PREDICTING THE FROZEN CEILING: A TRAIN-FREE DIAGNOSTIC FOR SELF-SUPERVISED RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

How far can frozen self-supervised features go for strict nearest-neighbor retrieval, and can that limit be known without training? We show that the retrieval ceiling collapses to a single sufficient statistic—the signal-to-spread ratio of the per-query log-distance ratio, $J = \mathbb{E}[\log(b/a)]/\sqrt{\mathrm{Var}(\log(b/a))}$—so that a closed-form diagnostic $\widehat{\mathrm{R@1}} = \Phi(J)$, computed from just five summary statistics with no training and no labels beyond one labelled query set, recovers observed strict R@1 to within 2.4 points (raw) and 5.3 points (readout) and ranks backbones at zero cost (ImageNet-100 Spearman 0.943). The diagnostic reveals that frozen DINOv2 features are linearly closed, and that the newer DINOv3 generation lands on the same plateau (0.687–0.696 across both generations): no linear post-processor improves retrieval, and only a supervised nonlinear readout opens the space (0.688→0.739). The ceiling is governed by the spread of the log-ratio, not by margins or dependence alone; raising dependence ρ alone is insufficient as shared query difficulty can drive $\rho \to 1$ without changing retrieval. The trained readout acts through regime-dependent mechanisms—ρ-driven on COCO, margin-driven on Stanford Dogs—and the theory-derived loss $\mathcal{L}_J$ performs comparably to SUPCON when trained, a difference we attribute to training moving the statistics that the closed form treats as fixed.

1 INTRODUCTION

Frozen self-supervised features have become the default front end for image retrieval: extract a backbone once, then rank images by cosine similarity (Oquab et al., 2024; Caron et al., 2020; He et al., 2022). The appeal is clear—a single feature extractor serves every downstream retrieval task at zero training cost. Yet a basic question remains unanswered: *how far can a frozen embedding actually go for strict retrieval, and can that limit be known before any training is spent?* Without an answer, practitioners cannot tell whether a post-processor is recovering hidden structure or merely exploiting annotation artefacts (Ermolov et al., 2021; Zbontar et al., 2021; Huang et al., 2024).

This gap matters because retrieval research spends its effort precisely at the boundary between a frozen embedding and its potential. Two broad families of remedy are routinely tried. The first *decorrelates* or whitens the embedding, motivated by the observation that self-supervised features are anisotropic and prone to dimensional collapse (Jing et al., 2022; He et al., 2024; Wang et al., 2024a); the second attaches a *learned readout*—a linear probe (Zhai et al., 2019; Kumar et al., 2022; El Banani et al., 2024) or a supervised-contrastive head (Khosla et al., 2020)—to lift retrieval. A parallel line even characterizes the geometric structure that contrastive objectives induce (Wang et al., 2024b; Jiang et al., 2025; Pan & Cao, 2026; Cui et al., 2025; Zhang et al., 2025a; Lin & Mei, 2025).

Both families share a blind spot. Their claims that an embedding is "closed", or that a cheap operation unlocks it, rest on *loose* evaluation protocols that permit cross-class leakage and treat multi-label or super-class structure as relevance (Shi et al., 2026). Under such protocols a post-processor can look essential when it merely harvests annotation artefacts, so the literature both over-claims ("whitening unlocks frozen features") and under-claims ("the plateau is an intrinsic limit of self-supervision") at once, with no train-free way to tell which.

We argue the confusion comes from conflating two distinct notions of *closure*: *linear closure* (whether any label-free transformation of an already-frozen embedding improves retrieval) and *space closure*

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.

(whether the embedding holds no further extractable, class-discriminative structure). A space can be linearly closed while wide open nonlinearly. To separate them we adopt a deliberately strict protocol, *strict primary-class retrieval*, in which a retrieval counts only if it shares the query's primary class, the query and gallery are the same split, self-matches are excluded, and super-class and multi-label structure are discarded.

Under it we find that frozen DINOv2 features are linearly closed at a capacity-independent plateau (0.688 R@1) that the newer DINOv3 generation reproduces (0.687–0.696; Table 2), that the label ceiling is 0.9997 and therefore not the bottleneck, and that the space is nevertheless nonlinearly and *supervisedly* open: a minimal SupCon head reaches 0.739 while an equally nonlinear label-free autoencoder does not. We then show the plateau is governed by a single sufficient statistic, give its closed form, and decompose *why* the readout opens the space. To summarize:

- **A strict-primary protocol and a controlled decomposition of closure** (Section 4.2). Opening the space requires *both* nonlinearity *and* a supervised contractive signal, while spectral, hierarchical, whitening, and nested-dimension post-processors are neutral or harmful. The plateau is capacity-independent and generation-invariant across the DINOv2 and DINOv3 backbones, yet family-sensitive (CLIP reaches only 0.422), verified across four datasets and six frozen backbones.
- **A closed-form, train-free diagnostic for retrieval trends** (Section 3.2). From heavy-tailed margin parameters and a Gaussian copula in log space the closed form recovers observed plateaus to within 2.4 points (raw) and 5.3 points (readout) and collapses to the sufficient statistic $J = \mathbb{E}[w]/\sqrt{\mathrm{Var}(w)}$ with $\widehat{\mathrm{R@1}} = \Phi(J)$. We prove that raising the dependence ρ alone is *not* sufficient, and the derived objective $\mathcal{L}_J$ performs comparably to SupCon under the same protocol (0.731 vs. 0.735); we attribute the difference to training moving the statistics that the closed form treats as fixed.
- **A mechanistic decomposition** (Section 4.5). The projector contracts the same-class margin on the training split while sharply raising log-space dependence on the test split (ρ: $0.42 \to 0.87$); a ρ-decomposition with margins fixed shows dependence alone reproduces the test gain, and a 2×2 counterfactual confirms dependence drives the test side while the margin drives the training side.

2 Related Work

This work touches four literatures, each of which we draw on differently. Self-supervised learning supplies the frozen feature families we diagnose, and the decorrelation tools a practitioner reaches for when those features look degenerate. Supervised contrastive and metric learning supplies the readouts whose effect we attribute to individual statistics. Nearest-neighbor margins, copulas, and extreme-value theory supply the distributional machinery behind our closed-form score. Finally, linear-probe studies of frozen features make the closure claims that motivate the strict protocol adopted here. We review these strands in turn.

Self-supervised learning and frozen features. Contrastive and non-contrastive objectives learn transferable embeddings (Chen et al., 2020; He et al., 2020; Caron et al., 2020; Zbontar et al., 2021; Grill et al., 2020; Bardes et al., 2022; Tian et al., 2020); masked modeling offers a complementary paradigm (He et al., 2022), and DINOv2 (Oquab et al., 2024) with DINOv3 (Siméoni et al., 2025) give features whose nearest-neighbor structure is salient without supervision, making them natural frozen families for retrieval. SSL features are anisotropic and prone to dimensional collapse (Jing et al., 2022; He et al., 2024; Wang et al., 2024a), so a parallel line studies whitening and decorrelation (Ermolov et al., 2021; Zbontar et al., 2021), and recent theory characterizes the geometry contrastive objectives induce (Wang et al., 2024b; Jiang et al., 2025; Pan & Cao, 2026; Cui et al., 2025; Zhang et al., 2025a; Lin & Mei, 2025). We evaluate whitening and related post-processors as candidate opening operations and find that they do not improve strict R@1. A separate literature asks how far frozen features go with linear readouts (Zhai et al., 2019; Kumar et al., 2022; El Banani et al., 2024; Huang et al., 2024; Mehra et al., 2024); LP++ (Huang et al., 2024) shows how much a well-designed linear head extracts. Claims that a space is "closed" have rested on spectral or hierarchical motivation; we argue these rest on a protocol artefact, because under strict evaluation the linear plateau is low and robust. Rather than explaining features in the attribution sense (Shi et al., 2026), we ask a quantitative

question: given only frozen features, what strict-retrieval rate a readout can reach, and which statistic sets it? **Supervised contrastive and metric learning.** Label-aware contrastive objectives pull in-class anchors closer (Khosla et al., 2020), metric losses do the same (Song et al., 2016; Schroff et al., 2015), and mean-field analysis characterizes the induced geometric structure (Furusawa, 2024). Our readout isolates supervision and nonlinearity *jointly* and attributes each objective's effect to a specific statistic rather than to a phenomenon. **Nearest-neighbor margins, copulas, and extremes.** Classical work bounds nearest-neighbor error from class-ratio and margin statistics (Cover & Hart, 1967; Fukunaga, 1990; Siffer et al., 2017). Copulas separate marginals from dependence and have entered representation learning through cross-modal alignment (Wu et al., 2025) and dependent censoring (Zhang et al., 2024); extreme-value theory underlies streaming anomaly detection (Siffer et al., 2017), Gumbel-prior set matching (Zhang et al., 2025b), and risk-sensitive policy optimisation (Gao et al., 2025). Because a, b are minima over large galleries, extreme-value theory *motivates* heavy-tailed log-margins (de Haan & Ferreira, 2006), and a single Gaussian copula (Nelsen, 2006) supplies the dependence parameter carrying the test-side mechanism; we use these as a working approximation and test them accordingly (Section 4.4). **Positioning.** The protocol we adopt differs from those used to evaluate whitening, clustering-derived relabeling, and contrastive retraining; Supplement B provides the comparison. Against that background we contribute a protocol-corrected measurement, a two-factor necessity result, a train-free closed-form score with an estimation bound, and a mechanistic decomposition; the practical payoff is train-free model selection for retrieval readouts.

3 Method: A Closed-Form, Train-Free Diagnostic

3.1 Problem Formulation and Modeling Assumptions

The margin abstraction. For a query q we define

$$a \;=\; \min_{g \neq q,\ \mathrm{label}(g)=\mathrm{label}(q)} d_\theta(q,\, g), \qquad b \;=\; \min_{g,\ \mathrm{label}(g)\neq\mathrm{label}(q)} d_\theta(q,\, g), \tag{1}$$

so retrieval succeeds iff $b > a$ and therefore

$$\mathrm{R@1}(\theta) = \mathbb{P}_q[b > a], \tag{2}$$

Recomputing observed R@1 from per-query margin indicators recovers reported values to within 0.001. This lets us separate *margin* (the relative location of a vs. b) from *dependence* (the correlation between $\log a$ and $\log b$) below.

Distributional assumptions. We use the closed form as a *calibrated diagnostic*; the downstream results (J, the ρ-insufficiency result) inherit this scope. Because a, b are minima over large galleries, extreme-value theory *motivates* modelling $\log a, \log b$ as heavy-tailed (Gumbel-type) (de Haan & Ferreira, 2006), joined by a single Gaussian copula (Nelsen, 2006); the misfit this entails is quantified in Section 4.4 and is benign for relative-ranking use.

3.2 A Closed-Form, Train-Free Diagnostic Model

Beyond *measuring* the plateau we ask whether it can be *predicted* from the frozen features alone, without training any readout. The predictor consumes only five statistics of the frozen features ($\mu_{\log a}, \sigma_{\log a}, \mu_{\log b}, \sigma_{\log b}, \rho$) and returns a closed-form calibrated score that tracks strict R@1. Under the Gaussian-copula approximation it recovers that score in rank and approximate magnitude, using no training and no labels beyond one labelled query set. As a *diagnostic*, it decomposes the observed rate into margin and dependence contributions and indicates whether the frozen ceiling is set by margins or by dependence.

Model. Recall that strict R@1 is $\mathbb{P}[b > a]$. We model $(\log a, \log b)$ with heavy-tailed margins joined by a Gaussian copula with single parameter $\rho = \mathrm{corr}(\log a, \log b)$ (Nelsen, 2006; Zhang et al., 2024; Wu et al., 2025). Then $\log b - \log a$ is Gaussian with mean $\mu_{\log b} - \mu_{\log a}$ and variance $\sigma^2_{\log a} + \sigma^2_{\log b} - 2\rho\sigma_{\log a}\sigma_{\log b}$, giving

$$\widehat{P} \;=\; \Phi\Big(\tfrac{\mu_{\log b} - \mu_{\log a}}{\sqrt{\sigma^2_{\log a} + \sigma^2_{\log b} - 2\rho\,\sigma_{\log a}\sigma_{\log b}}}\Big), \tag{3}$$

where Φ is the standard normal CDF; ρ enters only through the denominator. Because a, b are minima over large galleries, extreme-value theory *motivates* Gumbel-type margins (de Haan & Ferreira, 2006; Siffer et al., 2017); as stressed in Section 3.1, this is a modelling choice, not a claim of exact fit, and we quantify the misfit in Section 4.4. Setting $\rho = 0$ makes (3) a univariate probit score on the log-distance difference, i.e. logistic regression up to link reparametrization; fitted $\hat{\rho}$ is far from zero (0.22–0.87), so the dependence parameter carries real information the ρ=0 rung discards.

The ceiling is a signal-to-spread ratio. Writing $\Delta = \mu_{\log b} - \mu_{\log a}$ and $w = \log(b/a)$, (3) collapses to a single sufficient statistic:

$$\widehat{\mathrm{R@1}} \;=\; \Phi(J), \qquad J \;=\; \frac{\Delta}{s} \;=\; \frac{\mathbb{E}[w]}{\sqrt{\mathrm{Var}(w)}}, \tag{4}$$

since $s^2 = \mathrm{Var}(w)$. The ceiling is thus controlled by the *signal-to-spread ratio (SNR) of the log-ratio* (the mean of w in units of its own standard deviation), not by the margin or the dependence alone. On COCO test J rises $0.425 \to 0.531$ through the readout: SupCon buys only $+0.106$ units of SNR, exactly why the predicted ceiling moves $0.665 \to 0.702$ (observed $0.688 \to 0.739$).

Dependence elasticity. For $\Delta > 0$ the ceiling is strictly increasing in ρ and convex on the measured range, with elasticity $\partial\widehat{\mathrm{R@1}}/\partial\rho = \varphi(z(\rho)) \cdot \Delta\, \sigma_{\log a}\sigma_{\log b}/s(\rho)^3$ (derivation and discussion in Supplement A): the same ρ-step moves the ceiling most when the projected margins are spread out and strongly dependent.

Finite-sample accuracy. Within the Gaussian-copula model, the plug-in estimate $\hat{J} = J(\hat{\theta})$ from n i.i.d. queries is asymptotically normal, $\sqrt{n}(\hat{J} - J) \Rightarrow \mathcal{N}(0, \sigma_{\mathrm{IF}}^2)$ with σ_{IF}^2 the variance of its influence function; since Φ is $(2\pi)^{-1/2}$-Lipschitz (derivation in Supplement A),

$$\left|\widehat{\mathrm{R@1}} - \mathrm{R@1}_{\mathrm{model}}\right| \;\le\; \frac{1.96}{\sqrt{2\pi}} \cdot \frac{\sigma_{\mathrm{IF}}}{\sqrt{n}} \tag{5}$$

with probability ≥ 0.95 asymptotically. Under our four fitted COCO regimes $\sigma_{\mathrm{IF}} \in [1.09, 1.26]$, giving ± 0.007 to ± 0.008 at $n = 16{,}056$. Equation (5) bounds only the *estimation* error, which shrinks as $n^{-1/2}$; it says nothing about the *misspecification* error, an empirical bias that does not shrink with n (Section 4.4).

Raising dependence is not sufficient. In the Gaussian-copula parametrization, suppose $\log a = c_a + \kappa_a t + \varepsilon_a$ and $\log b = c_b + \kappa_b t + \varepsilon_b$ with a query-level difficulty t, $\mathrm{Var}(t) = V$, and $(\varepsilon_a, \varepsilon_b) \perp\!\!\!\perp t$. Then $s^2 = (\kappa_a - \kappa_b)^2 V + \mathrm{Var}(\varepsilon_a) + \mathrm{Var}(\varepsilon_b) - 2\mathrm{Cov}(\varepsilon_a, \varepsilon_b)$, which depends on (κ_a, κ_b) only through their *difference*; consequently, scaling the shared component ($\kappa_a = \kappa_b = \kappa \to \infty$) drives $\rho \to 1$ while s^2, and hence $\widehat{\mathrm{R@1}} = \Phi(\Delta/s)$, is unchanged (derivation in Supplement A): a shared difficulty shifts both distances equally, so raising ρ alone does not raise the ceiling under this model. A Monte-Carlo counterfactual on the fitted COCO test model confirms this: injecting a shared difficulty term of growing strength drives ρ from 0.420 to 0.998 while s, strict R@1, and even the per-query correctness pattern stay fixed (Supplement D.7); the dependence rise we measure on COCO test instead compresses s at fixed marginals (Section 4.5), which is why it moves the ceiling.

Scope of the sufficiency claim. J is sufficient for the ceiling *under the fitted model of a frozen feature distribution*: on synthetic data drawn exactly from the Gaussian copula, R@1=$\Phi(J)$ holds to 10^{-3} (Supplement D); the prediction–optimization gap in Section 4.3 marks the boundary of this static claim.

4 Experiments

4.1 Experimental Setup

Data and frozen features. We evaluate on COCO val2014 (Lin et al., 2014), using the $40{,}137$ validation images carrying at least one annotation: 80 primary classes, $24{,}081$ images for readout training and $16{,}056$ for evaluation. An image's primary label is its largest-area annotation; classes

Method	Strict R@1 ↑	vs. raw
Frozen cosine (raw)	0.688	—
ZCA whitening	0.672	−0.016
Power reweight (p=1.0)	0.672	−0.016
Drop top-20 PC	0.680	−0.008
LDA within-class whitening	0.685	−0.003
Nested truncation (MRL-style, best k)	0.682	−0.007

Table 1: No linear post-processor opens the frozen space: every method moves strict R@1 by at most −0.016 (ZCA: zero-phase component analysis).

are heavily imbalanced (the largest primary class, `person`, accounts for 7,694 gallery images; the smallest, 2). We treat primary labels as the sole relevance signal and the 12 supercategory groupings as *cross-class* structure. Frozen features come from a DINOv2 ViT-S/14 student (384-dim) and a DINOv2 ViT-B/14 teacher (768-dim) (Oquab et al., 2024; Caron et al., 2021; Dosovitskiy et al., 2021), used without fine-tuning. To test whether the findings survive a newer self-supervised generation, we add DINOv3 ViT-S/16 and ViT-B/16 (Siméoni et al., 2025) on the identical strict splits; CLIP ViT-B/16 (Radford et al., 2021) serves as a cross-family probe (pre-projection 768-dim CLS hidden state). All features are L_2-normalized.

Strict primary-class retrieval. A query is scored against relevance defined by the primary label only, with three anti-leakage policies: **single-label relevance** (multi-label and supercategory overlap is ignored), **self-exclusion** (the query and its duplicates are removed from the gallery), and **no label peek** (labels enter only through the relevance definition). The headline number is strict R@1, the fraction of queries whose nearest gallery item shares their primary label.

Label ceiling. Ranking each query's gallery by label identity before any feature is seen gives a ceiling of 0.9997 strict R@1, so labels are not the bottleneck: any deficit below this value is attributable to the features.

4.2 Linear Closure and the Two-Factor Decomposition

We first measure the frozen-feature plateau and test the two hypotheses a practitioner would entertain: that the space is closed to any post-processing, and that some nonlinear or supervised operator is required to open it. All post-processors are fitted on the gallery split only and applied without supervision of the test split.

Linear closure. Table 1 reports a battery of linear post-processors. Every one degrades strict R@1 slightly relative to raw cosine (0.688) and none lifts it. (The one cross-family exception—MAE on ImageNet-100, whose spectral precondition fails—is analyzed in Section 4.4.) Whitening and power reweighting each cost about 1.6 points, dropping the leading principal components is monotone harmful, and within-class LDA whitening costs 0.3 points. We additionally test the *nested-dimension* truncation of Matryoshka Representation Learning (MRL) (Kusupati et al., 2022): retaining only the first k coordinates gives 0.682 at k=192, 0.665 at k=96 and 0.634 at k=48 against 0.688 for the full 384-dim feature: monotone degradation, no nested slice beats the full feature. The frozen features are therefore *linearly closed*: no linear transformation of the embedding, including label-free dimensionality reduction, improves strict retrieval.

Capacity independence. Opening the same protocol with the strictly larger frozen teacher gives 0.689 against the student's 0.688 (statistically identical), and the newer DINOv3 generation lands on the same plateau (0.687 for ViT-S/16, 0.696 for ViT-B/16; Table 2), so the plateau is both capacity-independent and generation-invariant—an intrinsic geometric bound of the self-supervised DINO family rather than a capacity-limited artefact.

Nonlinearity is necessary but not sufficient. We next train two readouts on the gallery split: a **label-free nonlinear** map (a shallow autoencoder-like projector, reconstruction objective, no labels) and a **supervised nonlinear** projector (supervised-contrastive (Khosla et al., 2020) on primary labels, the width-1024, depth-1 head of Section 4.5). As Figure 1 shows, the nonlinear label-free readout

Frozen family	Dim	Strict R@1	ZCA
DINOv2 ViT-S/14 (student)	384	0.688	0.672
DINOv2 ViT-B/14 (teacher)	768	0.689	0.672
DINOv3 ViT-S/16	384	0.687	0.660
DINOv3 ViT-B/16	768	0.696	0.674
CLIP ViT-B/16	768	0.422	0.419

Table 2: The strict-retrieval plateau is capacity- and generation-invariant but family-sensitive: self-supervised DINO features sit at ~ 0.69 across two generations and two capacities, whereas text-contrastive CLIP reaches only 0.422 on the identical strict splits. ZCA whitening hurts every family.

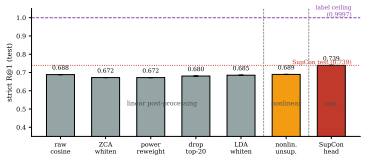

Figure 1: Closure decomposition on frozen DINO features: linear post-processors (at or below the plateau), nonlinear *unsupervised* autoencoder (0.689), and supervised-contrastive head (0.739; seed spread ± 0.001). Reference lines: label ceiling 0.9997, raw plateau 0.688.

Regime	Observed	Predicted	Gap
DINOv2 train, raw	0.695	0.678	-0.016
DINOv2 test, raw	0.688	0.665	-0.024
DINOv2 train, readout	0.957	0.904	-0.053
DINOv2 test, readout	0.739	0.702	-0.036
DINOv2 teacher, test	0.689	0.667	-0.022
CLIP, test	0.422	0.465	$+0.043$

Table 3: Predictor validation (gap = predicted − observed); full details in Supplement G.

reaches 0.689, identical to raw cosine, so nonlinearity alone does not open the space. Only the supervised nonlinear readout lifts strict R@1 to 0.739 (three seeds: 0.738 ± 0.001). The depth-2 projector used for the objective comparison (Table G.10) reaches 0.735 ± 0.002 (Supplement B). The factors are jointly necessary: nonlinearity alone gives no lift and linear supervised procedures (LDA) fail; only nonlinearity $\times$ supervision reaches 0.739.

Spectral-shaping and graded-hierarchical terms. Added to the SUPCON objective, these terms do not improve strict R@1 at any tested strength (Supplement G); gains reported for such terms in prior work may reflect protocol differences.

4.3 DIAGNOSTIC PREDICTION ACCURACY AND SCOPE

Validation. Figure 2 and Table 3 validate (3). DINO raw regimes are tracked to within 2.4 points and readout regimes to within 5.3; the teacher agrees (-2.2 points) and CLIP is over-predicted by $+4.3$ (a family-calibrated bias). The gap is a systematic under-prediction that grows once the readout amplifies non-Gaussian tail dependence. Within a family it is a *consistent-delta* predictor: it reproduces the ceiling and its change from raw to readout regime, so it calibrates—up to the diagnostic error discussed below—the gain a supervised head may deliver, before that head is trained.

Performance of the derived objective. Because J is what the ceiling depends on, the theory singles out

$$\mathcal{L}_J = -\hat{\Delta}/\hat{s} \tag{6}$$

as the principled training target. Under the same protocol the derived objective performs comparably to SUPCON (all rows use the depth-2 projector, whose SUPCON baseline is the 0.735 above rather than the depth-1 closure head's 0.739); the full comparison is in Supplement G.12.

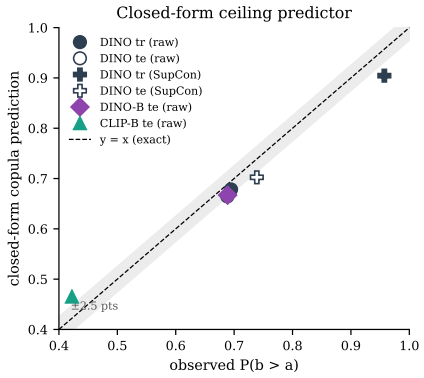

Figure 2: Predictor vs. observed strict recall (full table in Supplement G); shaded band ± 2.5 points.

We attribute this to training moving the statistics that the closed form treats as fixed: J is sufficient for the *model* but not for the optimization landscape, so the closed form is a predictor rather than an optimization objective (Supplement G.12).

Scope and limits of the Gaussian-copula diagnostic. Two terms make up the prediction error,

$$\widehat{\mathrm{R@1}} - \mathrm{R@1}_{\mathrm{true}} = \underbrace{\left(\widehat{\mathrm{R@1}} - \mathrm{R@1}_{\mathrm{model}}\right)}_{\text{estimation}} + \underbrace{\left(\mathrm{R@1}_{\mathrm{model}} - \mathrm{R@1}_{\mathrm{true}}\right)}_{\text{model misspecification}}. \tag{7}$$

Equation (5) controls the first term: it is at most ± 0.008 at $n = 16{,}056$ and shrinks as $n^{-1/2}$. The second is the disagreement between the Gaussian-copula model and the empirical margins that the goodness-of-fit tests in Section 4.4 flag; it does not shrink with n and it dominates. On the COCO benchmarks the misspecification bias is 0.02–0.05 (2–5 points; predicted below observed for DINOv2 and above for CLIP; Supplement G), and substantially larger on a fine-grained dataset, 0.07–0.19 on Stanford Dogs (Supplement F).

Within a family the misspecification term is systematic and one-signed, so it cancels in differences and ranking survives (Supplement G); we therefore restrict quantitative claims to *within-family* ranking, where the predictor acts as a calibrated, monotone screen indicating which backbones are usable.

4.4 Scope, limits, and using the predictor

The model is used as a diagnostic. We test the margins on the $n = 16{,}056$ COCO test queries with Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) statistics against fitted Normal, Gumbel and Logistic margins (full statistics in Supplement D). All three margins are rejected at the 5% level—with the EVT-motivated Gumbel fit being poorest—and the Gaussian copula likewise departs. With n=16,056 the tests hold enormous power, so even small gaps are flagged; this reflects the sample size rather than the practical utility of the ranking.

Why ranking survives the rejection. (i) With n=16,056 the tests flag arbitrarily small departures, so rejection indicates the *presence* of deviation, not its magnitude. (ii) The departures are structured: within a family the residual bias is one-signed and slowly varying (DINOv2 raw 1.6–2.4 points low, readout 3.6–5.3 low, teacher 2.2 low), so it cancels in differences and preserves orderings and signs of change; across families the sign flips (CLIP $+4.3$ high), which is why only within-family ranking is supported. (iii) Ranking is carried by the margin/dependence summary, not by exact distributional truth: on synthetic data drawn from the fitted model, $\mathrm{R@1}{=}\Phi(J)$ holds to 10^{-3} (Supplement D). Ranking utility needs the misspecification bias nearly constant within the comparison set, not the model exactly true—and this is what we observe.

Does closure survive multi-label relevance? Under multi-label relevance (sharing at least one of the 80 categories) absolute R@1 rises sharply and is not comparable to multi-label baselines, but closure survives in sign and ordering (Supplement G).

The spectral precondition and the MAE boundary case. Whitening hurts only when the precondition fails: the frozen covariance has no dominant direction for whitening to repair. DINOv2's spectrum is spread (leading eigenvalue $\lesssim 3\%$ of the trace on ImageNet-100; ranks in Supplement G), so whitening amplifies the low-variance tail, a uniform modest loss.

MAE is the exception. On ImageNet-100, MAE's raw features are nearly non-discriminative (R@1 0.125), ZCA *helps* ($0.125 \rightarrow 0.393$), and SUPCON restores discriminability (0.714), unlike the other families tested: its raw margin ratio $a/b = 1.44 > 1$ is exactly the regime where second-order repair works. On the Stanford Dogs proxy, MAE-B's leading eigenvalue carries 92.4% of the trace (effective rank 1.17 of 768) versus $\approx 6\%$ for DINOv2 (eigenvalue ranks in Supplement G). Measured on the *same* images, MAE is the sole family ZCA repairs because its precondition fails.

Using J for train-free model selection. Model selection here comes in two forms of decreasing ambition: *coarse-grained cross-family screening*—flagging severely degenerate representations such as raw MAE —and *fine-grained within-family ranking*, which is more reliable than cross-family comparison. The primary use is to rank frozen backbones by $\Phi(J)$ from frozen features alone. The six ImageNet-100 backbones (DINOv2 and DINOv3 at two capacities, CLIP, MAE) are tabulated in Supplement G. Ranking by $\Phi(J)$ against observed R@1 gives **Spearman** 0.943 (n=6, exact two-sided p=0.017; mean error 2.9 points on ImageNet-100) at zero training cost and flags the degenerate MAE features as unusable in their *raw* form—a property of raw MAE geometry, not of MAE as a model.

Limits of this use. This demonstration covers six backbones on one dataset; broader validation remains open. The 2.9-point figure is the ImageNet-100 error; on fine-grained Stanford Dogs absolute calibration degrades to 0.07–0.19 (Supplement F), so absolute calibration is regime-dependent. Practically: compute J from frozen features alone; flag representations whose raw margin ratio a/b exceeds 1 as degenerate; rank backbones *within* a family by $\Phi(J)$; and recalibrate per family for absolute calibration.

4.5 MECHANISM

The remaining question is which statistic the readout alters. Two levers are available: contracting the same-class margin (a/b) and tightening the dependence ρ between the two log-distances; which one a readout pulls is regime-dependent. After training the SUPCON head we recompute a/b and ρ.

The readout plays different roles on the training and test sets. On the *test* split (the headline metric), the margin ratio barely moves ($0.92 \rightarrow 0.89$) while ρ nearly doubles ($0.42 \rightarrow 0.87$), and the mapping concentrates the space (median distance down $\approx 40\%$). On the *training* split, by contrast, the readout compresses within-class margins and decorrelates classes, achieving near-perfect memory retrieval (full statistics in Supplement C). The decisive quantity is the spread of $\log(b/a)$: without the rise in dependence it would disperse to a standard deviation of 0.565 instead of 0.243, dropping predicted R@1 to 0.590 (*below* the plateau); the dependence reduces the spread of the ratio.

Mechanistic decomposition: dependence is the dominant test-side channel on COCO. To test whether the correlation is spurious, we decompose the coupled model while preserving the marginals and scanning only ρ: upgrading it from 0 to 0.87 moves the predicted R@1 from 0.590 to 0.702 (Figure 3). A 2×2 counterfactual confirms the disentanglement (full matrix in Supplement C): with the marginals frozen, the SUPCON dependence is the positive channel on test and, reversing marginals versus dependencies, it becomes negative on training. The test-side mechanism generalizes beyond DINO: the same signature—a near-constant margin ratio with a sharp rise in ρ—holds on CLIP (Supplement G.10).

Capacity: width is the lever, depth overfits. Widening the projector monotonically improves strict R@1 while adding depth past one layer degrades it; we use the width-1024, depth-1 head here (capacity details in Supplement G).

The mechanism is regime-dependent, and the boundary is determined empirically. On fine-grained Stanford Dogs (Figure 4) the positive channel inverts: contraction of the same-class margin (a/b: 0.53–$0.76 \rightarrow 0.18$–0.45) while ρ stays small or negative.

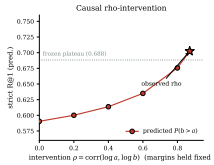

Figure 3: Mechanistic ρ-decomposition, margins held fixed: dependence alone raises predicted R@1 from 0.590 to 0.702 (on COCO test the dependence channel alone reproduces the full test gain).

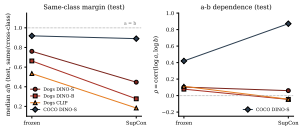

Figure 4: Regime-dependent mechanism (test split): on COCO the gain rides ρ (sharp rise, margins flat); on fine-grained Dogs it rides same-class margin contraction (a/b falls, ρ flat/negative).

A one-dimensional mechanism is refuted by our fourth dataset: CIFAR-100 shares COCO's ratio yet lacks the support and is margin-driven (Supplement G). What remains is two-dimensional: a dependence-driven regime needs a near-unity ratio *and* class support on the order of hundreds.

4.6 The prediction–optimization gap

The $\mathcal{L}_J$ result in Section 4.3 illustrates a principle that transcends this diagnostic: a statistic can be *sufficient for prediction*—the ceiling under a fixed feature distribution—yet *ineffective as an objective*, because training moves the very statistics it conditions on. The closed form is a predictor; its behavior as a training objective is a separate question.

5 Conclusion

We introduced a closed-form, train-free diagnostic that predicts the strict retrieval ceiling of frozen features from five frozen-feature statistics alone, collapsing to the sufficient statistic $J = \mathbb{E}[w]/\sqrt{\mathrm{Var}(w)}$ and, under a Gaussian-copula approximation, to $\widehat{\mathrm{R@1}} = \Phi(J)$ with no training and no label access. Using it we established three results that permissive protocols obscure. **First**, frozen features are *linearly closed* whenever their spectral precondition holds: no linear post-processor improves strict R@1, the plateau is capacity-independent and generation-invariant (DINOv2 and DINOv3), and opening the space takes *both* nonlinearity *and* a supervised contractive signal. **Second**, the ceiling is governed by the spread of the log-ratio rather than by margins or dependence alone: raising ρ is insufficient, because shared query difficulty inflates ρ while leaving $\log b - \log a$ unchanged. **Third**, the trained readout acts through regime-dependent mechanisms—dependence-driven on COCO, margin-driven on fine-grained Stanford Dogs—and the objective the theory singles out, $\mathcal{L}_J$, trails SupCon only slightly when trained.

The diagnostic supports *coarse-grained* cross-family screening of degenerate features (raw MAE) and *fine-grained* within-family ranking (Spearman 0.943, ≈ 2.9 points)—a zero-cost filter for whether frozen features are worth further investment; absolute calibration is family-specific, and multi-label closure survives in sign and ordering but not absolute value (Section 4.4).

Because the diagnostic is train-free and consumes only a single labelled query set, it can be applied before any compute is committed to a readout: the closed form $\widehat{\mathrm{R@1}} = \Phi(J)$ alone already segments families into degenerate versus openable backbones and ranks within a family at zero cost. The regime-dependent readout mechanism and the prediction–optimization gap identified here also suggest that extending the same closed-form analysis to vision–language backbones and to alternative retrieval objectives—including whether a probe should be replaced, or merely initialized, by the derived objective—is a promising direction, which we leave to future work. More generally, this pairing provides a principled stopping rule for how much supervision a frozen backbone merits.

AI USE STATEMENT

Large language models (LLMs) were employed exclusively to improve the quality of written English and for stylistic refinement, with attentive supervision by the authors. No scientific content, experimental design, data analysis, or any intellectual contribution were generated through LLMs, nor did such assistance influence any reported results.

ETHICS STATEMENT

All authors acknowledge and adhere to the ICLR Code of Ethics. This work analyses only public datasets (COCO val2014, Stanford Dogs, ImageNet-100, CIFAR-100) and publicly released pre-trained models. It introduces no new data collection, no human subjects, and no deployment, and therefore raises no ethical concerns beyond those inherent to standard benchmark use.

REPRODUCIBILITY STATEMENT

The main text and the appendix give a complete description of the experimental setup and the train-free diagnostic, including all summary statistics, fitting procedures, and the proof of the margin identity; full numeric tables, implementation details, derivations, and statistical diagnostics are collected in the same appendix. Upon acceptance, we will release anonymous code and data artifacts that reproduce every reported number.

REFERENCES

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

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1–9, 2022. doi: 10.18653/v1/2022.acl-short.1.

Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. FFF: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14172–14182, 2024.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, volume 11218 of *Lecture Notes in Computer Science*, pp. 139–156, 2018. doi: 10.1007/978-3-030-01264-9_9.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

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)*, pp. 1597–1607, 2020.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Jingyi Cui, Hongwei Wen, and Yisen Wang. An augmentation-aware theory for self-supervised contrastive learning. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer, 2006.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806, 2024.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning (ICML)*, pp. 3015–3024, 2021.

Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2nd edition, 1990. ISBN 978-0-08-047865-4.

Takuya Furusawa. Mean field theory in deep metric learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Shiqing Gao, Yihang Zhou, Shuai Shao, Haoyu Luo, Yiheng Bing, Jiaxin Ding, Luoyi Fu, and Xinbing Wang. Extreme value policy optimization for safe reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Jean-Bastien Grill, Florian Strub, Florent Altché, and et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 21271–21284, 2020.

Junlin He, Jinxiao Du, and Wei Ma. Preventing dimensional collapse in self-supervised learning via orthogonality regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A surprisingly strong linear probe for few-shot CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52733.2024.02244.

Ruijie Jiang, Thuan Nguyen, Shuchin Aeron, and Prakash Ishwar. Hard-negative sampling for contrastive learning: Optimal representation geometry and neural- vs dimensional-collapse. *Transactions on Machine Learning Research (TMLR)*, 2025.

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.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18661–18673, 2020.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kai Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Licong Lin and Song Mei. A statistical theory of contrastive learning via approximate sufficient statistics. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 141879–141923, 2025. doi: 10.52202/085713-4262.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Akshay Mehra, Yunbei Zhang, and Jihun Hamm. Understanding the transferability of representations via task-relatedness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 116513–116546, 2024.

Roger B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, 2nd edition, 2006. ISBN 978-0-387-28659-4. doi: 10.1007/0-387-28678-0.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, and et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Leyan Pan and Xinyuan Cao. Towards understanding neural collapse: The effects of batch normalization and weight decay. *Transactions on Machine Learning Research*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, and et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

Yingdong Shi, Zhiyu Yang, Changming Li, Jingyi Yu, and Kan Ren. MiCLIP: Learning to interpret representation in vision models. In *International Conference on Learning Representations (ICLR)*, 2026.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. Anomaly detection in streams with extreme value theory. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1067–1075, 2017.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, and et al. DINOv3. 2025. URL `https://arxiv.org/abs/2508.10104`.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 6827–6839, 2020.

Haoqi Wang, Tong Zhang, and Mathieu Salzmann. SINDER: Repairing the singular defects of DINOv2. In *European Conference on Computer Vision (ECCV)*, pp. 20–35, 2024a.

Yifei Wang, Kaiwen Hu, Sharut Gupta, Ziyu Ye, Yisen Wang, and Stefanie Jegelka. Understanding the role of equivariance in self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024b.

Ross Wightman. Pytorch image models. `https://github.com/huggingface/pytorch-image-models`, 2019. Accessed: 2026.

Feng Wu, Tsai Hor Chan, Fuying Wang, Guosheng Yin, and Lequan Yu. Cross-modal alignment via variational copula modelling. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

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)*, pp. 12310–12320, 2021.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4L: Self-supervised semi-supervised learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1476–1485, 2019.

Qi Zhang, Yifei Wang, and Yisen Wang. An augmentation overlap theory of contrastive learning. *Journal of Machine Learning Research (JMLR)*, 26(228):1–42, 2025a.

Weijia Zhang, Chun Kai Ling, and Xuanhui Zhang. Deep copula-based survival analysis for dependent censoring with identifiability guarantees. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 20613–20621, 2024.

Ziming Zhang, Fangzhou Lin, Haotian Liu, Jose Morales, Haichong Zhang, Kazunori Yamada, Vijaya B. Kolachalama, and Venkatesh Saligrama. GPS: A probabilistic distributional similarity with Gumbel priors for set-to-set matching. In *International Conference on Learning Representations (ICLR)*, 2025b.

APPENDIX

CONTENTS

A Proofs and derivations **16**

A.1 Dependence elasticity 16

A.2 Finite-sample accuracy 16

A.3 The ρ-insufficiency example 17

A.4 A two-class Gaussian model for the margin ratio and ρ 17

B Implementation and reproducibility **18**

B.1 Datasets, splits, and frozen features 18

B.2 The strict protocol, pseudocode 18

B.3 Protocol comparison with prior work 18

B.4 Trained readouts and baselines 19

B.5 Hardware and software 20

B.6 Code availability 20

C Complete experimental tables **20**

C.1 Full closure numbers (Block A and B) 20

C.2 Capacity sweep 20

C.3 Readout objectives: full diagnostics 21

C.4 Trainable baselines under the strict protocol 21

C.5 Protocol sensitivity: full matrix 21

C.6 Nested truncation and spectral statistics 22

C.7 MAE spectral degeneracy (Stanford Dogs) 22

C.8 Train-free model selection: full table 22

C.9 Mechanism statistics 22

C.10 Causal counterfactual: full table 23

D Statistical diagnostics, staircase, and calibration **23**

D.1 Goodness-of-fit diagnostics 23

D.2 Model staircase: heavy tails are necessary, dependence is sufficient 23

D.3 Consistency controls 24

D.4 Empirical calibration 24

D.5 Scope of the predictor 25

D.6 Synthetic control: J is sufficient where the model holds 25

D.7 Synthetic counterfactual: shared difficulty inflates ρ without touching retrieval . . . 25

E Cross-Family Generality **26**

E.1 The raw plateau is family-sensitive 26

E.2 The SUPCON opening is mechanism-general 26
E.3 Summary 27

F Cross-Dataset Validation on Fine-Grained Stanford Dogs **27**
F.1 Closure and opening generalize 27
F.2 The mechanism is the supervised contraction of the same-class margin 27
F.3 Per-regime calibration 28
F.4 Duplicate audit 29

G Experimental details deferred from the main text **29**
G.1 Predictor validation table 29
G.2 Spectral-shaping and graded-hierarchical terms 29
G.3 Capacity independence and the capacity sweep 29
G.4 Rank-order screening table 29
G.5 Multi-label relevance and spectral details 29
G.6 Goodness-of-fit detail 30
G.7 Mechanistic counterfactual decomposition 30
G.8 Phase-boundary thresholds 30
G.9 Derived objective: seed results 31
G.10 Mechanism on CLIP 31
G.11 Training dynamics of the derived objective 31
G.12 Full comparison of the derived objective 31

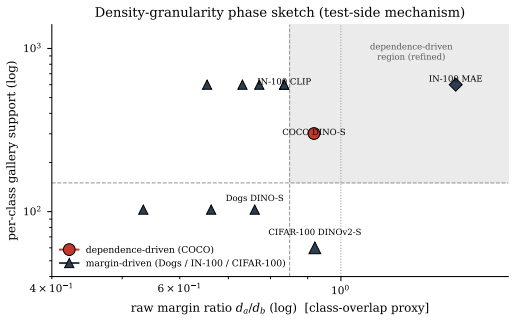

Figure A.1: Density-granularity phase diagram over the measured (dataset, class family) cells; axes and thresholds empirical. Shaded region: near-unity margin ratio *and* class support on the order of hundreds. COCO falls inside (dependence-driven); Dogs and CIFAR-100 outside (margin-driven), refuting a one-dimensional mechanism.

A Proofs and derivations

This section collects the derivations behind the closed-form diagnostic of the main text and derives the two-class Gaussian model that explains the sign and ordering of the phase sketch of Fig. A.1. Throughout, $u = \log a$, $v = \log b$, $\Delta = \mu_v - \mu_u$ and $s = \sqrt{\sigma_u^2 + \sigma_v^2 - 2\rho\sigma_u\sigma_v}$.

A.1 Dependence elasticity

Proof. Differentiating $z(\rho) = \Delta/s(\rho)$ and using $s'(\rho) = -\sigma_u\sigma_v/s(\rho)$ gives $z'(\rho) = \Delta\,\sigma_u\sigma_v/s(\rho)^3$, hence $\partial\widehat{\mathrm{R@1}}/\partial\rho = \varphi(z(\rho))\,\Delta\,\sigma_u\sigma_v/s(\rho)^3$, which is the stated form. For $\Delta > 0$ the sign of the derivative is that of $\sigma_u\sigma_v > 0$, so the ceiling is strictly increasing in ρ. Convexity on the measured range is by direct evaluation: on the readout margins the per-unit slope grows from 0.05 at ρ=0 to 0.37 at ρ=0.87, and the elasticity at ($\hat\rho$=0.87, projected) is 0.465 versus 0.116 at ($\hat\rho$=0.42, raw)—the same ρ-step moves the ceiling roughly four times more in the projected regime. The elasticity scaling follows from $s(\rho)^3$ in the denominator: it is amplified exactly when the projected margins are simultaneously spread out ($\sigma_u\sigma_v$ large) and strongly dependent (s small). □

A.2 Finite-sample accuracy

Influence functions. For a single query contributing (u, v) with $\theta = (\mu_u, \mu_v, \sigma_u, \sigma_v, \rho)$, the influence functions of the five statistics are

$$\begin{aligned}
\mathrm{IF}_{\mu_u} &= u - \mu_u, & \mathrm{IF}_{\mu_v} &= v - \mu_v,\\
\mathrm{IF}_{\sigma_u} &= \frac{(u-\mu_u)^2 - \sigma_u^2}{2\sigma_u}, & \mathrm{IF}_{\sigma_v} &= \frac{(v-\mu_v)^2 - \sigma_v^2}{2\sigma_v},\\
\mathrm{IF}_{\rho} &= \frac{(u-\mu_u)(v-\mu_v)}{\sigma_u\sigma_v} - \frac{\rho}{2}\left(\frac{(u-\mu_u)^2}{\sigma_u^2} + \frac{(v-\mu_v)^2}{\sigma_v^2}\right).
\end{aligned} \tag{A.1}$$

The gradient of $J = \Delta/s$ with respect to θ is

$$\nabla J = \left(-\tfrac{1}{s},\ \tfrac{1}{s},\ -\tfrac{\Delta(\sigma_u - \rho\sigma_v)}{s^3},\ -\tfrac{\Delta(\sigma_v - \rho\sigma_u)}{s^3},\ -\tfrac{\Delta\sigma_u\sigma_v}{s^3}\right), \tag{A.2}$$

so $\mathrm{IF}_J = \langle \nabla J, \mathrm{IF}_\theta \rangle$ and $\sigma_{\mathrm{IF}}^2 = \mathrm{Var}(\mathrm{IF}_J)$.

Proof. Each of the five statistics is a Hadamard-differentiable functional of the empirical distribution, with the influence functions listed above; the chain rule gives the influence function of J, and the central limit theorem gives $\sqrt{n}(\hat J - J) \Rightarrow \mathcal{N}(0, \sigma_{\mathrm{IF}}^2)$. Composing with Φ, whose Lipschitz constant is $\varphi(0) = (2\pi)^{-1/2}$, yields (5) in the main text. The constants are obtained by Monte Carlo evaluation (4×10^6 draws from each fitted Gaussian model): $\sigma_{\mathrm{IF}} = 1.123$ (student train), 1.088 (student test), 1.258 (readout train) and 1.202 (readout test); the 95% half-width $1.96\,\varphi(J)\,\sigma_{\mathrm{IF}}/\sqrt{n}$ evaluates to ± 0.0062, ± 0.0061, ± 0.0033 and ± 0.0064 at $n = 16{,}056$, and to ± 0.037 at $n = 500$ in

the readout-test regime. The bound is a first-order (delta-method) statement: it is asymptotic in n and ignores the $O(n^{-1})$ remainder, which is negligible at these sample sizes. □

A.3 THE ρ-INSUFFICIENCY EXAMPLE

Proof. Independence gives $\mathrm{Var}(\log a) = \kappa_a^2 V + \mathrm{Var}(\varepsilon_a)$, $\mathrm{Var}(\log b) = \kappa_b^2 V + \mathrm{Var}(\varepsilon_b)$ and $\mathrm{Cov}(\log a, \log b) = \kappa_a \kappa_b V + \mathrm{Cov}(\varepsilon_a, \varepsilon_b)$; subtracting yields $s^2 = (\kappa_a - \kappa_b)^2 V + \mathrm{Var}(\varepsilon_a) + \mathrm{Var}(\varepsilon_b) - 2\mathrm{Cov}(\varepsilon_a, \varepsilon_b)$, whose only κ-dependent term is $(\kappa_a - \kappa_b)^2 V$ and vanishes when $\kappa_a = \kappa_b$. Meanwhile $\rho = \mathrm{Cov}/(\sigma_{\log a}\sigma_{\log b}) \to \kappa^2 V/\kappa^2 V = 1$ as $\kappa \to \infty$. Since $\Delta = c_b - c_a + (\kappa_b - \kappa_a)\mathbb{E}[t]$ is likewise invariant under equal scaling, so is $\Phi(\Delta/s)$. □

Intuition. A query-level difficulty factor t that shifts both $\log a$ and $\log b$ by the same amount drives their correlation to 1 (they move in lockstep) while leaving their *difference*—the quantity that determines retrieval—unchanged. A high ρ can therefore be pure shared difficulty, which is why maximizing ρ is a ceiling-neutral target.

A.4 A TWO-CLASS GAUSSIAN MODEL FOR THE MARGIN RATIO AND ρ

The elasticity derivation above treats the five fitted statistics as given. This subsection asks the complementary question: *why* should ρ and the margin ratio co-vary across datasets, i.e. what generates the phase sketch of Fig. A.1? We answer in the simplest model that contains both effects. The model is deliberately idealized (two classes, isotropic Gaussian clouds, high-dimensional asymptotics); we use it to explain the *sign and ordering* of the measured effects, not to fit their magnitudes.

Model. Two classes with centres $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$ separated by $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = \Delta$. Each class contributes n samples $\mathbf{x} = \boldsymbol{\mu}_c + \tau\mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_d)$, where τ is the within-class scale. A query $\mathbf{q} = \boldsymbol{\mu}_1 + \tau\mathbf{z}_q$ belongs to class 1; let a, b be defined as in (1) and let $\mathrm{snr} = \Delta^2/(2\tau^2 d)$.

Proposition A.1 (Margin ratio and dependence in the two-class model)**.** *Assume $d \gg \log n$. Then, to relative order $O(\sqrt{\log n/d})$,*

1. ***(Distance concentration)*** *$a^2 = 2\tau^2 d$ and $b^2 = \Delta^2 + 2\tau^2 d$;*
2. ***(Closed-form margin ratio)*** *$a/b = (1 + \mathrm{snr})^{-1/2}$, depending on the model only through* snr *and decreasing monotonically from* 1 *(total overlap) towards* 0 *(perfect separation);*
3. ***(Common difficulty factor)*** *writing $\|\mathbf{z}_q\|^2 = d + s$ with $\mathbb{E}[s] = 0$, $\mathrm{Var}(s) = 2d$, the sensitivity of the two log-distances to the query's offset from its class centre is $\partial \log a/\partial\|\mathbf{z}_q\|^2 = 1/(4d)$ and $\partial \log b/\partial\|\mathbf{z}_q\|^2 = (1/(4d)) \cdot (1 + \mathrm{snr})^{-1}$, the latter strictly decreasing in* snr.

Proof. [Proof sketch] **(1)** For a same-class sample, $\|\mathbf{q} - \mathbf{x}_i\|^2 = \tau^2\|\mathbf{z}_q - \mathbf{z}_i\|^2$ with $\mathbf{z}_q - \mathbf{z}_i \sim \mathcal{N}(0, 2\mathbf{I}_d)$, so $\|\mathbf{z}_q - \mathbf{z}_i\|^2 \sim 2\chi^2_d$. The minimum over $n-1$ i.i.d. draws concentrates at $2d - 4\sqrt{d \log n}$, giving $a^2 = 2\tau^2 d$. For a cross-class sample with $\mathbf{e} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\Delta$, the cross term is $O(\Delta\tau\sqrt{\log n})$, dominated by $\Delta^2 + 2\tau^2 d$ whenever $d \gg \log n$, giving $b^2 = \Delta^2 + 2\tau^2 d$. **(2)** Immediate from (1). **(3)** Both squared distances contain $\tau^2\|\mathbf{z}_q\|^2$, so $\partial \log a/\partial\|\mathbf{z}_q\|^2 = \frac{1}{2}\tau^2/a^2$ and $\partial \log b/\partial\|\mathbf{z}_q\|^2 = \frac{1}{2}\tau^2/b^2$; substituting the concentrated values gives the stated forms. □

Consequence: the phase sketch becomes a model prediction. Under the additive difficulty factor of the ρ-insufficiency example, with t the query-level offset s above, ρ has the sign of $\kappa_a\kappa_b$ and, when independent noise dominates, is proportional to κ_b. Since κ_b decreases monotonically in snr while the margin ratio a/b also decreases monotonically in snr, the model predicts that *the margin ratio and ρ move in the same direction, both decreasing as class separation grows*. This is exactly the dichotomy of Fig. A.1: classes that overlap heavily (small snr, ratio near 1) leave only the dependence channel to the readout (the dependence-driven regime), while classes that are already well separated (large snr) have a large contractible margin and weak dependence (the margin-driven regime).

Scope and honesty about the model's limits. Three caveats keep the claim proportionate. (i) The model is isotropic; real frozen features are strongly anisotropic (the effective Fisher rank is $33/384$ on our DINOv2 features), so d should be read as an anisotropic effective dimension. (ii) The prediction is about ordering, not magnitudes: we do not fit τ or Δ. (iii) Within our measurements the ordering holds across datasets (Dogs CLIP: ratio 0.535, ρ=0.112; ImageNet-100 CLIP: 0.836, 0.412; COCO DINO-S: 0.918, 0.42) and within several families (DINOv2-S $0.772/0.381$ vs. DINOv2-B $0.654/0.325$; DINOv3-S $0.835/0.239$ vs. DINOv3-B $0.732/0.225$), but it is not exact point-by-point—e.g. across families DINOv3-S has a larger ratio than DINOv2-S yet a smaller ρ, which the isotropic model cannot explain and which we attribute to family-specific τ and anisotropy. We report the model as the mechanism behind the ordering, not as a quantitative law.

B Implementation and reproducibility

This section specifies every implementation detail needed to reproduce the main text, extending the setup description of Sec. 3.1.

B.1 Datasets, splits, and frozen features

COCO val2014 (Lin et al., 2014): $40{,}137$ validation images with at least one annotation; 80 primary classes; $24{,}081$ images for readout training, $16{,}056$ for evaluation. The primary label is the largest-area annotation; the 12 supercategories are treated as cross-class structure, never as relevance. Frozen features are extracted with timm (Wightman, 2019): DINOv2 ViT-S/14 (student, 384-d), DINOv2 ViT-B/14 (teacher, 768-d), and CLIP ViT-B/16 (768-d pre-projection CLS hidden state). All features are L_2-normalized. Strict R@1 and mAP@10 are computed by cosine similarity with self-exclusion and primary-label relevance.

Stanford Dogs (fine-grained, 120 classes) and **ImageNet-100** (medium-grained, 100 classes, $60{,}000$ gallery / $10{,}000$ query images) use the identical protocol; ImageNet-100 additionally covers DINOv3 ViT-S/16 and ViT-B/16 and MAE ViT-B/16. **CIFAR-100** (100 classes, $60/40$ per-class gallery/query split, $6{,}000$ gallery and $4{,}000$ query images) is used for the phase-boundary test of Sec. 4.5.

B.2 The strict protocol, pseudocode

1. Extract features $\mathbf{x}_i$ for all images; L_2-normalize.
2. Define relevance: $\mathrm{rel}(q, g) = 1$ iff $\mathrm{label}(q) = \mathrm{label}(g)$ and $g \neq q$ (self-exclusion; duplicates of the query are also removed).
3. For each query q, rank the gallery by cosine similarity, and compute strict R@K (is the top-1 relevant?) and mAP@10 (average precision over the top-10, relevance defined by step 2).
4. *Label ceiling*: rank by label identity before any feature is seen; this gives 0.9997 on COCO.
5. *Margin statistics*: per query compute $a = \min_{g:\mathrm{rel}(q,g)} d(q, g)$ and $b = \min_{g:\neg\mathrm{rel}(q,g)} d(q, g)$; fit $(\mu_{\log a}, \sigma_{\log a}, \mu_{\log b}, \sigma_{\log b}, \rho)$ from the per-query pairs. Recomputing R@1 from the margin indicators recovers the direct measurement to within 0.001. These five statistics are the only input of the closed-form predictor of Sec. 3.2: they require the one labelled query set that already defines relevance in step 2 (the predictor needs no additional labels), and no fitting and no per-backbone labelled readout.

B.3 Protocol comparison with prior work

Table B.1 compares our evaluation protocol with representative prior work. While several works study frozen features under various protocols, none enforces strict primary-class retrieval, self-exclusion, supercategory handling, or label-ceiling checks, all of which we show to be necessary for a correct assessment of whether the frozen space is closed. The characterizations are based on the protocols described in the respective publications; we mark a condition as satisfied only if the paper explicitly describes enforcing it.

Method	Strict primary	Self-excl.	Supercat.	Label ceiling	Tests closure
Bulat et al. (Bulat et al., 2024) (FFF)†	×	×	×	×	×
Caron et al. (Caron et al., 2018) (DeepCluster)	×	×	×	×	×
Chen et al. (Chen et al., 2020) (SimCLR)	×	×	×	×	×
Ermolov et al. (Ermolov et al., 2021) (Whitening)	×	×	×	×	×
Ours	✓	✓	✓	✓	✓

Table B.1: Comparison of evaluation protocols across related work as described in the respective publications. The last column records whether the work *tests* whether a frozen feature space is closed to post-processing. We find no prior work that does so under a strict protocol, and we do not attribute a closure claim to any of them. †FFF improves contrastive *pre-training* itself (false-negative reassignment, pseudo-caption augmentation, sigmoid loss) and reports retrieval gains on MSCOCO; it is listed because it reports retrieval on a corpus we also use, not because it claims frozen-space closure.

B.4 Trained readouts and baselines

All readouts are trained on the COCO gallery split only, with L_2-normalized frozen DINOv2-S features as input, and are evaluated on the test split.

SupCon head (Khosla et al., 2020): width 1024, one hidden block (depth 1), output dimension 384, LayerNorm–ReLU, 30 epochs of AdamW at learning rate 10^{-3} and weight decay 10^{-3}, batch 2048, temperature 0.1. Three seeds give 0.738 ± 0.001 strict R@1 (main text reports the 0.739 of seed 0). This single-block head is the one behind the closure, capacity, predictor-validation and mechanism results of the main text.

Five readout objectives (Sec. 3.2): these runs share the optimiser, epoch budget, width, batch size and temperature above, so that only the loss differs, but they use a *two-hidden-block* (depth-2) projector rather than the single-block head; their SupCon reference is therefore 0.7350 ± 0.0017 rather than 0.739, the same direction and a comparable magnitude as the depth ablation of Supplement C (at width 256, depth 2 costs 0.004 against depth 1). Per-seed strict R@1 on test:

Objective	seed 0	seed 1	seed 2	mean ± sd
SupCon	0.7361	0.7359	0.7331	0.7350 ± 0.0017
Triplet margin	0.7305	0.7309	0.7306	0.7307 ± 0.0002
SupCon + margin	0.7394	0.7359	0.7365	0.7373 ± 0.0019
$\mathcal{L}_J$	0.7329	0.7306	0.7285	0.7306 ± 0.0022
SupCon $+\mathcal{L}_J$	0.7351	0.7348	0.7306	0.7335 ± 0.0025

The $\mathcal{L}_J$ head optimizes $-\log \Phi\big((\hat{\mu}_b - \hat{\mu}_a)/\hat{s}\big)$ on in-batch log-distance pairs, the differentiable form of (6); composite objectives use $\lambda = 1$. The spread (0.0002–0.0025) implies the 0.0067 deficit of $\mathcal{L}_J$ against SupCon+margin is roughly three seed standard deviations and the 0.0044 deficit against plain SupCon roughly two: the ordering is not a seed artefact. Triplet margin is at once the most reproducible and the weakest objective, consistent with a well-defined but lower optimum.

LP and MLP probes (Sec. 4.4; full results in Supplement C): LP is an 80-way multinomial logistic regression with class-balanced sampling and logit adjustment (τ=1, the LP++-style recipe (Huang et al., 2024)), 60 epochs, batch 2048, lr 10^{-3}; retrieval embedding is the logit layer $W\mathbf{x}$ (bias-free cosine). MLP is a 384–1024–ReLU–80 adapter, cross entropy, same schedule; the better of hidden (1024-d) and logit (80-d) embeddings is reported (hidden wins).

BitFit and LoRA (Sec. 4.4): DINOv2 ViT-S/14 tuned end-to-end on the COCO train split under primary-label cross-entropy with logit adjustment: 6 epochs of AdamW at max lr 10^{-4}, one-cycle schedule, weight decay 10^{-4}, batch 48 at 224×224 with random-resized-crop and horizontal flip. BitFit (Ben Zaken et al., 2022) unfreezes every bias term ($51{,}456$ params, 0.23% of the backbone); LoRA (Hu et al., 2022) injects rank-8 adapters with α=16 into QKV and output projections of all 12 blocks ($221{,}184$ params, 0.99%). Features are re-extracted with the tuned backbone and scored under the identical strict protocol; we report the better of CLS and logit embeddings (CLS wins for both). Cost: ≈ 20 min/method on one RTX 3080 Laptop GPU versus ≈ 1 min for a frozen-feature readout.

Post-processors (Tab. 1): ZCA whitening fitted on the gallery split; power reweighting $x_i \mapsto x_i \cdot \|x_i\|^{p-1}$; dropping the leading 20 principal components fitted on the gallery; LDA within-class whitening; nested truncation to the first k coordinates ($k \in \{192, 96, 48, 24, 8\}$).

B.5 Hardware and software

New experiments in this revision (protocol-scope, spectral, nested-truncation, and PEFT runs) were executed on a single NVIDIA GeForce RTX 3080 Laptop GPU (16 GB VRAM) with PyTorch 2.6 (CUDA 12.4) and timm; the earlier COCO/Dogs/ ImageNet-100/CIFAR-100 measurements used the same public extractors (DINOv2, DINOv3, CLIP, MAE) and protocol, on a single RTX 4080 Laptop GPU (12 GB). All evaluations are deterministic given the frozen features; all readouts use three fixed seeds.

B.6 Code availability

The evaluation pipeline (protocol, margin statistics, readout training, and all baselines) will be released under a permissive license upon acceptance. The scripts are organized by block (closure, ceiling, mechanism, cross-dataset), and each produces the JSON artifacts cited in the tables above.

C Complete experimental tables

This section gives the full numeric tables behind the main text. Each table extends a condensed counterpart in the main text or reports a measurement discussed there in prose.

C.1 Full closure numbers (Block A and B)

Method	Strict R@1	mAP@10
Frozen cosine (raw)	0.688	0.582
ZCA whitening	0.672	0.566
Power reweight (p=1.0)	0.672	0.565
Drop top-20 PC	0.680	0.573
LDA within-class whitening	0.685	0.578
Nested truncation (best k=192)	0.682	0.574
AE (label-free nonlinear)	0.689	0.583
SupCon (w=1024)	0.739	0.675

Spectral and hierarchical terms added to SupCon: spectral-shaping $\lambda \in \{0.3, 1, 3\}$ gives 0.736, 0.733 ± 0.005, 0.734 against 0.738 for plain SupCon; a graded hierarchical term with $\alpha \in \{0.1, 0.2, 0.4, 0.8\}$ gives 0.736, 0.736, 0.732, 0.731.

C.2 Capacity sweep

Width (depth=1)	train	test
64	0.805	0.715
128	0.838	0.725
256	0.877	0.729
512	0.919	0.735
1024	0.957	0.739

Depth (width=256)	train	test
1	0.877	0.729
2	0.930	0.725
3	0.927	0.720
4	0.935	0.709

The discriminative subspace is small: the Fisher spectrum reaches 90% of its trace in 33 of 224 positive directions (Frobenius-normalized $\mathrm{tr}S_b/\mathrm{tr}S_w = 0.222$).

C.3 READOUT OBJECTIVES: FULL DIAGNOSTICS

Objective	$\hat{\Delta}$	$\hat{s}$	J	$\Phi(J)$	ρ	R@1	mAP@10
SUPCON	0.2635	0.4764	0.5530	0.7099	0.889	0.736	0.675
Triplet margin	0.2966	0.4820	0.6154	0.7308	0.533	0.731	0.665
SUPCON+margin	0.2645	0.4509	0.5867	0.7213	0.854	0.739	0.679
$\mathcal{L}_J$	0.2067	0.3569	0.5793	0.7188	0.337	0.733	0.662
SUPCON+$\mathcal{L}_J$	0.2404	0.4279	0.5619	0.7129	0.734	0.735	0.675

These objectives were trained with the two-hidden-block (depth-2) projector described in Supplement B; their SUPCON reference is 0.7350 ± 0.0017, slightly below the 0.739 of the single-block head used for the closure, capacity and mechanism results.

$\mathcal{L}_J$ attains the smallest denominator (s=0.357, the only objective that controls the term SUPCON ignores) and the largest plug-in prediction gain over SUPCON ($\Phi(J)$=0.719 vs. 0.710), yet realized R@1 does not follow: its numerator is the smallest (Δ=0.207) and it shrinks the tail dependence (ρ: $0.889 \rightarrow 0.337$), moving the Gaussian-plug-in calibration. Per-seed R@1 appears in Supplement B. The ordering by J and the ordering by R@1 disagree (most starkly for triplet margin, largest J and smallest R@1): the plug-in bias is objective-dependent.

C.4 TRAINABLE BASELINES UNDER THE STRICT PROTOCOL

Method	Tunable params	cls. acc.	R@1	mAP@10
Raw cosine (reference)	0	—	0.6886	0.5815
LP++-style linear probe	30,800	0.3792	0.6678	0.5578
MLP adapter (384+1024+80)	476,240	0.5196	0.6920	0.5968
BitFit (biases only)	51,456 (0.23%)	0.6811	0.7295	0.6403
LoRA r=8, α=16	221,184 (0.99%)	0.7261	0.7518	0.6738
SUPCON head (frozen feats)	0	—	0.7350 ± 0.0017	0.675
SUPCON+margin (frozen)	0	—	0.7373 ± 0.0019	0.679

The LoRA backbone-tuned embedding exceeds the frozen-feature readouts under the strict protocol; BitFit does not. Classification accuracy is decoupled from strict R@1 (BitFit's cls. acc. 0.681 far exceeds the linear probe's 0.379 while its R@1 is comparable to the frozen plateau), reinforcing the main text's decomposition argument. LoRA changes the embedding itself, so its ceiling is a different quantity from the frozen-embedding ceilings of the main text; we report it as a method-level reference, not as a contradiction of linear closure. The two frozen-feature rows at the foot of the table use the two-hidden-block (depth-2) projector of the five-objective block (Supplement B); the single-block head of the closure and capacity results reaches 0.739.

C.5 PROTOCOL SENSITIVITY: FULL MATRIX

Backbone	Post-processor	strict	multi-label	super-class
DINOv2-S	raw	0.6886	0.9564	0.7431
DINOv2-S	ZCA	0.6727	0.9405	0.7336
DINOv2-S	drop top-16	0.6806	0.9502	0.7365
DINOv2-S	drop top-64	0.6566	0.9282	0.7202
DINOv2-B	raw	0.6890	0.9672	0.7380
DINOv2-B	ZCA	0.6723	0.9488	0.7250
DINOv2-B	drop top-16	0.6845	0.9630	0.7345
DINOv2-B	drop top-64	0.6659	0.9479	0.7201

Relevance density (mean relevant items per query): 1,832 (strict), 5,498 (multi-label), 2,804 (super-class); 100% of queries gain relevant items under multi-label relevance. For every row the sign of the post-processor effect is identical across the three relevance definitions.

C.6 NESTED TRUNCATION AND SPECTRAL STATISTICS

k (first-k dims)	R@1	mAP@10
384	0.6886	0.5815
192	0.6818	0.5736
96	0.6653	0.5555
48	0.6340	0.5098
24	0.5399	0.4116
8	0.2744	0.1843

Statistic	DINOv2-S	DINOv2-B
dimension	384	768
top-1 eigenvalue share	2.93%	2.81%
top-10 share	18.5%	16.5%
effective rank (entropy)	232.6	391.3
participation ratio	142.6	190.3
dims for 90% variance	245	469

The spectra are spread, not degenerate: no dominant direction exists for whitening to repair, which is the precondition under which ZCA must hurt (see main text Sec. 4.4).

C.7 MAE SPECTRAL DEGENERACY (STANFORD DOGS)

To test the spectral-precondition claim directly, we measured the frozen-feature covariance spectrum for MAE-B and the two DINOv2 backbones on the 20,580 Stanford Dogs images, at each model's native input resolution, L2-normalizing the [CLS] token ($20{,}580 \times D$). Leading eigenvalue share is relative to the trace; effective rank is the participation ratio $(\sum_k \lambda_k)^2 / \sum_k \lambda_k^2$; coordinates-for-90% is the number of eigenvectors required to reach 90% of the variance.

Backbone (Stanford Dogs, n=20,580)	dim	leading eig. share	eff. rank	coords. for 90% var.
MAE-B	768	92.4%	1.17	1
DINOv2-S	384	5.8%	48.1	126
DINOv2-B	768	3.9%	84.5	259

MAE-B is spectrally degenerate: a single coordinate carries 92.4% of the trace and reaches 90% of the variance alone, whereas the DINOv2 spectra are spread (leading shares 3.9–5.8%; hundreds of coordinates for 90%). Because the contrast is measured on the *same* images and labels, it is a property of the backbone, not of the dataset. This confirms that MAE is the one family here that occupies the degenerate regime where second-order repair (ZCA) can help, matching its reversed sign in the main text.

C.8 TRAIN-FREE MODEL SELECTION: FULL TABLE

Model (ImageNet-100)	J	$\Phi(J)$	observed R@1	gap
DINOv2-B	0.7035	0.7591	0.7836	+0.0245
DINOv3-B	0.6971	0.7571	0.7900	+0.0329
DINOv2-S	0.5560	0.7109	0.7280	+0.0171
CLIP-B	0.5147	0.6966	0.7240	+0.0274
DINOv3-S	0.4989	0.6911	0.7204	+0.0293
MAE-B	−0.9542	0.1700	0.1250	−0.0450

Spearman rank correlation between $\Phi(J)$ and observed R@1: 0.943; mean absolute error 2.9 points; all residuals positive except MAE (systematically conservative). The predictor correctly flags MAE as unusable.

C.9 MECHANISM STATISTICS

	Frozen		SUPCON	
	train	test	train	test
a/b	0.92	0.92	0.53	0.89
ρ	0.52	0.42	−0.23	0.87
R@1	0.695	0.688	0.957	0.739

The test-side driver is the dependence rise at near-fixed margins; the train-side driver is margin contraction with dependence lost.

Variable	Distribution	KS	AD	AD 5% crit.
$\log a$	Normal	0.0644	134.7	0.787
$\log a$	Gumbel	0.1254	487.6	0.756
$\log a$	Logistic	0.0509	134.6	0.660
$\log b$	Normal	0.0705	192.5	0.787
$\log b$	Gumbel	0.1602	672.1	0.756
$\log b$	Logistic	0.0618	157.0	0.660

Table D.2: Goodness-of-fit of the marginals ($n = 16{,}056$; KS 5% critical value ≈ 0.011). All families are rejected at the 5% level; Logistic fits best and the EVT-motivated Gumbel worst. The Gaussian copula is separately rejected by a Rosenblatt KS test (statistic 0.048, $p < 10^{-15}$).

C.10 CAUSAL COUNTERFACTUAL: FULL TABLE

Counterfactual (test regime)	Predicted R@1
Frozen margins, frozen ρ (0.42)	0.665
Frozen margins, SUPCON ρ (0.87)	0.772
SUPCON margins, frozen ρ (0.42)	0.615
SUPCON margins, SUPCON ρ (0.87)	0.702
ρ-only gain	$+0.107$
margin-only gain	-0.049

On the training regime the order reverses (margins alone $0.678 \rightarrow 0.974$; dependence change negative): the two channels are the mechanistic signatures of the train/test asymmetry in the main text. The CIFAR-100 counterfactual gives a margin-only gain of $+0.054$ against a dependence-only gain of $+0.015$, establishing that cell as margin-driven.

D STATISTICAL DIAGNOSTICS, STAIRCASE, AND CALIBRATION

This section collects the statistical checks that support the "working approximation" framing of Sec. 3.1 and the calibration discussion of the main text: the model staircase, the empirical calibration factor, and the consistency controls.

D.1 GOODNESS-OF-FIT DIAGNOSTICS

Table D.2 reports the Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) statistics on the $n = 16{,}056$ COCO test queries against fitted Normal, Gumbel and Logistic margins; the main text quotes the headline numbers. The fitted log-margin quantile–quantile plots (Fig. D.2) show that the Normal approximation is reasonable in the central region but deviates in the tails, consistent with the AD statistic being amplified by tail sensitivity; the Gumbel fit is visibly worst. The Gaussian copula is rejected by a Rosenblatt-transform KS test (statistic 0.048, $p < 10^{-15}$); the empirical rank-based copula shifts the prediction by only 0.003 (see below), so the residual is not a marginal-family artefact.

D.2 MODEL STAIRCASE: HEAVY TAILS ARE NECESSARY, DEPENDENCE IS SUFFICIENT

Comparing the predictor against successively simpler models identifies which assumption does the work and what a logistic-regression-style score alone achieves:

Model	Strict R@1 (raw test)
Gaussian second-order ceiling	0.081
Lognormal margins, ρ=0 (independent)	0.631
Gaussian copula ($\hat{\rho}$=0.42)	0.665
Observed frozen plateau	0.688

The Gaussian baseline drops *both* the heavy tail and the dependence, so its failure (0.081) conflates the two factors; the staircase decouples them. Keeping the heavy-tailed (lognormal) margins but forcing

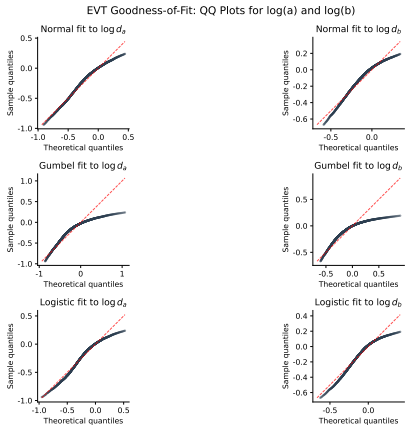

Figure D.2: Quantile–quantile diagnostics of fitted marginals for $\log a$ (top) and $\log b$ (bottom).

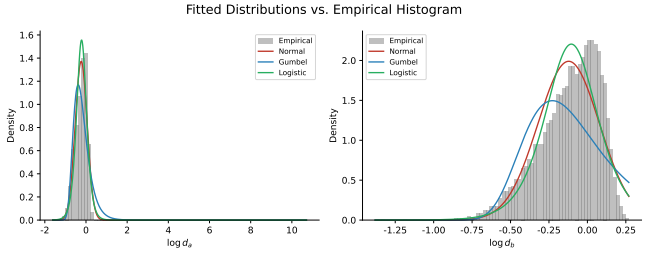

Figure D.3: Empirical histograms of $\log a, \log b$ against fitted Normal, Gumbel and Logistic densities.

independence (ρ=0, i.e. the univariate logistic/probit score on the log-distance difference) already reaches 0.631; adding the fitted copula dependence reaches 0.665. Both the heavy tail (margins) and the dependence are therefore necessary to reach the observed ceiling, and the fitted $\hat{\rho}$ is why the ρ=0 rung systematically under-predicts. On the readout regime the same staircase gives $0.590 \rightarrow 0.702$ against an observed 0.739 (Fig. 3), so the dependence rung carries the projected regime.

D.3 Consistency controls

A model-free control, which bins the empirical joint $(\log a, \log b)$ and computes the fraction of mass above the diagonal, reproduces the observed strict R@1 to within 10^{-4} (0.6884 vs. 0.6883): the gap between the parametric predictor and the control is attributable to the copula/marginal parametrization, not to the margin identity. Two further checks localize it. First, replacing the log-normal marginal parametrization by the *empirical* rank-based copula moves the prediction from 0.665 to 0.662 (a 0.003 shift), so the residual is not a marginal-family artefact. Second, a bootstrap 95% interval on the observed plateau has half-width ± 0.007, far smaller than the 0.024 residual: the gap is a genuine model limitation, not sampling noise. Both point to the same attribution: the two-parameter Gaussian copula cannot express the tail dependence that extreme queries carry, exactly where the heavy-tailed margins matter most.

D.4 Empirical calibration

Define $\delta \equiv \mathrm{R@1}_{\mathrm{obs}} - \widehat{\mathrm{R@1}}$ (positive when the copula under-predicts). Within each (family, coarseness) cell δ is nearly constant: for self-supervised features on coarse COCO, $\delta \approx +0.016$–0.024 in raw regimes and $+0.036$–0.053 in readout regimes (where non-Gaussian tail dependence is strongest); on ImageNet-100, $+0.017$–0.033 raw and $+0.035$–0.057 readout; on fine-grained Dogs, $+0.07$–0.19; for text-contrastive CLIP the sign flips ($\delta \approx -0.043$). The corrected estimate $\mathrm{R@1}_{\mathrm{cal}} = \widehat{\mathrm{R@1}} + \delta$ is train-free (the calibration is a measurement, not a learned head).

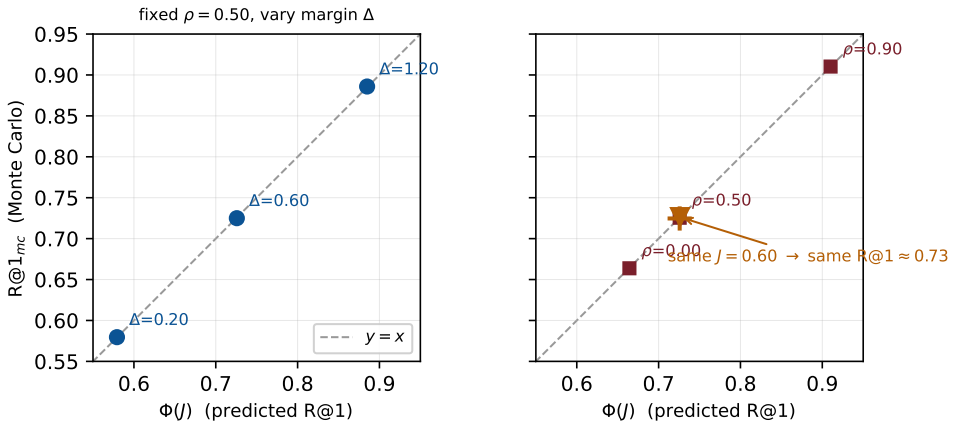

Figure D.4: Synthetic control: predicted $\Phi(J)$ against Monte-Carlo strict R@1 under the Gaussian-copula model. Left, fixed ρ=0.50 and margin $\Delta \in \{0.20, 0.60, 1.20\}$; right, fixed Δ=0.60 and $\rho \in \{0, 0.50, 0.90\}$. All points lie on the identity line R@1=$\Phi(J)$; the two settings sharing J=0.60 ($(\Delta$=0.60, ρ=0.50) and (Δ=0.85, ρ=0)) coincide, so J is sufficient whenever the model holds.

δ can also be estimated online from a small labeled query set drawn from the same cell: its estimator has standard error $\sqrt{p(1-p)/n} \approx 0.004\sqrt{16056/n}$ at the COCO operating point (p=0.688, matching the measured bootstrap half-width $\pm$0.007 at n=16,056). A few thousand labeled queries suffice: n=2000 gives a 95% interval of $\pm$0.020 on δ, and even n=500 gives $\pm$0.040—both far below the δ values themselves (0.016–0.19). This preserves the "train-free" property in the sense that matters: no gradient steps, no hyperparameters, no learned head.

D.5 Scope of the predictor

The closed-form predictor does not replace measurement: its residual δ does not shrink with n (Equation (5) bounds only the estimation term). The claim is narrower: within a (family, granularity) cell the predictor tracks *relative* ceilings (which head, which width, which family) at a cost of five moments on frozen features, and it yields the attribution and counterfactuals that measurement cannot. Where an absolute number is needed, δ supplies it at the cost of a labeled query set.

D.6 Synthetic control: J is sufficient where the model holds

To separate a theoretical defect of J from a violation of our modelling assumptions, we draw data *exactly* from the Gaussian-copula model and check whether the predictor remains sufficient. Fig. D.4 varies the margin at fixed dependence (left) and the dependence at fixed margin (right); in every case the Monte-Carlo strict R@1 sits on the identity line (R@1=$\Phi(J)$ to within 10^{-3}). Two configurations with the same J—(Δ=0.60, ρ=0.50) and (Δ=0.85, ρ=0)—land on the same point, so J, not the margin or the dependence alone, drives the ceiling. The real-data shortfall ($\mathcal{L}_J$ losing to SupCon, §5) is therefore attributable to training moving the very statistics the model fixes, not to J being theoretically insufficient.

D.7 Synthetic counterfactual: shared difficulty inflates ρ without touching retrieval

The ρ-insufficiency proposition of Section 3.2 is analytic; here we validate its construction directly on the fitted COCO test model ($\mu_{\log a} = -0.209$, $\sigma_{\log a} = 0.250$, $\mu_{\log b} = -0.110$, $\sigma_{\log b} = 0.164$, $\rho = 0.420$). Following the proposition, we inject a shared query-difficulty term κt ($t \sim \mathcal{N}(0,1)$) into *both* $\log a$ and $\log b$, keeping the residual pair at its fitted scale and correlation, and re-measure the correlation, the spread s, and Monte-Carlo strict R@1 (n=4 $\times$ 10^6 queries per row, seed 0; `scripts/rho_insufficiency_sim.py`).

The invariance is stronger than the aggregate numbers suggest: since $\log b - \log a$ contains no κt term, the *per-query correctness pattern* is bit-identical across all κ (verified empirically in addition to

shared difficulty κ	$\rho(\log a, \log b)$	s	Monte-Carlo R@1
0.00	0.420	0.2341	0.6647
0.25	0.754	0.2341	0.6647
0.50	0.909	0.2341	0.6647
1.00	0.974	0.2341	0.6647
2.00	0.993	0.2341	0.6647
4.00	0.998	0.2341	0.6647

Table D.3: Shared query difficulty drives $\rho \to 1$ while s and strict R@1 are unchanged. $\Delta/s = 0.425$ reproduces the COCO test J of Section 3.2, and $\mathrm{R@1} = \Phi(J)$ to Monte-Carlo precision.

Frozen family	Dim	Strict R@1	ZCA
DINOv2 ViT-S (student)	384	0.688	0.672
DINOv2 ViT-B (teacher)	768	0.689	0.672
DINOv3 ViT-S	384	0.687	0.660
DINOv3 ViT-B	768	0.696	0.674
CLIP ViT-B/16	768	0.422	0.419

Table E.4: The strict-retrieval plateau is family-dependent yet robust across model generations. Self-supervised DINO features (either capacity, and across the DINOv2/DINOv3 generations) sit in a tight plateau at ~ 0.69; CLIP, trained with text-contrastive supervision, reaches only 0.422 on the identical strict splits. ZCA hurts all five.

holding by construction). A shared difficulty shift therefore cannot be corrected by any reweighting of queries; it inflates the correlation statistic and nothing else. This is the empirical content of "high ρ may be nothing but shared difficulty": ρ can be moved arbitrarily by a factor that has no effect whatsoever on retrieval, which is why optimizing ρ is the wrong target. This is not in tension with the dependence-driven mechanism measured on COCO test (Section 4.5): there the ρ rise occurs at a fixed marginal scale and genuinely compresses the spread—holding the post-training marginals fixed, dropping ρ to 0 would let s disperse from 0.243 to 0.565—whereas the counterfactual inflates ρ by inflating a shared component, which leaves s untouched. The quantity the ceiling responds to is s, and ρ matters only through it.

E Cross-Family Generality

A diagnostic claim that holds only for one feature family is weak evidence for a mechanism. We repeat the two central measurements – the raw plateau and the effect of the SupCon readout – on CLIP ViT-B/16 (Radford et al., 2021), a fundamentally different feature family trained with a contrastive text-image objective, on the *same* strict splits as the DINO features.

E.1 The raw plateau is family-sensitive

Table E.4 shows a large gap between feature families: DINO features of either capacity – and across the DINOv2 and DINOv3 generations – sit in a tight plateau of 0.687–0.696 strict R@1, while CLIP reaches only 0.422. The plateau thus holds not only across capacities but across model generations: despite a substantially different self-supervised objective and architecture, DINOv3 features (Siméoni et al., 2025) land on the same raw-retrieval ceiling as DINOv2, with the same capacity-invariance (0.687 at 384-dim vs. 0.696 at 768-dim). Under the strict protocol, self-supervised object-centric features (DINO) are far better primed for primary-class retrieval than contrastive text-space features (CLIP). A loose protocol would wash out this information, which is why the strict protocol matters: it separates the plateau claim from any single feature family. In both families, ZCA whitening *hurts* strict retrieval (-0.016 and -0.003 for DINOv2; -0.027 and -0.022 for DINOv3).

E.2 The SupCon opening is mechanism-general

If the mechanism we identify in Section 4.2 is real, then the SupCon readout should open both families, with the larger opening where the raw base is lower. We train the identical supervised-contrastive head on each frozen family and measure strict R@1 before and after.

Frozen family	Raw	SupCon	Δ
DINOv2 ViT-S	0.688	0.735	+0.047
CLIP ViT-B/16	0.422	0.509	+0.087

Table E.5: Supervised-contrastive readout opens both frozen families. The opening is $+4.7$ points for DINO and $+8.7$ for CLIP: the mechanism is not family-specific, and the lower raw base requires a larger opening.

Frozen family	Dim	Raw	ZCA	AE (unsup.)	SupCon
DINOv2 ViT-S	384	0.813	0.732	0.794	0.853
DINOv2 ViT-B	768	0.853	0.817	0.850	0.883
CLIP ViT-B/16	768	0.905	0.867	0.905	0.918

Table F.6: Fine-grained Dogs: strict R@1. The same two-factor pattern as COCO reproduces: ZCA whitening hurts all three families; the label-free nonlinear autoencoder does *not* open the space (0.79–0.91, at or below raw); while the SUPCON head opens each family. Baselines are higher on Dogs because fine-grained classes are better separated at the raw-feature level.

Table E.5 shows the opening generalizes: DINO gains $+4.7$ points ($0.688 \to 0.735$) and CLIP gains $+8.7$ points ($0.422 \to 0.509$). The direction and the mechanism signature – margin contraction and decorrelation on the training split, and a sharp dependence rise on the test split (Section 4.5) – are present in both families, so the finding is not an artifact of DINO-specific geometry.

E.3 SUMMARY

The strict plateau and the SUPCON opening are robust across feature families, but their *values* are family-calibrated: CLIP starts lower and needs a larger opening (quantified in Section 3.2).

F CROSS-DATASET VALIDATION ON FINE-GRAINED STANFORD DOGS

The COCO corpus is coarse-grained (80 classes with $\sim$1 150 same-class neighbors per query) and heavily imbalanced. To test how much of our account – the closure decomposition, the closed-form ceiling, and the mechanism – is an artifact of that regime, we repeat the strict protocol on a *fine-grained* dataset: Stanford Dogs (Khosla et al., 2011) (120 breed classes, $\sim$171 images per class), with the same three frozen feature families and the identical 60/40 per-class gallery/test split as described in Section 3.1 (12,345 gallery / 8,235 eval). The two corpora differ sharply in both granularity and per-class support, so agreement between them is strong evidence that the findings are not tied to one dataset.

F.1 CLOSURE AND OPENING GENERALIZE

Table F.6 reproduces every element of the closure decomposition on the new dataset. As on COCO, (i) linear post-processing does *not* help (ZCA is uniformly harmful, -0.04 to -0.08), (ii) a nonlinear but *unsupervised* readout does not open the space (0.79–0.91, essentially the raw plateau), and (iii) the *supervised-contrastive* readout opens all three families. The raw plateaus are higher than on COCO (DINO 0.81–0.85 vs. 0.69; CLIP 0.91 vs. 0.42) because the fine-grained classes are less confusable at the level of a single same-class vs. cross-class neighbor. The *direction* of the decomposition, however, is identical: nonlinearity alone does nothing, and only the coupling of nonlinearity with supervision opens the space. This is the strongest cross-dataset confirmation of the two-factor claim.

F.2 THE MECHANISM IS THE SUPERVISED CONTRACTION OF THE SAME-CLASS MARGIN

Table F.7 repeats the margin/dependence diagnostics and the 2×2 counterfactual on the Dogs test split. The result is the sharpest and most informative finding: on the fine-grained corpus the mechanism that drives the SUPCON opening is the *contraction of the same-class margin*, not a rise in dependence.

Frozen family	a/b (raw→Sup)	ρ (raw→Sup)	Counterfactual
DINOv2 ViT-S	$0.76 \to 0.45$	$+0.10 \to +0.06$	margin $+0.12$; ρ -0.003
DINOv2 ViT-B	$0.66 \to 0.28$	$+0.08 \to -0.05$	margin $+0.14$; ρ -0.009
CLIP ViT-B/16	$0.53 \to 0.18$	$+0.11 \to -0.04$	margin $+0.14$; ρ -0.015

Table F.7: Mechanism on the Dogs test split. Unlike COCO – where the gain was a dependence rise with margins nearly frozen – on Dogs the positive channel is the same/cross-class margin ratio, which contracts sharply (a/b drops 26%–66%), while the dependence ρ stays small or slightly negative. The 2×2 counterfactual confirms: the margin-only gain is positive ($+0.12$ to $+0.14$) and the ρ-only gain is negligible or negative.

Dogs regime	Observed R@1	Copula pred.	Gap
DINO-S, raw	0.813	0.626	-0.187
DINO-S, sup	0.853	0.742	-0.111
DINO-B, raw	0.853	0.667	-0.186
DINO-B, sup	0.883	0.796	-0.087
CLIP, raw	0.905	0.724	-0.181
CLIP, sup	0.918	0.847	-0.071

Table F.8: Closed-form predictor on fine-grained Dogs. The predictor is a consistent *delta* within a family and dataset – it tracks the raw→sup change in every row – but its absolute calibration degrades relative to COCO, under-predicting by 0.07–0.19 (vs. $\sim$0.02 on COCO). We therefore restate the ceiling claim as per-regime.

The corrective to the COCO-centered mechanism in Section 4.5 is precise.[1] The COCO claim was that on the *test* split the opening is driven by a sharp rise in ρ with the margins nearly fixed. On Dogs the opposite happens: the margins contract strongly (a/b $0.76 \to 0.45$ for DINO-S, $0.66 \to 0.28$ for DINO-B, $0.53 \to 0.18$ for CLIP) while ρ does not rise (it slightly falls). In both datasets the identical supervised-contrastive objective is at work; the difference is which statistic moves enough to be measured. On COCO, the frozen margins are already near-optimal (raw a/b=0.92), so the readout can only add value by raising dependence; on Dogs, the raw margin is large and contractible (heavy fine-grained overlap), so the readout's main effect is to shrink a while leaving the dependence small. The universal mechanism is the supervised contraction of the same-class margin; the observable statistic through which it acts is regime-dependent.

A quantitative switch criterion. The two regimes can be separated by a single, measurable feature-statistic: the *raw* same/cross-class nearest-neighbor margin ratio a/b before any readout, which serves as a class-overlap proxy. When a/b is already close to 1 (COCO: 0.92), the same-class neighbor is barely closer than the cross-class neighbor, there is little margin left to contract, and the only available channel is the dependence rise; when a/b is well below 1 (Dogs: 0.53–0.76), the classes are strongly overlapped in feature space and a contractive readout has room to act on the margin. This yields a falsifiable prediction: datasets whose frozen margin ratio is near 1 should show the COCO mechanism (dependence-driven test gain), and datasets whose frozen margin ratio is well below 1 should show the Dogs mechanism (margin-driven gain). The three families within each dataset cluster by their raw margin ratio exactly as this criterion orders them: on Dogs the raw margins order as CLIP 0.53 < DINO-B 0.66 < DINO-S 0.76, all well below 1 (margin-driven); on COCO all families sit near 0.9 (dependence-driven). The measured mechanism in every row of Table F.7 follows the ordering. The relation is therefore an empirical ordering on the tested datasets; a broader multi-dataset sweep would test its boundary more precisely.

F.3 Per-regime calibration

Table F.8 records how the closed-form predictor behaves on Dogs. The predictor *qualitatively* tracks the opening – in every family the predicted R@1 rises when the observed R@1 rises (raw→sup), so the direction ("how much a supervised head will open") is recovered. But its *absolute* accuracy degrades sharply relative to COCO: it under-predicts by 0.07–0.19 rather than $\sim$0.02. The reason is

[1] We ran the identical pipeline on all three families for Dogs; Table F.7 reports the test statistics and the margin- and ρ-only gains from the 2×2 counterfactual.

the fine-grained regime: with only $\sim$102 gallery items per class (vs. 1,152 on COCO), the heavy-tailed extreme-value assumption (Section 3.2) is a poorer fit, so the predicted joint is too Gaussian and under-estimates the tail mass that a small margin advantage actually captures. The contrast with Table eftab:val of the main paper (COCO ± 0.02) defines the scope of the train-free predictor. We therefore restate the contribution in Section 3.2 as a *per-regime consistent-delta* predictor – accurate within a dataset and family, and reliable about the *sign/direction* of an opening; absolute calibration is therefore regime-dependent.

F.4 Duplicate audit

Because the opening is small on Dogs and the same-class margin contraction could in principle be manufactured from exact duplicate images, we apply the same duplicate audit as on COCO. The near-duplicate fraction inside the gallery is 0.004 on the training split, and the fraction of test queries whose nearest same-class neighbor is an exact duplicate is 0.001. The SupCon gain is therefore not an artifact of retrieval over copied examples; it reflects a genuine re-organization of the fine-grained same-class vs. cross-class margin structure, in agreement with the margin statistics in Table F.7.

Summary. Stanford Dogs confirms the two-factor closure claim (nonlinearity alone does not open; supervised nonlinearity does) and the supervised-contraction mechanism, while exposing exactly where the COCO picture must be corrected: the test-side driver is regime-dependent (dependence on coarse COCO, margin on fine-grained Dogs), and the closed-form predictor is a per-regime consistent-delta, not a tight universal bound.

G Experimental details deferred from the main text

This section collects the tables and quantitative details that support the main text but are deferred here to keep the body focused; no reported value is omitted.

G.1 Predictor validation table

The observed/predicted pairs behind Fig. 2 of the main text are summarized in Table 3 of the main paper (gap = predicted $-$ observed).

G.2 Spectral-shaping and graded-hierarchical terms

Added to the SupCon objective, a spectral-shaping term is neutral to mildly harmful at every strength (0.736, 0.733, 0.734 against the 0.738 baseline for λ=$0.3, 1, 3$), and a graded hierarchical term is harmful at every weight (0.736, 0.736, 0.732, 0.731 for α=$0.1, 0.2, 0.4, 0.8$, the deficit widening with α).

G.3 Capacity independence and the capacity sweep

Running the identical protocol with the strictly larger frozen teacher gives 0.689 strict R@1 against the student's 0.688 (statistically identical), so the plateau is capacity-independent. Sweeping readout capacity, widening the projector monotonically improves strict R@1 while adding depth past one layer degrades it; the discriminative subspace is small (the Fisher spectrum reaches 90% of its trace in 33 of 224 positive directions). We use the width-1024, depth-1 head for the closure, capacity, predictor-validation and mechanism results (Supplement B); the five-objective comparison of Sec. 4.3 uses a depth-2 projector of the same width.

G.4 Rank-order screening table

The six ImageNet-100 backbones behind the Spearman 0.943 screening claim:

G.5 Multi-label relevance and spectral details

Under multi-label relevance (a query is relevant if it shares *at least one* of the 80 categories), the task becomes roughly three times denser (mean relevant items per query $1{,}832 \rightarrow 5{,}498$); the full

Model (ImageNet-100)	J	$\Phi(J)$	observed R@1*	gap
DINOv2-B	0.7035	0.7591	0.7836	+0.0245
DINOv3-B	0.6971	0.7571	0.7900	+0.0329
DINOv2-S	0.5560	0.7109	0.7280	+0.0171
CLIP-B	0.5147	0.6966	0.7240	+0.0274
DINOv3-S	0.4989	0.6911	0.7204	+0.0293
MAE-B	−0.9542	0.1700	0.1250	−0.0450

*Spearman $\rho = 0.943$ (n=6, exact two-sided p=0.017); the sample is small, so the correlation is suggestive rather than definitive.

Table G.9: Preliminary rank-order screening on ImageNet-100 (six backbones; exploratory). All $\Phi(J)$ and observed R@1 values are for *raw frozen features*. Gap is observed − predicted, so a positive value means the predictor conservatively under-predicts; ranking by $\Phi(J)$ matches observed R@1 (Spearman 0.943, mean error 2.9 points) and flags the raw MAE features as unusable, serving as a screening demonstration.

correlation matrix lives in Supplement Section C. The spectral readouts behind the precondition claim: on ImageNet-100, DINOv2-S's leading eigenvalue carries 2.9% of the trace (effective rank 232 of 384; teacher 2.8%, rank 391 of 768). For MAE, the singular regime: raw R@1 0.125, ZCA $0.125 \to 0.393$, SUPCON restores 0.714, with raw margin ratio $a/b = 1.44 > 1$; on the Stanford Dogs proxy, MAE-B's leading eigenvalue carries 92.4% of the trace (effective rank 1.17 of 768) versus 5.8% and 3.9% for DINOv2-S/–B (effective ranks 48 of 384 and 85 of 768).

G.6 GOODNESS-OF-FIT DETAIL

Against the $n = 16{,}056$ COCO test queries, the KS 5% critical value is ≈ 0.011 and every fitted margin (Normal, Gumbel, Logistic) exceeds it by 4–15×. The EVT-motivated Gumbel is the worst fit (KS 0.125/ 0.160 against Logistic's 0.051/0.062); the Gaussian copula departs by a Rosenblatt-transform KS test (0.048, $p < 10^{-15}$). This sample size detects any deviation above 1.1% between the empirical and fitted CDFs.

G.7 MECHANISTIC COUNTERFACTUAL DECOMPOSITION

With the marginals frozen and only the dependence ρ scanned, the coupled model moves predicted R@1 from 0.590 (ρ=0) to 0.702 (ρ=0.87). The 2×2 counterfactual over the two channels:

- frozen marginals: the SUPCON dependence contributes +0.107 to predicted R@1 on test;
- frozen dependencies: the SUPCON marginal distribution costs −0.049 on test;
- on the training partition the order reverses: marginals move $0.678 \to 0.974$ (positive) while the dependence change becomes negative.

Hence on test the head opens the space by raising ρ under fixed marginals, while on training it contracts the marginals at the expense of dependence. For CIFAR-100 the original margin ratio is 0.921, within 0.001 of COCO's 0.920, yet there are only 60 samples per class; the one-dimensional reading predicts a dependence-driven cell, but counterfactual analysis yields a margin gain of +0.054 versus a dependence gain of +0.015, so CIFAR-100 is margin-driven.

G.8 PHASE-BOUNDARY THRESHOLDS

The phase-diagram axes are the raw margin ratio and per-class support (both log scales); the axes and the threshold dividing dependence- from margin-driven regimes are read directly from the data, not inferred from theory. The boundary, fit to four datasets and eleven cells, is $a/b \gtrsim 0.85$ together with per-class support $\gtrsim 150$. The one-dimensional reading fails on our fourth dataset, CIFAR-100: it has an original margin ratio 0.921, within 0.001 of COCO's 0.920, but only 60 samples per class, and counterfactual analysis shows it is margin-driven, not dependence-driven.

Objective	strict R@1 (3 seeds)
SupCon	0.735 ± 0.002
Triplet margin	0.731 ± 0.000
SupCon+margin	$\mathbf{0.737 \pm 0.002}$
$\mathcal{L}_J$	0.731 ± 0.002
SupCon+$\mathcal{L}_J$	0.734 ± 0.003

Table G.10: Performance of the derived objective on COCO test (depth-2 head; the depth-1 closure head reaches 0.739).

G.9 Derived objective: seed results

Over three seeds under the strict protocol, the derived objective $\mathcal{L}_J = -\hat{\Delta}/\hat{s}$ performs comparably to SupCon+margin: SupCon 0.735 ± 0.002, triplet margin 0.731 ± 0.000, SupCon+margin $\mathbf{0.737 \pm 0.002}$, $\mathcal{L}_J$ 0.731 ± 0.002, SupCon+$\mathcal{L}_J$ 0.734 ± 0.003. $\mathcal{L}_J$ attains the smallest denominator of any objective ($s = 0.357$ vs. 0.476 for SupCon+margin) yet also the smallest numerator ($\Delta = 0.207$) and shrinks the tail dependence (ρ: $0.889 \to 0.337$), because the interpolation bias is not invariant across targets.

G.10 Mechanism on CLIP

The test-side mechanism generalizes beyond DINO. Figure G.5 shows the same signature on CLIP: a near-constant margin ratio with a sharp rise in ρ.

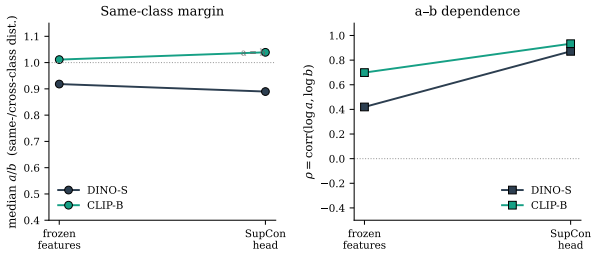

Figure G.5: Mechanism on the test split for DINO (dark) and CLIP (teal). Left: margin ratio a/b, essentially unchanged for both families. Right: dependence ρ, sharp rise for both families. The test-side mechanism is "near-unity relative margin + strong dependence", not margin contraction.

G.11 Training dynamics of the derived objective

The $\mathcal{L}_J$ result in Section 4.3 of the main paper illustrates a principle that transcends this diagnostic. A statistic can be *sufficient for prediction*—the ceiling under a fixed feature distribution—yet *ineffective as an objective*, because training moves the statistics it conditions on rather than holding them fixed. This mirrors the estimand-versus-optimization-landscape tension familiar from variational inference and score matching. The closed form is a predictor; its performance as a training objective is a separate question.

G.12 Full comparison of the derived objective

Table G.10 gives the five-objective comparison deferred from the main text; all rows use the depth-2, width-1024 projector, whose SupCon baseline is 0.735 rather than the depth-1 closure head's 0.739.

$\mathcal{L}_J$ attains the smallest denominator of any objective (s=0.357), yet also the smallest numerator (Δ=0.207). The gradient path of $\mathcal{L}_J$ shows why it is a predictor rather than an objective: it drives s down by attacking the cross-distance correlation (ρ: $0.889 \to 0.337$) that the ceiling rewards, so the quantity it minimises is not the one the ceiling depends on. Over three seeds SupCon+margin and SupCon+$\mathcal{L}_J$ show no significant difference from the head alone.