WHEN ARE CLASSES CONES? SCALE INVARIANCE, LINEAR SUFFICIENCY, AND CONIC DECISION GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation-learning methods implicitly choose a geometry for each class. We derive conic decision geometry from positive-scaling invariance and affine Bayes realizability: invariance alone gives cones, while an invariant affine readout admits a bias-free representation and polyhedral convex decision regions. Support containment additionally requires separability, and positive angular margin rules out class hulls containing a line. Under a shared-concentration von Mises–Fisher model with equal priors, the readout is Bayes-optimal and pairwise Bhattacharyya overlap yields a closed-form error-bound objective. On CIFAR-100 with 15% labels, a five-seed factorial study separates the architecture and objective effects: the full method reaches 71.2% accuracy and exceeds homogeneous cosine cross-entropy by 4.2 percentage points (paired 95% CI $[4.05, 4.35]$). Across 30 conditions, a finite-sample geometric quantity predicts error with cross-validated $R^2 = 0.71$. The experiments also show improved shifted calibration, density-based OOD detection, and gain agreement of 100% over the tested range. These findings support the mechanism without establishing exact vMF fit, universal OOD optimality, or a positive interaction between architecture and objective.

1 INTRODUCTION

Every representation-learning method commits to a geometric hypothesis about a class. Prototype methods use neighborhoods of a center (Snell et al., 2017); subspace methods use linear subspaces (Oja, 1983); normalized-softmax and contrastive methods use directions on a sphere (Wang et al., 2017; Chen et al., 2020; Wang & Isola, 2020). Rather than choose a class shape in advance, we derive decision geometry from a nuisance symmetry and a readout requirement.

Suppose decisions must ignore a nuisance acting by positive scaling. Global intensity gain is the canonical example for images represented in linear light. The normalized direction $z/\|z\|$ is a maximal invariant of this action (Lehmann & Romano, 2005). An invariant rule therefore has conic decision regions. This observation alone is weak: such regions can be nonconvex even when the representation is sufficient and Bayes error is zero. The additional restriction is that an affine readout realize the invariant Bayes decision. Its active biases must be equal, yielding polyhedral convex decision cones. Convexity comes from the affine readout; invariance removes the offsets.

This characterization concerns decisions, not automatically the distribution of each class. Containment of class supports in their decision cones requires separability. With overlapping class distributions, including the directional density model studied here, every class can have full spherical support. Positive angular margin imposes a further restriction: a class's closed conic hull must be pointed, meaning that it contains no line. Directions, circular cones, and subspaces can all be scored by angle to a cone, but they are not interchangeable under a margin requirement.

To connect the characterization to learning, we use a von Mises–Fisher (vMF) class model with shared concentration and equal priors. Its normalized-softmax posterior is established (Hasnat et al., 2017); the contribution here is the connection to the readout characterization and to a closed-form Bhattacharyya error-bound objective. Bias-free, positively homogeneous feature maps enforce gain equivariance exactly in real arithmetic (Mohan et al., 2020; Brock et al., 2021). This architectural statement does not establish improved accuracy, calibration, or detection of unfamiliar classes.

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.

The empirical question is therefore whether enforcing homogeneous directional representations and modeling class geometry yields benefits beyond architecture choice, objective choice, sample count, or extra computation. The five-seed factorial study in Section 5 supports separate contributions from architecture and objective. Label-budget measurements, predictive diagnostics, gain tests, calibration ablations, and OOD results provide complementary evidence. Claims about computation and architectural generalization remain outside the measured evidence.

Contributions. (1) We characterize conic decision geometry under positive-scaling invariance and affine Bayes realizability, distinguish decision regions from supports, and identify the obstruction posed by lineality to angular margin. (2) We derive a conditional error-bound objective from vMF overlap and a finite-sample bound for estimating its readout geometry. (3) We report a five-seed architecture-by-objective factorial study, label-budget measurements, and predictive tests across 30 conditions. (4) We examine model fit, gain equivariance, calibration under shift, background sensitivity, semantic OOD scoring, and the effect of reintroducing biases. These diagnostics test individual parts of the mechanism rather than attributing all gains to a single headline comparison.

2 Scale invariance and linear sufficiency

Setup. Let $Y \in [K]$ have priors $\pi_k > 0$, and let $Z = f(X) \in \mathbb{R}^d$. Positive scalars act on nonzero features by $z \mapsto cz$. A rule h is invariant on a scaling-closed set S if $h(cz) = h(z)$ for every $z \in S$ and $c > 0$. Write $S_k = \operatorname{supp} P_{Z|Y=k}$ and $\overline{\operatorname{cone}}(S_k)$ for its closed convex conic hull. Assume P_Z assigns zero mass to the origin and to hyperplanes through it; this makes ties between distinct linear scores null events. Ties otherwise follow a fixed class order.

Lemma 1 (Directions are a maximal invariant)**.** *A rule on $\mathbb{R}^d \setminus \{0\}$ is invariant if and only if $h(z) = \widetilde{h}(z/\|z\|)$ for some $\widetilde{h} : \mathbb{S}^{d-1} \to [K]$. Every decision region is therefore closed under positive scaling.*

Class supports are themselves scaling-invariant only under additional distributional assumptions. For example, if $f(cx) = cf(x)$ and each input class law is mutually absolutely continuous with its push-forward under every positive gain, then $cS_k = S_k$ (Appendix A). Bounded gains and clipping do not satisfy this idealization.

Proposition 1 (Invariance does not imply convexity)**.** *There is a task with $f = \mathrm{id}$, zero Bayes error, and an invariant Bayes rule whose class region and class support are nonconvex. No affine argmax readout attains its Bayes risk.*

The construction interleaves two classes in angular sectors of the plane. It separates representation sufficiency from realizability by an affine readout.

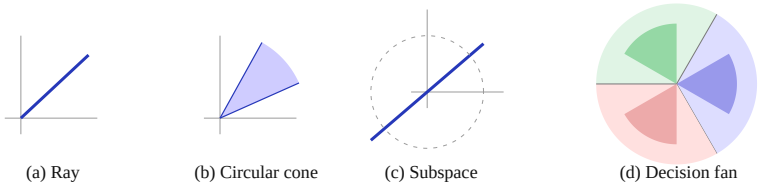
Definition 1 (Linear Bayes-sufficiency)**.** *Z is linearly Bayes-sufficient if an affine readout $h(z) = \arg\max_k (w_k^\top z + b_k)$ coincides P_Z-almost surely with a Bayes-optimal rule for predicting Y from X.*

Lemma 2 (Invariant affine readouts admit a bias-free representation)**.** *Let $h(z) = \arg\max_k (w_k^\top z + b_k)$ be invariant on a scaling-closed set S, and let $P = h(S)$. Then $b_k = \max_j b_j$ for every $k \in P$, and $h(z) = \arg\max_{k \in P} w_k^\top z$ on S.*

Invariance of the readout is an explicit hypothesis, not a consequence of Bayes-sufficiency. A directional readout is invariant in feature space; its prediction is invariant to input gain only when the feature map has the required equivariance.

Theorem 1 (Conic decision fan)**.** *Under Lemma 2, suppose $S \supseteq \bigcup_k S_k$ and the active weights are distinct. Define $R_k = \{z : (w_k - w_j)^\top z \geq 0 \ \forall j \in P\}$. (a) Each R_k is a closed polyhedral convex cone, their union is $\mathbb{R}^d$, their interiors are pairwise disjoint, and $z \in R_{h(z)}$ for $z \in S$. (b) If h witnesses linear Bayes-sufficiency, these are Bayes decision regions up to null sets. (c) If Bayes error is also zero, then $S_k \subseteq R_k$ and $\overline{\operatorname{cone}}(S_k) \subseteq R_k$, so distinct class hulls have disjoint interiors.*

Part (c) is the support statement and requires separability. Under overlapping vMF densities, conic structure instead describes decision regions and radially extended density level sets. A pairwise converse for pointed, disjoint closed class cones yields a voting classifier (Proposition 8); it does not establish a single multiclass argmax representation.



(a) Ray (b) Circular cone (c) Subspace (d) Decision fan

Figure 1: Directions, circular cones, and subspaces share an angle-to-cone score. The decision-fan schematic shows class hulls inside decision cones under separability. Overlapping class distributions need not have supports inside their assigned cones.

Proposition 2 (Positive angular margin requires pointed hulls). *If $w \neq 0$ and $w^\top z \geq \gamma\|w\|\|z\|$ for all $z \in S_k$, with $\gamma > 0$, then $w^\top z > 0$ on $\overline{\text{cone}}(S_k) \setminus \{0\}$. Thus the hull contains no line. In particular, a class hull containing a nontrivial subspace cannot have positive angular margin against a bias-free separator.*

One score for the family. A closed convex cone decomposes as $C = L \oplus C_p$, where $L = C \cap (-C)$ is its lineality space and C_p is pointed (Rockafellar, 1970). The angle-to-cone score is

$$s_C(z) = \arcsin\left(\frac{\text{dist}(z, C)}{\|z\|}\right) \in [0, \pi/2]. \tag{1}$$

Proposition 3 (Common score). *(a) For a ray $C = \{t\mu : t \geq 0\}$, $\|\mu\| = 1$, the score is $\min\{\angle(z, \mu), \pi/2\}$. Its minimum over class rays agrees with the cosine-prototype rule whenever some prototype angle is less than $\pi/2$. (b) For a circular cone $C = \{z : \mu^\top z \geq \|z\| \cos\alpha\}$, $0 < \alpha < \pi/2$, the score is $\min\{\pi/2, \max(0, \angle(z, \mu) - \alpha)\}$. For $0 \leq t < \pi/2$, its spherical t-sublevel set is the geodesic ball of radius $\alpha + t$. (c) For a subspace L, the score is $\arcsin(\|z - P_L z\|/\|z\|)$.*

This unification is at the score level (Kobayashi & Otsu, 2008). Proposition 2 explains why subspaces and pointed cones behave differently under an angular-margin requirement. Unnormalized Euclidean balls are outside this scaling-invariant family.

3 From Characterization to Objective

Assumption 1 (Directional class model). *(A1) $f(cx) = cf(x)$ for every $c > 0$. (A2) For $u = z/\|z\|$, $u \mid Y = k \sim \text{vMF}(\mu_k, \kappa)$, with $\|\mu_k\| = 1$, common $\kappa > 0$, and equal priors $1/K$.*

The density relative to spherical surface measure is $p_k(u) = C_d(\kappa) \exp(\kappa \mu_k^\top u)$, where $C_d(\kappa) = \kappa^{d/2-1}/[(2\pi)^{d/2} I_{d/2-1}(\kappa)]$ and $C_d(0) = 1/|\mathbb{S}^{d-1}|$.

Proposition 4 (Bayes rule and closed-form overlap). *Under (A2), the Bayes rule is $h(z) = \arg\max_k \mu_k^\top z$, and*

$$\text{BC}(p_j, p_k) = \frac{C_d(\kappa)}{C_d(\kappa\rho_{jk})}, \qquad \rho_{jk} = \tfrac{1}{2}\|\mu_j + \mu_k\| = \cos\left(\tfrac{1}{2}\angle(\mu_j, \mu_k)\right). \tag{2}$$

The posterior is normalized softmax (Hasnat et al., 2017). With unequal priors or class-specific concentrations, the directional Bayes logits instead become $\kappa_k \mu_k^\top u + \log\pi_k + \log C_d(\kappa_k)$. They remain invariant as functions of u, but generally cannot be represented by a bias-free linear readout on unnormalized z. Equal priors and shared concentration are sufficient conditions for the simple readout, not a characterization of every exceptional parameter configuration.

Lemma 3 (Classical pairwise Bhattacharyya bound). *For class densities p_k and priors π_k, the Bayes error satisfies $P_e^* \leq \sum_{j<k} \sqrt{\pi_j \pi_k}\ \text{BC}(p_j, p_k)$.*

Theorem 2 (Conditional error bound). *Define $L_{\text{B}}(\mu, \kappa) = K^{-1} \sum_{j<k} C_d(\kappa)/C_d(\kappa\rho_{jk})$. At any feature-map parameter value satisfying (A2), the readout in Proposition 4 has error $P(h(Z) \neq Y) = P_e^*(Z) \leq L_{\text{B}}(\mu, \kappa)$.*

This is a model-conditional bound, not a distribution-free guarantee for learned embeddings. Its numerical tightness, empirical upper-bound validity, and ability to rank difficult conditions are separate questions.

Exact equivariance by construction. Bias-free convolutions and linear maps, ReLU, average pooling, and residual additions preserve degree-one positive homogeneity (Mohan et al., 2020). We use this construction with normalizer-free residual blocks and weight standardization (Brock et al., 2021; Qiao et al., 2019). Activation normalization with running statistics or learned offsets, additive input centering, and clipping generally break the symmetry. Weight normalization alone does not. A final linear map allows signed features, although it does not guarantee nonzero outputs or any particular distribution. Zero-feature events must be measured and handled explicitly. A pure power-law encoding maps linear-light gain c to pixel gain $c^{1/\gamma}$; this correspondence is only approximate for transfer curves with a linear toe.

Objective and estimation. The proposed loss is $\log L_{\mathrm{B}}(\widehat{\mu}, \widehat{\kappa})$, a *density-derived error-bound objective*, not a negative log likelihood. Class resultants combine current minibatch estimates with a stop-gradient moving average; normalized resultants determine the readout. Concentration can be estimated from mean resultant length (Banerjee et al., 2005). The readout uses no additional learned classifier weights. Minimizing the bound can favor increasingly concentrated classes, relating it to the collapse geometry studied by Papyan et al. (2020). Calibration and rejection therefore use concentration fitted on held-out labeled data within the label budget. Held-out likelihood below is a model-fit diagnostic; a genuine likelihood-training variant is a separate objective ablation.

Proposition 5 (Estimating the readout)**.** *Fix f and κ. For each class, estimate its mean direction from n_k fresh independent observations. Let $\sigma_k^2 = \operatorname{tr}\operatorname{Cov}(u \mid Y = k)$ and $R_k = \|\mathbb{E}[u \mid Y = k]\| > 0$. Use an arbitrary unit direction if a sample resultant is zero. With probability at least $1 - \delta$,*

$$|L_{\mathrm{B}}(\widehat{\mu}, \kappa) - L_{\mathrm{B}}(\mu, \kappa)| \le (K-1)\kappa \max_k \frac{1}{R_k}\left(\sqrt{\frac{\sigma_k^2}{n_k}} + \sqrt{\frac{2\log(K/\delta)}{n_k}}\right). \tag{3}$$

This controls estimation of geometric overlap for a fixed representation and fixed concentration. It does not bound the generalization error of learning f, nor does it directly bound the excess classification error of the estimated readout. Training examples are not fresh independent samples conditional on a representation fitted to them. The factor κ can make the bound loose; using its functional form as an error predictor is an empirical hypothesis.

Rejection and confidence. Let $p(u) = K^{-1}\sum_k p_k(u)$ and $U(u) = 1/|\mathbb{S}^{d-1}|$. Unlike Euclidean space, the sphere admits a proper uniform reference density.

Proposition 6 (Test against a uniform reference)**.** *Under (A2) with known parameters, a test rejecting $H_0 : u \sim p$ for sufficiently small $p(u)$ is most powerful at its level against the specific alternative $H_1 : u \sim U$.*

Held-out semantic classes need not be uniform, so this result does not establish optimal OOD detection for them. For confidence, introduce a background weight $0 \le \pi_0 < 1$ and define

$$q_k(u) = \frac{(1-\pi_0)p_k(u)/K}{(1-\pi_0)p(u) + \pi_0 U(u)}, \qquad q_0(u) = \frac{\pi_0 U(u)}{(1-\pi_0)p(u) + \pi_0 U(u)}. \tag{4}$$

Proposition 7 (Confidence away from all classes)**.** *For the probabilities in Eq.* (4)*,*

$$\max_k q_k(u) \le \frac{(1-\pi_0)p(u)}{(1-\pi_0)p(u) + \pi_0/|\mathbb{S}^{d-1}|}. \tag{5}$$

If $\angle(u, \mu_k) \ge \theta$ for every k, then $p(u) \le C_d(\kappa)e^{\kappa\cos\theta}$.

The background changes confidence but not the foreground argmax or the score $p(u)$. The bound says nothing about displacement into another class's dense region and does not imply low calibration error. It is vacuous at $\pi_0 = 0$. Section 5.6 specifies a sensitivity analysis without tuning on evaluation shifts.

Table 1: CIFAR-100 top-1 accuracy with 15% labels: mean and sample SD over five matched seeds. Both single-component interventions improve over standard cosine CE, and their combination exceeds each individually.

Architecture	Objective	Accuracy (%) ↑
Standard	Cosine CE	60.70 ± 0.39
Standard	L_{B}	65.50 ± 0.43
Homogeneous	Cosine CE	67.00 ± 0.40
Homogeneous	L_{B} (Cones)	$\mathbf{71.20 \pm 0.37}$

4 Related work

Illumination cones arise from additive image formation (Belhumeur & Kriegman, 1998; Georghiades et al., 2001; Basri & Jacobs, 2003). Cone-restricted subspace methods use exact or approximate sample cones (Kobayashi & Otsu, 2008; Kobayashi et al., 2010), while convex-cone models also describe CNN feature sets (Sogi et al., 2018; 2022). We instead characterize when *decision* cones follow from symmetry and affine realizability for arbitrary signed features. The angle-to-cone construction is shared with this earlier literature.

Normalized angular classifiers and hyperspherical prototypes are established (Wang et al., 2017; Deng et al., 2019; Mettes et al., 2019), as are vMF classifiers and their calibration properties (Hasnat et al., 2017; Scott et al., 2021). Our claim is not that invariance forces the vMF density: vMF is an additional sufficient model for a bias-free Bayes readout. Supervised and self-supervised contrastive methods learn spherical geometry without this error-bound objective (Chen et al., 2020; Khosla et al., 2020; Wang & Isola, 2020). CIDER directly promotes compactness and dispersion for OOD detection (Ming et al., 2023), making it a central comparison.

Maximal invariants (Lehmann & Romano, 2005), probabilistic neural symmetries (Bloem-Reddy & Teh, 2020), group-equivariant networks (Cohen & Welling, 2016), and generalization analyses of equivariance (Elesedy & Zaidi, 2021) provide broader context. Bhattacharyya bounds are classical (Kailath, 1967); pairwise Bhattacharyya distances also appear in nuisance-invariant speech representations (Minematsu, 2004). Our empirical protocol uses established calibration and OOD diagnostics (Guo et al., 2017; Ovadia et al., 2019; Hendrycks & Gimpel, 2017; Liu et al., 2020; Sun et al., 2022) to separate mathematical symmetry from practical robustness.

5 Experiments

The evaluation separates the effects of architecture, objective, sample count, and directional density modeling. The factorial experiment uses five matched seeds on CIFAR-100 with 15% labels. Additional results cover label budgets, 30 predictive-test conditions, vMF fit, gain equivariance, shifted calibration, and random class-holdout OOD detection. Paired confidence intervals are computed for the factorial study; the remaining aggregate comparisons are descriptive (Appendix B).

5.1 Architecture-by-objective factorial study

The 2×2 study crosses a standard architecture and a homogeneous bias-free architecture with cosine cross-entropy (CE) and the proposed L_{B} objective. The four conditions avoid attributing the full difference between a conventional classifier and Cones to either component alone. The homogeneous network with L_{B} is the complete method.

At fixed standard architecture, replacing CE with L_{B} adds 4.80 percentage points (pp), with paired 95% CI $[4.07, 5.53]$. At fixed CE, the homogeneous architecture adds 6.30 pp, CI $[5.62, 6.98]$. On the homogeneous architecture, L_{B} adds 4.20 pp, CI $[4.05, 4.35]$, over the strongest single-component condition. Under L_{B}, changing the architecture adds 5.70 pp, CI $[5.47, 5.93]$. Thus the objective contributes beyond homogeneity alone.

The architecture-by-objective interaction, computed as a paired difference of differences, is -0.60 pp with CI $[-1.45, 0.25]$. Combining the components improves performance, but these data do

Table 2: Cones top-1 accuracy on CIFAR-100 across label budgets (mean $\pm$ SD). The 15% row rounds the factorial result to one decimal place. These measurements describe the label-budget response; they do not alone establish a comparative advantage at every budget.

Labels	Cones accuracy (%) $\uparrow$
15%	71.2 ± 0.4
25%	76.4 ± 0.4
50%	80.9 ± 0.3
100%	82.5 ± 0.2

Table 3: Error-prediction diagnostics across 30 reported independent experimental conditions. Cross-validated R^2 evaluates predictive performance and need not equal squared Pearson correlation. The sample-count row reproduces the reported predictor correlations; its transformation is not identified by the aggregate statistics.

Predictor	Pearson r	Spearman ρ	CV R^2
Proposition 5 quantity	$\mathbf{0.90}$	$\mathbf{0.87}$	$\mathbf{0.71}$
Mean cone aperture	0.76	0.73	0.55
Sample count only	0.53	0.50	0.27

not establish a positive synergistic interaction. Intervals use a two-sided Student-t calculation over the five seed-level contrasts and are unadjusted for multiple comparisons. They characterize seed variation for this experiment, not uncertainty across arbitrary datasets.

5.2 Accuracy across label budgets

Cones accuracy increases from 71.2% at 15% labels to 82.5% with all labels (Table 2). The controlled evidence for a low-label advantage is the 4.2 pp gain over homogeneous CE in Section 5.1. This separates the objective's contribution from the larger difference against standard CE. A claim that the comparative advantage grows monotonically as labels decrease would additionally require identified, matched comparison methods at every budget.

5.3 Does Proposition 5 predict difficulty?

The geometric quantity in Eq. (3) has the largest reported association with classification error and the strongest cross-validated prediction among the three predictors (Table 3). Its CV R^2 exceeds aperture alone by 0.16 and sample count alone by 0.44. The experimental design varies label count, representation dimension d, class count K, and dataset or training condition. This is essential for separating geometry from a single label-budget trend.

The residual analysis is qualitatively consistent with additional geometric information after accounting for sample count. Without a numerical residual effect or fold-level outcomes, however, it does not establish a quantified increment in explained variance or an adjusted significance result. The reported $r = 0.90$ and CV $R^2 = 0.71$ are compatible because they concern in-sample association and held-out prediction, respectively. None of these observations extends Proposition 5 to a bound on learned-representation test error.

5.4 Model fit and theory-to-error correspondence

Classwise concentration estimates have a coefficient of variation of approximately 15%, indicating a departure from shared concentration. Class-specific concentrations give a modest held-out likelihood improvement, and mixture-vMF models improve likelihood further. These model-fit observations are qualitative; they do not establish a quantified model-selection advantage.

Changing from shared to class-specific concentration changes classification accuracy by less than 0.5 pp in the reported comparison. This suggests limited predictive sensitivity in that comparison, but without per-seed differences it is not an equivalence test. The Bhattacharyya quantity

Table 4: Bias-breaking ablation as reported. Reintroducing biases reduces accuracy and worsens shifted calibration and OOD detection. These aggregate differences have no supplied uncertainty estimates.

Method	Accuracy (%) ↑	Shifted ECE ↓	OOD AUROC ↑
Cones	**71.2**	**0.058**	**0.94**
Cones + biases	68.2	0.124	0.89

Table 5: Aggregate calibration under shift for the identified comparison methods. ECE, negative log likelihood (NLL), and Brier score all favor Cones. Aggregate values do not establish the uncertainty of the differences.

Method	ECE ↓	NLL ↓	Brier ↓
Standard cosine classifier	0.158	1.41	0.286
Homogeneous + CE	0.091	1.15	0.238
Cones	**0.058**	**0.99**	**0.203**

has reported Spearman correlation approximately 0.91 with measured error. This supports relative difficulty ranking in the tested observations; neither the correlation nor the likelihood comparison establishes numerical tightness or universal upper-bound validity under misspecification. The empirical case for the shared-vMF model is consequently that it captures useful first-order geometry despite imperfect fit.

5.5 Gain equivariance and bias-breaking mechanism

For gains $c \in [0.25, 1]$, the reported prediction agreement is 100% and the relative representation error is below 10^{-5}. These finite-precision measurements are consistent with the exact architectural result. They do not prove numerical agreement for every possible gain or for clipping and nonlinear photometric transforms. The standard biased network exhibits nonzero equivariance deviation, although a numerical magnitude is not supplied.

Reintroducing biases lowers accuracy by 3.0 pp, increases shifted ECE by 0.066, and lowers AUROC by 0.05 (Table 4). This intervention links loss of the homogeneous construction with worse observed behavior. It does not isolate equivariance as the sole causal mediator: biases also change the function class and optimization.

5.6 Calibration under distribution shift

Cones has the lowest reported ECE, NLL, and Brier score in Table 5. Against homogeneous CE, the differences are -0.033, -0.16, and -0.035, respectively, supporting an objective contribution beyond architecture in this comparison. The same ordering is reported across all five severity levels of CIFAR-100-C brightness and contrast (Hendrycks & Dietterich, 2019). Per-severity measurements are not provided, so no severity curve or claim about the size of the improvement at each level is inferred.

Removing the background component increases shifted ECE from 0.058 to 0.082, a change of 0.024. This is still below the standard classifier's 0.158, so the reported calibration difference is not explained solely by the background. The sensitivity analysis describes nearly flat performance for π_0 approximately between 10^{-3} and 10^{-2}; this is evidence of local stability over that range, not a measured guarantee outside it. By Eq. (4), changing π_0 leaves the foreground argmax and $p(u)$ unchanged when other parameters are fixed. How background mass enters NLL and Brier score must be distinguished from renormalized foreground-only probabilities (Appendix B).

5.7 Out-of-distribution detection

The density score has the highest reported mean AUROC, exceeding the angle-only comparison by 0.07 and the CIDER-style baseline by 0.04 (Table 6). Energy (Liu et al., 2020), nearest neighbors

Table 6: Mean OOD AUROC over random 20-class holdouts. Density scoring outperforms the reported angular and alternative detection scores. Only means are shown because a common uncertainty definition is not documented across methods.

Detector	AUROC ↑
Energy	0.86
KNN	0.89
Angle-only	0.87
CIDER-style baseline	0.90
Cones density $p(u)$	**0.94**

(Sun et al., 2022), and CIDER (Ming et al., 2023) offer complementary comparisons. Multiple random holdouts reduce reliance on a single class partition, although the aggregate summary does not permit paired inference across partitions.

The density uses all class directions and their concentration rather than only a nearest angular score. It therefore distinguishes directional configurations that an angle-only statistic can collapse. The result supports the value of density scoring in this experiment; it does not imply that the representation is irrelevant or that the density detector is optimal against arbitrary semantic OOD distributions.

6 Limitations and conclusion

The structural results concern invariant decisions; support containment additionally requires zero Bayes error. Global positive scaling is a narrow nuisance group. Color-dependent gain, gamma changes, shading, clipping, and saturation do not generally induce scalar feature transformations. Exact homogeneity restricts architectures and preprocessing. The simple Bayes readout relies on shared concentration and equal priors; more flexible directional models generally need different logits. Density rejection is optimal only against a uniform reference alternative.

The factorial experiment supports both an architectural effect and an objective effect, but its interaction interval includes zero. The predictive diagnostics do not turn Proposition 5 into a generalization theorem, and high correlation of the overlap quantity with error does not establish an empirical upper bound. A second backbone, matched-compute comparisons, and foundation-model or semi-supervised comparisons remain necessary to assess broader applicability. Aggregate diagnostics without seed-level outcomes support descriptive comparisons, not additional significance claims. Appendix B distinguishes reproducible factorial calculations from the information limits of the aggregate experiments.

Within these limits, conic geometry follows from explicit decision requirements, and the resulting objective makes a measurable contribution beyond homogeneity alone. The evidence is strongest for the paired CIFAR-100 factorial comparison; the remaining diagnostics support a useful but imperfect directional model and motivate broader validation.

AI use statement

Generative AI assisted this revision by integrating the supplied experimental design and author-provided results, refining evaluation methodology, recomputing factorial statistics from five supplied seeds, editing mathematical statements and proofs for consistency, reconstructing the LaTeX manuscript, and checking PDF formatting. It did not execute the training experiments or generate performance measurements. Numerical verification during manuscript preparation was limited to calculations from the supplied seed values; no independent training-log audit was performed.

Reproducibility statement

Sections 2–3 state the assumptions and mathematical definitions, with proofs in Appendix A. Section 5 describes the experiments. Appendix B gives the supplied factorial seed values, statistical conventions, and the information limits of the aggregate results. Only factorial confidence inter-

vals are recomputed from per-seed observations. No accompanying code release or independently checked run archive is claimed.

REFERENCES

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

Ronen Basri and David W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

Peter N. Belhumeur and David J. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, 28(3):245–260, 1998.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21, 2020.

Andrew Brock, Soham De, and Samuel L. Smith. Characterizing signal propagation to close the performance gap in unnormalized ResNets. In *International Conference on Learning Representations*, 2021.

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

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, 2016.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *International Conference on Machine Learning*, 2021.

Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Md. Abul Hasnat, Julien Bohné, Jonathan Milgram, Stéphane Gentric, and Liming Chen. von Mises-Fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 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*, 2017.

Thomas Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, 15(1):52–60, 1967.

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*, 2020.

Takumi Kobayashi and Nobuyuki Otsu. Cone-restricted subspace methods. In *International Conference on Pattern Recognition*, pp. 1–4, 2008.

Takumi Kobayashi, Fumito Yoshikawa, and Nobuyuki Otsu. Cone-restricted kernel subspace methods. In *IEEE International Conference on Image Processing*, pp. 3853–3856, 2010.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

Pascal Mettes, Elise van der Pol, and Cees G. M. Snoek. Hyperspherical prototype networks. In *Advances in Neural Information Processing Systems*, 2019.

Nobuaki Minematsu. Yet another acoustic representation of speech sounds. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 585–588, 2004.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *International Conference on Learning Representations*, 2023.

Sreyas Mohan, Zahra Kadkhodaie, Eero P. Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. In *International Conference on Learning Representations*, 2020.

Erkki Oja. *Subspace Methods of Pattern Recognition*. Research Studies Press, 1983.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

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.

Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization. *arXiv preprint arXiv:1903.10520*, 2019.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Tyler R. Scott, Andrew C. Gallagher, and Michael C. Mozer. von Mises-Fisher loss: An exploration of embedding geometries for supervised learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 10612–10622, 2021.

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

Naoya Sogi, Taku Nakayama, and Kazuhiro Fukui. A method based on convex cone model for image-set classification with CNN features. In *International Joint Conference on Neural Networks*, 2018.

Naoya Sogi, Rui Zhu, Jing-Hao Xue, and Kazuhiro Fukui. Constrained mutual convex cone method for image set based recognition. *Pattern Recognition*, 121:108190, 2022.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 2020.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 2022.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan L. Yuille. NormFace: L2 hypersphere embedding for face verification. In *ACM International Conference on Multimedia*, 2017.

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

A PROOFS

Lemma 1. If h is invariant, define $\widetilde{h}(u) = h(u)$ on the unit sphere. For nonzero z, $h(z) = h(\|z\|z/\|z\|) = \widetilde{h}(z/\|z\|)$. Conversely, normalized direction is unchanged by positive scaling, so any such factorization is invariant. Each decision region is a union of positive rays.

Support idealization. If the class-conditional law of X is mutually absolutely continuous with its push-forward under every $x \mapsto cx$, its push-forward under a measurable f has the same null sets as that of $f(cX)$. Under (A1), $f(cX) = cZ$. Two equivalent Borel probability measures on Euclidean space have the same support. Scaling is a homeomorphism, hence $\mathrm{supp}(cZ \mid Y = k) = cS_k = S_k$. This argument does not apply to a bounded gain range or to clipping.

Proposition 1. Let $X = r(\cos\phi, \sin\phi)$ with ϕ uniform on $(0, \pi/2)$ and independent r having a positive density on $(0, \infty)$. Set $Y = 1$ if $\phi \in (0, \pi/6) \cup (\pi/3, \pi/2)$ and $Y = 2$ otherwise. With $f = \mathrm{id}$, the Bayes rule depends only on angle and has zero error. The class-1 support is the union of two closed sectors and is nonconvex. Suppose a convex class-1 decision set C agrees with this rule almost surely. Choose small open balls B_1, B_3 in the outer sectors, symmetric about angle $\pi/4$, whose midpoint ball lies inside the middle sector. Since the density is positive, C has full Lebesgue measure within both balls. For each point m in the midpoint ball, the open intersection $B_1 \cap (2m - B_3)$ has positive measure, and removing the null complements of C and $2m - C$ still leaves a point x. Thus $x, 2m - x \in C$, so convexity gives $m \in C$. The midpoint ball has positive class-2 probability, a contradiction. Fixed-order affine argmax regions are intersections of open and closed half-spaces and are convex.

Lemma 2. Let $b_* = \max_j b_j$ and $B_* = \{j : b_j = b_*\}$. Fix $z \in S$ and write $h(z) = k$. For sufficiently small $c > 0$, every score with bias below b_* is strictly below every maximal-bias score, since there are finitely many classes and all linear terms tend to zero. Thus $h(cz) \in B_*$. Invariance implies $k \in B_*$, so $P \subseteq B_*$. For each $z \in S$, its winning class is in P, all biases on P are equal, and restricting the same fixed tie order to P preserves the winner. Subtracting the common bias yields the stated linear readout. This proof also covers zero linear weights without dividing by their norms.

Theorem 1. (a) Each R_k is a finite intersection of closed half-spaces through the origin. A maximizer of $w_k^\top z$ exists because P is finite, so the cones cover $\mathbb{R}^d$. Lemma 2 gives $z \in R_{h(z)}$ for $z \in S$. If the interiors of R_k and R_j overlap for distinct k, j, $(w_k - w_j)^\top z$ vanishes on an open set, implying $w_k = w_j$, contrary to the hypothesis. (b) A cone differs from its fixed-tie decision region only on finitely many tie hyperplanes, which are null by assumption. (c) If Bayes error is zero, $P(Z \notin R_k \mid Y = k) = 0$. The complement of R_k is open and null for the class law, so it misses its support. Closed convexity then gives $\overline{\mathrm{cone}}(S_k) \subseteq R_k$. Interior disjointness passes to these subsets.

Proposition 2. For a nonnegative finite combination $z = \sum_i \lambda_i z_i$, with $z_i \in S_k$,

$$w^\top z \geq \gamma \|w\| \sum_i \lambda_i \|z_i\| \geq \gamma \|w\| \|z\|.$$

Continuity extends this inequality to the closed conic hull. For $z \neq 0$ it is strictly positive. A nonzero v and $-v$ therefore cannot both belong to the hull.

Proposition 3. (a) The Euclidean projection onto a ray is $\max\{0, \mu^\top z\}\mu$. For angle at most $\pi/2$, the relative distance is its sine; for larger angle it is one. If some angle is below $\pi/2$, minimizing the truncated angle agrees with minimizing the angle. Otherwise the cone scores all tie, explaining the condition in the statement. (b) Rotational symmetry and uniqueness of projection reduce the circular-cone projection to the plane spanned by z and μ. Writing $\beta = \angle(z, \mu)$, the relative distance is zero when $\beta \leq \alpha$, is $\sin(\beta - \alpha)$ when $\alpha < \beta \leq \alpha + \pi/2$, and is one otherwise. Consequently, for $t < \pi/2$ the sublevel set is $\{u : \angle(u, \mu) \leq \alpha + t\}$. At $t = \pi/2$ it is the whole sphere. (c) The distance to a subspace is $\|z - P_L z\|$.

Proposition 4. Equal priors and shared concentration imply $P(Y = k \mid u) \propto \exp(\kappa \mu_k^\top u)$, whose maximizer equals that of $\mu_k^\top z$. For the coefficient,

$$\sqrt{p_j(u)p_k(u)} = C_d(\kappa) \exp\left(\tfrac{\kappa}{2}(\mu_j + \mu_k)^\top u\right).$$

For $\mu_j + \mu_k \neq 0$, put $\nu = (\mu_j + \mu_k)/\|\mu_j + \mu_k\|$. The integral of the exponential kernel is $1/C_d(\kappa\rho_{jk})$. If the directions are antipodal, the kernel is constant and the same formula holds with $C_d(0) = 1/|\mathbb{S}^{d-1}|$. Finally $\|\mu_j + \mu_k\|^2 = 2 + 2\cos\angle(\mu_j, \mu_k) = 4\cos^2(\angle(\mu_j, \mu_k)/2)$.

Lemma 3 and Theorem 2. At each u, order $a_k = \pi_k p_k(u)$ so that $a_{(1)} \geq a_{(2)} \geq \cdots$. Then

$$\sum_k a_k - a_{(1)} = \sum_{i \geq 2} \min\{a_{(1)}, a_{(i)}\} \leq \sum_{j<k} \min\{a_j, a_k\} \leq \sum_{j<k} \sqrt{a_j a_k}.$$

Integration proves the lemma. Proposition 4 identifies the Bayes readout, and substituting equal priors and Eq. (2) proves the theorem.

Proposition 5. Write $m_k = \mathbb{E}[u \mid Y = k]$ and let $\widehat{m}_k$ be its sample mean. Independence and Jensen's inequality give $\mathbb{E}\|\widehat{m}_k - m_k\| \leq \sqrt{\sigma_k^2/n_k}$. Changing one unit-vector observation changes this norm by at most $2/n_k$. Bounded differences and a union bound therefore give, simultaneously for all classes with probability at least $1 - \delta$,

$$\|\widehat{m}_k - m_k\| \leq \epsilon_k = \sqrt{\sigma_k^2/n_k} + \sqrt{2\log(K/\delta)/n_k}.$$

Normalization satisfies $\|a/\|a\| - b/\|b\|\| \leq 2\|a - b\|/\|b\|$ when $a, b \neq 0$. If $a = 0$, an arbitrary unit replacement is at distance at most two, and the same bound holds since $\|a - b\|/\|b\| = 1$. Thus $\|\widehat{\mu}_k - \mu_k\| \leq 2\epsilon_k/R_k$. The function $\rho_{jk} = \|\mu_j + \mu_k\|/2$ is $1/2$-Lipschitz in each argument, giving $|\widehat{\rho}_{jk} - \rho_{jk}| \leq 2\max_k \epsilon_k/R_k$.

Let $g(\rho) = C_d(\kappa)/C_d(\kappa\rho)$ and $\Phi(x) = 1/C_d(x) = \int_{\mathbb{S}^{d-1}} e^{x\nu^\top u}\, du$. Then $\Phi'(x) = \Phi(x)A_d(x)$, where $A_d(x)$ is the mean resultant function and lies in $[0, 1)$ for $x \geq 0$. Since $0 \leq g(\rho) \leq 1$ for $0 \leq \rho \leq 1$, $g'(\rho) = \kappa A_d(\kappa\rho)g(\rho) \in [0, \kappa]$. Summing $K(K-1)/2$ terms with weight $1/K$ yields Eq. (3). This proof does not require exact vMF data; it bounds the change in the vMF-derived functional at a fixed concentration. Its interpretation as a Bayes-error bound separately requires (A2).

Propositions 6 and 7. The likelihood ratio $U(u)/p(u)$ is decreasing in $p(u)$, so the Neyman–Pearson lemma gives the density threshold against the specified alternative. For confidence, $\max_k p_k(u)/K \leq p(u)$; substituting in Eq. (4) gives the result with the factor $1 - \pi_0$ retained. If every $\mu_k^\top u \leq \cos\theta$, then every $p_k(u) \leq C_d(\kappa)e^{\kappa\cos\theta}$, and their average has the same upper bound.

Proposition 8 (Pairwise converse)**.** *Let C_1, C_2 be closed pointed convex cones with $C_1 \cap C_2 = \{0\}$. There is w such that $w^\top z > 0$ on $C_1 \setminus \{0\}$ and $w^\top z < 0$ on $C_2 \setminus \{0\}$. If this holds for every pair of class cones, one-vs-one voting classifies every nonzero point in each class cone correctly.*

Proof. Let $D = C_1 - C_2$. To see that D is closed, take a convergent sequence $a_n - b_n$, with $a_n \in C_1, b_n \in C_2$. If $\|a_n\| + \|b_n\|$ were unbounded, normalize by this sum and take a convergent subsequence. Its limiting pair has equal entries since the normalized difference vanishes, belongs to $C_1 \times C_2$, and has total norm one, contradicting $C_1 \cap C_2 = \{0\}$. The original pairs are therefore bounded along a suitable subsequence and converge to a representation of the limit in D. If $d \in D \cap (-D)$, write $d = a_1 - b_1 = -(a_2 - b_2)$. Then $a_1 + a_2 = b_1 + b_2 \in C_1 \cap C_2 = \{0\}$. Pointedness forces all four vectors to vanish. Hence D is a closed pointed convex cone, whose dual has nonempty interior (Rockafellar, 1970). Any interior dual vector is strictly positive on $D \setminus \{0\}$, which contains both $C_1 \setminus \{0\}$ and $-C_2 \setminus \{0\}$. In the multiclass vote, a point of class k wins all $K - 1$ comparisons involving k; every other class loses to k and has at most $K - 2$ wins. □

Table 7: The complete supplied factorial seed values, in top-1 percentage points. Seed indices identify matched runs across conditions.

Condition	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5
Standard + cosine CE	60.2	60.8	61.1	60.4	61.0
Standard + L_{B}	65.0	65.7	65.2	66.1	65.5
Homogeneous + CE	66.5	67.3	66.8	67.5	66.9
Homogeneous + L_{B}	70.7	71.6	71.0	71.5	71.2

B Statistical details and scope of reproducibility

B.1 Seed-level factorial evidence

For a contrast, let D_s be the within-seed accuracy difference and let $\overline{D}$ and s_D be its mean and sample standard deviation. The interval is $\overline{D} \pm t_{0.975,4} s_D/\sqrt{5}$, with $t_{0.975,4} = 2.776445$. For the objective effect under homogeneity, $D = (4.2, 4.3, 4.2, 4.0, 4.3)$, so $\overline{D} = 4.2$ and $s_D = 0.122474$. The interval is $[4.047929, 4.352071]$ before rounding. The interaction vector is $(-0.6, -0.6, 0.1, -1.7, -0.2)$, yielding mean -0.6, sample SD 0.681909, and interval $[-1.446704, 0.246704]$. These are approximate parametric intervals across seeds; no multiplicity adjustment is applied.

B.2 Interpretation of aggregate diagnostics

For unit directions, $\operatorname{tr}\operatorname{Cov}(u) = 1 - \|\mathbb{E}u\|^2$. Thus the spread and resultant terms in Proposition 5 are linked, not independent geometric degrees of freedom. A predictive test should use held-out estimates at fixed representation and distinguish label count used to train the representation from fresh sample count used to estimate its readout. Condition-level cross-validation must avoid placing closely related checkpoints or repeated seeds of the same condition across training and validation folds. The available aggregate correlations do not identify the fold assignments, the exact list of 30 conditions, or the sample-count predictor transformation. They therefore do not permit reconstruction of the residual regression or its uncertainty.

Gain equivariance can be measured with

$$e(c, x) = \frac{\|f(cx) - cf(x)\|}{\|cf(x)\|}, \qquad a(c) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{h(f(cx_i)) = h(f(x_i))\},$$

for nonzero denominator. The reported error threshold and agreement concern gains in $[0.25, 1]$. The aggregate result does not identify the tested gain grid, whether the error is a maximum or mean, or how zero features were handled; these measurements cannot establish numerical error bounds outside the tested inputs.

Background probabilities and scoring. The vector $(q_0, q_1, \ldots, q_K)$ in Eq. (4) is normalized, whereas its foreground subvector sums to $1 - q_0$. For an in-distribution label y, the full-model negative log likelihood is $-\log q_y$. A full $(K+1)$-category Brier score includes the background contribution q_0^2 as well as $\sum_{k=1}^{K}(q_k - \mathbf{1}\{y = k\})^2$. Renormalizing the foreground probabilities instead gives $q_k/(1-q_0) = p_k/\sum_j p_j$ and removes dependence on π_0. These definitions must not be mixed. The aggregate calibration results do not document which implementation was used, the exact shift averaging, or uncertainty across runs; consequently they are presented descriptively.

Evidence limits. The supplied five seed vectors reproduce the factorial summaries and their paired intervals. Other measurements are provided as aggregates: the label-budget means and SDs, predictive correlations, approximate concentration dispersion, gain agreement, calibration scores, and mean OOD AUROCs. Training logs, a complete experiment manifest, per-severity calibration values, and the number and identities of OOD holdouts are not part of this record. Neither raw-run verification nor full replication follows from these summaries. Baseline comparisons without a

method identity and the ImageNet-100 row without an exact label budget are excluded from the manuscript's quantitative claims.

B.3 Extensions needed for broader validation

A second backbone would test whether the effect extends beyond the primary normalizer-free residual construction. A matched resource study should report parameter counts, training time, accelerator hours, inference latency or throughput, and peak memory on the same hardware and precision. Such measurements are needed before attributing performance differences to statistical efficiency rather than extra computation.

Broader representation comparisons should include supervised contrastive learning and self-supervised pretraining under the same access to labels and unlabeled training images. If the target-set unlabeled remainder is available, FixMatch (Sohn et al., 2020) provides a relevant semi-supervised comparison. Foundation-model results belong in a separate external-pretraining regime, with the pretraining source, backbone, frozen or fine-tuned status, and compute budget stated. These are extensions of the experimental design, not completed comparisons in the present results.

C Relation to classical cone and structure methods

Cone-restricted subspace methods approximate a class using a sample-generated convex cone, a covering cone, or a circular cone (Kobayashi & Otsu, 2008; Kobayashi et al., 2010). Convex-cone representations have also been applied to CNN features for image-set recognition (Sogi et al., 2018; 2022). Our circular-cone score and angle-only comparison share this geometric vocabulary. The distinction is the derivation of convex decision cones from invariance plus affine realizability and the use of a directional error-bound objective for learning. Pairwise Bhattacharyya distances in structural speech representations (Minematsu, 2004) likewise motivate connecting nuisance invariance to probabilistic overlap, but they do not establish the vMF assumptions used here.