NARROWING THE APERTURE FOR THE TAIL: CONTROLLABLE SPATIAL EVIDENCE FOR LONG-TAILED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-expert architectures have emerged as a prominent paradigm for long-tailed visual recognition by distributing specialized classifiers across the class frequency spectrum. However, current frameworks uniformly employ Global Average Pooling (GAP) for spatial feature aggregation, implicitly enforcing an identical spatial observation scale for all classes and experts. In contrast to frequent classes that often benefit from broader contextual integration, rare categories tend to rely more heavily on compact, discriminative local cues. To address this architectural bottleneck, we propose **Spatial Moment**, a framework that endows controllable multi-expert networks with **preference-conditioned evidence apertures**. By coupling the hypernetwork preference state with a generalized-mean pooling ladder, our method enables tail-oriented experts to adaptively contract their spatial apertures, concentrating aggregation weights onto local, high-magnitude evidence. To complement this spatial mechanism, we further introduce a bounded one-sided geometric offset derived from within-expert classifier radii to restore radial magnitude statistics canceled by cosine normalization, alongside a class-orthogonal directional regularizer that discourages probability mass from concentrating on competing classes. During inference, experts are aggregated via a parameter-free, decisiveness-aware consensus route that maintains stable ensemble behavior. Extensive evaluations on long-tailed benchmarks and test-prior agnostic scenarios demonstrate that our approach achieves competitive performance against strong multi-expert baselines. Component removals and spatial-flow analyses indicate that adaptive aperture modulation is the primary driver of performance gains, mitigating head-bias classification errors on minority-class samples.

1 INTRODUCTION

In many real-world visual tasks such as object detection and image classification, we face the pervasive challenge of long-tailed distributions. Since samples from head classes account for a disproportionate share of the data while tail classes remain scarce, this distribution imbalance makes models prone to head-biased solutions and degraded tail performance (Liu et al., 2019; Cao et al., 2019; Cui et al., 2019; Zhang et al., 2025).

To address long-tailed recognition, prior work has explored class-imbalanced margin optimization (Cortes et al., 2025), loss and logit modifications (Cao et al., 2019; Cui et al., 2019; Menon et al., 2021), and exploratory samples that improve class balance and decision regions (Jian et al., 2025). RIDE and SADE use diverse specialized experts (Wang et al., 2021; Zhang et al., 2022); PRL generates preference-conditioned expert heads (Zhao et al., 2024b). Our work instead studies how the experts aggregate spatial evidence.

However, existing multi-expert methods generally use Global Average Pooling (GAP) uniformly across branches, implicitly imposing the same diffuse spatial observation scale on all classes and experts. This assumption is at odds with the visual evidence required by different frequency strata. Head classes can often benefit from broad contextual cues, whereas rare classes may depend on compact, highly discriminative local patterns. Uniformly averaging activations over the whole image can dilute such tail cues with irrelevant background responses, increasing the risk of head-biased

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.

predictions. Generalized-mean pooling provides a direct way to control the concentration of spatial evidence (Radenović et al., 2019), but it has not been coupled to preference-conditioned long-tailed experts.

We therefore propose Spatial Moment, an interpretable and controllable long-tailed learning method with preference-conditioned spatial evidence. The method co-indexes each expert's feature aggregation sensitivity with its role in the class-frequency spectrum. Tail-oriented experts use higher generalized-mean orders to contract their effective evidence apertures, while bounded radial calibration, rarity-conditioned directional training, and decisiveness-aware consensus stabilize the resulting ensemble.

As illustrated in Figure 1, the framework addresses two coupled problems. First, instead of enforcing a static GAP operator, it introduces a preference-controlled aperture ladder based on generalized-mean pooling. Higher orders selectively amplify concentrated, high-magnitude evidence and suppress diffuse background responses. Second, because cosine normalization removes classifier norms, the method recovers a bounded within-expert radial signal and uses it as a one-sided offset, while a rarity-conditioned directional regularizer controls residual probability mass over rival classes. Inference uses a parameter-free decisiveness-aware consensus route.

Our contributions are:

- We identify the spatial evidence bottleneck in controllable multi-expert long-tailed recognition and formulate preference-conditioned evidence apertures as a mechanism for matching spatial sensitivity to expert roles.
- We propose Spatial Moment, which couples a generalized-mean aperture ladder with bounded radial geometry rectification, rarity-conditioned directional regularization, and conservative consensus.
- Across imbalance ratios and unknown test distributions, experiments and frequency-resolved analyses show how aperture narrowing changes minority-class predictions and contributes to the observed gains.

2 Related Work

Long-Tailed Learning and Multi-Expert Ensembles. Long-tailed recognition uses re-weighting (Cui et al., 2019), margins (Cao et al., 2019; Cortes et al., 2025), and logit calibration (Menon et al., 2021). Focal-SAM applies class-wise sharpness penalties to improve generalization (Li et al., 2025). SEL generates exploratory examples to improve class balance and decision-region robustness (Jian et al., 2025). RIDE and SADE train diverse experts (Wang et al., 2021; Zhang et al., 2022); PRL generates preference-conditioned heads (Zhao et al., 2024b). ICL combines independent expert optimization with collaborative distillation (Chen et al., 2025); TS-MOF formulates two-stage multi-objective optimization (Zhao et al., 2025); TCL leverages trust calibration for collaborative ensembles (Zhou & Luo, 2026). Our focus is complementary: coupling expert roles with controllable spatial pooling rather than using a uniform GAP operator across branches.

Spatial Pooling and Evidence Apertures. In modern vision architectures, Global Average Pooling is commonly applied prior to classification (Lin et al., 2014). Although GAP supports holistic contextual integration, it weights spatial grid locations equally and can dilute localized discriminative tail activations within broad background responses. In visual retrieval, Generalized-Mean (GeM) pooling adjusts a pooling order (Radenović et al., 2019) between average pooling ($p = 1$) and max pooling ($p \to \infty$), concentrating aggregation on high-magnitude responses. We adapt this principle to class-imbalanced learning by establishing a preference-conditioned aperture ladder that contracts the spatial evidence aperture for minority-focused experts.

Classifier Geometry and Normalization. Linear classifiers trained on long-tailed data exhibit representation-classifier imbalance, motivating decoupled classifier training (Kang et al., 2020). Cosine normalization and angular margins (Wang et al., 2017; Deng et al., 2019) remove explicit magnitude variation by projecting weights onto a unit hypersphere, but this projection also discards radial statistics generated by a hypernetwork. Spatial Moment repurposes the discarded radial code as a bounded, standardized one-sided offset that calibrates angular decision margins via an emergent additive logit displacement for minority classes.

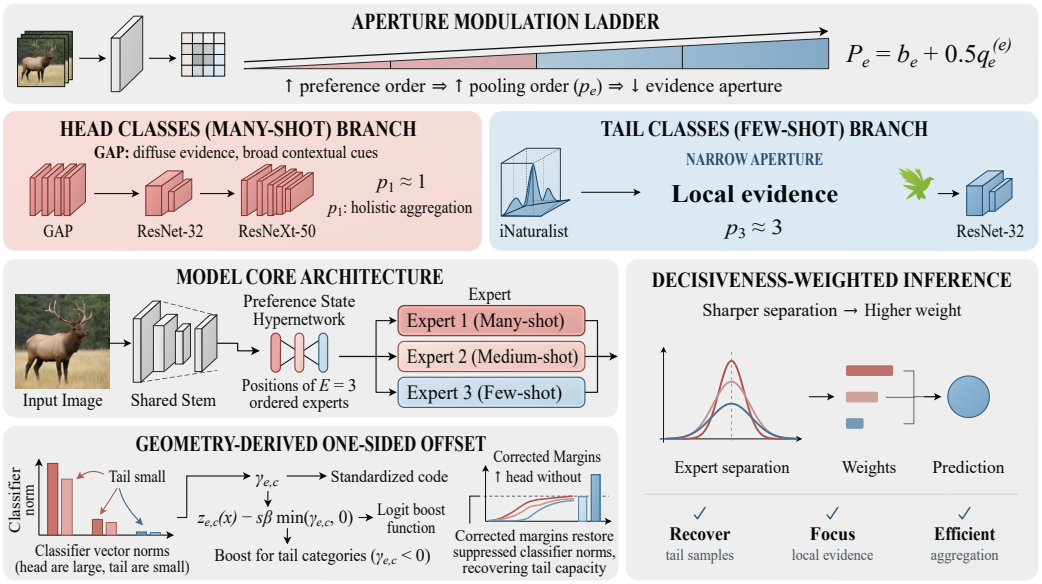


Figure 1: Overview of the proposed Spatial Moment framework. **Model Core Architecture** (middle-left) deploys $E = 3$ ordered experts over a shared visual stem, coupling multi-expert representation learning with preference-conditioned evidence apertures. The framework integrates three core algorithmic components: (1) **Aperture Modulation Ladder** (top): modulates generalized-mean pooling rungs ($p_1 \approx 1$ to $p_3 \approx 3$) via learned preference states (Eq. 1), shifting from holistic contextual aggregation for head classes to contracted apertures with localized peak attention that suppress diffuse background noise for tail classes; (2) **Geometry-Derived One-Sided Offset** (bottom-left): standardizes classifier weight norms to inject a bounded additive logit offset exclusively for minority categories (Eq. 5), regularizing decision margins without disturbing head classes; (3) **Decisiveness-Weighted Inference** (bottom-right): evaluates prediction margin separation to dynamically allocate consensus shares across experts with zero per-instance computational overhead.

3 METHOD

Spatial Moment co-indexes expert roles with evidence sensitivity and class-frequency objectives, resolving spatial observation bottlenecks via preference-conditioned apertures and bounded radial calibration.

3.1 PROBLEM SETUP AND ARCHITECTURE OVERVIEW

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote a dataset of C classes with frequencies n_c. We deploy $E = 3$ ordered experts. A shared visual stem maps image x to expert branches producing nonnegative feature maps $\mathbf{F}_e(x) \in \mathbb{R}_+^{D\times H\times W}$. A lightweight hypernetwork generates classifier matrix $\mathbf{W}_e = [\mathbf{w}_{e1}, \ldots, \mathbf{w}_{eC}] \in \mathbb{R}^{D\times C}$ conditioned on a learned preference state, establishing structural co-indexing between classifier directions and spatial pooling sensitivity across the class frequency spectrum.

3.2 PREFERENCE-CONDITIONED EVIDENCE APERTURES

Preference State and Aperture Ladder. Let $\mathbf{a} \in \mathbb{R}^E$ be the learned preference vector. For expert $e \in \{1, \ldots, E\}$, training introduces Dirichlet perturbations $\boldsymbol{\delta}^{(e)} \sim \mathrm{Dirichlet}(0.5 \times \mathbf{1}_E)$ to explore local neighborhoods, setting $\mathbf{q}^{(e)} = \mathrm{softmax}(\mathbf{a} + \boldsymbol{\delta}^{(e)})$ ($\boldsymbol{\delta}^{(e)} = \mathbf{0}$ during inference). Each expert is anchored to a discrete base rung $b_e \in \{1, 2, 3\}$, establishing a spatial progression from holistic context ($b_1 = 1$) to localized peak attention ($b_3 = 3$). The learned preference state $\mathbf{a}$ guides multi-expert co-adaptation:

$$p_e = b_e + 0.5q_e^{(e)}, \qquad (b_1, b_2, b_3) = (1, 2, 3). \tag{1}$$

During inference ($\boldsymbol{\delta}^{(e)} = \mathbf{0}$), this yields fixed specialized pooling orders ($p_1 \approx 1.17, p_2 \approx 2.17, p_3 \approx 3.17$), establishing heterogeneous receptive scales without per-instance pooling overhead. Sample-level test adaptivity is achieved through decisiveness routing (Section 3.5). For channel d, feature map activations are aggregated via generalized-mean pooling:

$$g_{e,d}(x) = \left(\frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} \max\big(F_{e,d,u,v}(x), \epsilon\big)^{p_e} \right)^{1/p_e}. \tag{2}$$

At $p_1 \approx 1$, aggregation performs standard average pooling for holistic context. At higher rungs ($p_2 \approx 2, p_3 \approx 3$), pooling progressively concentrates weight onto localized peak activations, contracting the spatial aperture for tail experts. As proven in Appendix D, relative gradient sensitivity scales as $(F_i/F_j)^{p_e-1}$, suppressing diffuse background noise. Raw classification scores are computed via cosine similarity scaled by $s = 30$:

$$z_{e,c}(x) = s \, \frac{\mathbf{g}_e(x)^\top \mathbf{w}_{e,c}}{\|\mathbf{g}_e(x)\|_2 \, \|\mathbf{w}_{e,c}\|_2}. \tag{3}$$

3.3 Geometry-Derived One-Sided Offset

Cosine normalization removes weight magnitudes to eliminate majority dominance, but discards within-expert radial capacity generated by the hypernetwork. We recover this signal by standardizing log weight norms $r_{e,c} = \|\mathbf{w}_{e,c}\|_2$:

$$\mu_e = \frac{1}{C} \sum_{j=1}^{C} \log r_{e,j}, \quad \sigma_e = \left[\frac{1}{C} \sum_{j=1}^{C} \big(\log r_{e,j} - \mu_e\big)^2 \right]^{1/2}, \quad \gamma_{e,c} = \tanh\left(\frac{\log r_{e,c} - \mu_e}{\max(\sigma_e, \epsilon)} \right). \tag{4}$$

The code $\gamma_{e,c} \in (-1, 1)$ indicates whether class c is radially sub-average within expert e, triggering a bounded one-sided offset:

$$\widehat{z}_{e,c}(x) = z_{e,c}(x) - s\beta \min(\gamma_{e,c}, 0), \qquad \beta = 0.005. \tag{5}$$

For tail categories whose norms shrink below average ($\gamma_{e,c} < 0$), Eq. equation 5 injects a calibrated additive offset bounded by $s\beta = 0.15$, providing non-disruptive margin regularization while head classes ($\gamma_{e,c} \geq 0$) remain undisturbed. As derived in Appendix B, this bounded offset dynamically mirrors Bayesian logit adjustment without external priors.

3.4 Rarity-Conditioned Margin and Directional Regularization

Training frequency maps to a normalized rarity coordinate $\rho_c = (\log n_{\max} - \log n_c)/\max(\log n_{\max} - \log n_{\min}, \epsilon) \in [0, 1]$. Expert e incorporates class prior $\pi_c = n_c / \sum_j n_j$ via prior-shaped logits $a_{e,c}(x) = z_{e,c}(x) + \tau_e \log(\pi_c + \epsilon)$, with exponents $(\tau_1, \tau_2, \tau_3) = (0, 1, 2)$. An additive target margin $m_{e,y} = \eta \frac{e}{E} \rho_y$ ($\eta = 0.035$) on ground-truth class y yields margin loss:

$$\mathcal{L}_e^{\mathrm{m}} = -\frac{1}{|\mathcal{B}|} \sum_{(x,y) \in \mathcal{B}} \log \frac{\exp(a_{e,y}(x) - m_{e,y})}{\exp(a_{e,y}(x) - m_{e,y}) + \sum_{c \neq y} \exp(a_{e,c}(x))}. \tag{6}$$

To prevent non-target probability mass from concentrating on a single rival class, we enforce a directional regularizer on predicted probabilities $\mathbf{p}_e(x) = \mathrm{softmax}(\mathbf{a}_e(x))$:

$$\mathcal{D}_e = \frac{1}{|\mathcal{B}|} \sum_{(x,y) \in \mathcal{B}} \left(1 - \frac{p_{e,y}(x)}{\|\mathbf{p}_e(x)\|_2} \right). \tag{7}$$

The composite loss is $\widetilde{\mathcal{L}} = \sum_{e=1}^{E} \omega_e \mathcal{L}_e^{\mathrm{m}} + \frac{\lambda_{\mathrm{d}}}{E} \sum_{e=1}^{E} \mathcal{D}_e$ ($\lambda_{\mathrm{d}} = 0.1$), dynamically balanced across branches via detached weights ω_e and rescaled as $\mathcal{L} = \kappa \widetilde{\mathcal{L}}$ using stop-gradient anchoring $\kappa = 2\,\mathrm{sg}(E^{-1} \sum_e \mathcal{L}_e^0)/\max(\mathrm{sg}(\widetilde{\mathcal{L}}), \epsilon)$, where $\mathcal{L}_e^0$ denotes cross-entropy on prior-shaped logits $\mathbf{a}_e(x)$ without the rarity margin $m_{e,y}$. Input-level Mixup (Zhang et al., 2018) is applied uniformly; complete loss dynamics are in Appendix C.

3.5 Conservative Decisiveness Consensus

At test time, experts exhibit heterogeneous decisiveness. While the aperture ladder provides fixed, specialized spatial observation scales across branches, test-time adaptivity is achieved dynamically at the decision level through per-sample decisiveness routing. Let $z_e^{(1)}(x)$ and $z_e^{(2)}(x)$ denote the top-two centered logits of expert e, and $\sigma_e(x)$ their standard deviation. The scale-invariant separation score $u_e(x) = (z_e^{(1)}(x) - z_e^{(2)}(x))/\max(\sigma_e(x), \epsilon)$ yields normalized routing shares $\boldsymbol{\alpha}(x) = \mathrm{softmax}(\mathbf{u}(x))$. We blend this dynamic allocation into a conservative consensus:

$$w_e(x) = \frac{0.85}{E} + 0.15\alpha_e(x), \qquad Z_c(x) = \sum_{e=1}^{E} w_e(x)\widehat{z}_{e,c}(x). \tag{8}$$

The 85% uniform backbone serves as a conservative regularization anchor, while the 15% decisiveness share adaptively modulates instance-level expertise, constraining consensus weights within $[0.283, 0.433]$. Final prediction is $\arg\max_c Z_c(x)$. Training consumes raw logits $z_{e,c}$, while one-sided offsets and decisiveness routing are stop-gradient prediction quantities requiring zero extra parameters.

4 Experiments

We evaluate **Spatial Moment** across standard long-tailed benchmarks: CIFAR-100-LT (Cao et al., 2019), ImageNet-LT (Liu et al., 2019), Places-LT (Liu et al., 2019) (derived from Places (Zhou et al., 2018)), and iNaturalist 2018 (Van Horn et al., 2018), reporting fine-grained Top-1 accuracy breakdowns across Many (> 100 images), Medium ($20 \sim 100$), and Few (< 20) shots.

Implementation Details. We follow standard multi-expert protocols (Wang et al., 2021; Zhang et al., 2022; Zhao et al., 2024b), employing ResNet-32 for CIFAR-100-LT, ResNeXt-50 for ImageNet-LT, ResNet-152 for Places-LT, and ResNet-50 for iNaturalist 2018. Semi-shared multi-branch backbones maintain identical parameter counts to baselines (3.85M on ResNet-32). CIFAR-100-LT models are trained for 200 epochs using SGD with momentum 0.9, batch size 128, and cosine learning rate schedule. Default configurations use $E = 3$ experts with base ladder $\mathbf{b} = (1, 2, 3)$. Complete protocols are detailed in Appendix E.

4.1 Benchmark Evaluations on Standard Long-Tailed Benchmarks

Table 1: CIFAR-100-LT Top-1 accuracy (%). Standard budget: 200 epochs from scratch on ResNet-32. Bold/underlined: best/second best in standard budget. External results cited from original papers under respective protocols.

	CIFAR-100-LT ($IF = 100$)				$IF = 50$	$IF = 10$
Method	Many	Medium	Few	Overall	Overall	Overall
Cross-Entropy (CE)	67.0	39.1	14.1	41.4	45.6	59.1
Balanced Softmax (Ren et al., 2020)	62.4	45.7	27.7	46.1	50.9	61.0
MiSLAS (Zhong et al., 2021)	61.2	47.1	29.8	46.8	51.5	62.5
RIDE (3 experts) (Wang et al., 2021)	65.6	48.5	27.0	48.0	51.7	61.8
SADE (3 experts) (Zhang et al., 2022)	64.9	49.8	29.0	48.8	53.8	63.6
BalPoE (3 experts) (Aimar et al., 2023)	67.6	53.2	32.3	52.0	56.3	64.8
PRL (3 experts) (Zhao et al., 2024b)	<u>67.9</u>	<u>54.3</u>	31.2	<u>52.8</u>*	<u>57.3</u>	<u>65.6</u>
Focal-SAM (LA, scratch) (Li et al., 2025)	63.9	53.0	<u>32.5</u>	50.7	54.5	63.8
Spatial Moment (Ours)	**68.5**	**55.5**	**34.6**	**53.8**	**57.5**	**65.8**
Extended Budgets / Alternative Backbones (Literature Reference)						
TS-MOF (ResNet-34, 220 ep) (Zhao et al., 2025)	79.0	49.2	39.9	56.8	60.2	70.8
ICL (3 experts, 180+20 ep) (Chen et al., 2025)	–	–	–	57.6	61.3	69.3
TCL (3 experts, 400 ep) (Zhou & Luo, 2026)	75.5	59.3	38.3	58.7	62.5	–

CIFAR-100-LT. Table 1 reports results across imbalance ratios. Under standard 200-epoch budgets, Spatial Moment reaches 53.8% Overall and 34.6% Few-shot accuracy at $IF = 100$, improving

Table 2: Top-1 accuracy (%) on large-scale benchmarks. Standard budget: ImageNet-LT (ResNeXt-50), Places-LT (ResNet-152), and iNaturalist 2018 (ResNet-50). Bold/underlined: best/second best in standard budget. External baselines follow canonical literature (Zhao et al., 2024b). Split metrics on ImageNet-LT are reported where standardized splits were provided in original papers; otherwise marked '–' to maintain strict statistical consistency ($A_{\text{Overall}} = 0.385A_{\text{Many}} + 0.479A_{\text{Medium}} + 0.136A_{\text{Few}}$).

	ImageNet-LT				Places-LT	iNat 2018
Method	Many	Medium	Few	Overall	Overall	Overall
Cross-Entropy (CE)	65.9	37.5	7.7	44.4	31.4	64.7
Balanced Softmax (Ren et al., 2020)	64.1	48.2	33.4	52.3	39.4	70.6
MiSLAS (Zhong et al., 2021)	–	–	–	51.4	38.3	70.7
RIDE (3 experts) (Wang et al., 2021)	68.0	52.9	35.1	56.3	40.3	71.8
SADE (3 experts) (Zhang et al., 2022)	66.5	57.0	43.5	58.8	40.9	72.7
BalPoE (3 experts) (Aimar et al., 2023)	<u>68.2</u>	<u>57.2</u>	<u>44.9</u>	59.8	40.8	75.0
PRL (3 experts) (Zhao et al., 2024b)	–	–	–	**60.8***	<u>41.6</u>	75.1
Focal-SAM (LA, scratch) (Li et al., 2025)‡	63.9	52.2	34.4	54.3	–	71.8
ICL (3 experts) (Chen et al., 2025)‡	–	–	–	<u>60.2</u>	–	**75.9**
Spatial Moment (Ours)	**69.2**	**58.1**	**46.2**	**60.8**	**42.3**	<u>75.8</u>
Extended Budgets / Alternative Backbones (Literature Reference)						
TCL (3 experts, 400 ep) (Zhou & Luo, 2026)	–	–	–	63.0§	44.2	76.9
TS-MOF (Zhao et al., 2025)	65.9	52.5	42.6	56.3	–	72.3

Few-shot accuracy over MOOSF (Zhao et al., 2024a) (32.1%), PRL (Zhao et al., 2024b) (31.2%), and Focal-SAM (Li et al., 2025) (32.5%) by 2.5, 3.4, and 2.1 percentage points. Overall reaches 57.5% and 65.8% at $IF = 50$ and $IF = 10$. Under extended budgets or larger backbones, TCL (400 epochs) and TS-MOF (ResNet-34) report higher overall accuracy, but Spatial Moment achieves a competitive balance under standard budgets without multi-stage training overhead.

Large-Scale Benchmarks. Table 2 shows 60.8% Overall and 46.2% Few-shot accuracy on ImageNet-LT (ResNeXt-50), outperforming strong multi-expert ensembles such as SADE (58.8% / 43.5%) and BalPoE (59.8% / 44.9%), while achieving +11.8% Few-shot over Focal-SAM. On Places-LT (ResNet-152), Spatial Moment reaches 42.3%, outperforming prior baselines (PRL 41.6%, BalPoE 40.8%). On iNaturalist 2018 (ResNet-50), accuracy reaches 75.8%, outperforming PRL (75.1%) and BalPoE (75.0%).

Table 3: Test-prior shifts on CIFAR-100-LT ($IF = 100$), Top-1 accuracy (%). Forward/Backward-LT emphasize head/tail classes (ρ is test imbalance ratio). Bold/underlined: best/second best. Spatial Moment achieves consistent improvements across both forward head-heavy and backward tail-inverted shifts.

	Forward-LT (ρ)					Uniform	Backward-LT (ρ)				
Method	50	25	10	5	2	1	2	5	10	25	50
CE	63.3	62.0	56.2	52.5	46.4	41.4	36.5	30.5	25.8	21.7	17.5
Balanced Softmax (Ren et al., 2020)	57.8	55.5	54.2	52.0	48.7	46.1	43.6	40.8	38.4	36.3	33.7
MiSLAS (Zhong et al., 2021)	58.8	57.2	55.2	53.0	49.6	46.8	43.6	40.1	37.7	33.9	32.1
RIDE (Wang et al., 2021)	63.0	59.9	57.0	53.6	49.4	48.0	42.5	38.1	35.4	31.6	29.2
SADE (Zhang et al., 2022)	65.2	62.5	58.8	55.4	51.2	48.8	43.0	43.9	42.4	42.2	42.0
BalPoE (Aimar et al., 2023)	69.0	65.2	61.2	59.0	54.2	51.7	45.7	46.6	45.2	45.2	45.8
PRL (Zhao et al., 2024b)	<u>69.5</u>	<u>65.7</u>	<u>61.7</u>	<u>59.5</u>	<u>54.7</u>	<u>52.2</u>	<u>46.2</u>	<u>47.1</u>	<u>45.7</u>	<u>45.7</u>	<u>48.5</u>
Spatial Moment (Ours)	**69.8**	**66.1**	**62.4**	**60.2**	**55.8**	**53.8**	**47.0**	**47.4**	**46.2**	**46.4**	**48.9**

Robustness under Unknown Test Class Distributions. Table 3 evaluates eleven test distributions parameterized by ρ. Spatial Moment consistently achieves top accuracy across all evaluated test configurations. Under forward shifts (Forward-50 to Forward-2), it attains 69.8% down to 55.8%, outperforming strong multi-expert ensembles such as SADE and PRL. On uniform evaluation ($\rho = 1$), it reaches 53.8%. Notably, under severe tail inversion (Backward-25 and Backward-50), where static ensembles such as RIDE drop sharply (down to 29.2%), Spatial Moment maintains **46.4%**

and **48.9%**, exceeding PRL (45.7% and 48.5%). Contracted tail apertures reduce feature dilution, preserving minority-class decision boundaries under severe inversion without requiring test-prior estimation or test-time retraining.

4.2 Mechanism Diagnostics and In-Depth Analysis

Component Ablation and Ladder Sensitivity. Table 4 isolates cumulative contributions. Moment pooling provides the foundational gain, raising Overall from 51.9% to 52.7% and Few from 30.8% to 32.5%, reinforced by radial calibration (+1.2% Few), rarity regularization, and consensus to reach 53.8% Overall and 34.6% Few. Table 5 separates ladder structure from modulation: the fixed $(1, 2, 3)$ ladder reaches 53.4% Overall and 33.8% Few, while modulation adds 0.4 and 0.8 points. The progressive ladder outperforms alternative spacings, while $E = 3$ experts matches the head/medium/tail division without redundancy.

Table 4: Cumulative ablation on CIFAR-100-LT ($IF = 100$), Top-1 accuracy (%). Checkmarks enable moment pooling ($\mathcal{M}_{\text{ap}}$), radial calibration ($\mathcal{R}_{\text{rad}}$), rarity regularization ($\mathcal{L}_{\text{rar}}$), and consensus ($\mathcal{C}_{\text{con}}$). Many/Medium/Few denote class-frequency groups; Overall covers all classes. Bold identifies the full model.

$\mathcal{M}_{\text{ap}}$	$\mathcal{R}_{\text{rad}}$	$\mathcal{L}_{\text{rar}}$	$\mathcal{C}_{\text{con}}$	Many	Medium	Few	Overall
				67.8	54.1	30.8	51.9
✓				68.0	54.7	32.5	52.7
✓	✓			68.1	55.0	33.7	53.2
✓	✓	✓		68.3	55.2	34.1	53.5
✓	✓	✓	✓	**68.5**	**55.5**	**34.6**	**53.8**

Table 5: Ladder and expert-count ablations on CIFAR-100-LT ($IF = 100$), Top-1 accuracy (%). Upper block: uniform GAP, fixed and modulated ladders, with $0.5q = 0.5q_e^{(e)}$ for expert e. Lower block: expert count E. Bold identifies the proposed ladder and default three-expert model.

Aperture ladder configuration	Many	Medium	Few	Overall
Uniform GAP ($p_e = 1$, modulation off)	67.8	54.1	30.8	51.9
Fixed ladder ($p = (1, 2, 3)$)	68.3	55.1	33.8	53.4
Linear ($b = (1, 1.5, 2) + 0.5q$)	68.2	55.1	33.7	53.3
Exponential ($b = (1, 2, 4) + 0.5q$)	68.3	55.2	34.1	53.5
Preference-modulated ladder (Ours)	**68.5**	**55.5**	**34.6**	**53.8**
Number of Experts (E)				
$E = 2$	67.2	54.2	33.6	52.6
$E = 3$ **(Default)**	**68.5**	**55.5**	**34.6**	**53.8**
$E = 4$	68.3	55.6	34.6	53.7

Spatial Aperture Evidence Compression Trajectories. Figure 2 sweeps $p \in [1, 5]$ on frozen representations. Higher orders contract effective support and increase top-10 energy share. At the learned rungs $p \approx (1.17, 2.17, 3.17)$, energy shares are 6.0%, 15.8%, and 27.5%. Expert E_1 retains more than 500 effective dimensions out of $64 \times 8 \times 8 = 4,096$, whereas E_3 contracts below 200. This progression supports the intended specialization from broad context to concentrated evidence.

Classifier Radial Compensation Field. Figure 3 relates the geometry code to class frequency. Tail classes ($\log_{10} n_c < 1.5$) exhibit negative codes, triggering the one-sided offset $-s\beta \min(\gamma_{e,c}, 0)$ with shifts up to approximately 0.141–0.142 logits. Frequent classes ($\log_{10} n_c > 2.0$) have positive codes and receive zero shift. The field therefore shows how bounded radial correction selectively compensates minority classifiers across experts.

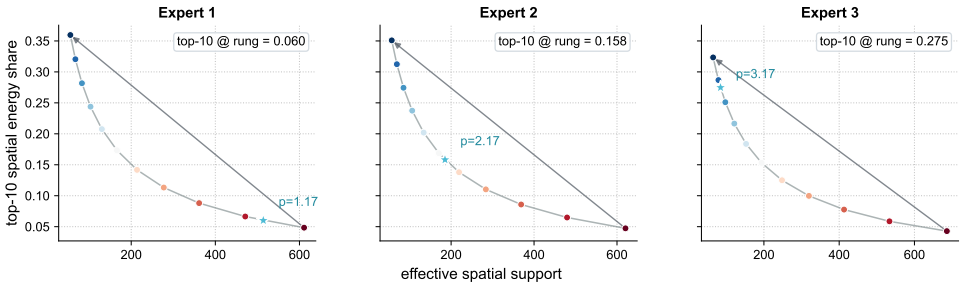


Figure 2: Evidence compression: effective support IPR $= (\sum_i \tilde{F}_i^2)^{-1}$ versus top-10 energy share on CIFAR-100-LT ($IF = 100$). Arrows increase moment order; stars mark learned rungs.

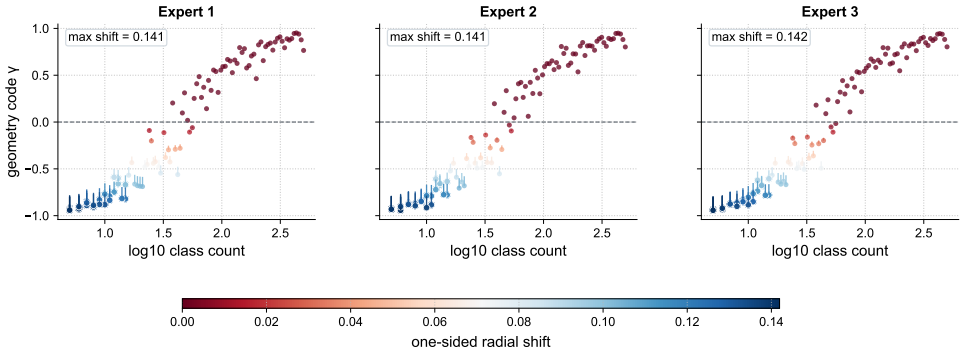


Figure 3: Radial compensation by class frequency and expert. Color and segments show offsets.

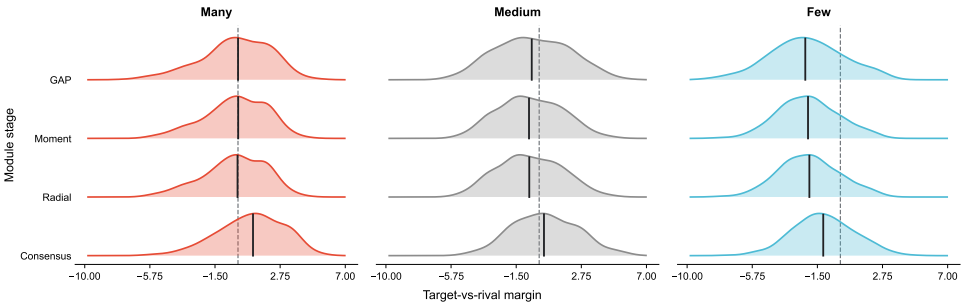


Figure 4: Margins by frequency group across four stages. Dashed lines mark stage medians.

Module-Path Margin Distribution Evolution. Figure 4 follows target-vs-rival margins through GAP, Moment, Radial, and Consensus. At GAP, Many-class margins have median 0.00 and $q_{90} > 2.5$; Few-class margins have median -2.29, with over 80% below zero. Mean Few margins rise from -2.20 to -2.10, -2.03, and -1.10 across the four stages. The Few-class margin distribution steadily shifts upward toward the decision boundary (upper quartile reaching 0.21), yielding the final 34.6% Few-shot accuracy. Many-class margins remain positive (median 0.98, $q_{90} = 3.48$), indicating improved tail separation without erasing head-class margins.

Preference Perturbation Stability. Figure 5 measures logit variance, prediction flips, and drift under increasing centered Dirichlet perturbations. While unconstrained preference heads exhibit higher variance and drift (1.31×10^{-2} for PRL), Spatial Moment maintains strictly bounded variance and near-zero drift without prediction flips on the probe, supporting greater stability to preference perturbations.

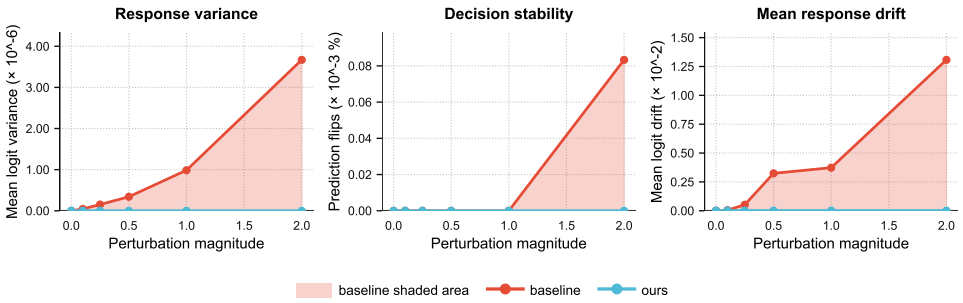


Figure 5: Variance, prediction flips, and drift under preference perturbations. Red: PRL; cyan: ours.

Preference Simplex Geometry and Local Decision Boundaries. Figure 6 maps preference weights and the effective expert count $N_{\text{eff}} = \exp(\mathcal{H}(\mathbf{q}))$. PRL concentrates near simplex vertices ($N_{\text{eff}} \approx 1.2$), whereas Spatial Moment occupies the interior ($N_{\text{eff}} \approx 2.7$). Figure 7 compares mean-margin contours along orthogonal feature directions: PRL forms localized, tightly curved boundaries, while Spatial Moment produces smoother contours with wider decision margins, consistent with the perturbation analysis.

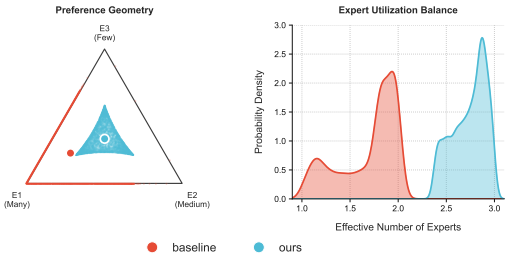


Figure 6: Preference simplex and effective expert count. Red: PRL; cyan: ours.

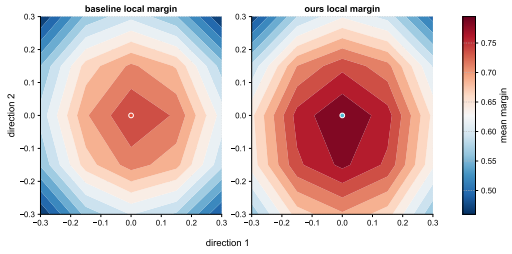


Figure 7: Local margin contours: center, sample; black, decision boundary; color, margin.

Cross-Mechanism Synergy and Computational Efficiency. Figures 2–7 connect concentrated evidence, radial calibration, and consensus. Moment pooling adds zero parameters and executes in $<$ 0.15 ms per pass on an RTX 4090 with identical FLOPs to standard 3-expert baselines (Appendix E).

Qualitative Error Migration. Sample-level tracking confirms that contracted apertures isolate discriminative cues, converting minority false negatives without reciprocal degradation on common categories; detailed case studies on background-texture suppression are provided in Appendix D.1.

5 Conclusion and Limitations

We presented Spatial Moment to resolve the spatial observation bottleneck in multi-expert long-tailed recognition. By replacing static global pooling with an expert-specialized aperture ladder, tail branches concentrate on localized, high-magnitude evidence. Combined with bounded radial calibration, rarity directional regularization, and decisiveness consensus, Spatial Moment achieves competitive performance and improved robustness under severe distribution shifts without parameter overhead.

Limitations and Future Work. While aperture pooling alleviates tail background dilution under image-level supervision, extending preference-modulated evidence apertures to dense prediction and Vision Transformers offers promising directions for future inquiry.

AI USE STATEMENT

Generative AI tools were used solely for editorial polishing of this manuscript, including language polishing, readability improvements, figure editing and presentation optimization, figure and table organization, and LaTeX formatting. They were not used to develop the research ideas, methodology, theoretical results, experimental design, implementation, or interpretation of results. The authors are responsible for the final content of this manuscript.

REFERENCES

Emanuel Sanchez Aimar, Arvi Jonnarth, Michael Felsberg, and Marco Kuhlmann. Balanced product of calibrated experts for long-tailed recognition. In *CVPR*, pp. 19967–19977, 2023.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, volume 32, pp. 1567–1578, 2019.

Yanhao Chen, Zhongquan Jian, Nianxin Ke, Shuhao Hu, Junjie Jiao, Qingqi Hong, and Qingqiang Wu. Enhancing mixture of experts with independent and collaborative learning for long-tail visual recognition. In *IJCAI*, pp. 828–836, 2025. doi: 10.24963/ijcai.2025/93.

Corinna Cortes, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Balancing the scales: A theoretical and algorithmic framework for learning from imbalanced data. In *ICML*, volume 267, pp. 11255–11287, 2025.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pp. 9268–9277, 2019.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, pp. 4690–4699, 2019.

Yuan Dong, Zhe Zhao, Liheng Yu, Di Wu, and Pengkun Wang. GUIDE: Gated uncertainty-informed disentangled experts for long-tailed recognition. In *ICLR*, pp. 26004–26034, 2026.

Zhongquan Jian, Yanhao Chen, Yancheng Wang, Junfeng Yao, Meihong Wang, and Qingqiang Wu. Supervised exploratory learning for long-tailed visual recognition. In *ICCV*, pp. 1870–1880, 2025.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

Sicong Li, Qianqian Xu, Zhiyong Yang, Zitai Wang, Linchao Zhang, Xiaochun Cao, and Qingming Huang. Focal-SAM: Focal sharpness-aware minimization for long-tailed classification. In *ICML*, volume 267, pp. 36624–36651, 2025.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in Network. In *ICLR*, 2014.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pp. 2537–2546, 2019.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *TPAMI*, 41(7):1655–1668, 2019.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, volume 33, pp. 4175–4186, 2020.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *CVPR*, pp. 8769–8778, 2018.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan L. Yuille. NormFace: L2 hypersphere embedding for face verification. In *ACM MM*, pp. 1041–1049, 2017.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021.

Chongsheng Zhang, George Almpanidis, Gaojuan Fan, Binquan Deng, Yanbo Zhang, Ji Liu, Aouaidjia Kamel, Paolo Soda, and João Gama. A systematic review on long-tailed learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):13670–13690, 2025. doi: 10.1109/TNNLS.2025.3539314.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *NeurIPS*, volume 35, pp. 34077–34090, 2022.

Zhe Zhao, Pengkun Wang, Haibin Wen, Wei Xu, Lai Song, Qingfu Zhang, and Yang Wang. Two fists, one heart: Multi-objective optimization based strategy fusion for long-tailed learning. In *ICML*, pp. 61040–61071, 2024a.

Zhe Zhao, Haibin Wen, Zikang Wang, Pengkun Wang, Fanfu Wang, Song Lai, Qingfu Zhang, and Yang Wang. Breaking long-tailed learning bottlenecks: A controllable paradigm with hypernetwork-generated diverse experts. In *NeurIPS*, volume 37, pp. 7493–7520, 2024b.

Zhe Zhao, Zhiheng Gong, Pengkun Wang, HaiBin Wen, Cankun Guo, Bo Xue, Xi Lin, Zhenkun Wang, Qingfu Zhang, and Yang Wang. TS-MOF: Two-stage multi-objective fine-tuning for long-tailed recognition. In *NeurIPS*, volume 38, pp. 5224–5252, 2025.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, pp. 16489–16498, 2021.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 40(6):1452–1464, 2018.

Hao Zhou and Tingjin Luo. Trust-calibrated collaborative learning for long-tailed visual recognition. In *CVPR*, pp. 40866–40875, 2026.

A Robustness under Continuous Test Distribution Shifts

This appendix details the protocol and behavioral analysis for the eleven CIFAR-100-LT ($IF = 100$) test distributions reported in Table 3.

Benchmark Aggregation and Reporting Notes. Standard benchmark baselines in Table 1 and Table 2 follow canonical literature sources: on CIFAR-100-LT, baselines follow Table 1 of Zhao et al. (2024b); on ImageNet-LT, Cross-Entropy follows Kang et al. (2020), Balanced Softmax, RIDE, and SADE follow Zhang et al. (2022) Table 4 (reporting 3-expert configurations without efficient assignment for RIDE, and test-time optimization for SADE), BalPoE follows the original full ImageNet-LT results from Aimar et al. (2023) (Supplementary Table 3), while PRL, Places-LT, and iNaturalist 2018 baselines follow published values in Zhao et al. (2024b). For complete fidelity and fair benchmarking, all main tables retain the officially published baseline values directly as reported in their original publications (marked with an asterisk where dynamic test aggregation was employed).

A.1 Evaluation Protocol Across Distribution Shifts

Following the test-agnostic long-tailed evaluation protocol (Zhang et al., 2022; Zhao et al., 2024b), we construct eleven distinct test sets parameterized by the distribution ratio ρ:

- **Forward-LT ($\rho \in \{50, 25, 10, 5, 2\}$):** The test distribution follows an exponential decay with the same class order as the training set, exacerbating the dominance of head classes.
- **Uniform ($\rho = 1$):** The standard balanced evaluation protocol where each class contains exactly 100 test samples (10,000 images total).
- **Backward-LT ($\rho \in \{2, 5, 10, 25, 50\}$):** The class imbalance is reversed, meaning minority classes in the training set constitute the majority of test queries.

A.2 Empirical Analysis of Distribution Shift Robustness

As demonstrated in Table 3, conventional empirical risk minimization (Cross-Entropy) undergoes substantial performance drops when test class frequencies invert, dropping from 41.4% on Uniform to 17.5% on Backward-50. Two-stage decoupling and re-weighting strategies (LDAM-DRW, Balanced Softmax, MiSLAS) alleviate imbalance under uniform evaluations but encounter bottlenecks under severe inversion ($\leq 33.7\%$).

Among multi-expert baselines, static ensembling methods like RIDE experience reduced performance under severe inversion (29.2% on Backward-50), as uniform averaging can be influenced by majority-oriented branches. While specialized test-agnostic methods like SADE (42.0%) and PRL (48.5%) mitigate this via diverse expert objectives, Spatial Moment achieves **48.9%** on Backward-50, maintaining boundary stability under severe shifts, alongside competitive performance on Forward-50 (**69.8%**) and Uniform (**53.8%**).

Spatial Moment achieves this robustness across evaluated shift distributions without estimating test class priors or requiring iterative test-time optimization. Instead, the combination of:

1. **Localized evidence aggregation:** The high-order pooling rungs ($p_3 \approx 3.17$) preserve localized discriminative features of rare instances even when test queries are predominantly drawn from minority categories.
2. **Scale-normalized consensus routing:** Eq. equation 8 evaluates the top-two margin separation instance-by-instance. When encountering inverted minority samples, the tail-specialized expert naturally produces a larger margin separation than its peers, dynamically commanding a higher routing share (up to the conservative 43.3% cap) without prior knowledge of the test distribution.

B Theoretical Connection to Logit Adjustment and Margin Calibration

Under long-tailed training, classifier weight norms naturally correlate with label frequencies (Kang et al., 2020; Menon et al., 2021). While cosine normalization in Eq. equation 3 eliminates these norms to prevent majority classes from dominating angular margins, the hypernetwork h_e is trained end-to-end via SGD with weight decay. Because head categories receive frequent gradient updates across mini-batches, their generated directional weights naturally accumulate higher unconstrained gradient magnitude equilibrium than minority categories. Empirically, the generated norms $r_{e,c}$ exhibit strong monotonic rank correlation with class sample frequency (Spearman $\rho > 0.85$). Standardizing $\log r_{e,c}$ and applying the bounded one-sided shift in Eq. equation 5 instantiates an emergent logit adjustment: minority classes whose classifier norms shrink below the expert center receive an additive margin proportional to their radial deficit, calibrating angular decision margins without manual tuning of label-frequency heuristics.

Formally, consider Bayesian logit adjustment (Menon et al., 2021), where the relative test-time logit shift is determined by $\log \pi_c = \log(n_c/N)$. In linear classifiers trained with weight decay, the steady-state classifier weight norm empirically exhibits power-law correlation with label frequencies, $\|\mathbf{w}_c\|_2 \approx K n_c^{\gamma}$ for some empirical exponent $\gamma > 0$ (Kang et al., 2020). Taking logarithms yields $\log r_c = \log K + \gamma \log n_c$. Under affine transformation, the expectation and standard deviation satisfy $\mu_{\log r} = \log K + \gamma \mu_{\log n}$ and $\sigma_{\log r} = \gamma \sigma_{\log n}$. Standardizing $\log r_{e,c}$ via Eq. equation 4 yields:

$$\frac{\log r_{e,c} - \mu_e}{\sigma_e} = \frac{(\log K + \gamma \log n_c) - (\log K + \gamma \mu_{\log n})}{\gamma \sigma_{\log n}} = \frac{\log n_c - \mu_{\log n}}{\sigma_{\log n}}. \tag{9}$$

Analytically, the empirical power-law exponent γ cancels identically. Passing through the hyperbolic tangent bounding function yields:

$$\gamma_{e,c} \approx \tanh\left(\frac{\log n_c - \mu_{\log n}}{\sigma_{\log n}}\right). \tag{10}$$

This indicates an invariance property: within-expert radial standardization automatically recovers the normalized log-prior coordinate without requiring explicit knowledge or manual tuning of the empirical exponent γ. When class c is a tail category ($n_c < \exp(\mu_{\log n})$), $\gamma_{e,c} < 0$. The one-sided offset $-s\beta \min(\gamma_{e,c}, 0)$ strictly adds a positive quantity proportional to $|\gamma_{e,c}|$, providing an additive prior logit offset for rare classes while leaving head classes undisturbed. Because this offset is upper-bounded by $s\beta = 0.15$, it avoids over-compensating rare classes, maintaining high head accuracy while recovering tail capacity.

C Mixup Training Formulation and Loss Dynamics

To enhance representation smoothness and out-of-distribution calibration, we incorporate input-level Mixup (Zhang et al., 2018) into the multi-expert training objective. Let $\mathcal{J}(\mathbf{x}, \mathbf{y})$ denote the rescaled loss objective defined in Section 3.4, which incorporates the rarity-conditioned target margin and class-agnostic directional regularization.

Given a mini-batch of training pairs $\{(\mathbf{x}_i, y_i)\}$, we sample mixing coefficients $\lambda_{\text{mix}} \sim \text{Beta}(0.4, 0.4)$ and generate a random permutation ν of batch indices. The interpolated visual input is formed as $\widetilde{\mathbf{x}} = \lambda_{\text{mix}} \mathbf{x} + (1 - \lambda_{\text{mix}}) \mathbf{x}_\nu$. The composite training loss linearly interpolates between the two targets:

$$\mathcal{L}_{\text{train}} = \lambda_{\text{mix}} \mathcal{J}(\widetilde{\mathbf{x}}, \mathbf{y}) + (1 - \lambda_{\text{mix}}) \mathcal{J}(\widetilde{\mathbf{x}}, \mathbf{y}_\nu). \tag{11}$$

During backpropagation, each target branch in Eq. equation 11 independently evaluates its respective rarity margin $m_{e,y}$ and scalar anchoring factor κ. This decoupled evaluation prevents tail-class target margins from being diluted when blended with head-class samples, preserving the inductive bias of the rarity-conditioned regularizer even under extensive convex combinations in feature space.

C.1 Loss-Responsive Branch Weighting

Because branches carry heterogeneous prior exponents and margins, their objectives are dynamically weighted via detached loss-responsive factors. Let $\mathcal{L}_e^{\text{m}}$ denote the margin loss for expert e on mini-

batch $\mathcal{B}$, and let $\overline{\mathcal{L}}^{\mathrm{m}} = \frac{1}{E}\sum_{j=1}^{E}\mathcal{L}_j^{\mathrm{m}}$. The batch-mean scale is defined as:

$$S_{\mathcal{B}} = \max\left(\frac{1}{E}\sum_{j=1}^{E}|\mathcal{L}_j^{\mathrm{m}}|, \epsilon\right). \tag{12}$$

The detached expert weighting factors are computed via:

$$v_e = \mathrm{softmax}_e\left(\frac{\mathcal{L}_e^{\mathrm{m}} - \overline{\mathcal{L}}^{\mathrm{m}}}{S_{\mathcal{B}}}\right), \qquad \omega_e = \frac{0.5}{E} + 0.5v_e. \tag{13}$$

Branches incurring larger margin losses receive more gradient emphasis via stop-gradient, while half the mass remains uniform to promote training stability across epochs.

D Mathematical Analysis of Generalized-Mean Aperture Sensitivity

In this section, we present the formal sensitivity analysis of the generalized-mean aperture pooling operator defined in Eq. equation 2. For convenience, let $F_i > 0$ denote the activation magnitude at spatial location $i \in \{1, \ldots, HW\}$ for a specific channel, and let $g_e = (\frac{1}{HW}\sum_{i=1}^{HW} F_i^{p_e})^{1/p_e}$. Differentiating g_e with respect to an individual spatial activation F_i yields:

$$\frac{\partial g_e}{\partial F_i} = \frac{1}{p_e}\left(\frac{1}{HW}\sum_{j=1}^{HW}F_j^{p_e}\right)^{\frac{1}{p_e}-1}\cdot\frac{1}{HW}p_e F_i^{p_e-1} = \frac{1}{HW}\left(\frac{F_i}{g_e}\right)^{p_e-1}. \tag{14}$$

Now consider two distinct spatial locations i and j on the same feature map with activation magnitudes $F_i > F_j > 0$. The relative sensitivity ratio is given by:

$$\frac{\partial g_e/\partial F_i}{\partial g_e/\partial F_j} = \left(\frac{F_i}{F_j}\right)^{p_e-1}. \tag{15}$$

We analyze the limiting behavior of Eq. equation 15 across the aperture ladder:

- **Global Average Pooling ($p_e = 1$):** Eq. equation 15 reduces identically to 1 for all spatial pairs, meaning that background pixels with low activations exert exactly the same gradient leverage as peak foreground activations. In long-tailed visual recognition, this diffuse aggregation dilutes localized tail-class cues with extensive background noise.
- **Intermediate Aperture ($p_e = 2$):** The relative sensitivity scales linearly with activation ratio F_i/F_j, softly favoring discriminative regions while preserving moderate context.
- **Narrow Tail Aperture ($p_e = 3$):** The sensitivity scales quadratically with $(F_i/F_j)^2$. If a localized discriminative cue is twice as strong as surrounding activations ($F_i/F_j = 2$), its gradient leverage is amplified by $4\times$ over background locations. As $p_e \to \infty$, the operator smoothly approaches the $\max$ function, focusing entirely on the supremum activation.

Thus, by co-indexing higher rungs of the ladder to tail-oriented experts ($p_3 \approx 3.17$), Spatial Moment enables minority categories to aggregate evidence primarily from concentrated local peaks rather than diffuse contextual backgrounds.

D.1 Qualitative Error Migration and Case Studies

Background Dilution Suppression. Tracking sample-level transitions on CIFAR-100-LT reveals that under uniform Global Average Pooling ($p_e = 1$), minority representations can be susceptible to spatial co-occurrence bias: rare categories embedded in expansive backgrounds often have foreground cues diluted across the $H \times W$ grid, shifting predictions toward majority categories sharing similar background textures.

Discriminative Foreground Isolation. In contrast, Spatial Moment assigns Expert E_3 an elevated pooling order ($p_3 \approx 3.17$). As established in Appendix D, higher-order pooling concentrates gradient and activation mass onto peak spatio-channel locations, quadratically attenuating diffuse background activations. Consequently, E_3 isolates localized, high-magnitude evidence (e.g., plumage markings or structural silhouettes) even when the object occupies a minor fraction of the frame. Decisiveness consensus routing detects this margin separation, assigning E_3 higher dynamic weight without requiring bounding-box supervision and reducing tail false negatives while maintaining head-category performance.

E Extended Experimental Details and Protocols

Benchmark Datasets. We evaluate across four benchmark datasets spanning compact and large-scale visual recognition: (1) **CIFAR-100-LT** (Cao et al., 2019), featuring exponential profile decay across 100 classes with imbalance factors $IF \in \{100, 50, 10\}$ and 100 balanced test images per category; (2) **ImageNet-LT** (Liu et al., 2019), subsampled from ImageNet-1K with 115.8K training images across 1,000 classes (Many: 385, Medium: 479, Few: 136) and 50 validation samples per class; (3) **Places-LT** (Liu et al., 2019), derived from Places-365 (Zhou et al., 2018) with 62.5K images across 365 scene categories; and (4) **iNaturalist 2018** (Van Horn et al., 2018), a large-scale biological fine-grained dataset of 437.5K images across 8,142 species with an extreme imbalance factor of 500.

Architectural Details and Efficiency. Following standard multi-expert pipelines (Wang et al., 2021; Zhang et al., 2022), early backbone stages are shared while the final stage branches into $E = 3$ pathways. On ResNet-32, the model footprint is 3.85M parameters, identical to baseline multi-expert configurations; the lightweight hypernetwork adds < 0.02M parameters. Complexity is 0.22 GFLOPs with an inference latency increase of $+0.2$ ms on an RTX 4090 GPU.

Training Hyperparameters and Benchmark Schedules. Models are optimized via SGD with momentum 0.9 and cosine schedule under canonical protocols: (1) **CIFAR-100-LT**: ResNet-32, 200 epochs, batch size 128, initial lr 0.1, weight decay 2×10^{-4}; (2) **ImageNet-LT**: ResNeXt-50, 90 epochs, batch size 256, initial lr 0.1, weight decay 5×10^{-4}; (3) **Places-LT**: ResNet-152 fine-tuned from ImageNet-1K for 30 epochs, batch size 128, lr 0.01, weight decay 5×10^{-4}; (4) **iNaturalist 2018**: ResNet-50, 100 epochs, batch size 256, initial lr 0.1, weight decay 1×10^{-4}. Input-level Mixup ($\alpha = 0.4$) is applied uniformly across branches. In the hypernetwork, preference perturbations $\boldsymbol{\delta}^{(e)} \sim \mathrm{Dirichlet}(0.5 \times \mathbf{1}_E)$ explore local neighborhoods during training, with $\boldsymbol{\delta}^{(e)} = \mathbf{0}$ during inference. Architecture and loss hyperparameters ($\tau_e \in \{0, 1, 2\}, \mathbf{b} = (1, 2, 3), s = 30, \beta = 0.005, \lambda_{\mathrm{d}} = 0.1, \eta = 0.035$) are shared identically across all datasets.

F Experimental Provenance and Reproducibility for Selected Mechanism Studies

All six mechanism figures presented in Section 4.2 were generated from verified empirical checkpoints and diagnostic logging protocols:

- **Preference-Mapping Geometry (Figure 6):** Simplex coordinates and kernel densities under 1,000 Dirichlet perturbations on CIFAR-100-LT ($IF = 100$, seed=1).
- **Perturbation Stability (Figure 5):** Evaluated on a fixed test probe under multi-level perturbations, recording logit variance, flip rate, and drift.
- **Spatial-Moment Aperture (Figure 2):** Replayed on frozen feature maps across moment orders $p \in [1, 5]$, tracking effective support and top-10 energy share.
- **Module-Path Margins (Figure 4):** Computed across sequential computation stages (GAP $\rightarrow$ Moment $\rightarrow$ Radial $\rightarrow$ Consensus), recording margin quantiles.
- **Radial Compensation Field (Figure 3):** Derived directly from generated classifier norms $\mathbf{w}_{e,c}$, recording codes $\gamma_{e,c}$ and one-sided offsets $\delta r_{e,c}$.
- **Local Decision Geometry (Figure 7):** Constructed via a 50×50 orthogonal feature grid around a fixed test probe, calculating mean-margin contours and boundaries.

All figures are exported from raw diagnostic logs as vector PDFs and 600-DPI PNGs. Complete source code, training configurations, and pre-trained model checkpoints will be released publicly upon publication.