A GEOMETRIC DIAGNOSTIC FOR FROZEN FOUNDATION FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

ABSTRACT

Classification on domain-specific data often pairs frozen foundation-model features with a lightweight classifier, yet uneven per-class performance is usually attributed to class imbalance rather than representation geometry. On PVEL-AD, a balanced linear probe on frozen DINOv2 features recalls 94% of one defect class from 56 labeled examples but only 26% of another from 96, showing that class frequency alone does not explain the failure pattern. We use Neural Collapse as a geometric reference and show that anisotropy imposes a strictly positive lower bound on deviation from the simplex equiangular tight frame (ETF), with a median 99% of the measured deviation across our benchmark accounted for by this floor. We introduce the Geometric Collapse Score (GCS), the mean pairwise cosine of L2-normalized class centroids, as a measure of how tightly an encoder compresses a domain. Across eleven datasets and four encoders, dataset ordering by GCS is nearly invariant, with pairwise Spearman correlations from 0.96 to 1.00. Across 55 classes from seven specialized datasets, per-class GCS strongly correlates with recall ($\rho = -0.785$), while class frequency does not ($\rho = -0.162$). Frozen GCS also separates two empirical regimes of adaptation response. LoRA improves balanced accuracy on eight of nine specialized datasets and reduces it on both natural-image controls, although the boundary is an ordering rather than a universal threshold. An ETF-alignment objective improves PVEL-AD balanced accuracy by 3.3 points over an identical-pipeline control while reducing cone alignment, with class-level gains remaining within seed noise and no meaningful transfer to ResNet-50. Together, these results identify anisotropic feature geometry as a measurable failure mode of frozen representations and position GCS as a practical pre-adaptation diagnostic for when the representation itself warrants intervention.

1 INTRODUCTION

Frozen foundation models are the default source of visual representations wherever labeled data are scarce. The standard recipe extracts features from a pretrained encoder such as DINOv2 (Oquab et al., 2023) or CLIP (Radford et al., 2021) and trains only a linear classifier on top. When per-class performance comes out uneven, the long-tailed recognition literature diagnoses the cause as class imbalance and applies the remedy at the classifier, through re-weighting (Cui et al., 2019), post-hoc logit adjustment (Menon et al., 2020), or decoupled training (Kang et al., 2019). These methods act at the classifier while leaving the representation fixed. They therefore cannot directly correct failures in the geometry of frozen features. Such failures are particularly relevant on specialized inspection data whose imaging physics and category systems differ substantially from web-scale pretraining content.

These methods act at the classifier while leaving the representation fixed. They therefore cannot directly correct failures in the geometry of frozen features. Such failures are particularly relevant on specialized inspection data whose imaging physics and category systems differ substantially from

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.

web-scale pretraining content. On such data, class frequency no longer predicts the recall of the model. On PVEL-AD, a solar-cell electroluminescence benchmark with 13 classes and a 5,677:1 imbalance ratio (Chen et al., 2020), a balanced linear probe trained at a 2,000-label budget on frozen DINOv2 ViT-S/14 features recalled 94.0% of ShortCircuit from 56 labeled examples, yet only 25.9% of Crack from 96 and 49.0% of Normal, the most frequent class, from 1,308. This pattern is not explained by class frequency alone, which motivates examining how the frozen encoder arranges the classes.

The representation is precisely what the standard pipeline never inspects, and that omission carries a cost. A practitioner choosing among collecting more labels, adapting the encoder, and probing it frozen has no basis for the decision until each option has been trained and evaluated. What is missing is a measurement of the geometry of the frozen features, available before any training, that identifies where the representation will fail and indicates whether adapting it is warranted.

Papyan et al. (2020) described the arrangement that measurement would have to be read against. A network trained under supervision past zero error on balanced data develops four co-occurring properties in its last-layer representations, and the second of them, NC2, fixes the arrangement of the class means. Under NC2 the class means converge to a simplex equiangular tight frame (ETF), the configuration that separates C classes with maximal and equal pairwise angles, in which every pair of class-mean directions attains cosine similarity exactly $-1/(C-1)$. Such a configuration exists only when the ambient dimension satisfies $d \geq C - 1$, which holds for every encoder and dataset we study. Collapse is established for networks trained under supervision on the target classes to the terminal phase of training, and frozen foundation features satisfy none of these conditions. Whether the arrangement is merely unattained on such features, or cannot be attained at all, has not been asked.

For the anisotropic frozen representations we study, we show that positive cone alignment places a nonzero lower bound on deviation from the ETF arrangement. This constraint belongs to the representation rather than to the classifier trained on it. Foundation-model features do not spread out over the representation sphere. They concentrate in a narrow cone around a single dominant direction, and since a class centroid is an average of features from that class, the centroids are confined to the same cone. Anisotropy of this kind is a general property of learned representations, but its severity is domain-dependent, and that dependence is what we measure. Directions packed into a narrow cone cannot point away from one another, so their pairwise cosines are forced positive, while the ETF requires them to be negative. Under positive cone alignment the ETF target is unreachable without changing the representation, so classifier-side corrections alone cannot remove this geometric constraint. The quantity worth measuring is therefore not whether frozen features deviate from the ETF, which they must, but how tightly an encoder compresses a domain into that cone.

We propose the Geometric Collapse Score (GCS), a diagnostic statistic that locates the failure and converts directly into an objective that repairs it. GCS is the mean pairwise cosine of L2-normalized class centroids, the quantity the cone forces above the ETF value. Because it is defined on pairs of classes rather than on a dataset as a whole, its individual terms identify which classes collide rather than only how badly a representation is arranged overall, and because it is differentiable in the centroids, the quantity that diagnoses the failure can be minimized during fine-tuning to correct it. Scalar summaries of embedding structure provide neither. We evaluate GCS on eleven datasets and four encoders, spanning six application domains and including two long-tailed natural-image controls, and we use LoRA fine-tuning (Hu et al., 2021) at matched label budgets to test whether reshaping the geometry changes recall.

Under a single protocol spanning eleven datasets, six application domains, and four encoders, our experiments establish three findings. First, the departure from the ETF is set by the data domain and not by the encoder. Frozen GCS ranges from 0.64 to 0.99 on the nine specialized benchmarks and falls to 0.31 and 0.11 on the controls, and the four encoders order the datasets alike at Spearman 0.96 to 1.00, even when an ImageNet-supervised ResNet-50 (He et al., 2016) is set against three vision transformers. Second, GCS predicts per-class recall where class frequency does not, reaching $\rho = -0.79$ over 55 classes against a statistically insignificant -0.16 for training-set size. Third, the size of the departure separates two regimes of adaptation response. LoRA improved balanced accuracy on eight of the nine specialized datasets by roughly 2 to 20 points and underperformed frozen probing by 5.1 and 6.5 points on the controls, a boundary we report as an ordering rather than a threshold.

This paper makes four contributions.

1. **Infeasibility of the ETF arrangement under anisotropy.** We derive a strictly positive lower bound on the deviation from the ETF value, and show that a median 99% of the observed deviation lies at that bound.
2. **A domain-structural departure across four encoders** differing in architecture family, pretraining objective and corpus, which locates the departure in the data domain.
3. **A pair-resolved diagnostic of per-class failure,** which orders classes by recall where frequency does not and resolves failure to individual class pairs, as no scalar validity index does.
4. **An ETF-alignment objective and its operating range,** which improves accuracy in the geometry-failure regime by widening the cone, with ablations establishing its encoder and weighting limits.

2 Related work

This section covers four bodies of work: Neural Collapse and the conditions under which its predicted geometry emerges, training objectives that impose that geometry, statistics that predict downstream behavior from embedding structure, and long-tailed recognition together with the adaptation methods it relies on.

Neural Collapse and its boundary conditions. Papyan et al. (2020) established that networks trained past zero error on balanced data develop four co-occurring properties in their last-layer representations, among them the convergence of class means to a simplex equiangular tight frame. Later work characterized when that geometry does and does not emerge. Under class imbalance the arrangement deforms and minority centroids merge (Fang et al., 2021), and the ETF generalizes to the SELI geometry (Thrampoulidis et al., 2022). Fixing the classifier to an ETF restores collapse under imbalance (Yang et al., 2022). Kothapalli & Tirer (2024) argue that unconstrained-features analyses are data-blind, and show through kernel approximations that data structure controls the extent of collapse. This literature asks under what training conditions the ETF geometry fails to emerge, whereas we ask whether frozen features can reach it at all (Section 4).

Training objectives that target the ETF. A number of methods impose the ETF arrangement rather than waiting for it to appear. Fixed ETF classifiers induce collapse under imbalance (Yang et al., 2022), pre-assigned ETF prototypes anchor few-shot class-incremental learning (Yang et al., 2023), and ETF structure has been used to score out-of-distribution detection (Ben Ammar et al., 2024). These objectives apply the target uniformly, wherever the model is trained. Our alignment loss belongs to the same family, and differs in being coupled to a diagnostic that decides in advance whether to apply it at all (Section 6).

Predicting downstream behavior from embedding structure. A separate line of work predicts task performance from the structure of a representation. Cluster-validity indices such as the Silhouette score weigh within-class compactness against between-class separation. Spectral imbalance ties per-class error to disparities in the spectra of class-conditional covariances (Kaushik et al., 2024), and geometric complexity tracks transfer quality over the course of pretraining (Munn et al., 2024). Huh et al. (2024) document the convergence of independently trained models toward shared representations. Each of these statistics summarizes a representation with a single number per class or per dataset, and on our benchmark several of them predict pooled per-class recall as well as or better than GCS (Section 5). GCS differs in being defined on pairs of classes and differentiable in the centroids, which is what makes it both pair-resolved and usable as a training objective.

Long-tailed recognition and the adaptation of foundation models. Long-tailed recognition corrects classifier bias on top of a feature space taken to be adequate. Re-weighting (Cui et al., 2019), margin-based and decoupled training (Cao et al., 2019; Kang et al., 2019), and post-hoc logit adjustment (Menon et al., 2020) all act at the classifier, and the adjustment has been extended to foundation-model features (Zhu et al., 2023) and calibrated against an assumed collapsed geometry (Hasegawa & Sato, 2025). On the adaptation side, LoRA reshapes features at a small parameter cost (Hu et al., 2021), PaCo adds frequency-aware contrastive structure (Cui et al., 2021), and label-free

continued pretraining adapts encoders without target labels (Rodas et al., 2025). Both groups presuppose the geometry or reshape it without first measuring it, and our experiments delimit where the premise that the feature space is adequate actually holds (Section 5).

The next section fixes the benchmark, the encoders, and the statistic on which every subsequent claim rests.

3 THE GEOMETRIC COLLAPSE SCORE

Under NC2, the second Neural Collapse property, the class means of a fully collapsed representation converge to a simplex ETF in which every pairwise cosine equals the ETF value $s^* = -1/(C-1)$ (Papyan et al., 2020). We measure class geometry against that reference with a single statistic, the mean pairwise cosine of L2-normalized class centroids. Let $f(x) \in \mathbb{R}^d$ denote the frozen encoder output for image x in a task with C classes. The *class mean* is the average feature of class c over its n_c training images, $\boldsymbol{\mu}_c = n_c^{-1} \sum_{i:y_i=c} f(x_i)$, and the *class centroid* $\hat{\boldsymbol{\mu}}_c = \boldsymbol{\mu}_c / \|\boldsymbol{\mu}_c\|_2$ is that mean scaled to unit length, so it records the average *direction* of the class. We average raw features and normalize only afterwards, which keeps the centroid at the class mean rather than at the mean of already-normalized features. The Geometric Collapse Score is then

$$\mathrm{GCS} = \frac{2}{C(C-1)} \sum_{i<j} \hat{\boldsymbol{\mu}}_i^\top \hat{\boldsymbol{\mu}}_j. \tag{1}$$

The deviation $\Delta_{\mathrm{ETF}} = \mathrm{GCS} - s^*$ vanishes for a fully collapsed representation and grows as the centroid directions concentrate. We keep the individual cosines, which Section 5 uses to resolve failure to class pairs.

The benchmark is eleven datasets from six application domains. Nine are specialized inspection tasks and two are natural-image controls. Three electroluminescence benchmarks cover solar-cell defect inspection: PVEL-AD (Chen et al., 2020), BenchmarkEL (Pratt et al., 2023), and ELPV (Deitsch et al., 2019). Three infrared collections cover photovoltaic thermal inspection: binary and 12-class anomaly benchmarks (Millendorf et al., 2020), and THED-PV (Ansari, 2025). NEU-CLS covers hot-rolled steel-surface defects (Song & Yan, 2013), HAM10000 dermoscopic skin-lesion diagnosis, and EuroSAT Sentinel-2 land-use classification. The two controls, CIFAR-100-LT (Krizhevsky et al., 2009; Cui et al., 2019) and Places-LT (Liu et al., 2019), are long-tailed natural-image benchmarks close to web-scale pretraining content, so they keep label scarcity and imbalance without domain shift. Table 2 in Appendix A gives the full composition, from class counts to imbalance ratios and label budgets.

We use four frozen encoders, so that we can tell whether the geometry we measure belongs to the network or to the data. They are DINOv2 ViT-S/14 and ViT-B/14 (Oquab et al., 2023), CLIP ViT-B/16 (Radford et al., 2021), and an ImageNet-supervised ResNet-50 (He et al., 2016), covering two architecture families and three pretraining regimes. ViT-S/14 is the reference backbone for every number not otherwise attributed.

One protocol produces every accuracy in the paper, so probes, adaptation, and the statistic are always compared at the same label budget. Per-class recall, the quantity the statistic is asked to predict, comes from balanced logistic-regression probes trained on L2-normalized frozen features, with inverse regularization strength 0.316 and inverse-frequency class weights. Every dataset carries a fixed label budget denoted L4, which is 2,000 labels on PVEL-AD, and we use official splits and three seeds wherever budget subsampling applies. We train probes on raw features alongside and score adaptation against the stronger of the two. Adaptation is rank-8 LoRA (Hu et al., 2021) on the attention projections at the same label budgets, and on convolutional layers for ResNet-50. Five of PVEL-AD's 13 classes have fewer than 10 test images, so their recalls are reported for completeness rather than inference (Appendix A).

GCS is available before any fine-tuning but does require labels, with a bootstrap confidence interval of width 0.05 taking a median of 20 labeled examples per class and up to 50 on the hardest datasets (Appendix F). The next section shows that on such features the ETF reference is not merely unmet but unreachable.

4 THE ETF ARRANGEMENT IS INFEASIBLE FOR FROZEN FEATURES

On frozen features, most of the similarity measured between class centroids is not attributable to class structure. Under frozen DINOv2 ViT-S/14, two unrelated PVEL-AD images share a mean cosine of 0.753, where an isotropic representation in $d = 384$ would give 0.000 ± 0.051, placing the observed value roughly fifteen standard deviations above chance. Two further measurements agree. The leading eigenvalue of the uncentered second-moment matrix carries 78.7% of the total energy, and the median image lies $19°$ from the corresponding direction rather than the $90° \pm 3°$ that isotropy implies (Figure 1a). Against that background the thirteen class centroids have mean pairwise cosine 0.727, the GCS of Section 3, so two centroids are no more similar than two random images. PVEL-AD is the extreme case among the nine datasets where we measured the background, which elsewhere falls below the class value by factors of 1.3 to 4.1. GCS on frozen features therefore reports the anisotropy of the representation, not a class arrangement contaminated by it, and Section 5 shows that the anisotropy is what carries the predictive signal.

We quantify anisotropy through a cone axis estimated without labels. Let $\boldsymbol{m}$ be the leading eigenvector of the uncentered second-moment matrix of the L2-normalized features, which summarizes the representation's dominant direction, and for each class define the cone alignment $\gamma_c = \hat{\boldsymbol{\mu}}_c^\top \boldsymbol{m}$, the residual $\boldsymbol{v}_c = \hat{\boldsymbol{\mu}}_c - \gamma_c \boldsymbol{m}$ with $\boldsymbol{v}_c \perp \boldsymbol{m}$, and its magnitude $s_c = \|\boldsymbol{v}_c\| = \sqrt{1-\gamma_c^2}$. That label-free choice is load-bearing. Were $\boldsymbol{m}$ taken instead to be the mean centroid direction, the residuals would satisfy $\sum_c \boldsymbol{v}_c = 0$ identically and the bound below would hold with equality by construction, which is true but vacuous. Per-class γ_c and GCS predict recall about equally well and each retains independent signal when the other is partialled out, so the two divide the labor rather than substituting for one another (Appendix C). We measure γ on centroids rather than images, and the two are related without being interchangeable, since averaging within a class can leave centroids far more concentrated than the images they summarize. THED-PV pairs a median image angle of $45°$, against PVEL-AD's $19°$, with the highest centroid alignment of any multiclass dataset we test.

Anisotropy places a hard floor under the mean pairwise centroid cosine.

Proposition 1 (Anisotropy floor). *Let $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_k \in \mathbb{R}^d$ be the unit centroids of the k classes present in a sample ($k = C$ when every class occurs), let $\boldsymbol{m} \in \mathbb{R}^d$ be any unit vector, which in our measurements is the label-free cone axis defined above, and write $\gamma_i = \hat{\boldsymbol{\mu}}_i^\top \boldsymbol{m}$, $\boldsymbol{v}_i = \hat{\boldsymbol{\mu}}_i - \gamma_i \boldsymbol{m}$, $s_i = \|\boldsymbol{v}_i\| = \sqrt{1-\gamma_i^2}$. Then the mean pairwise centroid cosine of equation 1, computed over these k classes, obeys*

$$\mathrm{GCS} \;\geq\; \mathrm{GCS}_{\min} \;=\; \frac{2}{k(k-1)} \sum_{i<j} \gamma_i \gamma_j \;-\; \frac{1}{k(k-1)} \sum_{i=1}^{k} s_i^2, \tag{2}$$

with equality if and only if $\sum_{i=1}^{k} \boldsymbol{v}_i = \boldsymbol{0}$.

Proof. Since $\hat{\boldsymbol{\mu}}_i^\top \hat{\boldsymbol{\mu}}_j = \gamma_i \gamma_j + \boldsymbol{v}_i^\top \boldsymbol{v}_j$, expanding $0 \leq \|\sum_i \boldsymbol{v}_i\|^2 = \sum_i s_i^2 + 2\sum_{i<j} \boldsymbol{v}_i^\top \boldsymbol{v}_j$ gives $\sum_{i<j} \boldsymbol{v}_i^\top \boldsymbol{v}_j \geq -\frac{1}{2}\sum_i s_i^2$, and substituting into $\mathrm{GCS} = \frac{2}{k(k-1)} \sum_{i<j} (\gamma_i \gamma_j + \boldsymbol{v}_i^\top \boldsymbol{v}_j)$ yields equation 2, with equality if and only if $\sum_i \boldsymbol{v}_i = \boldsymbol{0}$. □

The bound is tight in the common-alignment case. When $\gamma_i = \gamma$ and $d - 1 \geq k - 1$, a simplex arrangement of the residual directions attains equality (construction in Appendix B). For heterogeneous measured alignments, equation 2 remains a valid lower bound.

Corollary 1 (Minimum attainable ETF deviation). *If $\gamma_i = \gamma$ for all i, then $\mathrm{GCS} \geq \gamma^2 - (1-\gamma^2)/(k-1)$, and since the ETF value on k classes is $s^* = -1/(k-1)$, the minimum attainable deviation is*

$$\min \Delta_{\mathrm{ETF}} \;=\; \gamma^2 \, \frac{k}{k-1} \;>\; 0 \qquad \textit{for every } \gamma > 0,$$

while $\gamma = 0$ recovers the ETF value exactly.

For any positive cone alignment the ETF arrangement is therefore unreachable, and with all C classes present the floor is the $\gamma^2 C/(C-1)$ of Section 1. Rearranged, the corollary gives a condition that can be checked directly against a measurement. The floor is negative only when $\gamma < 1/\sqrt{C}$,

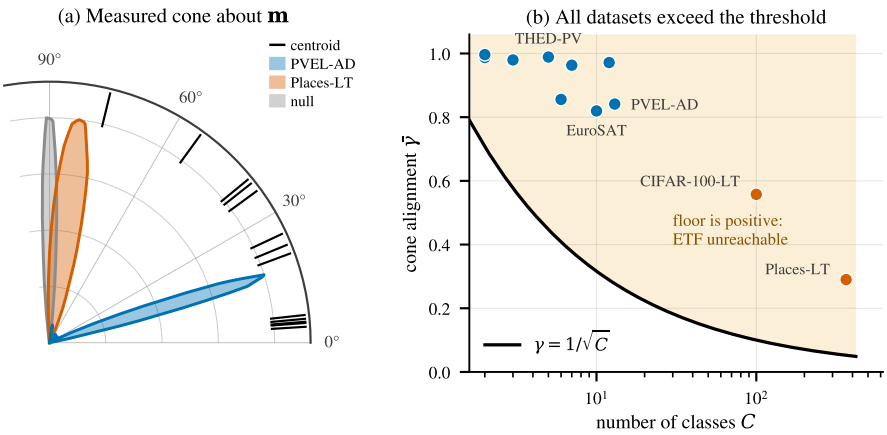


Figure 1: **The cone is measured, and it places the ETF out of reach on every dataset. (a)** Density of image features by angle to the label-free axis $\boldsymbol{m}$, each curve rescaled to peak at one, with the thirteen PVEL-AD centroids as rays. **(b)** Measured $\bar{\gamma}$ against the feasibility threshold $\gamma = 1/\sqrt{C}$. All eleven datasets lie where $\mathrm{GCS}_{\min} > 0$, so no centroid arrangement attains $s^* < 0$.

and no dataset in the benchmark approaches that (Figure 1b). PVEL-AD would need γ below 0.277 and measures 0.841, and Places-LT, the widest cone we observe, would need 0.052 and measures 0.290. Numerical checks confirm the bound and its tightness (Appendix B).

Measured on our benchmark, the floor is not merely valid but nearly saturated. The $\bar{\gamma}$ and $\mathrm{GCS}_{\min}$ columns of Table 1 (Section 5) evaluate equation 2 with the measured per-class alignments γ_c on frozen ViT-S/14 features. The bound holds on all eleven datasets under all three ViT encoders, and a median 99% of the observed deviation $\Delta_{\mathrm{ETF}} = \mathrm{GCS} - s^*$ is forced by the cone. The largest excess over the floor is 0.044 (PVEL-AD) and the smallest forced fraction 0.774 (Places-LT). Mean alignment $\bar{\gamma}$ falls from 0.963–0.996 on the six most tightly coned specialized datasets to 0.820–0.856 on NEU-CLS, PVEL-AD and EuroSAT, and to 0.558 and 0.290 on the two in-distribution controls, which is the ordering a domain-distance reading predicts.

Two further results locate the predictive content of GCS in the cone term rather than the residual class arrangement. First, the excess $\mathrm{GCS}-\mathrm{GCS}_{\min}$ does not predict recall, reaching $\rho = +0.108, +0.128$ and $+0.054$ on the three ViT encoders (all $p > 0.3$) over the 55 classes of the seven multiclass specialized datasets, against -0.785, -0.806 and -0.762 for raw GCS. Second, subtracting each dataset's background cosine destroys the signal rather than sharpening it, taking those coefficients to -0.309, -0.461 and -0.069, and whitening fares no better (Table 4, Appendix B). GCS predicts recall *because* it measures how tightly the encoder compresses the domain, which makes anisotropy the mechanism to model rather than a background to subtract.

NC2 therefore cannot transfer to frozen features as a matter of geometry rather than of optimization (Papyan et al., 2020), because no classifier built on them reaches an arrangement that does not exist inside the cone. The violation is domain-structural, its size set by how tightly the encoder cones the domain, so encoders that compress a domain alike must agree on the deviation. Only a representation-level intervention lowers the floor, because only such an intervention moves γ, so post-hoc corrections assuming a collapsed geometry act on one that cannot be present (Appendix H) and any intervention that works must widen the cone. Whether this compression matters for recognition is an empirical question, and the next two sections answer it.

5 Geometry separates two failure regimes

The compression that Section 4 showed to be forced is also the quantity that predicts recognition failure, since within datasets GCS orders classes by recall where class frequency does not. Pooled over the 55 classes of the seven multiclass specialized datasets, the per-class statistic GCS_c, the mean cosine of $\hat{\boldsymbol{\mu}}_c$ to the remaining centroids, correlated with the recall of the balanced probes of Section 3 at $\rho = -0.785$ (bootstrap 95% CI $[-0.849, -0.677]$), whereas training-set size reached only -0.162 ($p = 0.24$, Figure 2). Three controls left the correlation intact: partialling out training-

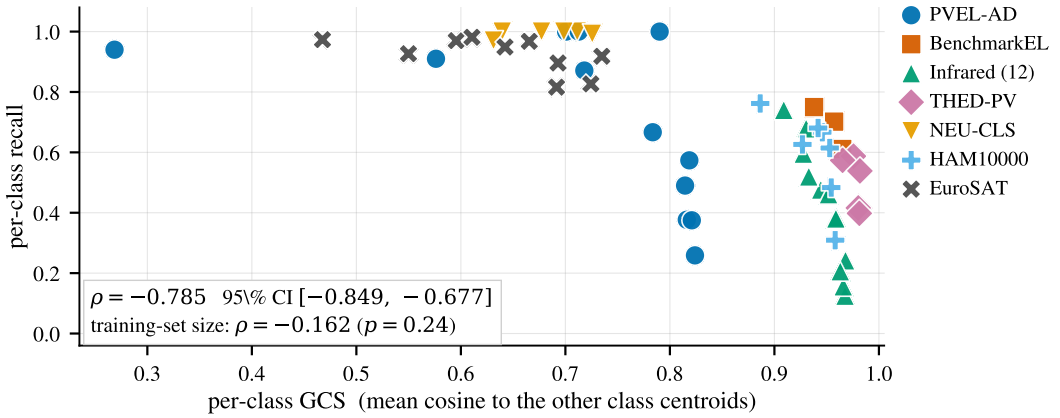


Figure 2: **Geometry orders classes by recall where class frequency does not.** Per-class GCS against recall over the 55 classes of the seven specialized datasets (frozen ViT-S/14, balanced probes, L4 budget), with one marker style per dataset and a running mean over 17 neighbouring classes.

set size (-0.783), computing centroids from training data only (-0.782), and restricting to the 50 classes with at least 10 test images (-0.800). It replicated on DINOv2 ViT-B/14 (-0.806) and CLIP ViT-B/16 (-0.762). Because pooling mixes GCS scales that differ across domains, the pooled coefficient summarizes a within-dataset ordering, not a transportable scale.

The correlation is not an artifact of test-set redundancy. Dropping the five duplicate-heavy classes moved it to -0.802 ($n = 50$) rather than weakening it, and duplicate load predicted neither recall nor GCS (Appendix E).

GCS resolves failure to individual class pairs. On the eight multiclass datasets with supported pairs, the pairwise cosine $\hat{\boldsymbol{\mu}}_i^\top \hat{\boldsymbol{\mu}}_j$ correlated with the symmetrized confusion mass of the pair at median $\rho = +0.574$, significant after Benjamini–Hochberg correction on seven of the eight and reaching $+0.630$ on PVEL-AD. Because 35–94% of these pairs carry exactly zero confusion, the statistic is better described as a strong detector of whether a pair confuses than as a ranker of how much, a decomposition Appendix D reports.

Scatter-aware statistics predict pooled recall as well as or better than GCS, yet the predictive content lives in the between-class term GCS measures. Raw within-class scatter (NC1) was a null pooled predictor ($\rho = +0.046$, $p = 0.74$) and became predictive only once normalized by between-class centroid distance (-0.904), while the Fisher ratio ($+0.897$) and Silhouette ($+0.857$) matched or exceeded pooled GCS (-0.785). Centroid–classifier self-duality (NC3) reproduced the specialized-versus-control split, corroborating the geometric reading. Routing by angle alone still carries a cost, since 6.8% to 8.2% of the 513 classes with at least 10 test images combine below-median GCS, above-median scatter and recall below 0.5, which GCS alone would misattribute to sample scarcity (Appendix C).

The magnitude of the frozen deviation separates two regimes of adaptation response (Table 1). Against linear probes on L2-normalized and on raw frozen features, rank-8 LoRA fine-tuning (Hu et al., 2021) at the same label budgets improved balanced accuracy on eight of the nine specialized datasets, by 1.8 to 20.4 points over the stronger of the two probes. The ninth, NEU-CLS, is saturated. On the two controls the same procedure underperformed the stronger probe, by 5.1 and 6.5 points, consistent with adaptation sacrificing pretrained structure where the encoder has already spread the classes. The two largest seed spreads, ± 0.063 on PVEL-AD and ± 0.093 on BenchmarkEL, are comparable to those datasets' own gains, so the regime separation rests on the sign pattern across all eleven datasets rather than on any single margin.

The boundary between the regimes is an ordering, not a threshold. Scoring frozen GCS as a binary predictor of whether LoRA helps, and counting saturated NEU-CLS as a dataset it does not help, the best in-sample accuracy is 10 of 11, attained by every cut in $[0.472, 0.704]$, and leave-one-dataset-out accuracy is 9 of 11. The two errors, EuroSAT (0.637, helped) and NEU-CLS (0.681, not), sit in inverted GCS order, which no refinement of the cut can separate. We therefore describe each regime by its observed range and make no prediction between them.

Table 1: **The cone forces essentially all of the observed deviation, and its magnitude separates two regimes.** Mean cone alignment $\bar{\gamma}$, frozen GCS, the floor $\mathrm{GCS}_{\min}$ of equation 2, probes on L2-normalized and raw frozen features, LoRA (mean $\pm$ std, three seeds), and Δ, the gap to the stronger probe before rounding. DINOv2 ViT-S/14 at the L4 budget, ordered by GCS, rule separating specialized datasets from controls.

Dataset	$\bar{\gamma}$	GCS	$\mathrm{GCS}_{\min}$	LP (norm.)	LP (raw)	LoRA BAcc	Δ
Infrared (binary)	0.996	0.986	0.986	0.748	0.793	0.826 ± 0.010	$+3.3$
THED-PV	0.989	0.977	0.972	0.503	0.614	0.817 ± 0.005	$+20.4$
ELPV	0.987	0.959	0.950	0.793	0.828	0.845 ± 0.006	$+1.8$
BenchmarkEL	0.980	0.953	0.940	0.688	0.697	0.781 ± 0.093	$+8.4$
Infrared (12-class)	0.972	0.948	0.939	0.414	0.408	0.534 ± 0.007	$+11.9$
HAM10000	0.963	0.938	0.916	0.592	0.557	0.705 ± 0.022	$+11.3$
PVEL-AD	0.841	0.727	0.683	0.705	0.628	0.753 ± 0.063	$+4.8$
NEU-CLS	0.856	0.681	0.679	0.994	1.000	0.995 ± 0.005	-0.5
EuroSAT	0.820	0.637	0.635	0.922	0.947	0.983 ± 0.001	$+3.6$
CIFAR-100-LT	0.558	0.307	0.304	0.579	0.533	0.528 ± 0.016	-5.1
Places-LT	0.290	0.106	0.082	0.392	0.293	0.327 ± 0.008	-6.5

An ordering this strong still establishes only correlation. If the geometry is implicated in the failure rather than merely co-occurring with it, then making the geometry an explicit training objective should move recall. Section 6 tests exactly that, with the identical-pipeline controls the claim requires.

6 From diagnostic to objective

The deviation statistic that Section 5 showed to predict recall converts directly into a differentiable auxiliary objective. During fine-tuning we penalize the squared deviation of every pairwise centroid cosine from the ETF target,

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{CE}} \;+\; \lambda \,\frac{2}{C(C-1)} \sum_{i<j} \big(\cos(\mathbf{c}_i, \mathbf{c}_j) - s^*\big)^2, \tag{3}$$

the per-pair analogue of the deviation Δ_{ETF} of Section 3. The centroids $\mathbf{c}_i$ are exponential moving averages of within-batch class means with decay $\alpha = 0.1$, which keeps the penalty defined for classes absent from a batch. We add the term during rank-8 LoRA fine-tuning at the budgets of Section 3, with $\lambda = 0.1$ and inverse-frequency-weighted cross-entropy (Cui et al., 2019) unless stated otherwise.

Against a control trained in an identical pipeline, the ETF term raised balanced accuracy on PVEL-AD by 3.3 points, and its gain on the most confused class overlapped with what re-weighting already provided. All four cells of Table 11 come from one training script, crossing the cross-entropy weighting with the presence of the ETF term. Under weighted cross-entropy the term moved balanced accuracy from 0.711 ± 0.066 to 0.744 ± 0.062 while moving Crack recall by only 0.8 points, and under unweighted cross-entropy it added 0.5 points of balanced accuracy but 2.8 points of Crack recall. The geometric term and inverse-frequency re-weighting therefore do overlapping work on the confused class, while the accuracy benefit is one re-weighting does not supply. Seed spreads (up to 0.066) are comparable to the 3.3-point gain, so we call the improvement modest rather than decisive.

The intervention widens the cone, which is the mechanism Proposition 1 requires. The bound leaves adaptation two handles: reduce the cone alignment γ, which lowers the floor, or rearrange centroids within the cone, which lowers the excess. Across eight datasets, vanilla LoRA reduced γ on eight of eight (mean -0.152), and its effect on GCS was dominated by the floor (mean -0.251) while the excess rose slightly ($+0.036$). Adding the ETF term reduced γ further on five of eight (mean -0.036), again predominantly through the floor (-0.028 against $+0.009$ of excess). Adaptation widens the cone rather than rearranging classes inside a fixed one, which is consistent with the mechanism identified by the theory (adapted cosine matrices in Appendix G).

The intervention transfers across pretraining objectives but not, on our evidence, across architecture families Appendix H). On CLIP ViT-B/16 the ETF term improved PVEL-AD balanced accuracy by 2.9 points and reduced GCS on all three datasets tested, with accuracy flat on THED-PV and Infrared (12-class). On ResNet-50 it moved balanced accuracy by at most 0.6 points, left GCS

nearly unchanged, and reduced Crack recall from 0.353 to 0.294. We report this as a scope limit: among the encoders tested the objective works on the ViT-family representations and not on the ImageNet-supervised CNN.

No alternative objective in our ablations matched the ETF term's combination of overall accuracy and confused-class recall, and the design choices are not delicate. The imbalance-aware SELI target (Thrampoulidis et al., 2022), a focal loss, a supervised contrastive term and a penalty restricted to the confused-cluster pairs each either traded balanced accuracy for confused-class recall or failed to improve it, LoRA ranks 2 through 16 were indistinguishable at three seeds, and at a fixed epoch count balanced accuracy is flat in the EMA decay α while GCS falls monotonically (Appendix G). The intervention therefore yields a modest, controlled gain in the geometry-failure regime, accompanied by geometric changes consistent with the theory and bounded by the tested range of encoders and weighting schemes.

7 Discussion and scope

Interpreted as a measure of representational compression rather than of inter-class crowding, GCS accounts for two regularities that would otherwise be unexplained. If the statistic reflects how tightly an encoder compresses a domain, encoders differing in architecture family, pretraining objective and corpus should order domains alike, as they do at pairwise Spearman 0.96 to 1.00, sharing the ordering but not the scale. This extends the convergence of independently trained models toward shared representations (Huh et al., 2024) from the geometry of success to the geometry of failure. It also accounts for post-hoc corrections that presuppose a collapsed geometry, since logit adjustment (Menon et al., 2020) underperformed a balanced probe on frozen features and largely recovered once the backbone was adapted (Appendix H). Because geometry did not quantitatively predict the magnitude of that degradation once class count was controlled, we read the result as delimiting the applicability of such methods rather than predicting their failure.

In practice the diagnostic requires one forward pass through the frozen encoder and roughly twenty labels per class, and returns three orderings: datasets by geometric risk, classes by expected recall, and class pairs by likelihood of confusion. The first informs whether to adapt the backbone, the second where to allocate annotation effort, and the third which confusions to audit, subject to the routing cost quantified in Section 5.

Three limits bound the scope of these claims. GCS is a within-dataset instrument, so its scale does not transport across datasets and the regime boundary is an ordering rather than a threshold, with the separating interval unidentified (Section 5). Adaptation results are evaluated over the three official splits, with balanced-accuracy spreads reaching 0.093. We therefore treat the +3.3-point balanced-accuracy gain as directional and the +0.8-point per-class gain as within seed noise, overlapping with what inverse-frequency re-weighting already supplies. The geometric account is established on the ViT family: γ is measured for those three encoders only, ResNet-50 enters through its GCS ordering alone, and the intervention does not meaningfully transfer to it (Section 6). The Neural Collapse panel is observational throughout, since we measured NC1 and NC3 but intervened only on centroid geometry, and ImageNet-LT and iNaturalist are left to future work.

Whether γ predicts the sample efficiency of adaptation, and whether a domain-level normalization can make the GCS scale transport, are the natural next questions.

8 Conclusion

Across the specialized datasets studied, uneven per-class recall tracks representation geometry more strongly than class frequency. Positive cone alignment imposes a nonzero floor on deviation from ETF geometry, and a median 99% of the measured deviation is accounted for by this floor. Dataset ordering by GCS is highly consistent across encoders, although the absolute scale remains encoder-dependent. An ETF-alignment objective produces a modest balanced-accuracy gain on ViT representations while reducing cone alignment. These results position geometry as a practical diagnostic for identifying when frozen representations may benefit from adaptation.

AI USE STATEMENT

Generative AI tools were used for proofreading and language refinement, assistance with schematic figure creation, and limited support in organizing the paper structure. All scientific content, methodology, experiments, results, mathematical claims, and conclusions were developed and verified by the authors. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work studies representation geometry using existing publicly released benchmark datasets and involves no new human-subject data collection. HAM10000 is used only as a benchmark for representation analysis, and we make no clinical or diagnostic claims. The intended use of GCS is to diagnose representation quality and guide model adaptation, not to support high-stakes decisions without task-specific validation.

REPRODUCIBILITY STATEMENT

All experiments use fixed dataset splits, matched label budgets, and the evaluation protocols described in Sections 3–6. Linear probes, LoRA adaptation, the GCS computation, and the ETF-alignment objective are fully specified in the main text, with dataset composition, hyperparameters, ablations, seed-level variation, and additional diagnostics reported in the appendices. Adaptation results are evaluated over the three official splits, and all reported comparisons use the same preprocessing and training pipelines within each experiment.

REFERENCES

Mohammed Yaqoob Mohammed Shoeb Ansari. THED-PV: A high-resolution multi-perspective thermal imaging dataset for photovoltaic homography estimation, 2025. URL `https://doi.org/10.5281/zenodo.17404247`.

Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. In *International Conference on Learning Representations*, volume 2024, pp. 1182–1211, 2024.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Haiyong Chen, Yue Pang, Qidi Hu, and Kun Liu. Solar cell surface defect inspection based on multispectral convolutional neural network. *Journal of Intelligent Manufacturing*, 31(2):453–468, 2020.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 695–704. IEEE, 2021.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9260–9269. IEEE, 2019.

Sergiu Deitsch, Vincent Christlein, Stephan Berger, Claudia Buerhop-Lutz, Andreas Maier, Florian Gallwitz, and Christian Riess. Automatic classification of defective photovoltaic module cells in electroluminescence images. *Solar Energy*, 185:455–468, 2019.

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

Naoya Hasegawa and Issei Sato. Multiplicative logit adjustment approximates neural-collapse-aware decision boundary adjustment. In *International Conference on Learning Representations*, volume 2025, pp. 97630–97662, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Chiraag Kaushik, Ran Liu, Chi-Heng Lin, Amrit Khera, Matthew Y Jin, Wenrui Ma, Vidya Muthukumar, and Eva L Dyer. Balanced data, imbalanced spectra: Unveiling class disparities with spectral imbalance. *arXiv preprint arXiv:2402.11742*, 2024.

Vignesh Kothapalli and Tom Tirer. Can kernel methods explain how the data affects neural collapse? *arXiv preprint arXiv:2406.02105*, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2532–2541. IEEE, 2019.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Matthew Millendorf, Edward Obropta, and Nikhil Vadhavkar. Infrared solar module dataset for anomaly detection. In *Proc. Int. Conf. Learn. Represent*, volume 3, 2020.

Michael Munn, Benoit Dherin, and Javier Gonzalvo. The impact of geometric complexity on neural collapse in transfer learning. *Advances in Neural Information Processing Systems*, 37:66713–66742, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Vardan Papyan, Xiao 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.

Lawrence Pratt, Jana Mattheus, and Richard Klein. A benchmark dataset for defect detection and classification in electroluminescence images of pv modules using semantic segmentation. *Systems and Soft Computing*, 5:200048, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Bryan Rodas, Natalie Montesino, Jakob Ambsdorf, David Klindt, and Randall Balestriero. Diet-cp: Lightweight and data efficient self supervised continued pretraining. *arXiv preprint arXiv:2509.06990*, 2025.

Kechen Song and Yunhui Yan. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 285:858–864, 2013.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in neural information processing systems*, 35:37991–38002, 2022.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv preprint arXiv:2302.03004*, 2023.

Beier Zhu, Kaihua Tang, Qianru Sun, and Hanwang Zhang. Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models. *Advances in Neural Information Processing Systems*, 36:64663–64680, 2023.

A Dataset composition and test support

The benchmark spans six application domains and four orders of magnitude in imbalance ratio, with two long-tailed natural-image controls chosen deliberately for their proximity to web-scale pretraining data. Table 2 lists the composition. Three electroluminescence datasets cover solar-cell defect inspection (Chen et al., 2020; Pratt et al., 2023; Deitsch et al., 2019), three infrared datasets cover photovoltaic thermal anomalies (Millendorf et al., 2020; Ansari, 2025), and the remaining specialized sets cover steel-surface defects (Song & Yan, 2013), dermoscopy, and satellite land use. CIFAR-100-LT (Krizhevsky et al., 2009; Cui et al., 2019) and Places-LT (Liu et al., 2019) are the controls. Class counts, image totals, and imbalance ratios follow the dataset registry used by the extraction pipeline, and the budget column gives the matched label budget (L4) at which every probe and every LoRA run in the paper was trained.

Table 2: The benchmark spans six application domains and four orders of magnitude in imbalance, and the two datasets below the rule are in-distribution controls. Modality abbreviations: EL electroluminescence, IR infrared. Budget is the matched label budget (L4) shared by linear probes and LoRA fine-tuning.

Dataset	Modality	Classes	Images	Imbalance	Budget	Description
PVEL-AD	EL	13	17,393	5,677:1	2,000	Solar-cell EL defect detection (Chen et al., 2020)
BenchmarkEL	EL	3	2,354	7.6:1	432	Solar-cell EL criticality triage (Pratt et al., 2023)
ELPV	EL	2	2,624	1:1	800	Cell-level defect grading from EL (Deitsch et al., 2019)
Infrared (binary)	IR	2	20,000	1:1	6,976	PV-module thermal anomaly detection (Millendorf et al., 2020)
Infrared (12-class)	IR	12	20,000	57:1	4,000	PV thermal anomaly categorization (Millendorf et al., 2020)
THED-PV	IR	5	12,052	1.1:1	4,000	Thermal PV imagery, five acquisition days (Ansari, 2025)
NEU-CLS	Visible	6	1,800	1:1	400	Hot-rolled steel surface defects (Song & Yan, 2013)
HAM10000	Dermoscopy	7	10,015	58:1	2,000	Dermoscopic skin-lesion diagnosis
EuroSAT	Satellite	10	27,000	1.5:1	8,000	Sentinel-2 land-use classification
CIFAR-100-LT	Visible	100	20,847	100:1	1,000	Long-tailed natural-image control (Krizhevsky et al., 2009; Cui et al., 2019)
Places-LT	Visible	365	69,800	996:1	20,000	Long-tailed scene-recognition control (Liu et al., 2019)

Test support on PVEL-AD concentrates in a minority of classes, and five of the 13 classes have fewer than 10 test images. Table 3 reports the per-class composition of the 2,000-label training draw and of the held-out test split of 1,748 images. The five classes in boldface have between 0 and 8 test images each, and Corner has none, so its recall is undefined. We report recalls for these five classes for completeness only, and no claim in the paper rests on them. The headline pooled correlation is robust to this support structure, since restricting to classes with at least 10 test images leaves $\rho = -0.800$ over 50 classes (Section 5).

Table 3: Five of the 13 PVEL-AD classes (boldface) have fewer than 10 test images and Corner has none, so their recalls are reported for completeness only. Per-class composition of the L4 training draw and the test split.

Class	Train (L4)	Test
Normal	1,308	1,136
Finger	126	111
Crack	96	85
BlackCore	115	100
ThickLine	78	69
HorizDislocation	30	28
ShortCircuit	56	50
Scratch	177	155
VertDislocation	3	4
StarCrack	8	8
PrintError	1	1
Corner	1	0
Fragment	1	1
Total	2,000	1,748

B PROOF DETAILS AND NUMERICAL VERIFICATION OF THE FEASIBILITY FLOOR

Equality in Proposition 1 holds exactly when the tangential components of the centroids cancel. Writing each unit centroid as $\hat{\boldsymbol{\mu}}_i = \gamma_i \boldsymbol{m} + \boldsymbol{v}_i$ with $\boldsymbol{v}_i \perp \boldsymbol{m}$ and $\|\boldsymbol{v}_i\| = s_i = \sqrt{1-\gamma_i^2}$, the pairwise sum obeys the identity

$$2\sum_{i<j} \boldsymbol{v}_i^\top \boldsymbol{v}_j \;=\; \Big\|\sum_i \boldsymbol{v}_i\Big\|^2 \;-\; \sum_i s_i^2,$$

and the proof of Proposition 1 discards only the term $\|\sum_i \boldsymbol{v}_i\|^2 \geq 0$. The bound in equation 2 is therefore attained if and only if $\sum_i \boldsymbol{v}_i = \mathbf{0}$.

The floor is attained, not merely approached, whenever $d-1 \geq k-1$. The tangential components live in the $(d-1)$-dimensional orthogonal complement of $\boldsymbol{m}$. In the common-alignment case $\gamma_i = \gamma$, choosing the unit directions $\boldsymbol{v}_i/s$ as the k vertices of a simplex equiangular tight frame inside that complement gives $\sum_i \boldsymbol{v}_i = 0$ and pairwise cosines $\gamma^2 - (1-\gamma^2)/(k-1)$ for every pair, which is exactly $\mathrm{GCS}_{\min}$. That construction needs k unit vectors summing to zero, which is where $d-1 \geq k-1$ comes from. Setting $\gamma = 0$ recovers the ETF itself and $\mathrm{GCS}_{\min} = s^* = -1/(k-1)$, so the bound degrades gracefully to the isotropic reference.

The bound is informative only because the cone axis m is estimated without labels, as the leading eigenvector of the uncentered second-moment matrix of the L2-normalized features. If m were instead derived from the class centroids themselves, for instance as their mean direction, the tangential components would satisfy $\sum_c \boldsymbol{v}_c = 0$ by construction, the inequality would hold with equality identically, and the floor would merely restate the observed GCS.

We verified Proposition 1 numerically before measuring it on data. Over $C \in \{3, 6, 13, 30\}$ crossed with heterogeneous alignment profiles $\{\gamma_i\}$, randomly drawn unit-centroid configurations constrained to the prescribed alignments produced zero violations of the general bound in equation 2. In the homogeneous case the bound was attained to numerical precision, and at $\gamma = 0$ the attained value equals s^* exactly. On the eleven datasets the measured bound holds in all eleven cases (Table 1).

Scalar background subtraction is not a viable alternative to the floor correction. Table 4 compares the pooled 55-class GCS–recall correlation under three treatments of anisotropy. The raw statistic is the reference. The centered treatment subtracts from each per-class GCS its dataset's background similarity, the mean cosine of random feature pairs, a scalar shift that leaves within-dataset orderings unchanged while recalibrating datasets against one another. The whitened treatment recomputes GCS after removing the dataset mean direction from every feature, which is the standard isotropy fix. Centering collapses the pooled correlation from -0.785 to -0.309 on ViT-S/14 and to a non-significant -0.069 on CLIP, and whitening removes the ambient cone almost entirely yet destroys most of the signal with it (-0.357/-0.305/-0.393). The predictive content of GCS therefore lives largely in the cone term itself. That is why the principled correction is the feasibility floor of Propo-

sition 1, and why, as Section 4 reports, the excess over that floor carries no predictive signal of its own.

Table 4: Background subtraction destroys the pooled GCS–recall correlation, leaving the floor correction of Proposition 1 as the principled alternative. Spearman ρ between per-class GCS and per-class recall, pooled over the 55-class pool, for three treatments of anisotropy. Centered subtracts each dataset's mean random-pair cosine from its per-class GCS, and whitened removes the dataset mean direction from every feature before recomputing GCS. All entries $p < 0.05$ except where noted.

Treatment	DINOv2 ViT-S/14	DINOv2 ViT-B/14	CLIP ViT-B/16
Raw GCS	-0.785	-0.806	-0.762
Centered	-0.309	-0.461	-0.069 (n.s., $p = 0.62$)
Whitened	-0.357	-0.305	-0.393

C THE FULL NEURAL COLLAPSE PANEL

Angle statistics and scatter statistics divide the predictive labor, and raw within-class scatter alone predicts nothing. Table 5 reports the pooled Spearman correlation between six per-class statistics and per-class recall over the 55-class pool for the three ViT-family encoders. Raw NC1 within-class scatter is a null predictor on every encoder, at $+0.046$, -0.078, and $+0.129$ with all $p > 0.3$. It becomes strongly predictive only when normalized by between-class distance (-0.904 on ViT-S/14), so the predictive content of the ratio is carried by its between-class term. That result is a principled defense of measuring the between-class geometry (NC2), which is what GCS does. We reiterate that the scatter-aware indices predict pooled recall as well as or better than GCS (Section 5), and that GCS's case rests on pair resolution and differentiability rather than pooled strength.

Table 5: Raw within-class scatter is a null pooled predictor of recall on all three encoders and becomes predictive only when normalized by between-class distance. Pooled Spearman ρ with per-class recall over the 55-class pool. All entries $p < 0.05$ except where a p-value is shown.

Per-class statistic	ViT-S/14	ViT-B/14	CLIP ViT-B/16
GCS (NC2; ours)	-0.785	-0.806	-0.762
NC1 within-class scatter (raw)	$+0.046$ (p=0.74)	-0.078 (p=0.57)	$+0.129$ (p=0.35)
NC1 scatter / between-class distance	-0.904	-0.894	-0.754
Centroid norm	$+0.195$ (p=0.15)	$+0.315$	-0.274
Silhouette	$+0.857$	$+0.728$	$+0.811$
Fisher discriminant ratio	$+0.897$	$+0.896$	$+0.744$

NC3 self-duality replicates the specialized-versus-control split that NC2 shows. Table 6 reports, for the nine multiclass datasets in the panel (the two binary datasets are excluded, since a two-class GCS is a single pair), the mean cosine between the balanced probe's row-normalized weights and the L2-normalized class centroids on frozen ViT-S/14 features. Weight–centroid alignment is low on every specialized domain (0.087 to 0.496) and high on both controls (0.653 and 0.825). Where the cone crowds the centroids, the discriminative directions the probe learns cannot align with them, so NC3 fails on exactly the domains where NC2 does.

The NC1 trace ratio does not split the two groups, and we do not claim it does. The global ratio of within-class to between-class scatter spans 0.598 to 62.0 across the specialized datasets, straddling the controls' values of 2.40 and 3.03. PVEL-AD (0.598) and NEU-CLS (1.48) fall below both controls. Among the global NC statistics we measured, only NC3 reproduces the domain split cleanly.

A joint diagnostic quantifies what scatter adds to the angle statistic. A class is counted as misrouted when its GCS sits below its dataset median (so the angle statistic reads it as unproblematic) while its recall falls below 0.5 and its within-class scatter sits above the dataset median. That combination is the signature of a failure GCS alone would assign to sample scarcity when the scatter shows overlap. GCS-alone routing misroutes $35/513 = 6.8\%$ of classes on ViT-B/14 and $42/513 = 8.2\%$ on CLIP across the nine panel datasets, and adding scatter to the angle statistic raises the pooled rank-R^2 for

Table 6: NC3 self-duality splits specialized domains from controls cleanly (all specialized ≤ 0.496, both controls ≥ 0.653), whereas the NC1 trace ratio does not, since its specialized values straddle the controls. Global statistics on frozen DINOv2 ViT-S/14 features for the nine multiclass panel datasets.

Dataset	Frozen GCS	NC3 weight–centroid cosine	NC1 trace ratio
THED-PV	0.977	0.087	62.0
BenchmarkEL	0.953	0.146	14.5
Infrared (12-class)	0.948	0.167	6.18
HAM10000	0.938	0.183	6.78
PVEL-AD	0.727	0.406	0.598
NEU-CLS	0.681	0.496	1.48
EuroSAT	0.637	0.475	2.22
CIFAR-100-LT	0.307	0.653	2.40
Places-LT	0.106	0.825	3.03

recall from 0.690 to 0.810 on ViT-B/14 (permutation $p < 2 \times 10^{-4}$) and from 0.740 to 0.842 on CLIP. The joint instrument is the natural successor to GCS alone (Section 7).

D Pair-level resolution, decomposed

Confusion matrices on these benchmarks are heavily zero-inflated, with 35 to 94% of supported pairs (both classes having at least 10 test images) carrying exactly zero confusion mass. A single rank correlation therefore mixes two questions, and we report them separately. As a detector of whether a pair confuses at all, the pairwise cosine $\hat{\boldsymbol{\mu}}_i^\top \hat{\boldsymbol{\mu}}_j$ reached median AUC 0.900 on DINOv2 ViT-B/14 and 0.855 on CLIP ViT-B/16. As a ranker of magnitude, restricted to the non-zero pairs alone, it reached median $\rho = +0.499$ and $+0.377$ on the same two encoders. The one dataset not significant after Benjamini–Hochberg correction is THED-PV, whose five classes supply only 10 pairs ($\rho = +0.43$).

Cone alignment against GCS at the class level. At the class level GCS is not reducible to cone alignment. Per-class γ_c and GCS predicted recall about equally well (-0.809 against -0.785 on ViT-S/14) and each retained independent signal when the other was partialled out, so the two divide the labor rather than substituting for one another: cone alignment carries the feasibility theorem and the domain-structural explanation, GCS the pair resolution and the objective of Section 6.

E Near-duplicate audit

Near-duplicate contamination is confined to two datasets. Only PVEL-AD (8.75%) and EuroSAT (3.11%) exceed a 1% flag rate, and the remaining seven panel datasets sit at or below 0.75%. The audit behind these rates is a nearest-neighbor test. For every test image we computed the cosine similarity to its nearest training neighbor in frozen ViT-S/14 feature space and flagged it as duplicate-adjacent above a threshold of 0.98. A looser threshold of 0.95 is uninformative on this feature space, because the ambient cone places most same-domain images above that value, so 0.95 flags 83.9% of PVEL-AD test images. That behavior is itself an illustration of the anisotropy of Section 4.

Duplicate contamination does not drive the headline correlation, and removing duplicate-heavy classes strengthens it. Table 7 reports the pooled GCS–recall correlation under successively stricter exclusions. Dropping classes with more than 10% duplicate-adjacent test images moves ρ from -0.785 to -0.802 ($n = 50$), and combining that filter with the test-support filter gives -0.798 ($n = 46$). Per-class duplicate load predicts neither recall ($\rho = -0.125$, $p = 0.36$) nor GCS ($\rho = -0.090$, $p = 0.52$), so the diagnostic is not a disguised duplication meter and the correlation is not test-set leakage.

Table 7: Removing duplicate-heavy classes slightly strengthens the pooled GCS–recall correlation, so the correlation is not driven by test-set leakage. Pooled Spearman ρ (frozen ViT-S/14, 55-class pool) under successive exclusions. The duplicate fraction is the share of a class's test images with nearest-training-neighbor cosine above 0.98.

Class filter	n classes	ρ
None (all classes)	55	-0.785
$n_{\text{test}} \geq 10$	50	-0.800
Duplicate fraction $\leq 50\%$	54	-0.784
Duplicate fraction $\leq 25\%$	53	-0.780
Duplicate fraction $\leq 10\%$	50	-0.802
Duplicate fraction $\leq 10\%$ and $n_{\text{test}} \geq 10$	46	-0.798

F LABELS REQUIRED TO ESTIMATE GCS

Estimating GCS reliably requires roughly twenty labeled examples per class, up to fifty on the hardest dataset. For each dataset we subsampled n labeled examples per class over the grid $n \in \{2, 3, 5, 10, 20, 50, 100\}$, recomputed the dataset GCS on 500 bootstrap draws per grid point, and recorded the width of the resulting 95% confidence interval. A CI width of at most 0.05 requires a median of $n = 20$ per class across datasets, with the worst dataset requiring 50. Tightening to a width of ≤ 0.02 requires 50 per class and 100 in the worst case, and ≤ 0.01 requires 75 and 100. The practical reading used throughout the paper is that roughly twenty labels per class make the diagnostic available before any fine-tuning. The diagnostic is inexpensive, but it is not free of labels, and we list this cost as a limitation in Section 7.

G HYPERPARAMETER SWEEPS AND ALTERNATIVE OBJECTIVES

Only the ETF target improved balanced accuracy and Crack recall together against the identical-pipeline weighted-CE control of Section 6, which every comparison below uses as the uniform baseline. Table 8 compares training objectives on PVEL-AD under rank-8 LoRA at the L4 budget. Against the control (BAcc 0.711 ± 0.066, Crack 0.651 ± 0.058), the ETF target gained $+3.3$ points of balanced accuracy and $+0.8$ of Crack recall. SELI alignment (Thrampoulidis et al., 2022) traded 4.0 points of balanced accuracy for $+2.7$ of Crack, and a restricted cluster-pair penalty matched the ETF term's Crack recall at a 0.8-point accuracy cost. Focal loss and supervised contrastive training did not improve Crack recall at all (0.522 and 0.545 against 0.651). The per-class deltas here sit within seed spreads of comparable size, so we read the table as ruling alternatives out rather than certifying the ETF term's margin. Figure 3 resolves the ETF row's GCS reduction to pair level, where the pairwise centroid cosines of PVEL-AD fall across the whole class structure and not only on the flagged pairs.

Table 8: Against the identical-pipeline weighted-CE control, only the ETF target improves balanced accuracy and Crack recall together. PVEL-AD, rank-8 LoRA on DINOv2 ViT-S/14 at the L4 budget, mean $\pm$ std over three seeds. GCS is measured on the adapted features (frozen reference 0.727).

Objective	BAcc	Crack recall	GCS
Weighted CE, no auxiliary (control)	0.711 ± 0.066	0.651 ± 0.058	0.497 ± 0.058
ETF alignment, $\lambda = 0.1$	0.744 ± 0.062	0.659 ± 0.033	0.441 ± 0.032
Cluster-pair penalty, $\lambda = 0.1$	0.703 ± 0.071	0.659 ± 0.048	0.474 ± 0.057
SELI alignment, $\lambda = 0.1$	0.671 ± 0.023	0.678 ± 0.024	0.536 ± 0.029
Focal loss	0.725 ± 0.070	0.522 ± 0.131	0.579 ± 0.071
Supervised contrastive	0.749 ± 0.036	0.545 ± 0.075	0.507 ± 0.048

The EMA coefficient α does not matter for accuracy once training length is controlled, and an earlier sweep that suggested otherwise was an early-stopping artifact. In the original sweep the canonical $\alpha = 0.1$ run had trained for 46 epochs while every alternative stopped at 19, so the apparent peak at $\alpha = 0.1$ measured training length, not α. Table 9 reports the corrected sweep, in which every cell trains for exactly 50 epochs. Balanced accuracy is flat within seed noise (spread 3.4 points against per-cell standard deviations up to 6.7), $\alpha = 0.1$ is not optimal, and only GCS responds

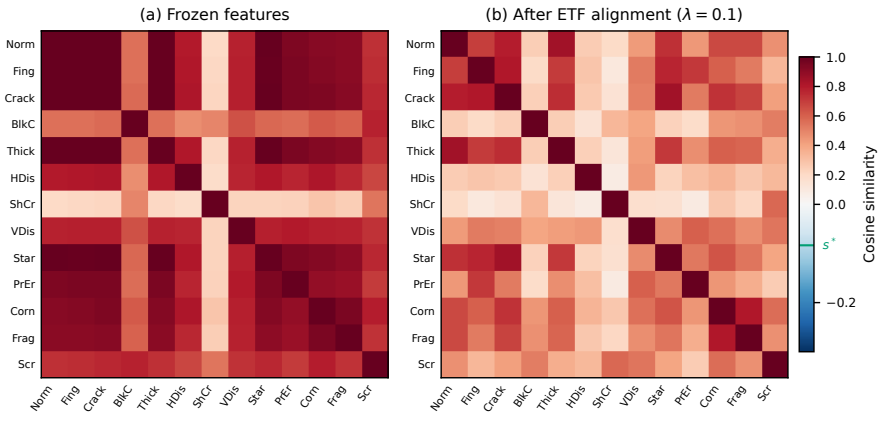


Figure 3: ETF-aligned adaptation reduces pairwise centroid cosines across the class structure of PVEL-AD, not only on the flagged pairs. The mean pairwise cosine falls from 0.727 on frozen ViT-S/14 features to 0.441 after ETF-LoRA at $\lambda = 0.1$. Pairwise centroid-cosine matrices before (left) and after (right) adaptation.

systematically, falling monotonically from 0.459 to 0.363 as α grows. The geometric effect of the loss is real at every α, although the earlier accuracy story was not.

Table 9: At matched training length (50 epochs for every cell), balanced accuracy is flat in the EMA coefficient α while GCS falls monotonically. PVEL-AD, ETF alignment at $\lambda = 0.1$, rank 8, mean $\pm$ std over three seeds. This table supersedes the early-stopped sweep, whose $\alpha = 0.1$ cell had trained 46 epochs against 19 for all others.

α	BAcc	Crack recall	GCS
0.01	0.758 ± 0.052	0.643 ± 0.058	0.459 ± 0.049
0.05	0.777 ± 0.026	0.651 ± 0.049	0.438 ± 0.034
0.1	0.748 ± 0.067	0.647 ± 0.054	0.428 ± 0.039
0.3	0.744 ± 0.060	0.631 ± 0.006	0.392 ± 0.042
0.5	0.743 ± 0.062	0.643 ± 0.011	0.363 ± 0.043

No LoRA rank between 2 and 16 is distinguishable at three seeds. In the sweep of Table 10 on PVEL-AD at $\lambda = 0.1$, balanced accuracy spans 0.744 to 0.767 and Crack recall 0.600 to 0.659, with per-configuration standard deviations up to 0.062 and 0.058 respectively, so the canonical rank 8 is a convention rather than an optimum.

Table 10: No LoRA rank between 2 and 16 is distinguishable from rank 8 at three seeds. PVEL-AD, ETF alignment at $\lambda = 0.1$, EMA $\alpha = 0.1$, mean $\pm$ std over three seeds.

Rank	BAcc	Crack recall	GCS
2	0.767 ± 0.009	0.600 ± 0.058	0.484 ± 0.051
4	0.745 ± 0.044	0.616 ± 0.011	0.436 ± 0.035
8	0.744 ± 0.062	0.659 ± 0.033	0.441 ± 0.032
16	0.760 ± 0.053	0.612 ± 0.058	0.449 ± 0.023

Two caveats bound the remaining sweep evidence. First, an earlier λ sweep of these runs carries the same early-stopping confound as the original α sweep (its $\lambda = 1.0$ rows stopped at 19 epochs), so we do not reproduce it here. The THED-PV $\lambda = 1.0$ result quoted in Section 6 survives because that ETF run early-stopped *before* its same-script control, which makes the reported $+0.9$-point gain conservative. Second, the stored SELI $\lambda = 0.01$ and ETF $\alpha = 0.01$ runs coincide exactly. Both are near-vanilla configurations whose auxiliary gradients are numerically insignificant, and two runs with identical seeds on identical hardware follow the same optimization path. Reproducibility reruns on mixed hardware break the tie and stay within seed variance (0.708 ± 0.063 fresh against 0.670 ± 0.023 stored), so the coincidence is benign.

Table 11: **The ETF term's accuracy gain survives an identical-pipeline control, and its confused-class gain overlaps with re-weighting.** All four adapted cells come from one script (PVEL-AD, ViT-S/14, rank-8 LoRA, $\lambda = 0.1$, 2,000 labels, mean $\pm$ std over three seeds). GCS is recomputed on the adapted features, and the frozen probe is shown for scale.

CE weighting	ETF term	BAcc	Crack recall	GCS
inverse-frequency	no	0.711 ± 0.066	0.651 ± 0.058	0.497 ± 0.058
inverse-frequency	yes	0.744 ± 0.062	0.659 ± 0.033	0.441 ± 0.032
uniform	no	0.692 ± 0.018	0.533 ± 0.077	0.495 ± 0.016
uniform	yes	0.698 ± 0.020	0.561 ± 0.093	0.464 ± 0.035
frozen probe (reference)		0.705	0.259	0.727

Behavior at a binding penalty weight. At a penalty weight large enough to bind, the term drives the adapted geometry most of the way to the ETF target. At $\lambda = 1.0$ on THED-PV it reduced adapted GCS from 0.800 to 0.119 while improving balanced accuracy from 0.817 ± 0.005 to 0.826 ± 0.003 over the same-script control, a comparison made conservative by the ETF run early-stopping sooner. At $\lambda = 0.1$ the effect was null ($+0.1$ to $+0.2$ points), since a geometry starting at 0.800 needs a heavier penalty to bind.

H POST-HOC BASELINES AND BACKBONE INVARIANCE

All reported logit-adjustment results use the standard correction of Menon et al. (2020), with τ swept uniformly over $\{0, 0.25, 0.5, 0.75, 1, 1.5, 2\}$ and the best value selected consistently for each dataset on both frozen and LoRA-adapted features.

Post-hoc calibration assumes a geometry that frozen specialized features do not have, and it recovers once the backbone is adapted. Table 12 compares corrected logit adjustment at its best τ against the balanced linear probe. On frozen features the correction damages balanced accuracy on the geometry-failure datasets, taking PVEL-AD to 0.572 from 0.705 and HAM10000 to 0.526 from 0.592, because the prior-based shift repairs a frequency-shaped failure that these datasets do not have. On LoRA-adapted features the damage largely vanishes on the specialized data (HAM10000 $-0.066 \rightarrow -0.003$, EuroSAT $-0.002 \rightarrow 0.000$, NEU-CLS $+0.001 \rightarrow 0.000$) and persists on the many-class controls (CIFAR-100-LT $-0.224 \rightarrow -0.221$, Places-LT $-0.150 \rightarrow -0.154$), where the failure is frequency-shaped but the method still underperforms a balanced probe at this label budget.

Geometry does not quantitatively predict the size of the post-hoc damage once class count is controlled, and we make no failure-prediction claim. Class count and ETF deviation are strongly confounded across the panel ($\rho(C, \Delta_{\mathrm{ETF}}) = -0.883$, $n = 9$ datasets), and the partial correlation of GCS with damage given class count is null ($\rho = -0.128$, $p = 0.74$, $n = 9$). The defensible statement is categorical rather than quantitative. The assumptions of prior-based post-hoc correction are absent on frozen specialized features and reappear after adaptation.

Table 12: Corrected logit adjustment damages balanced accuracy on most frozen features, the damage largely vanishes on specialized data once the backbone is adapted, and it persists on the many-class controls. Balanced accuracy of the balanced linear probe (LP) and of corrected logit adjustment at its best τ (LA), with their difference Δ, on frozen and LoRA-adapted DINOv2 ViT-S/14 features over three seeds. The PVEL-AD adapted cell was not run.

	Frozen			LoRA-adapted		
Dataset	LP	LA	Δ	LP	LA	Δ
PVEL-AD	0.705	0.572	-0.133	—	—	—
BenchmarkEL	0.688	0.609	-0.079	0.787	0.757	-0.030
Infrared (12-class)	0.414	0.355	-0.059	0.543	0.522	-0.022
THED-PV	0.503	0.503	0.000	0.797	0.797	0.000
NEU-CLS	0.994	0.995	$+0.001$	0.998	0.998	0.000
HAM10000	0.592	0.526	-0.066	0.735	0.733	-0.003
EuroSAT	0.922	0.921	-0.002	0.980	0.980	0.000
CIFAR-100-LT	0.579	0.355	-0.224	0.625	0.405	-0.221
Places-LT	0.392	0.242	-0.150	0.401	0.246	-0.154

An ImageNet-supervised ResNet-50 (He et al., 2016) orders the eleven datasets by frozen GCS nearly identically to the three ViT-family encoders, which is the architecture-family test of the

Table 13: **The intervention transfers to CLIP and does not meaningfully work on ResNet-50.** Each row pairs the no-auxiliary control with the ETF run from the same script, at rank-8 LoRA, $\lambda = 0.1$, and matched budgets, with mean $\pm$ std over three seeds and ΔBAcc in points computed before rounding. Crack recall is reported for PVEL-AD, the dataset with the diagnosed confused class, where it rose on CLIP ($0.525 \rightarrow 0.553$) and fell on ResNet-50 ($0.353 \rightarrow 0.294$).

Backbone	Dataset	BAcc (control)	BAcc (ETF)	ΔBAcc	ΔGCS
CLIP ViT-B/16	PVEL-AD	0.743 ± 0.027	0.771 ± 0.036	$+2.9$	-0.051
CLIP ViT-B/16	THED-PV	0.810 ± 0.005	0.809 ± 0.007	-0.2	-0.041
CLIP ViT-B/16	Infrared (12-class)	0.537 ± 0.013	0.540 ± 0.014	$+0.3$	-0.009
ResNet-50	PVEL-AD	0.601 ± 0.039	0.607 ± 0.041	$+0.6$	-0.010
ResNet-50	THED-PV	0.688 ± 0.004	0.691 ± 0.003	$+0.3$	-0.011
ResNet-50	Infrared (12-class)	0.452 ± 0.009	0.454 ± 0.009	$+0.2$	$+0.003$

domain-structural claim. Table 14 lists its per-dataset values next to the ViT-S/14 reference. The Spearman correlations of its ordering against ViT-S/14, ViT-B/14 (Oquab et al., 2023), and CLIP ViT-B/16 (Radford et al., 2021) are 0.964, 0.982, and 1.000, and the three ViT pairs agree at 0.991, 0.964, and 0.982 (Figure 4). Absolute scale remains encoder-dependent, since ResNet-50 reads CIFAR-100-LT at 0.596 against 0.307 for ViT-S/14 and both CLIP and ResNet-50 shift the controls upward. The invariance claim is therefore about ordering only, never about transportable values.

Table 14: An ImageNet-supervised ResNet-50 reproduces the ViT encoders' ordering of the datasets by frozen GCS (Spearman 0.964–1.000 against the three ViTs) while its absolute scale differs, most visibly on the controls. Frozen GCS per dataset, with rows ordered by the ViT-S/14 value.

Dataset	DINOv2 ViT-S/14	ResNet-50
Infrared (binary)	0.986	0.994
THED-PV	0.977	0.989
ELPV	0.959	0.968
BenchmarkEL	0.953	0.965
Infrared (12-class)	0.948	0.977
HAM10000	0.938	0.929
PVEL-AD	0.727	0.668
NEU-CLS	0.681	0.735
EuroSAT	0.637	0.617
CIFAR-100-LT	0.307	0.596
Places-LT	0.106	0.487

Cone alignment is at least as backbone-invariant as GCS, consistent with its role as the more fundamental quantity in Proposition 1. Table 15 reports the mean cone alignment $\bar{\gamma}$ per dataset for the three ViT-family encoders, whose pairwise Spearman correlations over the γ orderings are 1.000 (S/14 vs. B/14), 0.982 (S/14 vs. CLIP), and 0.982 (B/14 vs. CLIP). We measured γ for the ViT-family encoders only. No ResNet-50 cone alignment was measured, so ResNet-50 enters the invariance evidence through its GCS ordering alone.

I Compute

The retained experimental program consumed approximately 46 GPU-hours, with a median run time of roughly 6 minutes, and the revision program added 86 GPU tasks totaling approximately 12 GPU-hours (estimate). All runs executed on a SLURM cluster on NVIDIA L40S, A100, and H100 GPUs. The evidence-base loader underlying every v2 analysis was verified against all 33 of its 33 stored artifacts before any result in this paper was computed.

Table 15: Cone alignment orders the datasets almost identically across the three ViT-family encoders (pairwise Spearman 1.000 / 0.982 / 0.982). Mean cone alignment $\bar{\gamma}$ of the L2-normalized class centroids with the label-free cone axis, per dataset. γ was not measured for ResNet-50.

Dataset	ViT-S/14	ViT-B/14	CLIP ViT-B/16
Infrared (binary)	0.996	0.995	0.999
THED-PV	0.989	0.989	0.998
ELPV	0.987	0.988	0.996
BenchmarkEL	0.980	0.973	0.992
Infrared (12-class)	0.972	0.973	0.995
HAM10000	0.963	0.952	0.986
NEU-CLS	0.856	0.839	0.950
PVEL-AD	0.841	0.825	0.932
EuroSAT	0.820	0.818	0.935
CIFAR-100-LT	0.558	0.440	0.909
Places-LT	0.290	0.249	0.780

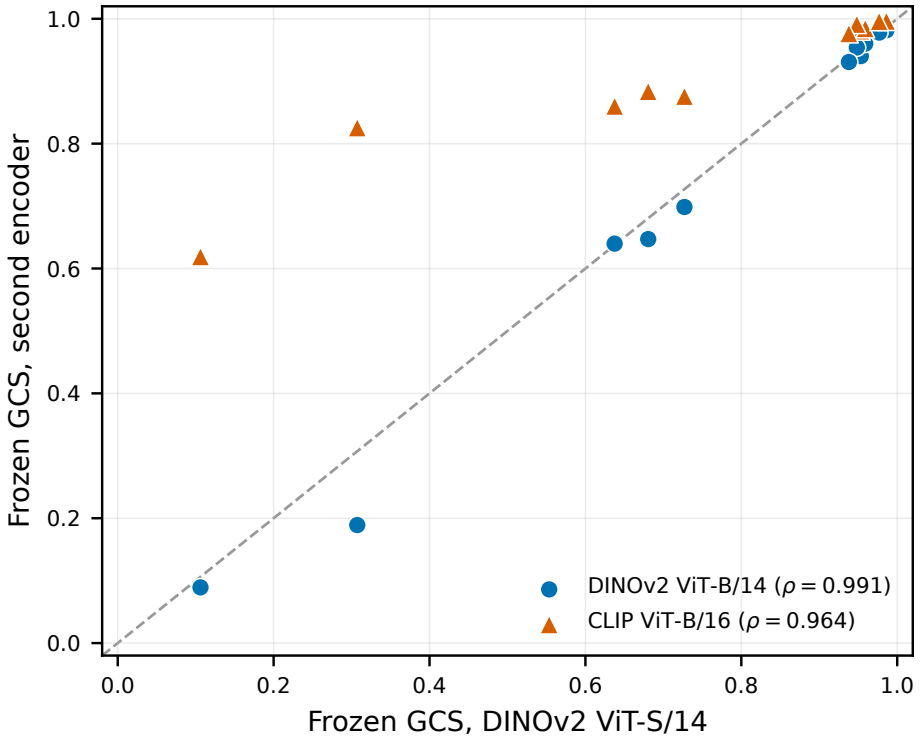


Figure 4: The three ViT-family encoders order the eleven datasets by frozen GCS nearly identically (pairwise Spearman 0.991 / 0.964 / 0.982), showing that the dataset ordering is strongly domain-structured across encoders even though absolute GCS values remain encoder-dependent. ResNet-50 appears in Table 14.