fan's principle ranks, and the two coincide only under the assumption just used.

Lemma 2 makes the budget metaphor exact: π_k is the fraction of a slot factor k occupies, fractional occupancy is feasible for the relaxed matrix A, and total occupancy is conserved at D; Horn's theorem is what guarantees every feasible π is in turn realized by some actual projector, not only by matrices in $\mathcal{F}_D$ that are not themselves projections. Two idealizations are worth flagging before going further, since both matter later. First, SIGReg is trained as a penalty, not enforced exactly; Proposition 3 below addresses what changes if the target covariance is merely a different exact one, though not the fully penalized case. Second, SIGReg's actual target is the isotropic Gaussian, a strictly stronger condition than isotropic covariance. Solving under $\mathrm{Cov}(z) = I_D$ alone is therefore a genuine relaxation, and a relaxation's optimum need not be feasible for the original problem: a perfectly view-stable but non-Gaussian direction (the finite-support bits of a planted key, say) can be optimal here while being exactly the kind of structure a full distributional test is built to penalize. What follows is consequently a theory of a covariance surrogate for SIGReg, and its selection conclusion is not guaranteed to transfer to the true, Gaussian-targeting constraint; the paper's empirical ρ-spectrum (Section 6) should be read as testing whether it does, not as illustrating an already-settled point.

Proposition 2 (Hard top-D selection under isotropy)**.** *With $K > D$ as in Lemma 2, let $\rho_{(1)} \geq \cdots \geq \rho_{(K)}$ be the factor ICCs in decreasing order with directions $v_{(1)}, \dots, v_{(K)}$. Then:*

(i) The minimum predictive loss equals $\frac{V-1}{V}\big(D - \sum_{i=1}^{D} \rho_{(i)}\big)$: the attainable ICC total is the sum of the D largest ICCs and nothing more.

(ii) If $\rho_{(D)} > \rho_{(D+1)}$*, every minimizer has readout subspace* $\mathrm{row}(W) = \mathrm{span}\{v_{(1)}, \ldots, v_{(D)}\}$*, unique up to an orthogonal rotation within the subspace, and the allocation is a hard* $0/1$ *selection,* $\pi_{(i)} = \mathbf{1}[i \le D]$*. If instead* $\rho_{(D)} = \rho_{(D+1)}$*, every minimizer's subspace contains* $\mathrm{span}\{v_{(i)} : \rho_{(i)} > \rho_{(D)}\}$ *in full, every direction strictly above the tied cutoff is retained outright, and fills its remaining dimensions, however many directions sit strictly above the cutoff, subtracted from* D*, with any subspace of the tied eigenspace at exactly* $\rho_{(D)}$*; fractional* $\pi_k \in (0, 1)$ *within that tied block are optimal as well, but no optimum drops a direction strictly above the cutoff to make room for one at or below it.*

Proof. By Lemma 2 the problem is $\max_{A \in \mathcal{F}_D} \langle A, S \rangle$. By Ky Fan's maximum principle (Fan, 1949) this equals $\sum_{i=1}^{D} \lambda_i(S)$, giving (i). Under the factor-basis assumption, $\lambda_i(S) = \rho_{(i)}$ with eigenvector $v_{(i)}$. When $\rho_{(D)} > \rho_{(D+1)}$ the top-D eigenspace of S is exactly D-dimensional and unique, giving the first case of (ii); pulling back through $M = RU_D^\top$ for $R \in O(D)$ and $\Sigma_T^{-1/2}$'s diagonality in $\{v_k\}$ preserves the span. When $\rho_{(D)} = \rho_{(D+1)}$, Ky Fan's maximum principle forces every maximizer to act as the identity on the span of eigenvalues strictly above $\lambda_D(S)$ and as zero on the span of eigenvalues strictly below it, since moving any weight out of the strictly-above span can only lower the trace; freedom remains only within the eigenspace tied at $\lambda_D(S)$ itself, where any rank-matching subspace attains the same trace $\sum_{i=1}^{D} \lambda_i(S)$, giving the second case. □

The mechanism producing hardness is worth isolating explicitly, since three separate ingredients are at work and only one is what "isotropy" names: $\mathrm{rank}(W) \le D$ supplies the budget $\sum_k \pi_k \le D$; any full-rank constraint on $\mathrm{Cov}(z)$ forces that budget to bind at equality rather than allowing the encoder to shrink it voluntarily; and it is the linearity of the objective in π that forces an optimum to exist at a vertex of the polytope, i.e., at a $0/1$ allocation, though as Proposition 2(ii) shows, a tie can admit non-vertex optima alongside it. Isotropy's specific contribution is making the bound bind at exactly D and, per Proposition 3 below, standing between the top-D solution and complete rank-one collapse onto the single most view-stable factor.

A second consequence is the point at which this result departs from Xue et al. rather than restating it under a different budget. The selection statistic $\rho_k = b_k/t_k$ is invariant to rescaling v_k's coordinate; Xue et al.'s discriminativeness statistic is not, and their minimum-norm minimizer keeps exactly the features with the largest raw variance, zeroing every feature whose variance is not among the p largest regardless of whether it is class-relevant. A feature's fate there is decided by raw amplitude directly: a class feature can be suppressed simply for being quiet, and an irrelevant feature can survive simply for being loud. Under an isotropy constraint, amplitude plays no role in the decision at all: whitening rescales every candidate direction to unit variance before ICC is compared, so a feature's only property under selection is view-stability, independent of how faint or loud it was in the raw signal.

Proposition 3 (Selection under a prescribed positive covariance spectrum)**.** *Replace the exact-isotropy assumption with* $\mathrm{Cov}(z) = \Lambda$*,* Λ *diagonal with* $\Lambda_{11} \ge \cdots \ge \Lambda_{DD} > 0$*,* $\mathrm{tr}\,\Lambda = D$*, held exactly. Under the same non-tie condition as Proposition 2, the minimizer still has* $\mathrm{row}(W) = \mathrm{span}\{v_{(1)}, \ldots, v_{(D)}\}$*, with row* j *matched to* $v_{(j)}$ *in decreasing order; only the within-subspace weighting changes.*

Proof. With $M = \Lambda^{-1/2} W \Sigma_T^{1/2}$ the objective is $\langle M^\top \Lambda M, S \rangle$, where $A' = M^\top \Lambda M$ has eigenvalues $\{\Lambda_{jj}\} \cup \{0\}^{d-D}$. By von Neumann's trace inequality (von Neumann, 1937) this is at most $\sum_{j \le D} \Lambda_{jj} \lambda_j(S)$, with equality exactly when A' and S are simultaneously diagonalizable in matched decreasing order. □

This is a statement about an exactly enforced, merely non-isotropic covariance target, not about the penalized objective SIGReg actually optimizes in training. Real training minimizes $L(W) + \mu\, R_{\mathrm{SIGReg}}(Wx)$ for a distributional test statistic R_{SIGReg} that targets the full isotropic Gaussian, not merely a covariance value, and is enforced only approximately. Proposition 3 does not by itself establish that the top-D selection survives that penalized, distributional relaxation; it establishes only that the selection is insensitive to which particular positive-definite covariance is imposed, exactly, in place of isotropy. Extending this to the soft, approximately-enforced case is an open step, not

a corollary of what is proved here. Two further consequences follow from the proposition itself, one reassuring and one not, both still scoped to an exactly enforced target: the top-D *selection* does not depend on which positive-definite covariance is imposed, only how the retained subspace is weighted does; but if the constraint on $\mathrm{Cov}(z)$ is relaxed to a trace constraint alone, the same computation gives an optimum of $D \cdot \rho_{(1)}$, attained by putting all mass on the single highest-ICC direction: complete rank-one collapse, unique whenever $\rho_{(1)} > \rho_{(2)}$ and, at a tie, achieved by any direction in the tied top eigenspace. Isotropy's role is therefore best read not as what forces hardness, but as what stands between the top-D solution and that collapse. With no scale constraint on $\mathrm{Cov}(z)$ at all, the invariance loss alone is minimized by $W = 0$: some constraint fixing a positive scale, isotropy, a prescribed spectrum, or even a bare trace, is what keeps the problem from being trivial in the first place.

Corollary 1 (Feature suppression under isotropy)**.** *Let $\rho_{(D)} > \rho_{(D+1)}$, and let v_{k^*} be class-relevant with ρ_{k^*} not among the D largest. At every population optimum, $Wv_{k^*} = 0$ exactly, so $\mathrm{Cov}(z, c_{k^*}) = 0$ and the best linear predictor of c_{k^*} from z is constant.*

Proof. By Proposition 2(ii), $\mathrm{row}(W) = \mathrm{span}\{v_{(1)}, \ldots, v_{(D)}\}$, orthogonal to v_{k^*}; hence $Wv_{k^*} = 0$, and $\mathrm{Cov}(z, c_{k^*}) = W\Sigma_T v_{k^*} = t_{k^*} W v_{k^*} = 0$. □

This says nothing about statistical independence or about nonlinear decodability of c_{k^*} from z; it is a statement about the best affine least-squares predictor only, matching the linear probes reported throughout the paper.

This is the isotropy analogue of Xue et al.'s Theorem 5.1, and the two proofs differ in a way worth stating plainly: their loss is degenerate at the min-norm solution, any p-dimensional subspace of the relevant column space is a global minimizer, and their conclusion needs an appeal to gradient descent's implicit bias to break the tie. The predictive loss here is not degenerate: it strictly prefers high-ICC directions, so Corollary 1 follows from the optimum itself, with no appeal to an optimizer's inductive bias. The cost of that strengthening is that our constraint is enforced exactly, where their global minimizer is degenerate and their own conclusion needs gradient descent's implicit bias, not the loss alone, to break the tie in practice.

Corollary 2 (Competitive eviction is a one-for-one swap)**.** *Lemma 2 set $K = d$, so $v_1, \ldots, v_K$ already span the whole ambient space and no direction orthogonal to all of them can exist inside it; introducing a new competing feature is a statement about a larger ambient space, $\mathbb{R}^{K+1}$, in which the original K factors and their covariance structure sit unchanged as a block. Assume, in that enlarged space, the newly introduced competing feature occupies a fresh direction $v_{K+1} \perp \mathrm{span}\{v_1, \ldots, v_K\}$, uncorrelated with every existing factor at both levels, leaving b_k, t_k for $k \le K$ unchanged. Assume also, for $D > 1$, that the retained factors are not tied against each other either, $\rho_{(D-1)} > \rho_{(D)}$, so the marginal retained factor $v_{(D)}$ is itself uniquely determined before any new factor is introduced. Let the new factor have ICC $\rho_{\mathrm{new}} > \rho_{(D)} > \rho_{(D+1)}$, distinct from every existing ρ_k. Then the retained set changes by exactly $|\mathcal{S}' \triangle \mathcal{S}| = 2$: $v_{(D)}$ is evicted, the new factor enters, and every other retained factor is unchanged. If instead $\rho_{\mathrm{new}} < \rho_{(D)}$, the retained set is unchanged and the new factor is not represented at all.*

Proof. Inserting one element strictly greater than the D-th order statistic of a sorted list shifts the old D-th element to position $D+1$ and leaves positions $1, \ldots, D-1$ fixed; the added gap $\rho_{(D-1)} > \rho_{(D)}$ ensures the pre-injection D-th element is a single, uniquely identified factor rather than one member of a tied block, so that "the marginal retained factor" is well defined. Apply Proposition 2 to the resulting multiset. □

At $\rho_{\mathrm{new}} = \rho_{(D)}$ exactly, the retained set is genuinely undetermined by the optimum (Proposition 2(ii)): this is the same discontinuity Xue et al. report empirically (their Figure 4 and Remark 5.2), where the trained alignment to either of two competing features diverges across random seeds at the corresponding threshold; we take that as precedent rather than a new observation.

B.1 Discreteness is a claim about the idealized optimum, not about training

Corollary 2's one-for-one swap holds exactly at the population optimum of the linearized problem under an exactly enforced constraint. Five distinct effects smooth it in practice, and none should be

read as already ruled out: finite-sample noise in $\hat{\rho}_k$ randomizes the argmax whenever $|\rho_{\text{new}} - \rho_{(D)}|$ is within sampling error, exactly as in Xue et al.'s own seed variation; near-ties are ill-conditioned, with the retained subspace's sensitivity to perturbation scaling as $1/(\rho_{(D)} - \rho_{(D+1)})$; SIGReg is enforced only approximately in training (measured $\lambda_{\max}(\mathrm{Cov}(z)) = 2.24$ against the idealized 1) – Proposition 3 shows the top-D selection is stable across any exactly enforced positive-definite covariance target, which is suggestive but does not by itself cover a penalized, distributional target enforced only approximately; a nonlinear encoder can host superposed factors at some distortion cost, so the budget binds smoothly rather than as a hard constraint; and nothing here describes whether gradient descent reaches this optimum at all, only what the optimum is. The paper's claims should read accordingly: the isotropy budget predicts a graded exchange that sharpens toward a discrete swap as the constraint is enforced more strictly, not a step function observed directly in any single run.

B.2 What the theory does not cover: the projector-probe gap

Every proposition above concerns the space on which SIGReg and the predictive loss act, the projector output. Every reported probe number in this paper, linear accuracy and key-tracking alike, is measured on the backbone output, one layer earlier. If the projector were linear and full row rank, the backbone representation would retain a null space of dimension $\dim(h) - D$ (128 of 384 at our ViT-S/16, $D = 256$ setting) on which Corollary 1 says nothing: a factor excluded from the projector's top-D could in principle survive untouched one layer back. The paper's projector is a nonlinear MLP, so this is at best a local, Jacobian-level picture rather than a single global null space, and we use it as an informal guide to scale, not a computed quantity.

This is not a corollary of anything proved above; it is a gap the theory leaves open. The paper's proj_dim sweep (Section 6) speaks to it, since raising D shrinks this linearized null space in the informal sense above, but raising D also changes the optimization itself, not only the size of what is excluded, so the sweep cannot isolate the null-space mechanism on its own. Separately, under the additional, plausible but unproven assumption that the class-relevant feature's rank among the retained factors does not worsen as D grows, Corollary 2 suggests the displacement gap should shrink as D increases; this is an empirical prediction resting on that added assumption, not a direct consequence of the corollary itself, which describes a single insertion at one fixed D rather than the effect of varying D, and a non-monotonic counterexample is easy to construct. A null result on the sweep would weaken the case for the mechanism under that added assumption; it would not by itself falsify Corollary 2.

C Training Recipes and Hyperparameters

Every LeJEPA variant (base and ours) shares one objective, $L = 0.05 \cdot \mathrm{SIGReg} + 0.95 \cdot L_{\text{pred}}$, AdamW, and a projector of width 2048–2048–256; runs differ only in architecture, view protocol, dataset scale, and whether the penalty (Section 5) is enabled. Table 3 gives the full recipe for the clean-data control (Table 1); the TileOnImageNet and DigitOnImageNet experiments (Tables 2, 4, 5) use the same objective and optimizer at the ImageNet-100 scale: ViT-S/16, 8 plain views (no multi-crop), batch 256, 100 epochs, projector dimension 256.

Probe protocol: every reported number uses a linear probe on frozen features, standardized using train-split statistics only, with the best of three learning rates (0.3, 1, 3) selected on a held-out validation split, identical across systems and conditions.

D Additional Results

Table 4 reads two signals off the DigitOnImageNet stimulus (Chen et al., 2021), using the same encoders as Table 2: how much classification accuracy the digit costs, and how well its features still decode the digit itself. A mechanism that suppresses through the route Section 4 describes should move both together; Section 6 discusses the result.

Table 5 extends Table 1's two representation-health measures to a fuller set, read on DigitOnImageNet under three sharing conditions for the competing feature: private (each glyph sits on 2–3 images), randomly-shared (100 glyphs, 1,300 images each, uncorrelated with class), and class-shared

Table 3: Training recipe for the clean-data control (Table 1), ViT-B/16 on ImageNet-1k.

config	value
optimizer	AdamW
epochs	100
learning rate	5e-4 $\rightarrow$ 5e-7 (linear warmup, cosine decay)
warmup	10 epochs
weight decay	5e-2 (constant)
batch size	512 (global)
views	2 global @ 224^2, 6 local @ 96^2 (multi-crop)
crop scale	global $(0.4, 1.0)$, local $(0.05, 0.4)$
patch size	16
projector	width 2048, 2 hidden layers, output dim 256
SIGReg weight	0.05
penalty weight μ (ours only)	auto-scaled, see Section 6

Table 4: Linear-probe top-1 accuracy (%) on the DigitOnImageNet stimulus (Chen et al., 2021), using the same encoders as Table 2. This stimulus has no control condition (Section 3), so we instead read two signals off a single stamped-trained checkpoint: how much classification accuracy the digit costs, and how well its features still decode the digit itself.

System	Clean	Stamped	MNIST decode (%)
I-JEPA	51.32	51.82	81.40
I-JEPA + VICReg	34.98	34.36	65.88
LeJEPA	63.82	63.50	90.44
LeJEPA (ours)	62.12	62.22	54.00

(100 glyphs, one per class, so every image of a class carries the same glyph). Higher RankMe and stable rank indicate a richer representation (Garrido et al., 2023a); an α_{ReQ} nearer 1.0 indicates a better-conditioned spectrum; a more negative uniformity indicates a better spread on the hypersphere (Wang & Isola, 2020); a lower top-1 variance fraction indicates less single-feature dominance.

Table 5: Representation health on DigitOnImageNet's three sharing conditions, frozen encoder, clean validation features.

Condition	System	RankMe	Stable rank	α_{ReQ}	Top-1 var.	Uniformity
Private	LeJEPA	190.25	2.09	2.034	0.045	-1.984
Private	LeJEPA (ours)	215.49	3.30	1.939	0.036	-2.677
Random	LeJEPA	189.26	2.03	2.038	0.042	-1.936
Random	LeJEPA (ours)	214.26	3.24	1.941	0.035	-2.650
Class	LeJEPA	205.60	4.21	2.063	0.046	-2.917
Class	LeJEPA (ours)	221.72	5.39	1.960	0.043	-3.139

Ours improves every measure in all three conditions. The one exception, not in this table, is colour-invariance alignment (the cosine similarity between a positive pair's embeddings under colour augmentation alone), which drops by about 0.02–0.03 consistently. A reasonable cost: colour jitter is itself a shared-but-noisy augmentation, not a class-relevant feature, and the penalty was never built to protect it.

Table 6 extends Table 1 on the same clean ImageNet-1k run, with the measures that table left out. The picture there is mixed rather than a clean sweep: ours lowers the top-1 variance fraction, but its alignment distance is worse, not better (0.0055 to 0.0095), and α_{ReQ} moves slightly away from 1.0 rather than toward it. We report this plainly rather than selecting only the favourable measures.

Figure 6 reads the same two capacity measures over the course of training, on TileOnImageNet's quadruple. Both measures sit well below their own control throughout training for every system, the same suppression signature Section 4 reports in accuracy; ours sits above the baseline in the test condition specifically, where the competing feature is occupying part of the budget, and close to it in the control condition, where there is little room to improve.

Table 6: Representation health on clean ImageNet-1k (ViT-B/16, 100 epochs), extending Table 1.

System	Eff. rank (cov.)	α_{ReQ}	Top-1 var.	AB-cosine	Alignment
LeJEPA	211.88	1.605	0.053	0.997	0.0055
LeJEPA (ours)	197.48	1.744	0.048	0.995	0.0095

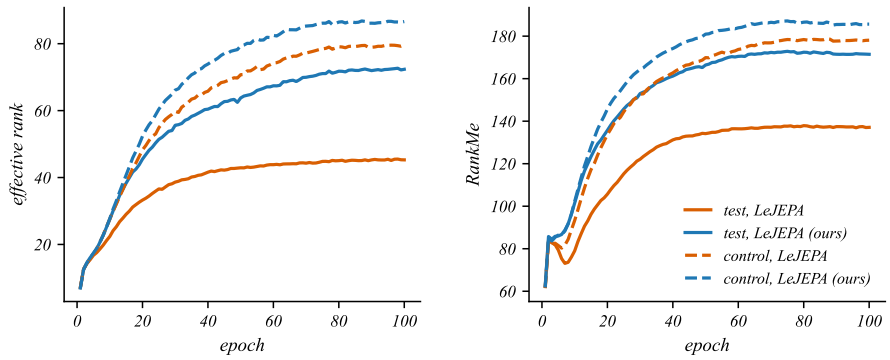

Figure 6: Effective rank and RankMe of the encoder representation, over training, TileOnImageNet's test and control conditions.

Three further results are still owed and not yet run: a bit-count sweep quantifying allocation as a function of the planted feature's information content; additional backbones and latent widths, including the wider-embedding mitigation at every width tested, the proj_dim sweep flagged in Section 6 as the highest-priority remaining experiment; and per-seed breakdowns behind the aggregated curves reported in the main text, which currently rest on a single seed per condition.

E Additional Representation Analyses

Figure 7 shows the training-loss crossover Section 4 describes: LeJEPA's test-condition invariance loss drops below its control condition early and stays below it, while masked I-JEPA's two conditions track one another for the whole run.

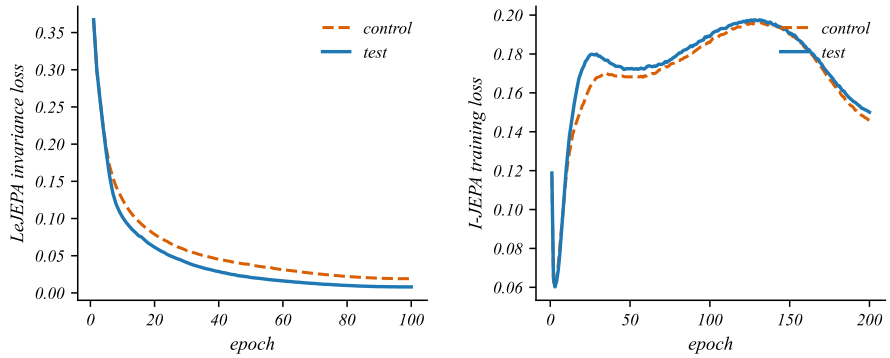

Figure 7: Training loss under the test and control conditions. LeJEPA's test-condition loss drops below its control condition early and stays below it; masked I-JEPA's two conditions track one another for the whole run.

Figure 8 unpacks Figure 2's wasted fraction into the spectrum it sums: the positive eigenvalues of $\Delta = C_{cc} - \mathrm{sg}(C_{cn})$ under our penalty, sorted by rank, at an early and a late checkpoint of the same run. Training pushes the largest private directions down fastest – the leading eigenvalue falls by roughly a factor of three – while the long tail of small positive directions barely moves, exactly the pattern a penalty whose gradient is proportional to each eigenvalue predicts. The count of directions still positive falls only modestly, from 231 to 210: the penalty narrows how much any single private direction can win, more than it shrinks how many directions still compete for one.

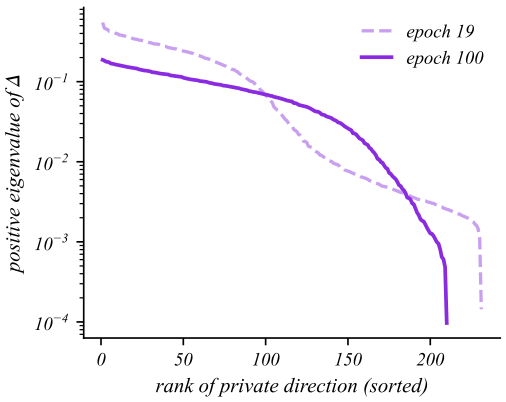

Figure 8: Sorted positive eigenvalues of $\Delta = C_{cc} - \mathrm{sg}(C_{cn})$ under our penalty, DigitOnImageNet's private-glyph condition, at an early and a late checkpoint of the same run.

A layer-wise similarity analysis, showing where in the encoder a competing feature takes hold, is not yet possible: it needs a per-epoch pair-similarity trajectory, same-image pairs against same-pattern pairs over training, and only the endpoint values exist today (Section 6).

F Extended Limitations and Negative Results

Wasted capacity does not predict accuracy across sharing conditions, only within one. The class-shared condition (Appendix D) has barely more wasted capacity at convergence than the private condition, 0.100 against 0.095, yet scores 44 points lower on stamped accuracy: the class-shared glyph is corroborated by every other image of its class, so it survives the isotropy budget on genuinely shared grounds, not by evading the penalty. Within a single condition, less wasted capacity tracks better accuracy in the class-shared case but very slightly worse accuracy in the private and randomly-shared cases, a reminder that the wasted-fraction measure diagnoses what the penalty removes, not what removing it is worth; the two only line up once a competing feature is genuinely private rather than shared.