FROM INVERTED CODES TO ANISOTROPIC ENCODERS: A SINGLE RATIO PREDICTS FEW-SHOT OUT-OF-DISTRIBUTION ACCURACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot out-of-distribution (OOD) classification requires pre-trained encoders to generalize to unseen classes and novel corruptions. While performance variations across models are often attributed to scale or data, we demonstrate that this generalization is fundamentally a property of latent space geometry. We introduce the anisotropy ratio, a single metric quantifying how strongly an encoder contracts within-class (nuisance) variation relative to how strictly it preserves between-class (semantic) separation. Theoretically, we prove a margin law guaranteeing that one-shot nearest-centroid classification succeeds whenever the task margin adequately exceeds this ratio. Furthermore, we establish an isometry floor showing that invertible representations cannot improve this ratio beyond the raw input geometry, offering a geometric explanation for the performance gap between generative and contrastive models. Practically, we demonstrate that this ratio can be estimated from unlabeled, in-distribution data alone-requiring no gradients or access to the shifted test set. Extensive experiments across synthetic geometries and four image benchmarks, spanning seven encoder families and the CIFAR-10-C robustness suite, confirm that this label-free estimator strongly predicts actual OOD accuracy. Finally, all theoretical results are fully machine-checked in the Lean 4 proof assistant, ensuring mathematical rigor.

1 INTRODUCTION

A frozen encoder is a geometry. It takes a cloud of images and returns a cloud of vectors, and everything a downstream few-shot learner can do is decided by the shape of that second cloud. When the downstream task is the hardest kind - classes never seen during training, presented under a corruption never seen during training, with only a handful of labelled examples per class - the shape is essentially all that is left, because there is no data with which to repair a bad one.

Which shapes work? The intuition is old and simple: a good code should collapse the variation *within* a class while preserving the separation *between* classes (Papyan et al., 2020; Khosla et al., 2020). What has been missing is a statement precise enough to be false. This paper supplies one. We define the *anisotropy ratio* of an encoder as the ratio of its directional Lipschitz constant along within-class (nuisance) directions to its directional lower Lipschitz constant along between-class (semantic) directions, and we prove a margin law: one-shot nearest-centroid classification is correct whenever the task's normalised margin exceeds four times this ratio (Theorem 1). The constant four is not decorative - it is the exact price of the fact that both the query and the single support example carry their own nuisance, and we prove no smaller constant works in general (Proposition 3), while also proving that it halves to two when the encoder maps semantic and nuisance directions to orthogonal latent subspaces (Proposition 2).

The margin law immediately explains a dichotomy that is otherwise puzzling. Any bijective encoder - an orthogonal map, a normalising flow (Dinh et al., 2017), an invertible residual network (Behrmann et al., 2019) - has an anisotropy ratio of at least one, because it cannot contract any pair of points at all (Proposition 5, Figure 1b). Such a code preserves whatever margin the raw inputs already had and can never manufacture more. A learned, non-invertible encoder faces no such floor, and a contrastive

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

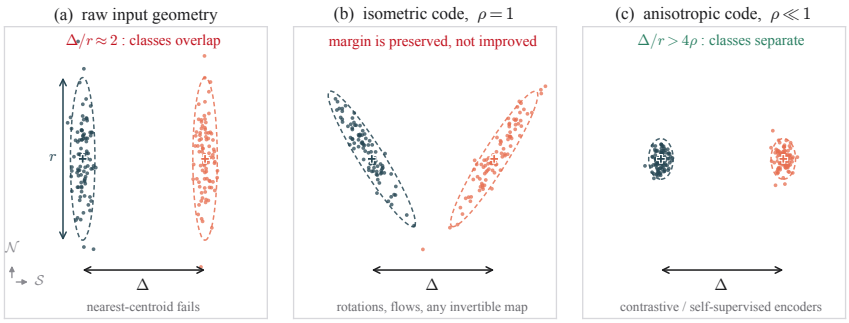

Figure 1: The mechanism. **(a)** Raw inputs: class centres are separated by Δ in the semantic subspace $\mathcal{S}$ but each class spreads to radius r along the nuisance subspace $\mathcal{N}$; with $\Delta/r \approx 2$ the clouds overlap and one-shot nearest-centroid fails. **(b)** An isometric code - a rotation, a normalising flow, any invertible map - moves the clouds but cannot change their relative shape: the anisotropy ratio is pinned at $\rho = 1$ (Proposition 5), so the margin is preserved, never improved. **(c)** An anisotropic code contracts $\mathcal{N}$ while holding $\mathcal{S}$; once $\Delta/r > 4\rho$ the classes are provably separable (Theorem 1).

objective drives exactly the quantity in the numerator towards zero. The gap between inverted codes and anisotropic encoders is therefore not empirical folklore but a theorem.

The second half of the paper asks a harder question. The ratio, as defined, is a property of the *test* geometry: it refers to the very nuisance directions we have not seen. Can it be estimated from what we do have? We prove it can, under one assumption we make explicit - that the unseen nuisance subspace is contained in the subspace the encoder already contracts, which we call N-coverage or (A4). Under (A4) the encoder's directional gains transfer verbatim from seen to unseen nuisances (Theorem 8), which turns the margin law into a *label-free predictor*: three numbers - prototype separation, within-class scatter, and the latent displacement caused by a held-out corruption - each computable from seen-class images with forward passes alone, combine into an estimate of the ratio that should track OOD few-shot accuracy (Figure 2).

We test this on a controlled synthetic geometry in which every quantity of the theory is an exact knob, and on MNIST, CIFAR-10, CIFAR-100 and the full 75-point CIFAR-10-C benchmark, with an encoder zoo running from raw pixels, a VAE, SimCLR and supervised contrastive learning to an ImageNet ResNet-18, CLIP and DINOv2. The result is real but has a shape worth stating up front, because it is easy to overclaim: pooled across encoders the raw correlation is close to zero, and the predictive signal lives *within* an encoder and *across* corruptions. Once encoder identity is controlled for, the estimator reaches a partial correlation of -0.49 over all 525 encoder-corruption blocks and -0.88 over the 375 blocks whose encoders are not pinned at chance, with $p < 10^{-3}$. Under the same protocol it is the strongest predictor on the non-degenerate subset, though not on every slice - a self-supervised rotation head is stronger on the full pooled fixed-effects analysis, and softmax entropy is far stronger on the raw pooled correlation. We report all of this.

Finally, we formalised the theory. Every definition, theorem, proposition, corollary and lemma in Section 3 is stated and proved in Lean 4 against Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020); the development contains no `sorry` and no result depends on any axiom beyond Lean's three standard ones. This was not ceremony. An earlier draft of this work asserted the 4ρ margin law in the general bi-Lipschitz setting; the proof assistant refused it, and correctly so - in that setting the sharp constant is $2(1+\rho)$ (Theorem 4), and 4ρ holds only in the decomposed model that the Jacobian estimator actually instantiates. Both statements now appear, each with its own regime.

Contributions. **(i)** The anisotropy ratio and a margin law for one-shot nearest-centroid classification with the sharp constant 4, a strictly better constant 2 under latent orthogonality, and a constant $2(1+\rho)$ in the assumption-free metric setting (Section 3.2). **(ii)** An isometry floor showing invertible codes cannot beat the raw input geometry, and a mixture result showing region-wise specialists localise the ratio while a single encoder with a bounded contraction budget provably leaks (Sections 3.3 and 3.4). **(iii)** The N-coverage theorem and the label-free estimator it licenses (Section 3.5). **(iv)** A complete Lean 4 formalisation of all seventeen results (Section 3.7, Section C). **(v)** Validation on a controlled geometry and on four image benchmarks with seven encoder families and 525 encoder-corruption blocks, against four label-free baselines under an identical protocol (Section 4).

2 Related work and background

We situate the work here and expand every comparison, with the same citations, in Section I.

Feature geometry. Neural collapse (Papyan et al., 2020) describes the endpoint of the process our ratio measures, and Hermann & Lampinen (2020) the suppression mechanism that produces it. Contrastive theory (Arora et al., 2019; HaoChen et al., 2021; Wang & Isola, 2020) bounds *linear-probe* error on *seen* classes via augmentation-graph or alignment/uniformity quantities; we bound few-shot *nearest-centroid* error on *unseen* classes under *unseen* nuisances, by a single scalar that can be both proved about and measured.

Margins, and predicting generalisation without labels. Margin bounds and Lipschitz-constrained architectures (Bartlett et al., 2017; Neyshabur et al., 2017; Sokolić et al., 2017; Cisse et al., 2017) control the *global* Lipschitz constant; the anisotropy ratio is deliberately directional and two-sided, which is why an isometry - optimal by the global criterion - is the worst case by ours (Proposition 5). A large literature estimates accuracy from unlabelled data (Jiang et al., 2020; Deng & Zheng, 2021; Guillory et al., 2021; Garg et al., 2022; Deng et al., 2021; Sun et al., 2020; Corneanu et al., 2020), but almost all of it reads the *target* distribution; our estimator reads seen classes only and is derived from a theorem rather than fitted. We benchmark against the strongest of these families plus Mahalanobis scoring (Lee et al., 2018) and maximum-softmax/entropy (Hendrycks & Gimpel, 2017; Gidaris et al., 2018).

Shift, few-shot, mixtures. CIFAR-10-C (Hendrycks & Dietterich, 2019) supplies our corruptions, with the documented caveat that synthetic robustness transfers weakly to natural shift (Taori et al., 2020; Recht et al., 2019). Domain-adaptation theory (Ben-David et al., 2010) bounds target error by a distribution divergence; our (A4) plays an analogous role but is a statement about the *encoder*, which is what makes it measurable without target data. Nearest-centroid classification (Snell et al., 2017; Vinyals et al., 2016) is our protocol and Tian et al. (2020) argue the embedding does the work - the claim our ratio makes quantitative - while Proposition 6 gives a geometric reason for the advantage of mixtures of experts (Jacobs et al., 1991; Shazeer et al., 2017).

3 Method

3.1 Setup and the anisotropy ratio

Inputs live in $\mathbb{R}^D$ with the Euclidean norm; an encoder is a map $E : \mathbb{R}^D \to \mathbb{R}^d$. A labelling $c(\cdot)$ assigns each input a class. The downstream task is N-way one-shot nearest-centroid classification: given one support point s_c per class, a query x is assigned $\arg\min_c \|E(x) - E(s_c)\|$.

Definition 1 (Anisotropy ratio)**.** For an encoder E let

$$L_{\mathrm{intra}}(E) = \sup_{\substack{x \neq x' \\ c(x)=c(x')}} \frac{\|E(x) - E(x')\|}{\|x - x'\|}, \qquad \ell_{\mathrm{inter}}(E) = \inf_{\substack{x \neq x' \\ c(x)\neq c(x')}} \frac{\|E(x) - E(x')\|}{\|x - x'\|}, \tag{1}$$

and define the *anisotropy ratio* $\rho(E) = L_{\mathrm{intra}}(E)/\ell_{\mathrm{inter}}(E)$.

Values $\rho < 1$ mean the encoder contracts within-class variation more than it contracts between-class separation; this is the geometry nearest-centroid classification exploits. Definition 1 is metric and assumption-free, which makes it robust but loose. For sharper statements - and because it is what a Jacobian estimator actually measures - we also use a decomposed model.

Decomposed model. Write $\mathbb{R}^D = \mathcal{S} + \mathcal{N}$ for a *semantic* subspace $\mathcal{S}$ and a *nuisance* subspace $\mathcal{N}$ (not assumed orthogonal, not assumed to intersect trivially), and put

$$\eta \;=\; \sup_{0 \neq v \in \mathcal{N}} \frac{\|Ev\|}{\|v\|}, \qquad \beta \;=\; \inf_{0 \neq v \in \mathcal{S}} \frac{\|Ev\|}{\|v\|}, \qquad \rho \;=\; \eta/\beta. \tag{2}$$

We use four assumptions, stated in full in Section A. **(A1)** class centres $\mu_c \in \mathcal{S}$ are pairwise separated by at least Δ; **(A2)** every observation is $x = \mu_{c(x)} + n$ with $n \in \mathcal{N}$ and $\|n\| \leq r$; **(A3)** E is linear on

the scale of interest (for a smooth encoder, E is replaced by its Jacobian J_E at the probe point, which is what the estimator of Section 3.6 evaluates); **(A4)** the nuisance subspace encountered at test time is contained in $\mathcal{N}$. (A4) is used only in Section 3.5.

3.2 THE MARGIN LAW

Theorem 1 (Margin law)**.** *Assume* (A1)-(A3) *with* $\eta, \beta, r \geq 0$*. If*

$$4\,\eta\,r \;<\; \beta\,\Delta \qquad \Big(\textit{equivalently } \Delta/r > 4\rho \textit{ when } \beta, r > 0\Big), \tag{3}$$

then for every query x *of class* c *and every choice of one support point per class,* $\|E(x) - E(s_c)\| < \|E(x) - E(s_{c'})\|$ *for all* $c' \neq c$*; that is, one-shot nearest-centroid classification is correct.*

The proof (Section B) is three lines of triangle inequality, and the constant is entirely transparent: the query and its own support each carry a nuisance of norm up to r, contributing $2\eta r$ to the numerator, and the query and the *competing* support each carry one too, eating another $2\eta r$ out of the semantic gap $\beta\Delta$. Neither $2\eta r$ can be removed, and this is exactly why the constant cannot be improved.

Remark (The bound is stated for one shot but holds for K shots). Theorem 1 places no structure on the support point beyond *(A2)*. A K-shot prototype is $p_c = \mu_c + \bar{n}_c$ with $\bar{n}_c = \frac{1}{K}\sum_i n_{c,i}$, and $\bar{n}_c \in \mathcal{N}$ because $\mathcal{N}$ is a subspace, with $\|\bar{n}_c\| \leq \max_i \|n_{c,i}\| \leq r$. The theorem therefore applies verbatim to K-shot nearest-centroid, and does so conservatively: averaging shrinks $\|\bar{n}_c\|$ well below r in practice, so the $K = 5$ image experiments satisfy a strictly easier condition than the one we prove. The same is true of the Lean statement, whose support nuisances are arbitrary elements of $\mathcal{N}$ of norm at most r.

Proposition 2 (Latent orthogonality halves the constant)**.** *If in addition* $\langle Eu, Ev\rangle = 0$ *for all* $u \in \mathcal{S}$*,* $v \in \mathcal{N}$*, then the conclusion of Theorem 1 holds under the weaker hypothesis* $2\,\eta\,r \;<\; \beta\,\Delta$*, i.e.* $\Delta/r > 2\rho$*.*

Proposition 3 (Sharpness)**.** *For every* $c \in (2, 4)$ *there is an instance satisfying* (A1)-(A3) *and* $c\,\eta\,r \leq \beta\,\Delta$ *on which one-shot nearest-centroid classification is* incorrect*. Hence the constant* 4 *in Theorem 1 cannot be lowered.*

Proposition 2 is a design recommendation as much as a theorem: an encoder that merely shrinks nuisance directions pays the factor 4, whereas one that also *routes* them into a latent subspace orthogonal to the semantic one pays only 2, doubling the range of corruption severities it tolerates at fixed ρ.

Dropping the decomposition entirely costs a slightly weaker constant.

Theorem 4 (Metric margin law)**.** *Let* E *satisfy* $\|E(x) - E(y)\| \leq L_{\text{intra}}\|x - y\|$ *for same-class pairs and* $\ell_{\text{inter}}\|x - y\| \leq \|E(x) - E(y)\|$ *for different-class pairs, with* $L_{\text{intra}}, \ell_{\text{inter}} \geq 0$*; let every point lie within* r *of its class centre and let centres be* Δ*-separated. If* $2L_{\text{intra}}r < \ell_{\text{inter}}(\Delta - 2r)$ *- equivalently* $\Delta/r > 2\,(1 + \rho)$ *with* $\rho = L_{\text{intra}}/\ell_{\text{inter}}$ *- then one-shot nearest-centroid classification is correct.*

3.3 THE ISOMETRY FLOOR

Proposition 5 (Isometry floor)**.** *Let* E *satisfy* $\|x - y\| \leq \|E(x) - E(y)\|$ *for all* x, y *(in particular any isometry, orthogonal map, or invertible flow with non-contracting inverse). Then* $L_{\text{intra}}(E) \geq 1$*. If* E *is an isometry then additionally* $\ell_{\text{inter}}(E) \leq 1$*, so* $\rho(E) \geq 1$*.*

Combined with Theorem 1, Proposition 5 says an invertible code inherits the raw input margin and can never improve it: its best case is $\Delta/r > 4$, the condition the pixels themselves already had to satisfy (Figure 1b). A learned non-invertible encoder is under no such constraint, and a supervised-contrastive objective directly minimises the numerator of ρ.

3.4 MIXTURES LOCALISE THE RATIO

Proposition 6 (Mixtures localise ρ)**.** *Partition the input space into regions* $1, \ldots, m$ *and let* ρ_i *be the anisotropy ratio of the specialist assigned to region* i*, measured on that region. If every specialist is*

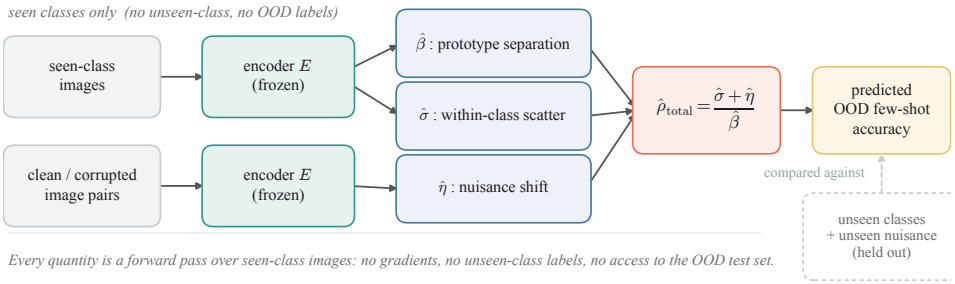

Figure 2: The label-free estimator. All three ingredients are forward passes over *seen*-class images; the unseen classes and the unseen nuisance (dashed) are used only to score the prediction, never to form it.

no worse than the global encoder on its own region - $L^{(i)}_{\text{intra}} \le L_{\text{intra}}$ and $\ell_{\text{inter}} \le \ell^{(i)}_{\text{inter}}$ - then the mixture's effective ratio satisfies $\rho_{\text{eff}} = \max_i \rho_i \le \rho_{\text{global}}$.

Proposition 6 is an inequality, not a benefit: restricting to a region can never hurt, but on a homogeneous partition it also does not help, and that null is what we observe on real images (Section 4, I3). The benefit needs a reason for a single encoder to be unable to serve all regions at once, and the following gives one.

Proposition 7 (A contraction budget forces leakage)**.** *Let E be linear and suppose everything it contracts below gain γ lies in a subspace U, i.e. $\|Ev\| < \gamma\|v\| \Rightarrow v \in U$. If a nuisance subspace $\mathcal{N}$ satisfies $\dim U < \dim \mathcal{N}$, then some $v \in \mathcal{N}$ has $\|Ev\| \ge \gamma\|v\|$: the encoder necessarily leaks along a nuisance direction.*

A single encoder facing the *union* of two region-specific nuisance subspaces therefore leaks whenever its budget covers only one of them, while per-region specialists each spend their budget on their own subspace. This is the mechanism experiment T4 realises exactly.

3.5 N-COVERAGE: MAKING THE RATIO LABEL-FREE

Theorem 8 (N-coverage)**.** *Assume* (A1)-(A3) *and* (A4)*: the test-time nuisance subspace $\mathcal{N}_{\text{test}}$ satisfies $\mathcal{N}_{\text{test}} \subseteq \mathcal{N}$. Then the gain bound transfers, $\|Ev\| \le \eta\|v\|$ for all $v \in \mathcal{N}_{\text{test}}$, and one-shot nearest-centroid classification on* unseen *classes corrupted by that* unseen *nuisance is correct whenever $4\rho_{\text{OOD}} < \Delta/r$, where $\rho_{\text{OOD}} = \eta/\beta$ is measured on seen data alone. The certified corruption radius $\Delta/(4\rho_{\text{OOD}})$ is antitone in ρ_{OOD}.*

This is the content of the paper's second half: (A4) converts a statement about the test distribution into a measurement on the encoder. It is also falsifiable, and Section 4 (T3) falsifies it in the controlled setting by dialling the overlap between the unseen nuisance and the contracted subspace from 0 to 1 and watching ρ_{OOD} and OOD accuracy move in exact anti-phase.

3.6 LABEL-FREE ESTIMATORS

On natural images the pixel metric is not the geometry the theory assumes, so we replace it by the encoder's own ℓ_2-normalised latent geometry. Writing $\hat{E}$ for the normalised encoder, p_c for the prototype of seen class c, and T for a held-out nuisance transform, we estimate

$$\hat{\beta} = \operatorname*{mean}_{c \ne c'} \|p_c - p_{c'}\|, \quad \hat{\sigma} = \operatorname*{mean}_{x} \|\hat{E}(x) - p_{c(x)}\|, \quad \hat{\eta} = \operatorname*{mean}_{x} \|\hat{E}(Tx) - \hat{E}(x)\|, \tag{4}$$

and form two predictors,

$$\hat{\rho}_{\text{OOD}} = \hat{\eta}/\hat{\beta}, \qquad \hat{\rho}_{\text{total}} = (\hat{\sigma} + \hat{\eta})/\hat{\beta}. \tag{5}$$

$\hat{\rho}_{\text{OOD}}$ is the direct finite-sample version of ρ_{OOD} in Theorem 8 and is the *within-encoder* predictor: which corruptions break a given encoder. It omits the baseline scatter $\hat{\sigma}$ and therefore mis-ranks encoders whose class *means* separate while their per-class clouds overlap - an unsupervised VAE is the canonical case - so for *cross-encoder* comparison we fold that term back in as $\hat{\rho}_{\text{total}}$. The two coincide after removing encoder fixed effects, because $\hat{\sigma}/\hat{\beta}$ is constant within an encoder; this is why the fixed-effects columns of Table 1 are identical for both.

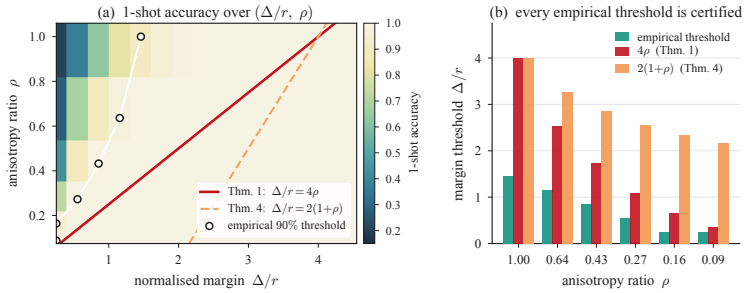

Figure 3: T1, the margin law. (a) One-shot accuracy over the $(\Delta/r, \rho)$ plane; the red line is the sufficient condition of Theorem 1 and the orange dashed line that of Theorem 4. The empirical 90%-accuracy contour (white) lies strictly inside both, as a sufficient condition requires. **(b)** The same comparison numerically: at every ρ the measured threshold falls below 4ρ, and 4ρ is the tighter of the two envelopes throughout the regime $\rho < 1$ that a useful encoder occupies.

Baselines. We compare against four label-free predictors computed under an identical protocol on identical rows: softmax *entropy* of a seen-class linear probe (Hendrycks & Gimpel, 2017), *agreement* between two independently seeded probes, *Mahalanobis* distance to seen-class Gaussians (Lee et al., 2018), and the accuracy of a self-supervised *rotation* head (Gidaris et al., 2018; Deng et al., 2021). Definitions are in Section F.

3.7 Machine-checked proofs

All seventeen results - Theorems 1, 4 and 8 and Propositions 2, 3 and 5 to 7 together with the supporting lemmas - are formalised in Lean 4 against Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020). The mapping from paper statement to Lean identifier to the experiment that tests it is given in Section C.

4 Experiments

Protocol. Every encoder is trained or fitted on *seen* classes only. All evaluation is N-way K-shot nearest-centroid on *unseen* classes, either clean or corrupted by a nuisance never used in training, with $K = 5$ for the image experiments and $K = 1$ for the controlled geometry of Section 4.1. Section 3.2 explains why Theorem 1 covers both. Estimators see seen-class images only. Splits are 0-4 / 5-9 for MNIST and CIFAR-10 (5-way) and 0-49 / 50-99 for CIFAR-100 (10-way); accuracies are means over 400 episodes. Encoder zoo: raw pixels, a convolutional VAE (Kingma & Welling, 2014), SimCLR (Chen et al., 2020), supervised contrastive (Khosla et al., 2020), an ImageNet ResNet-18 (He et al., 2016; Deng et al., 2009), CLIP ViT-B/32 (Radford et al., 2021) and DINOv2-S (Oquab et al., 2024). Full details - architectures, optimisers, seeds, hardware, corruption parameters - are in Section E.

4.1 Controlled geometry: the mechanism in isolation (T1-T4)

We build $\mathbb{R}^{48} = \mathcal{S} \oplus \mathcal{N}$ with $\dim \mathcal{S} = 6$ and inject within-class variation into a chosen block of $\mathcal{N}$, so that Δ, r, ρ and the seen/unseen nuisance overlap are exact knobs rather than estimates (Section D).

T1 (Figure 3) sweeps a diagonal encoder that contracts $\mathcal{N}$ by a factor a, giving $\rho \approx a$, over a grid of margins. The measured 90%-accuracy thresholds are $1.46, 1.16, 0.86, 0.55, 0.25, 0.25$ against $4\rho = 4.00, 2.54, 1.73, 1.09, 0.66, 0.35$: the law is sufficient at every ρ, and progressively tight as ρ falls. **T2** (Figure 4a) exhibits the isometry floor: two bijections pinned at $L_{\text{intra}} = 1$ versus a contrastive encoder at $L_{\text{intra}} = 0.044$. **T3** (Figure 4b) is the decisive test of (A4). **T4** (Figure 4c) realises Proposition 7 exactly: the single encoder's per-region ratios are 0.25 and 1.00 - it contracts one region's nuisance and leaks on the other's, precisely as the dimension count predicts.

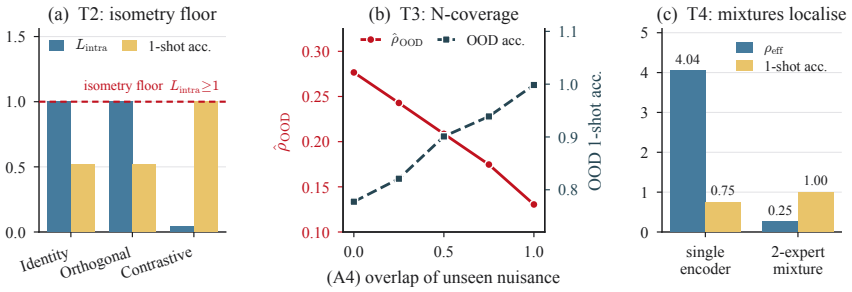

Figure 4: T2-T4. (a) Identity and a random orthogonal map both sit exactly on the isometry floor $L_{\text{intra}} = 1$ and reach 0.52 accuracy; a contrastive MLP reaches $L_{\text{intra}} = 0.044$ and 1.00. **(b)** As the unseen nuisance subspace is dialled into the contracted one, $\hat{\rho}_{\text{OOD}}$ falls from 0.277 to 0.130 while OOD accuracy rises from 0.777 to 0.998 - exact anti-phase, with the estimator never seeing an unseen label. **(c)** On two regions with orthogonal 8-dimensional nuisance blocks, a single encoder with a contraction budget of 8 of the 16 union dimensions has $\rho_{\text{global}} = 4.04$ and 0.75 accuracy; the nuisance-aligned two-expert mixture has $\rho_{\text{eff}} = 0.25$ and 1.00.

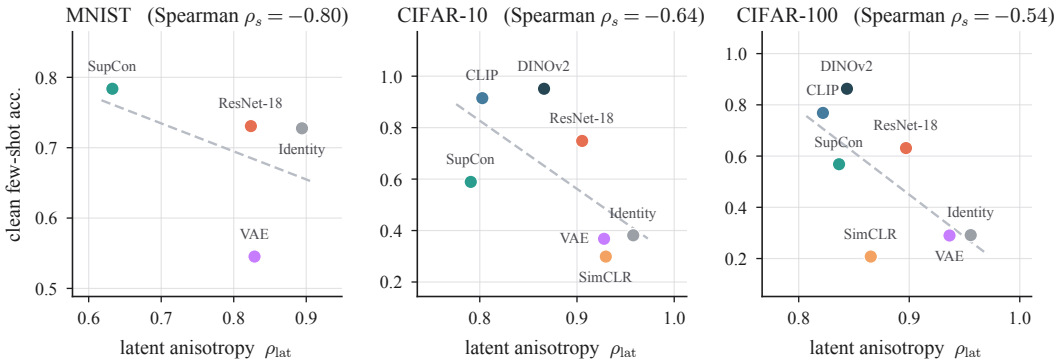

Figure 5: Experiment I1. Latent anisotropy versus clean few-shot accuracy on unseen classes. The ordering is right on all three datasets (Spearman -0.80, -0.64, -0.54) but the sample is seven encoders, so no single-dataset correlation is individually significant; the claim these panels support is directional, and the quantitative test is I2.

4.2 Images: does ρ order encoders? (I1)

Figure 5 plots latent anisotropy against clean unseen-class few-shot accuracy. The relation is negative on all three datasets, as Theorem 1 requires. Two honest caveats. First, the *pixel-referenced* ratio of Definition 1 does *not* order encoders on natural images (Spearman $+0.80$, -0.32, -0.11 on MNIST, CIFAR-10, CIFAR-100): the Euclidean pixel metric is simply not the input geometry the theory assumes, which is why Section 3.6 works in latent coordinates. Second, with seven points per panel these correlations are suggestive, not conclusive. The decisive experiment is the next one, which has 525 rows.

4.3 Images: does the label-free estimator predict OOD accuracy? (I2)

Figure 6 and Table 1 are the paper's central empirical result, and they have a specific shape. Pooled across encoders the raw correlation is essentially zero on CIFAR-10-C. This is not a failure of the theory but a consequence of it: ρ_{OOD} is defined relative to a fixed encoder, and each family has its own scale of $\hat{\eta}$ and its own accuracy band, so pooling mixes seven different regressions (Figure 6a). Removing encoder identity - the standard fixed-effects correction, and the analysis the theory actually licenses - recovers a partial correlation of -0.49 over all 525 blocks and -0.88 over the 375 blocks belonging to encoders that are not pinned at chance, with an OLS slope of -0.059 ± 0.005 and $p < 10^{-3}$ (Figure 6b). Within individual encoders the relation is very strong: -0.97 for supervised contrastive, -0.93 for the ImageNet trunk, -0.89 for CLIP, -0.88 for DINOv2 (Figure 6c). The two encoders where it fails, raw pixels (-0.31) and SimCLR (-0.02), are exactly those whose OOD accuracy never moves off chance across all 75 corruption blocks, leaving almost no variance to

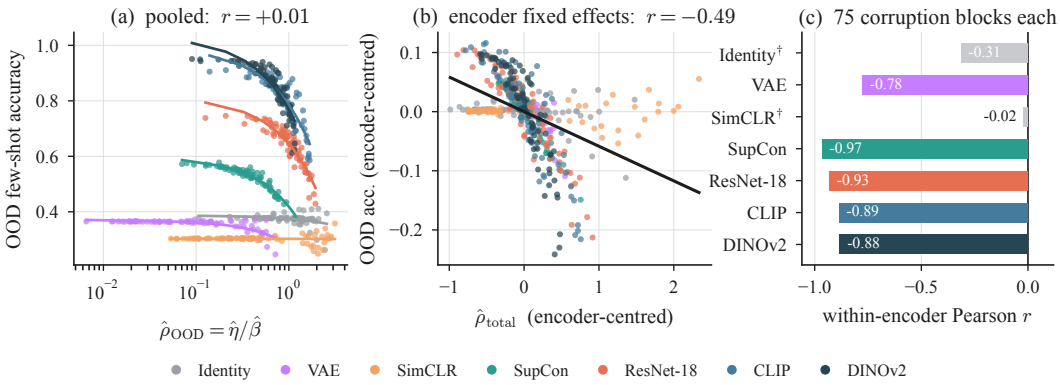

Figure 6: CIFAR-10-C, 525 encoder × corruption blocks. (a) Pooled, the estimator is uncorrelated with accuracy ($r = +0.01$): each encoder family occupies its own band, and pooling destroys the signal - a textbook Simpson's paradox. **(b)** Centring both variables within each encoder recovers a clean negative relation ($r = -0.49$, $p < 10^{-3}$). **(c)** Within-encoder correlations for each of the seven families; † marks the two encoders whose OOD accuracy is flat at chance across all corruptions, so that there is almost no accuracy variance for any predictor to explain.

Table 1: The label-free estimator against OOD few-shot accuracy. "raw" is the pooled Pearson correlation of each predictor; "z-scored" standardises $\hat{\rho}$ within each encoder before pooling; "partial" is the correlation after removing encoder fixed effects by OLS; "non-deg." repeats that after dropping encoders whose OOD accuracy is flat at chance. The two predictors of Equation (5) differ by a per-encoder constant, so they share the fixed-effects columns exactly.

		raw pooled r					OLS acc $= \alpha_{\text{enc}} + b\,\hat{\rho}$	
Dataset	n	$\hat{\rho}_{\text{OOD}}$	$\hat{\rho}_{\text{total}}$	z-scored	partial	non-deg.	b (SE)	p
MNIST	20	$+0.10$	$+0.16$	-0.22	-0.37	-0.55	-0.054 (0.036)	0.16
CIFAR-10	35	-0.19	-0.37	-0.22	-0.27	$\mathbf{-0.92}$	-0.031 (0.022)	0.16
CIFAR-100	35	-0.17	-0.40	-0.18	-0.34	$\mathbf{-0.92}$	-0.038 (0.021)	0.08
CIFAR-10-C	525	$+0.01$	-0.33	-0.18	$\mathbf{-0.49}$	$\mathbf{-0.88}$	$\mathbf{-0.059}$ $\mathbf{(0.005)}$	$< \mathbf{10^{-3}}$

predict. The same pattern, with the same sign, appears on the five held-out transforms of MNIST, CIFAR-10 and CIFAR-100 (Table 1; per-dataset scatter plots in Section H).

I3. Class-stratified partitions of real images give $\rho_{\text{eff}} \approx \rho_{\text{global}}$ on all three datasets (gaps -0.023, -0.008, -0.013; Section H). This is the null Proposition 6 predicts for a homogeneous partition, not evidence for mixtures; the localisation *benefit* requires region-dependent nuisance and is demonstrated in T4.

4.4 Comparison with label-free baselines

Table 2 and Figure 7 put the estimator against four standard label-free predictors. The honest summary is that no method sweeps. Softmax entropy has by far the best raw pooled correlation (-0.845), but most of that is between-encoder variance - it collapses to -0.156 once $\hat{\rho}$-style within-encoder standardisation is applied, and to -0.420 under fixed effects. The rotation head is the strongest predictor under full pooled fixed effects (-0.621 against our -0.489); it also requires training an extra self-supervised head per encoder, whereas Equation (4) needs only forward passes. On the non-degenerate subset - the 375 blocks where accuracy actually varies, and therefore the only ones where a difficulty predictor is meaningful - $\hat{\rho}_{\text{total}}$ is the strongest at -0.879 against rotation's -0.795 and entropy's -0.576. Mahalanobis distance fails throughout, and probe agreement is degenerate. We regard the theoretical derivation, not the margin, as the contribution: $\hat{\rho}$ is the only entry in the table that comes with a proof of *why* it should work and a stated assumption, (A4), under which it must.

Table 2: Label-free OOD-difficulty predictors on CIFAR-10-C, all computed on the same 525 rows under the same protocol. No predictor wins every column. Entropy dominates the raw pooled correlation; the rotation head dominates the full fixed-effects analysis; $\hat{\rho}_{\text{total}}$ dominates once the two chance-level encoders are removed, and is the only entry derived from a theorem rather than fitted. Agreement is constant (the two probes agree on essentially every input) and so has no variance to correlate.

Predictor	raw pooled r	z-scored r	partial (encoder FE)	non-deg. partial
$\hat{\rho}_{\text{total}}$ (ours)	-0.333	$\mathbf{-0.184}$	-0.489	$\mathbf{-0.879}$
entropy	$\mathbf{-0.845}$	-0.156	-0.420	-0.576
rotation	-0.673	-0.181	$\mathbf{-0.621}$	-0.795
Mahalanobis	-0.094	$+0.004$	$+0.000$	-0.137
agreement	n/a	n/a	0.000	0.000

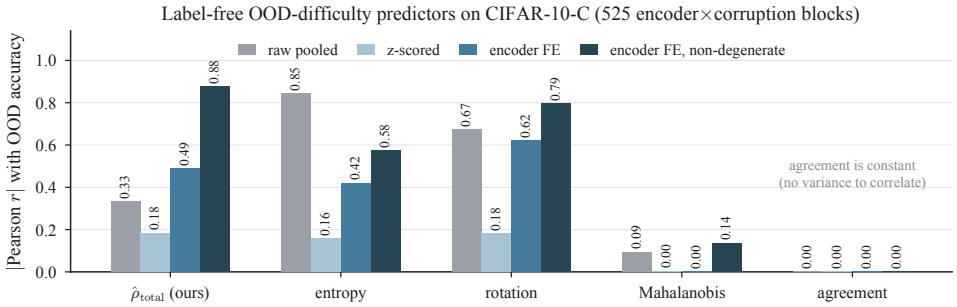

Figure 7: The same comparison as Table 2, by absolute correlation. Which predictor looks best depends entirely on which column is read, which is the reason we report four.

5 Limitations

Six caveats; each is expanded, in full, in Section J. **(i) Scale and shift realism.** Everything runs on MNIST, CIFAR-10, CIFAR-100 and CIFAR-10-C with synthetic corruptions, and synthetic robustness is known to transfer weakly to natural shift (Taori et al., 2020; Recht et al., 2019); ImageNet-C, DomainNet (Peng et al., 2019) and WILDS (Koh et al., 2021) are untested and we make no claim about them. **(ii) The pooled correlation is weak.** The raw pooled correlation on CIFAR-10-C is $+0.01$; only the fixed-effects and non-degenerate analyses are strong. A practitioner who wants to compare two *different* encoders with one unadjusted number will not get it from $\hat{\rho}_{\text{OOD}}$, and $\hat{\rho}_{\text{total}}$ only partly repairs this. **(iii) (A4) is an assumption, not a measurement**: T3 shows graceful degradation as it is violated, but nothing certifies it for a corruption we have not seen. **(iv) Oracle routing in T4**: Propositions 6 and 7 say what a mixture can achieve, not that a gate can be learned to achieve it. **(v) Linearisation**: (A3) replaces the encoder by its Jacobian, and $\hat{\eta}$ measures a secant at the corruption's own magnitude, not a derivative. **(vi) Two degenerate encoders**: raw pixels and SimCLR sit at chance on unseen CIFAR classes, which is why both the full and the non-degenerate splits appear in Tables 1 and 2 rather than only the flattering one.

6 Conclusion

Few-shot out-of-distribution success is not mysterious; it is a ratio. An encoder that contracts within-class directions harder than it contracts between-class ones buys margin, and the exchange rate is exactly four to one - provably no better in general, provably two to one if the encoder also routes nuisance into an orthogonal latent subspace, and provably unavailable to any invertible code. Because the ratio is a property of the encoder rather than of the test distribution, it can be estimated from seen data alone, and that estimate tracks OOD few-shot accuracy across seven encoder families and every corruption of CIFAR-10-C once the analysis respects the per-encoder nature of the quantity. The theory is machine-checked, which is how we discovered that one constant we had believed was wrong. The obvious next steps are ImageNet-scale shift, learned rather than oracle routing, and an encoder trained to minimise $\hat{\rho}$ directly.

IMPACT STATEMENT

This work is foundational: it studies when a representation supports few-shot recognition under distribution shift, and supplies a diagnostic for that capability. The most plausible positive impact is safety-adjacent - the estimator lets a practitioner flag, before deployment and without labelled target data, that an encoder is likely to degrade under a particular class of corruption, which is currently often discovered only after failure. The corresponding risk is misplaced confidence: $\hat{\rho}$ is a *prediction under a stated assumption* (A4), it is validated on synthetic corruptions of small images, and a low value must not be read as a safety certificate for a deployed system. We have therefore been explicit in Section 5 about where the signal is weak and where the assumption is untested. The experiments use only MNIST, CIFAR-10, CIFAR-100 and CIFAR-10-C, which contain no personal data; the pretrained encoders (ResNet-18, CLIP, DINOv2) carry whatever representational biases their training corpora encode, and the ratio we measure says nothing about those biases - an encoder can have an excellent anisotropy ratio and still be unfair. We see no additional dual-use concern beyond those already present in representation learning generally.

AI USE STATEMENT

In this work, generative AI tools were used to (i) create or modify visual figures based on our experimental results, (ii) edit English grammar, phrasing, and vocabulary, and (iii) assist in producing Lean formalizations for checking selected mathematical equations. The tools were not used to generate data, formulate the research questions or theoretical claims, design or run experiments, analyze or interpret experimental results, or make scientific decisions. They were also not used to replace the authors' judgment about the scientific content.

We reviewed all AI-assisted work. The experimental data and results were checked against the original runs and stored result files; AI-assisted figures were checked against those results; and the Lean formalizations were compiled and checked by the authors to ensure that they corresponded to the equations stated in the paper. AI-assisted language edits were manually reviewed for meaning and accuracy. The authors take responsibility for the final content of this work, including all text, claims, figures, and formalizations produced with AI assistance.

REFERENCES

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning (ICML)*, 2019.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning (ICML)*, 2019.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

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

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2017.

Ciprian A. Corneanu, Sergio Escalera, and Aleix M. Martinez. Computing the testing error without a testing set. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2677–2685, 2020.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *International Conference on Automated Deduction (CADE)*, pp. 625–635, 2021.

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

Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15069–15078, 2021.

Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? In *International Conference on Machine Learning (ICML)*, pp. 2579–2589, 2021.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations (ICLR)*, 2017.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations (ICLR)*, 2022.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1134–1144, 2021.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

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

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations (ICLR)*, 2020.

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)*, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, 2019.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

The mathlib Community. The Lean mathematical library. In *ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP)*, pp. 367–381, 2020.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, 2020.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

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

APPENDIX

The appendix contains, in order: the assumptions in full (Section A); complete proofs of every result in Section 3 (Section B); the Lean 4 formalisation and its axiom audit (Section C); the controlled-geometry construction (Section D); image-experiment details (Section E); baseline definitions (Section F); the full statistical tables (Section G); additional figures (Section H); the extended related work and limitations sections condensed in the main text (Sections I and J).

A ASSUMPTIONS IN FULL

Let $\mathcal{S}, \mathcal{N} \subseteq \mathbb{R}^D$ be linear subspaces (the *semantic* and *nuisance* subspaces). We do *not* assume $\mathcal{S} \perp \mathcal{N}$, nor $\mathcal{S} \cap \mathcal{N} = \{0\}$, nor $\mathcal{S} + \mathcal{N} = \mathbb{R}^D$; Proposition 2 adds an orthogonality hypothesis explicitly, and Proposition 3 exploits the fact that the general statement permits $\mathcal{S} = \mathcal{N}$.

Assumption 1 (Semantic separation)**.** The class centres $\{\mu_c\}_{c \in \mathcal{C}}$ satisfy $\mu_c \in \mathcal{S}$ for every c, and $\|\mu_c - \mu_{c'}\| \geq \Delta$ for every $c \neq c'$.

Assumption 2 (Bounded nuisance)**.** Every observation of class c has the form $x = \mu_c + n$ with $n \in \mathcal{N}$ and $\|n\| \leq r$. In particular both the query and every support point are of this form, with independent nuisance components.

Assumption 3 (Linearity / local linearisation)**.** E is linear. For a differentiable non-linear encoder, E is replaced throughout by its Jacobian $J_E(x_0)$ at a probe point x_0, and all statements hold on the neighbourhood on which the first-order approximation is valid. The quantities η, β of Equation (2) are then the maximal gain of J_E on $\mathcal{N}$ and its minimal gain on $\mathcal{S}$.

Assumption 4 (N-coverage)**.** The nuisance subspace realised at test time satisfies $\mathcal{N}_{\text{test}} \subseteq \mathcal{N}$, where $\mathcal{N}$ is the subspace on which the encoder's gain bound η has been established.

Assumption 4 is the only assumption that refers to the test distribution, and it is the one that makes the estimator label-free: it says the corruption we will meet lives inside the space the encoder has already learned to suppress. It is falsifiable, and experiment T3 traces exactly what happens as it is violated by degrees.

B PROOFS

Throughout, $\|\cdot\|$ is the Euclidean norm on the relevant space and we use the reverse triangle inequality in the form $\|X\| - \|Y\| \leq \|X + Y\|$.

B.1 PROOF OF THEOREM 1

Let $x = \mu_c + n_q$ be the query and $s_{c'} = \mu_{c'} + n_{c'}$ the support of class c', with $n_q, n_{c'} \in \mathcal{N}$ and $\|n_q\|, \|n_{c'}\| \leq r$ by Assumption 2.

Correct side. By linearity (Assumption 3), $E(x) - E(s_c) = E(n_q - n_c)$, and $n_q - n_c \in \mathcal{N}$ because $\mathcal{N}$ is a subspace. Hence, using $\eta \geq 0$,

$$\|E(x) - E(s_c)\| \;\leq\; \eta\|n_q - n_c\| \;\leq\; \eta\big(\|n_q\| + \|n_c\|\big) \;\leq\; 2\eta r. \tag{6}$$

Competing side. Again by linearity, $E(x) - E(s_{c'}) = E(\mu_c - \mu_{c'}) + E(n_q - n_{c'})$, where $\mu_c - \mu_{c'} \in \mathcal{S}$ and $n_q - n_{c'} \in \mathcal{N}$. By Assumption 1 and $\beta \geq 0$, $\|E(\mu_c - \mu_{c'})\| \geq \beta\|\mu_c - \mu_{c'}\| \geq \beta\Delta$, while $\|E(n_q - n_{c'})\| \leq 2\eta r$ as in Equation (6). The reverse triangle inequality gives

$$\|E(x) - E(s_{c'})\| \;\geq\; \beta\Delta - 2\eta r. \tag{7}$$

The hypothesis $4\eta r < \beta\Delta$ is exactly $2\eta r < \beta\Delta - 2\eta r$, so Equations (6) and (7) give $\|E(x) - E(s_c)\| < \|E(x) - E(s_{c'})\|$ for every $c' \neq c$. Since the correct prototype is strictly closest, the $\arg\min$ is unique and equals c. When $\beta > 0$ and $r > 0$, dividing by βr turns the hypothesis into $\Delta/r > 4\eta/\beta = 4\rho$. □

B.2 Proof of Proposition 3

Take $D = d = 1$, $\mathcal{S} = \mathcal{N} = \mathbb{R}$, $E = \mathrm{id}$, so $\eta = \beta = 1$ and $\rho = 1$. Fix $c \in (2, 4)$, set $r = 1$ and $\Delta = c$, and put $\mu_0 = 0$, $\mu_1 = c$, $n_q = +1$, $n_0 = n_1 = -1$. Then Assumptions 1 to 3 hold, $\|\mu_0 - \mu_1\| = c = \Delta$, all nuisance norms equal r, and the hypothesis $c\,\eta\,r \le \beta\Delta$ holds with equality. The query is $x = \mu_0 + n_q = 1$ and the supports are $s_0 = \mu_0 + n_0 = -1$ and $s_1 = \mu_1 + n_1 = c - 1$, so

$$|E(x) - E(s_0)| = 2, \qquad |E(x) - E(s_1)| = |1 - (c-1)| = |2 - c| = c - 2 < 2, \tag{8}$$

using $2 < c < 4$. The wrong support is strictly closer, so the classification is incorrect. Since c was an arbitrary element of $(2, 4)$, no constant below 4 suffices. (The restriction $c > 2$ is not a gap: for $c \le 2$ the hypothesis allows $\Delta \le 2r$, so the class clouds intersect and failure is immediate.) □

Remark. The construction uses $\mathcal{S} = \mathcal{N}$, which is legitimate because Theorem 1 does not assume the subspaces are independent. This is informative rather than pathological: it locates precisely what extra structure buys a better constant, and Proposition 2 supplies it.

B.3 Proof of Proposition 2

The correct side is unchanged: $\|E(x) - E(s_c)\| \le 2\eta r$. On the competing side write $X = E(\mu_c - \mu_{c'})$ and $Y = E(n_q - n_{c'})$. By hypothesis $\langle X, Y \rangle = 0$, so by the Pythagorean identity

$$\|X + Y\|^2 = \|X\|^2 + 2\langle X, Y\rangle + \|Y\|^2 = \|X\|^2 + \|Y\|^2 \;\ge\; \|X\|^2, \tag{9}$$

and taking square roots (both norms being non-negative) $\|E(x) - E(s_{c'})\| = \|X + Y\| \ge \|X\| \ge \beta\Delta$. The nuisance no longer erodes the semantic gap at all, so $2\eta r < \beta\Delta$ suffices. □

B.4 Proof of Theorem 4

Since $L_{\mathrm{intra}}, r \ge 0$ we have $2L_{\mathrm{intra}} r \ge 0$, so the hypothesis $2L_{\mathrm{intra}} r < \ell_{\mathrm{inter}}(\Delta - 2r)$ forces $\ell_{\mathrm{inter}}(\Delta - 2r) > 0$ and hence, with $\ell_{\mathrm{inter}} \ge 0$, both $\ell_{\mathrm{inter}} > 0$ and $\Delta > 2r$.

Correct side. With $m = \mu_{c(x)}$, $\|x - s\| = \|(x - m) - (s - m)\| \le \|x - m\| + \|s - m\| \le 2r$, so $\|E(x) - E(s)\| \le L_{\mathrm{intra}}\|x - s\| \le 2L_{\mathrm{intra}} r$.

Competing side. Write $m' = \mu_{c(s')}$. Two applications of the triangle inequality give $\|m - m'\| \le \|m - x\| + \|x - s'\| + \|s' - m'\| \le r + \|x - s'\| + r$, and $\|m - m'\| \ge \Delta$, so $\|x - s'\| \ge \Delta - 2r$. Therefore $\|E(x) - E(s')\| \ge \ell_{\mathrm{inter}}\|x - s'\| \ge \ell_{\mathrm{inter}}(\Delta - 2r) > 2L_{\mathrm{intra}} r \ge \|E(x) - E(s)\|$.

For the ratio form, $2L_{\mathrm{intra}} r < \ell_{\mathrm{inter}}\Delta - 2\ell_{\mathrm{inter}} r$ is equivalent to $2r(L_{\mathrm{intra}} + \ell_{\mathrm{inter}}) < \ell_{\mathrm{inter}}\Delta$, i.e. to $\Delta/r > 2(L_{\mathrm{intra}} + \ell_{\mathrm{inter}})/\ell_{\mathrm{inter}} = 2(1 + \rho)$. □

B.5 Proof of Proposition 5

Let $x \ne y$ with $c(x) = c(y)$. Then $\|x - y\| > 0$ and $\|x - y\| \le \|E(x) - E(y)\| \le L_{\mathrm{intra}}\|x - y\|$, so dividing by $\|x - y\| > 0$ gives $L_{\mathrm{intra}} \ge 1$. If E is an isometry, take any $u \ne v$ with $c(u) \ne c(v)$; then $\ell_{\mathrm{inter}}\|u - v\| \le \|E(u) - E(v)\| = \|u - v\|$ gives $\ell_{\mathrm{inter}} \le 1$, and $\rho = L_{\mathrm{intra}}/\ell_{\mathrm{inter}} \ge 1/1 = 1$. □

B.6 Proof of Proposition 6

Fix a region i. Restricting the supremum defining L_{intra} to pairs inside region i can only decrease it, and restricting the infimum defining ℓ_{inter} can only increase it; formally, for $A \subseteq B \subseteq \mathbb{R}$ with A non-empty and B bounded, $\sup A \le \sup B$ and $\inf B \le \inf A$. Hence $L^{(i)}_{\mathrm{intra}} \le L_{\mathrm{intra}}$ and $\ell_{\mathrm{inter}} \le \ell^{(i)}_{\mathrm{inter}}$ for the global encoder, and the same holds by hypothesis for the specialist. With $L^{(i)}_{\mathrm{intra}} \ge 0$ and $\ell_{\mathrm{inter}} > 0$,

$$\rho_i = \frac{L^{(i)}_{\mathrm{intra}}}{\ell^{(i)}_{\mathrm{inter}}} \;\le\; \frac{L^{(i)}_{\mathrm{intra}}}{\ell_{\mathrm{inter}}} \;\le\; \frac{L_{\mathrm{intra}}}{\ell_{\mathrm{inter}}} = \rho_{\mathrm{global}}, \tag{10}$$

the first step because $\ell_{\mathrm{inter}} \le \ell^{(i)}_{\mathrm{inter}}$ and the numerator is non-negative, the second because $L^{(i)}_{\mathrm{intra}} \le L_{\mathrm{intra}}$. Taking the maximum over the finitely many regions preserves the bound, giving $\rho_{\mathrm{eff}} = \max_i \rho_i \le \rho_{\mathrm{global}}$. □

B.7 Proof of Proposition 7

Suppose for contradiction that $\|Ev\| < \gamma\|v\|$ for every $v \in \mathcal{N}$. Then by hypothesis every $v \in \mathcal{N}$ lies in U, i.e. $\mathcal{N} \subseteq U$, and monotonicity of dimension gives $\dim \mathcal{N} \leq \dim U$, contradicting $\dim U < \dim \mathcal{N}$. □

B.8 Proof of Theorem 8

By Assumption 4, $\mathcal{N}_{\text{test}} \subseteq \mathcal{N}$, so for every $v \in \mathcal{N}_{\text{test}}$ we have $v \in \mathcal{N}$ and therefore $\|Ev\| \leq \eta\|v\|$: the gain bound transfers verbatim. Consequently Assumption 2 holds with $\mathcal{N}$ replaced by $\mathcal{N}_{\text{test}}$ and the same η, and Theorem 1 applies with ρ replaced by $\rho_{\text{OOD}} = \eta/\beta$, giving correctness whenever $4\rho_{\text{OOD}} < \Delta/r$. For the antitonicity claim, if $0 < \rho_1 \leq \rho_2$ and $\Delta > 0$ then $4\rho_1 \leq 4\rho_2$ and both are positive, so $\Delta/(4\rho_2) \leq \Delta/(4\rho_1)$. □

C The Lean 4 formalisation

paper result	*Lean 4 identifier*	*empirical test*
Thm. 1 margin law $4\eta r < \beta\Delta$	`margin_law_linear` `margin_law_ratio`	T1 heat map (Fig. 3a)
Prop. 2 orthogonal refinement $\Delta/r > 2\rho$	`margin_law_orthogonal`	theory only
Prop. 3 sharpness of the 4 no constant < 4 works	`margin_law_four_is_sharp`	Fig. 15
Thm. 4 metric margin law $\Delta/r > 2(1+\rho)$	`margin_law_lipschitz` `margin_law_lipschitz_ratio`	T1 envelope (Fig. 3b)
Prop. 5 isometry floor $L_{\text{intra}} \geq 1,\ \rho \geq 1$	`isometry_floor` `isometry_rho_ge_one`	T2 (Fig. 4a), I1
Prop. 6 mixtures localise $\max_i \rho_i \leq \rho_{\text{global}}$	`mixture_localizes`, `rho_mono` `sSup/sInf_restrict`	T4 (Fig. 4c), I3
Prop. 7 contraction budget a too-small budget leaks	`budget_forces_leakage`	T4 single encoder
Thm. 8 N-coverage under (A4) $4\rho_{\text{OOD}} < \Delta/r$	`A4_transfer`, `ncoverage` `ncoverage_radius_antitone`	T3, I2, CIFAR-10-C

Every row is machine-checked in Lean 4 / Mathlib. No result depends on any axiom beyond propext, Classical.choice and Quot.sound; the development contains no sorry and no additional hypothesis is admitted without proof.

Figure 8: Every result of Section 3, the Lean identifier that proves it, and the experiment that tests it. Numbering follows the paper; the Lean file proves seventeen statements in total, including the supporting lemmas (`nearest_centroid_correct`, `sSup_restrict_le`, `sInf_restrict_ge`, `rho_mono`, `A4_transfer`, `ncoverage_radius_antitone`) not separately numbered above.

Contents. The development lives in `lean/ARTheory/Basic.lean` (394 lines) and is organised in the same order as Section 3: the decision rule; the margin law in the decomposed model and its ratio form; the orthogonality refinement; the metric margin law and its ratio form; the isometry floor; the restriction, mixture and budget lemmas; the N-coverage theorem and its antitone corollary; and the sharpness construction. All statements are over an arbitrary real normed space (`NormedAddCommGroup V` with `NormedSpace ℝ V`); the orthogonality refinement additionally requires an `InnerProductSpace`, and the budget lemma requires `FiniteDimensional`.

Reproducing the check. With `elan` installed:

```
cd lean && lake build && lake env lean Verify.lean
```

The toolchain is pinned to `leanprover/lean4:v4.33.1` with Mathlib at the matching tag, recorded in `lean/lean-toolchain` and `lean/lake-manifest.json`. `lake build` reports *Build completed successfully*; `Verify.lean` runs `#print axioms` on all seventeen results.

Axiom audit. The output of `Verify.lean` is seventeen lines of the form

```
'ARTheory.margin_law_linear' depends on axioms:  [propext,
Classical.choice, Quot.sound]
```

These are Lean's three standard axioms, which are consistent with ZFC and are used throughout Mathlib. No occurrence of `sorryAx` appears, which is the mechanical certificate that no proof is incomplete.

Statement fidelity. Formalisation is only meaningful if the Lean statement says what the paper says. Two choices deserve comment. First, Theorem 1 is formalised with η and β as *hypothesised bounds* - the Lean hypothesis reads "for all v in $\mathbb{N}$, $\|Ev\| \le \eta\|v\|$" - rather than as suprema and infima; this is strictly more general, since any valid bounds may be substituted, and it is what the finite-sample estimator supplies. Second, Proposition 3 is formalised as an existential over an explicit witness rather than as a negation of a universally quantified statement, which is the constructively stronger form. The full Lean source is included in the supplementary material.

D THE CONTROLLED GEOMETRY (T1-T4)

Construction. Draw $Q \in \mathbb{R}^{48\times 48}$ from a QR factorisation of a Gaussian matrix; set $\mathcal{S} = \mathrm{span}(Q_{:,1:6})$ and $\mathcal{N} = \mathrm{span}(Q_{:,7:48})$. For n_{cls} classes, sample unit vectors $u_c \in \mathbb{R}^6$ and place centres at $\mu_c = \Delta\,(u_c Q_{:,1:6}^{\top})$. Within-class points are $\mu_c + (w \odot \mathrm{rad})B^{\top}$ where B is a chosen k-column block of $\mathcal{N}$, w is uniform on the unit sphere of $\mathbb{R}^k$, and $\mathrm{rad} = r\,U^{1/k}$ with $U \sim \mathrm{Unif}[0,1]$, giving a uniform distribution on the ball of radius r inside the chosen nuisance block. This realises Assumptions 1 and 2 exactly.

Encoders. The diagonal encoder used in T1 is $E_a(x) = x\,(SS^{\top} + a\,NN^{\top})^{\top}$, which preserves $\mathcal{S}$ and scales $\mathcal{N}$ by a, so $\eta = a$, $\beta = 1$, $\rho = a$ analytically. The measured ρ (via $99\%/1\%$ quantiles of the pairwise ratio over 8000 pairs) is $1.000, 0.636, 0.432, 0.273, 0.164, 0.088$ for $a = 1.0, 0.6, 0.4, 0.25, 0.15, 0.08$ - quantile estimation of a supremum biases ρ slightly upward at small a, which makes the test of Theorem 1 conservative rather than favourable. T2 and T3 use a three-layer MLP ($48 \to 128 \to 128 \to 16$, ReLU) trained with the supervised contrastive loss ($\tau = 0.1$) for 400 (T2) or 600 (T3) Adam steps at learning rate 10^{-3}, batch 256. T4 uses two linear encoders that each contract one 8-dimensional nuisance block by $a = 0.05$ while preserving everything else.

Settings. T1: 6 classes, $r = 1$, Δ swept over 20 points in $[0.25, 6]$, 80 points per class, 6-way 1-shot with 10 queries, 120 episodes. T2: 8 classes, $\Delta = 2$, $r = 3$, 120 per class, 8-way 1-shot, 200 episodes. T3: 8 seen classes, $\Delta = 3$, $r = 6$, unseen nuisance blocks overlapping the contracted 12-dimensional block by $0, 25, 50, 75, 100\%$, 6-way 1-shot, 150 episodes; $\hat{\rho}_{\mathrm{OOD}}$ is a finite-difference Jacobian probe ($\varepsilon = 10^{-2}$) along the 6 semantic basis directions and the 12 unseen-nuisance directions. T4: 8 classes, $\Delta = 3$, $r = 6$, 160 per class per region, 8-way 1-shot with one support per region, 300 episodes. Table 7 gives T1 numerically.

E IMAGE-EXPERIMENT DETAILS

Data and splits. MNIST (LeCun et al., 1998) and CIFAR-10/100 (Krizhevsky, 2009) from `torchvision`; CIFAR-10-C (Hendrycks & Dietterich, 2019) streamed from the HuggingFace parquet mirror `robro/cifar10-c-parquet` (the script-based mirrors are incompatible with current `datasets`). Seen/unseen splits are $\{0\ldots4\}/\{5\ldots9\}$ for MNIST and CIFAR-10 (5-way) and $\{0\ldots49\}/\{50\ldots99\}$ for CIFAR-100 (10-way). Up to 1000 training and 300 evaluation images per class. For CIFAR-10-C the code asserts that the corrupted blocks are in canonical CIFAR-10 test order before pairing clean with corrupted images, and aborts otherwise; $\hat{\eta}$ is computed on those aligned pairs.

Encoders. *Identity* flattens pixels. *VAE*: convolutional encoder ($3 \to 64 \to 128 \to 128$, stride 2), 64-dimensional posterior mean. *SupCon* and *SimCLR*: the same four-layer CNN trunk ($64, 64, 128, 128$ channels, BatchNorm, ReLU, global pool, 128-d feature) with a two-layer projection

Table 3: Encoder zoo (I1): pixel-referenced constants of Definition 1, latent anisotropy, and clean few-shot accuracy on *unseen* classes ($\pm$ s.d. over 400 episodes). The pixel-referenced ρ does not order encoders on natural images; the latent ratio does.

Dataset	Encoder	L_{intra}	ℓ_{inter}	ρ (pixel)	ρ_{lat}	few-shot acc.
MNIST	Identity (pixels)	1	1	1	0.894	0.728 ± 0.063
	VAE	0.485	0.145	3.34	0.829	0.545 ± 0.067
	SupCon	0.319	0.0925	3.45	0.633	0.784 ± 0.054
	ResNet-18 (ImageNet)	3.96	1.17	3.39	0.824	0.731 ± 0.063
CIFAR-10	Identity (pixels)	1	1	1	0.958	0.382 ± 0.064
	VAE	0.334	0.121	2.77	0.928	0.368 ± 0.062
	SimCLR	0.367	0.0445	8.25	0.930	0.299 ± 0.061
	SupCon	0.334	0.0843	3.96	0.791	0.589 ± 0.064
	ResNet-18 (ImageNet)	6.13	1.58	3.87	0.905	0.748 ± 0.058
	CLIP ViT-B/32	0.688	0.247	2.78	0.803	0.915 ± 0.034
	DINOv2-S	5.43	1.96	2.76	0.866	0.951 ± 0.028
CIFAR-100	Identity (pixels)	1	1	1	0.956	0.291 ± 0.054
	VAE	0.377	0.156	2.43	0.936	0.290 ± 0.057
	SimCLR	0.577	0.0448	12.9	0.865	0.208 ± 0.045
	SupCon	0.366	0.0952	3.85	0.837	0.568 ± 0.068
	ResNet-18 (ImageNet)	5.9	1.4	4.2	0.897	0.631 ± 0.069
	CLIP ViT-B/32	0.695	0.216	3.22	0.822	0.769 ± 0.073
	DINOv2-S	5.44	1.78	3.06	0.844	0.863 ± 0.049

Table 4: CIFAR-10-C: within-encoder correlation between the label-free estimator and OOD few-shot accuracy, over the 75 corruption blocks (15 corruptions $\times$ 5 severities) of each encoder. † marks encoders whose OOD accuracy is flat at chance across all blocks. Mean OOD accuracy and its spread are given for context.

Encoder	n	Pearson r	Spearman ρ_s	mean OOD acc.	s.d. of acc.
Identity (pixels)†	75	-0.314	-0.407	0.374	0.020
VAE	75	-0.776	-0.530	0.356	0.021
SimCLR†	75	-0.020	$+0.090$	0.302	0.015
SupCon	75	-0.966	-0.948	0.525	0.041
ResNet-18 (ImageNet)	75	-0.931	-0.942	0.642	0.073
CLIP ViT-B/32	75	-0.886	-0.879	0.802	0.082
DINOv2-S	75	-0.882	-0.936	0.853	0.080

head to 64 dimensions, trained with the supervised contrastive loss ($\tau = 0.1$) and the NT-Xent loss ($\tau = 0.5$) respectively. *ResNet-18*: ImageNet-pretrained trunk, inputs resized to 64 and ImageNet-normalised. *CLIP ViT-B/32* and *DINOv2-S (ViT-S/14)*: frozen, inputs resized to 224, grayscale repeated to three channels. Trained encoders use Adam at 10^{-3}, batch 256, 30 epochs (MNIST) or 60 epochs (CIFAR), global seed 0.

Corruptions. The five held-out transforms for MNIST/CIFAR-10/CIFAR-100 are `rotate:30`, `rotate:60`, `blur:1.5` (Gaussian, $\sigma = 1.5$), `translate:0.15` (fraction of image width) and `brightness:1.5` (multiplicative). None is used during training, so the shift is genuine. CIFAR-10-C supplies its own 15×5 grid.

Evaluation. N-way K-shot nearest-centroid on ℓ_2-normalised embeddings with $K = 5$ ($N = 5$ for MNIST, CIFAR-10 and CIFAR-10-C; $N = 10$ for CIFAR-100), 15 queries per class, 400 episodes; reported accuracies are episode means with the episode standard deviation. $\hat{\eta}$ uses 512 probe images. Statistics: bootstrap confidence intervals use 2000 resamples; the fixed-effects partial correlation is the Pearson correlation between the residuals of $\hat{\rho}$ and of accuracy after regressing each on encoder indicator variables; the OLS model is $\mathrm{acc} = \alpha_{\mathrm{enc}} + b\,\hat{\rho}$.

Hardware. CIFAR-10/100 and CIFAR-10-C runs on a single NVIDIA GPU (Kaggle T4 / rented A-series); MNIST and the toy experiments run on CPU or Apple-Silicon MPS in minutes. The full suite is a few GPU-hours.

Table 5: Per-encoder Pearson correlation between $\hat{\rho}_{\mathrm{OOD}}$ and OOD few-shot accuracy on the five held-out transforms (`rotate:30`, `rotate:60`, `blur:1.5`, `translate:0.15`, `brightness:1.5`); $n = 5$ rows per cell, so individual cells are noisy and only the sign pattern is meaningful. "–" means the encoder was not run on that dataset.

Encoder	MNIST	CIFAR-10	CIFAR-100
Identity (pixels)	-0.41	-0.95	$+0.34$
VAE	-0.91	-0.19	-0.91
SimCLR	–	$+1.00$	$+0.99$
SupCon	-0.88	-0.96	-0.95
ResNet-18 (ImageNet)	$+0.38$	-0.98	-0.95
CLIP ViT-B/32	–	-0.95	-0.98
DINOv2-S	–	-0.97	-0.98

Table 6: Pooled correlations with bootstrap 95% confidence intervals on CIFAR-10-C ($n = 525$). Both anisotropy predictors and all four baselines are computed on the identical rows.

Predictor	Pearson r [95% CI]	Spearman ρ_s [95% CI]	z-scored r [95% CI]	OLS p
$\hat{\rho}_{\mathrm{OOD}}$	$+0.007$ $[-0.07, +0.09]$	$+0.192$ $[+0.10, +0.28]$	-0.184 $[-0.26, -0.11]$	$< 10^{-3}$
$\hat{\rho}_{\mathrm{total}}$	-0.333 $[-0.39, -0.27]$	-0.216 $[-0.29, -0.14]$	-0.184 $[-0.26, -0.11]$	$< 10^{-3}$
entropy	-0.845 $[-0.86, -0.83]$	-0.871 $[-0.89, -0.85]$	-0.156 $[-0.23, -0.07]$	$< 10^{-3}$
rotation	-0.673 $[-0.75, -0.60]$	-0.750 $[-0.80, -0.69]$	-0.181 $[-0.25, -0.11]$	$< 10^{-3}$
Mahalanobis	-0.094 $[-0.16, -0.03]$	$+0.501$ $[+0.42, +0.57]$	$+0.004$ $[-0.08, +0.08]$	0.992
agreement	n/a	n/a	n/a	$< 10^{-3}$

F Baseline predictors

All four baselines are computed on the same rows as $\hat{\rho}$, from seen-class data plus the corrupted *unlabelled* images, and are scored with the identical statistics.

Entropy (Hendrycks & Gimpel, 2017): fit a linear probe on frozen seen-class features, then report the mean softmax entropy of that probe on the corrupted unseen images. **Agreement**: fit two probes with different seeds and report the fraction of corrupted images on which their $\arg\max$ agrees. **Mahalanobis** (Lee et al., 2018): fit class-conditional Gaussians with a shared covariance on seen features and report the mean minimum Mahalanobis distance of the corrupted images. **Rotation** (Gidaris et al., 2018; Deng et al., 2021): train a four-way rotation head on frozen seen features and report its accuracy on the corrupted images; higher rotation accuracy is expected to indicate an easier shift, which is why its correlation with OOD accuracy is negative once expressed on the same "difficulty" orientation as the other predictors.

Agreement is degenerate here: with frozen high-quality features the two probes agree on essentially every input, so the predictor is constant and no correlation is defined. We report this rather than dropping it, because it is a real failure mode of agreement-based predictors in the frozen-encoder regime.

G Full statistics

Table 3 gives the encoder zoo and the I1 measurements; Table 4 the within-encoder CIFAR-10-C correlations; Table 5 the per-encoder correlations on the five held-out transforms of each dataset; Table 6 the pooled correlations with bootstrap intervals for every predictor. Two patterns are worth naming. First, in Table 5 the sign is negative for every encoder that has accuracy variance to explain, and positive only for SimCLR and (on CIFAR-100) raw pixels - the two chance-level encoders. Second, in Table 6 the Spearman and Pearson coefficients of the Mahalanobis baseline have opposite signs ($+0.50$ against -0.09), which is a symptom of its extreme right-skewed scale and a reason to distrust its pooled numbers entirely.

H Additional figures

Table 7: T1 in numbers. For each contraction factor a of the diagonal encoder we report the measured anisotropy ratio, the empirical margin at which one-shot accuracy first reaches 90%, and the two sufficient thresholds. Both theorems are satisfied at every ρ; 4ρ is the tighter envelope for $\rho < 1$.

a	measured ρ	empirical 90% threshold	4ρ (Theorem 1)	$2(1+\rho)$ (Theorem 4)	slack
1.00	1.000	1.46	4.00	4.00	+2.54
0.60	0.636	1.16	2.54	3.27	+1.39
0.40	0.432	0.86	1.73	2.86	+0.87
0.25	0.273	0.55	1.09	2.55	+0.54
0.15	0.164	0.25	0.66	2.33	+0.41
0.08	0.088	0.25	0.35	2.18	+0.10

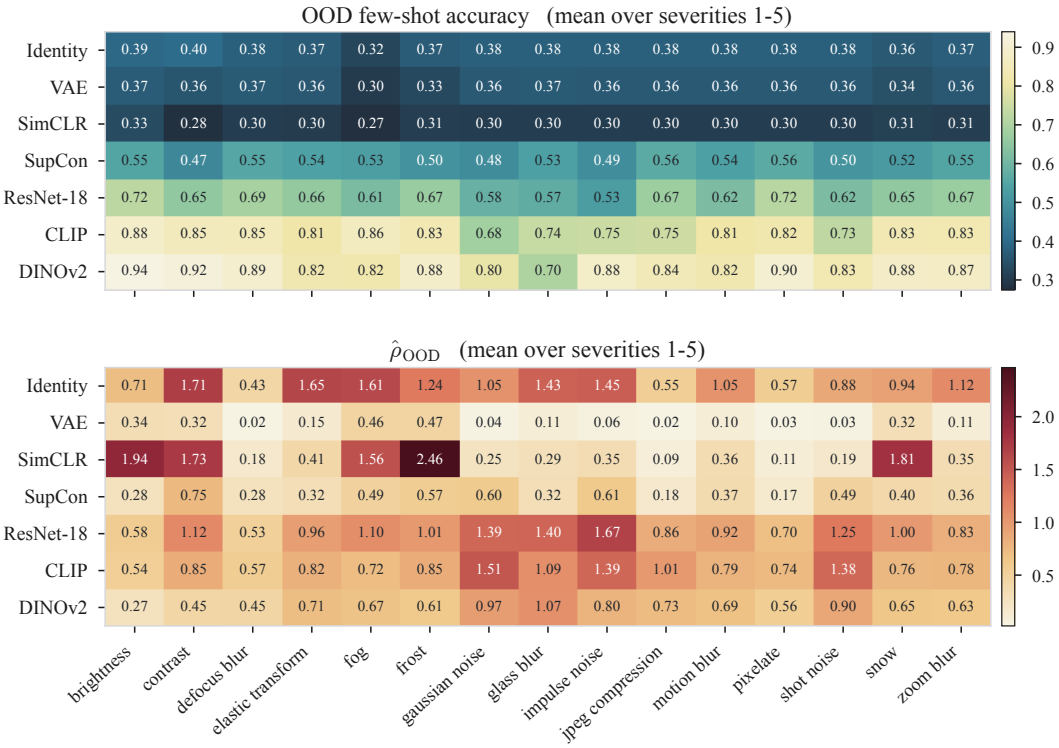

Figure 9: CIFAR-10-C broken out by corruption. Top: OOD few-shot accuracy; bottom: the label-free estimator. The estimator's largest values fall on the corruptions that most damage each encoder - contrast, elastic transform and fog for raw pixels; impulse, Gaussian and glass-blur noise for the pretrained trunks - and its smallest on the ones those encoders shrug off, brightness and defocus blur. For DINOv2, for instance, brightness gives the lowest $\hat{\rho}$ (0.27) and the highest accuracy (0.94), while glass blur gives the second-highest $\hat{\rho}$ (1.07) and the lowest accuracy (0.70).

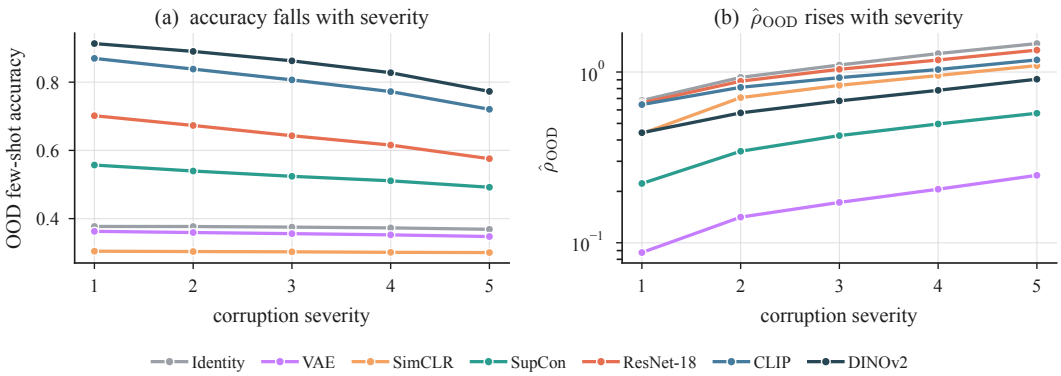

Figure 10: Severity sweeps. Accuracy falls and $\hat{\rho}_{\mathrm{OOD}}$ rises monotonically with severity for every encoder, which is the within-encoder trend Theorem 8 predicts. Note that the two chance-level encoders (Identity, SimCLR) have flat accuracy curves while their $\hat{\rho}$ still rises: this is the variance-free regime in which no difficulty predictor can succeed.

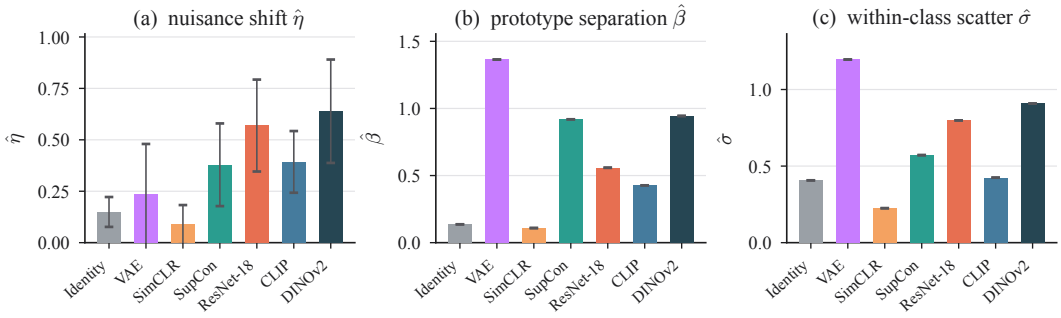

Figure 11: The three ingredients of Equation (4), per encoder, averaged over all 525 blocks (error bars: s.d. across blocks). The VAE illustrates exactly why $\hat{\rho}_{\mathrm{total}}$ is needed for cross-encoder comparison: it has the largest prototype separation $\hat{\beta}$ of any encoder *and* the largest within-class scatter $\hat{\sigma}$, so $\hat{\eta}/\hat{\beta}$ alone makes it look deceptively robust.

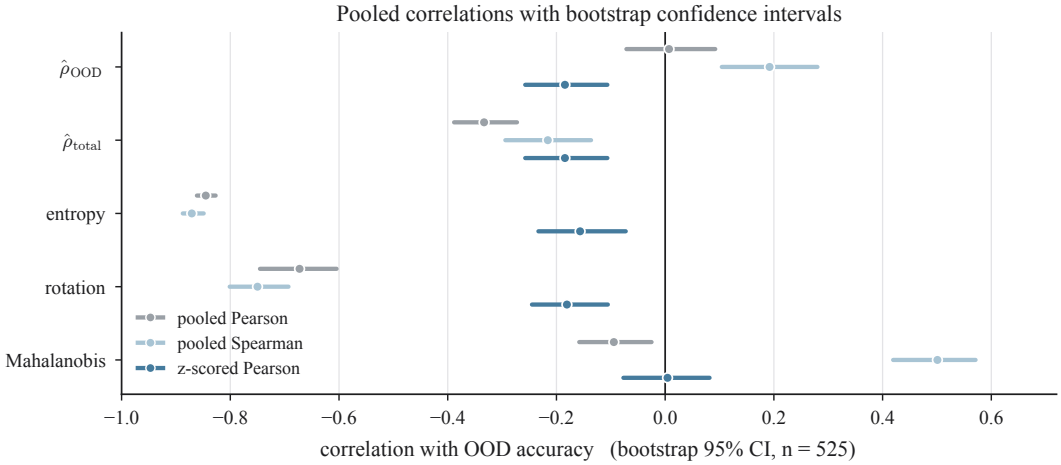

Figure 12: Pooled correlations with bootstrap 95% intervals (cf. Table 6). The z-scored column is the fair pooled comparison: after within-encoder standardisation every predictor lands in the same narrow band, and entropy's apparent dominance in the raw column disappears.

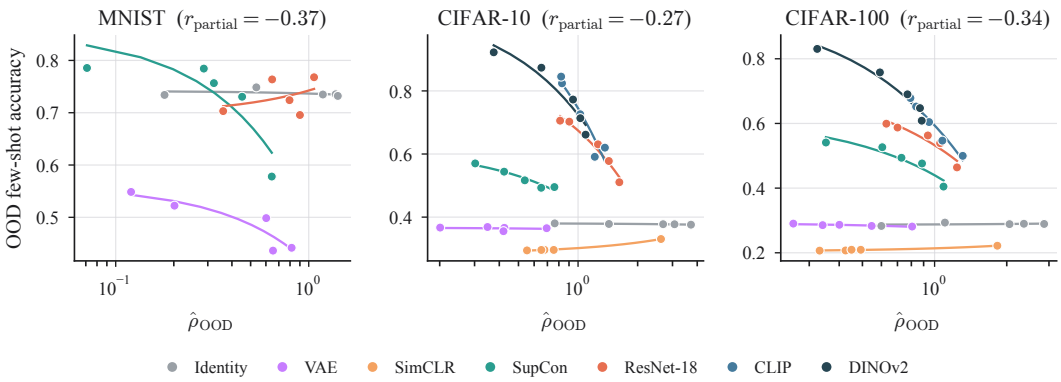

Figure 13: I2 on the five held-out transforms of each dataset, with per-encoder regression lines. The same structure as CIFAR-10-C appears: clean negative within-encoder slopes for the encoders with accuracy variance, flat lines for the chance-level ones.

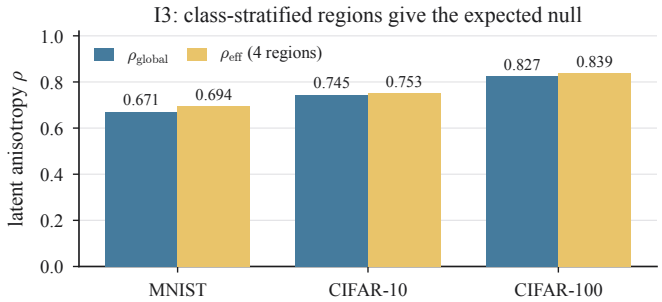

Figure 14: I3. Class-stratified regions give ρ_{eff} within 0.02 of ρ_{global} on all three datasets. This is the null Proposition 6 predicts for an exchangeable partition; it is reported so that the T4 result is not mistaken for a claim about arbitrary mixtures.

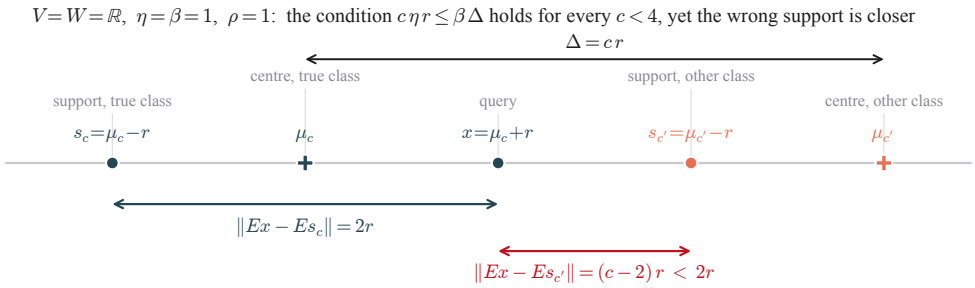

Figure 15: The witness of Proposition 3, drawn for $c = 3$. Both class clouds satisfy Assumptions 1 and 2 and the margin condition holds with the constant c, yet the query sits closer to the competing support than to its own. Formalised as `margin_law_four_is_sharp`.

I EXTENDED RELATED WORK

This section states in full the comparisons condensed in Section 2.

The geometry of learned features. Neural collapse (Papyan et al., 2020) documents that, late in training, within-class features concentrate on their class means and the means arrange into a simplex - a description of the endpoint of the process our ratio measures. Hermann & Lampinen (2020) show that networks preferentially represent one of several redundant features and partially suppress the rest, which is the mechanism by which a nuisance direction gets contracted at all. Theory for contrastive learning (Arora et al., 2019; HaoChen et al., 2021; Wang & Isola, 2020) bounds downstream linear-probe error in terms of augmentation-graph or alignment/uniformity quantities. We differ in target and in form: we bound few-shot *nearest-centroid* error on *unseen classes under unseen nuisances*, in terms of a single scalar that is an explicit ratio of directional Lipschitz constants, and can therefore be both proved about and measured.

Margins and Lipschitz control. Margin-based generalisation bounds (Bartlett et al., 2017; Neyshabur et al., 2017; Sokolić et al., 2017) and Lipschitz-constrained architectures (Cisse et al., 2017) also relate norms of the map to classification robustness. Those results control the *global* Lipschitz constant and bound in-distribution generalisation. The anisotropy ratio is deliberately *directional* and *two-sided*: what matters is not how much the encoder can stretch, but the contrast between what it does to nuisance directions and what it does to semantic ones. A globally 1-Lipschitz map can have any ratio at all, and an isometry - optimal by the global criterion - is the *worst* possible case by ours (Proposition 5).

Predicting generalisation without labels. A large literature estimates test accuracy from unlabelled data: complexity measures fitted across training runs (Jiang et al., 2020), distribution statistics of the unlabelled target set (Deng & Zheng, 2021; Guillory et al., 2021), thresholded confidence (Garg et al., 2022), rotation-prediction accuracy (Deng et al., 2021; Sun et al., 2020), and internal-representation topology (Corneanu et al., 2020). Almost all of these read the *target* distribution. Our estimator does not: it is computed on seen classes and a held-out nuisance applied to seen images, and it is derived from a theorem rather than fitted. We benchmark against the two strongest of these families (entropy (Hendrycks & Gimpel, 2017) and rotation (Gidaris et al., 2018; Deng et al., 2021)) plus Mahalanobis scoring (Lee et al., 2018) and probe agreement, under one protocol.

Robustness benchmarks and domain-adaptation theory. CIFAR-10-C (Hendrycks & Dietterich, 2019) supplies the 15 corruptions $\times$ 5 severities we use; Taori et al. (2020) and Recht et al. (2019) document how weakly synthetic robustness transfers to natural shift, a caveat we inherit and state in Section 5. Classical domain-adaptation theory (Ben-David et al., 2010) bounds target error by source error plus a divergence between distributions. Our (A4) plays an analogous role but is a statement about the *encoder* rather than about the pair of distributions, which is what makes it measurable without target data.

Few-shot classification and mixtures. Nearest-centroid (prototypical) classification (Snell et al., 2017; Vinyals et al., 2016) is our evaluation protocol; Tian et al. (2020) argue the embedding, not the meta-learner, does the work - exactly the claim our ratio makes quantitative. Proposition 6 concerns mixtures of experts (Jacobs et al., 1991; Shazeer et al., 2017) and gives a geometric reason for their advantage under region-dependent nuisance.

J EXTENDED LIMITATIONS

This section states in full the caveats condensed in Section 5.

Scale and shift realism. Everything here runs on MNIST, CIFAR-10, CIFAR-100 and CIFAR-10-C. The corruptions are synthetic, and synthetic robustness is known to transfer weakly to natural distribution shift (Taori et al., 2020; Recht et al., 2019). Whether the estimator survives ImageNet-C, DomainNet (Peng et al., 2019) or WILDS (Koh et al., 2021) is open, and we do not claim it.

The pooled correlation is weak. We want to be unambiguous: the raw pooled correlation on CIFAR-10-C is $+0.01$, and only the fixed-effects and non-degenerate analyses are strong. This is

what the theory predicts - the ratio is defined per encoder - but a practitioner who wants to compare two *different* encoders with a single unadjusted number will not get it from $\hat{\rho}_{\mathrm{OOD}}$, and $\hat{\rho}_{\mathrm{total}}$ only partly repairs this.

(A4) is an assumption, not a measurement. Theorem 8 is conditional on the unseen nuisance lying inside the contracted subspace. T3 shows what happens as that condition is violated - accuracy degrades exactly as $\hat{\rho}$ says it should - but we have no test that certifies (A4) holds for a corruption we have not seen. The estimator is a prediction under an assumption, not a guarantee.

Oracle routing in T4. The mixture experiment uses nuisance-aligned oracle routing. Propositions 6 and 7 say what a mixture can achieve, not that a gate can be learned to achieve it; learning the router is orthogonal work and we did not do it.

Linearisation. (A3) replaces a smooth encoder by its Jacobian. The estimator of Equation (4) uses finite differences at the corruption's own magnitude rather than an infinitesimal probe, which is the honest choice for severity-5 corruptions but means $\hat{\eta}$ measures a secant, not a derivative.

Two degenerate encoders. Raw pixels and SimCLR sit at chance on unseen CIFAR classes, which is why we report the non-degenerate split separately. Reporting only that split would be cherry-picking; reporting only the full set would hide that the estimator is being asked to explain variance that is not there. Both are in Tables 1 and 2.