BEYOND STANDALONE ACCURACY: CLASS-FREQUENCY CORRESPONDENCE SHAPES CONDITIONAL EXPERT GAIN IN LONG-TAILED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating an additional expert in multi-expert long-tailed recognition requires accounting for the correctness already covered by the existing expert set. Standalone accuracy cannot distinguish whether an auxiliary expert repeats existing coverage or recovers residual errors. We therefore define **conditional expert gain** to measure correctness contributed beyond existing coverage and establish a set-conditioned view of complementary expert value. Using this view, we identify **class-frequency correspondence** as a key factor shaping residual specialization, the distribution of newly recovered correctness across the class-frequency spectrum. Controlled correspondence interventions reveal an accuracy–gain inversion: reversing the class-to-frequency assignment raises standalone accuracy while sharply reducing conditional gain and shifting recovery away from rare classes. Class-accuracy-preserving analysis further shows that this gain shift is not explained by class-wise accuracy allocation alone, revealing an additional within-class residual-overlap contribution. Guided by this insight, we propose **Frequency-Structured Auxiliary Learning (FSAL)**, which uses a shared correspondence to orient frequency-dependent auxiliary scores within the inherited two-stage structure and expert budget, enforcing one class-wise frequency assignment across the auxiliary objective. On CIFAR-100-LT, FSAL achieves **59.5%** Overall Top-1 accuracy and **44.1%** Few-class accuracy at IR=100, with the highest Overall Top-1 among the compared methods at all three imbalance ratios. The same correspondence-dependent recovery shift extends to ImageNet-LT, while results on iNaturalist 2018 further support transfer under natural class imbalance. Code is available at **Supplement**.

1 INTRODUCTION

Long-tailed recognition (LTR) remains challenging because data-rich head classes dominate learning while tail classes receive less supervision. Existing approaches address this imbalance through frequency-aware objectives, representation learning, and label-distribution modeling (Cui et al., 2021; 2023; Hong et al., 2021; Wang et al., 2023). Multi-expert methods offer a complementary approach by encouraging experts to specialize across class regions or target distributions (Cai et al., 2021; Wang et al., 2020; Zhang et al., 2022; Aimar et al., 2023; Yang et al., 2024; Ru et al., 2024; Dong et al., 2026). Existing multi-expert work primarily studies how experts specialize or how their predictions are combined. Once an expert set is formed, however, an additional expert's value also depends on which residual errors it corrects beyond existing coverage.

Standalone accuracy counts all correct predictions, while prediction disagreement captures behavioral differences; neither reveals which residual errors an expert recovers beyond existing coverage. We therefore define **conditional expert gain** to measure this complementary correctness. As illustrated in Figure 1(a), candidates with identical standalone accuracy can provide different conditional gain when their correct predictions overlap differently with existing coverage. Under fixed common references, 17.46–26.59% of eligible cross-method orderings reverse, showing that accounting for existing coverage can change expert ordering.

Conditional gain quantifies how much complementary correctness an auxiliary expert contributes, but not where that correctness emerges across the long-tailed distribution. We refer to the distribution

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

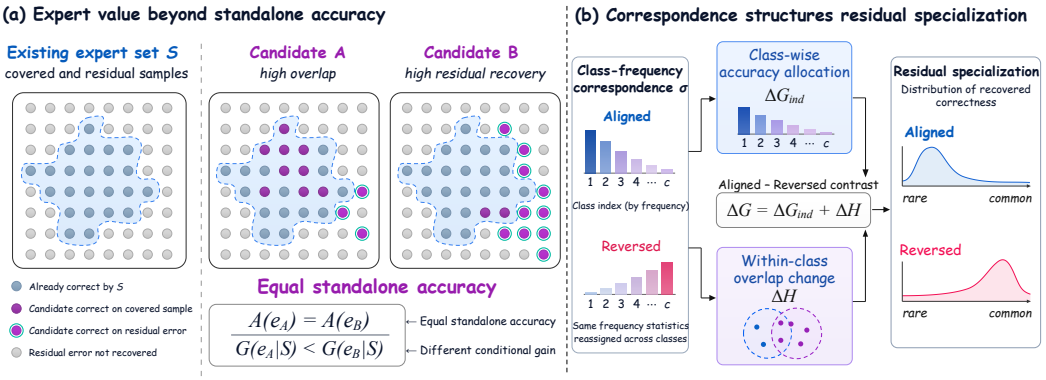


Figure 1: **Expert value and residual specialization beyond standalone accuracy.** (a) Equal-accuracy candidates can differ in conditional gain because their correctness overlaps differently with existing coverage. (b) Reassigning the same class-frequency statistics changes residual specialization; the aligned–reversed contrast separates class-wise accuracy allocation from within-class residual overlap.

of newly recovered correctness across the class-frequency spectrum as *residual specialization*. Class frequency determines the amount of supervision available to each class and therefore provides a natural structural axis for this specialization. We call the assignment between class identities and their frequency statistics *class-frequency correspondence*, and identify it as a structural factor shaping residual specialization. With the reference system fixed, reversing this correspondence raises auxiliary standalone accuracy while reducing conditional gain from **7.78% to 0.86%**; residual recovery simultaneously shifts from rarer toward more frequent classes. Graded correspondence interventions show a progressive redistribution of residual recovery as the assignment changes, while class-accuracy-preserving analysis shows that class-wise accuracy allocation alone cannot explain the gain shift, revealing an additional within-class residual-overlap contribution. Together, these results establish our central insight: **class-frequency correspondence shapes where complementary correctness emerges across the long-tailed spectrum**, rather than simply how often an auxiliary expert is correct.

Guided by this insight, we propose **Frequency-Structured Auxiliary Learning (FSAL)**, which uses a shared correspondence to orient frequency-dependent auxiliary scores within the inherited two-stage structure and expert budget, enforcing one class-wise frequency assignment throughout Stage II. Graded interventions show that residual specialization changes progressively with correspondence strength and persists across imbalance ratios, independently trained representations, and ImageNet-LT. FSAL attains the highest Overall Top-1 among the compared methods across all three CIFAR-100-LT imbalance ratios; relative to TS-MOF, the largest class-group differences occur on Medium and Few classes, while iNaturalist 2018 further supports transfer under natural class imbalance.

We summarize our contributions as follows.

- ❶ **We establish a set-conditioned view of expert value in multi-expert LTR.** Conditional expert gain measures newly contributed correctness beyond existing coverage and reveals a mismatch with standalone accuracy.
- ❷ **We identify class-frequency correspondence as a key factor shaping residual specialization in multi-expert LTR.** Controlled interventions reveal an accuracy–gain inversion and progressive redistribution across the long-tailed spectrum; class-accuracy-preserving analysis shows that class-wise accuracy allocation alone cannot explain the effect.
- ❸ **We propose FSAL to structure auxiliary-expert specialization through class-frequency correspondence.** Within the inherited two-stage framework, FSAL uses a shared correspondence to orient frequency-dependent auxiliary scores under one class-wise assignment; the complete configuration attains the highest Overall Top-1 among the compared methods across all CIFAR-100-LT imbalance ratios, with transfer to ImageNet-LT and iNaturalist 2018.

2 Related Work

► **Long-tailed recognition and frequency-aware learning.** Existing LTR methods address class imbalance through loss re-balancing, frequency-aware score adjustment, representation learning, and decoupled training. Re-weighting, margin-based losses, and prior-aware objectives use class-frequency statistics to rebalance optimization or reshape decision scores (Cui et al., 2019; Lin et al., 2017; Cao et al., 2019; Ren et al., 2020; Menon et al., 2020). PaCo, GPaCo, BCL, and ProCo improve long-tailed representations through parametric, class-balanced, or probabilistic contrastive learning (Cui et al., 2021; 2023; Zhu et al., 2022; Du et al., 2024), while ConCutMix combines contrastive learning with semantically corrected mixed-sample augmentation (Pan et al., 2024). LADE separates source-label statistics from model predictions to accommodate target-distribution shifts (Hong et al., 2021). Decoupled training separates representation learning from classifier adaptation (Kang et al., 2019), while weight balancing and unified analyses provide complementary treatments of classifier bias and frequency-aware objectives (Alshammari et al., 2022; Wang et al., 2023). Across these lines, class-frequency information is primarily used to rebalance learning or adjust predictions. Our work instead studies the class-to-frequency assignment itself and shows how it shapes residual specialization relative to an existing expert set.

► **Multi-expert specialization and complementarity.** Multi-expert LTR methods improve recognition by assigning distinct roles to multiple experts. ACE and PECEL explicitly promote complementary expert learning (Cai et al., 2021; Ru et al., 2024), whereas RIDE learns distribution-aware experts with an expert-routing mechanism (Wang et al., 2020). SADE, BalPoE, and DirMixE specialize experts for different target distributions or label-distribution variations (Zhang et al., 2022; Aimar et al., 2023; Yang et al., 2024), while PRL uses hypernetworks to generate preference-controllable experts across target distributions (Zhao et al., 2024b). MDCS promotes expert diversity, while GUIDE disentangles representation, policy, and optimization across experts (Zhao et al., 2023; Dong et al., 2026). Strategy-fusion methods instead coordinate heterogeneous expert objectives, with TS-MOF further separating representation learning from multi-objective Stage-II adaptation (Zhao et al., 2024a; 2026). These methods primarily study how experts are specialized, diversified, or combined. Our work instead evaluates expert value relative to existing coverage and shows how class-frequency correspondence shapes residual specialization.

► **Ensemble diversity and set-conditioned value.** Classical ensemble methods use disagreement, double-fault, and Q-statistic to characterize predictive diversity (Kuncheva & Whitaker, 2003; Brown et al., 2005), while neural ensembles commonly induce diversity through different models or training procedures (Krogh & Vedelsby, 1994; Lakshminarayanan et al., 2017). Beyond diversity, classifier-valuation work has used Shapley values to average the marginal contribution of a classifier across possible ensemble coalitions (Rozemberczki & Sarkar, 2021). Conditional expert gain instead measures newly covered ground-truth correctness relative to one specified existing expert set S, rather than averaging marginal contributions across possible coalitions. We use this set-conditioned measure to study how class-frequency correspondence shapes residual specialization in multi-expert LTR.

3 Methodology

Existing expert coverage defines the context of auxiliary-expert value, while *class-frequency correspondence* shapes residual specialization along the long-tailed spectrum. We first define **conditional expert gain** to measure complementary correctness relative to a specified expert set, and then introduce **Frequency-Structured Auxiliary Learning (FSAL)**, which uses a shared class-frequency correspondence to structure auxiliary-expert training.

3.1 Set-Conditioned Expert Value

► **Conditional gain.** Let (x, y) follow the evaluation distribution. For an expert e, define its correctness set as $\mathcal{C}_e = \{(x, y) : e(x) = y\}$. For an existing expert set S, let $\mathcal{C}_S = \bigcup_{s \in S} \mathcal{C}_s$ denote the samples already covered by at least one expert in S, and let $B(S) = \Pr(\mathcal{C}_S)$ denote the corresponding coverage.

Definition 1 (Conditional expert gain)**.** *The conditional gain of a candidate expert e relative to an existing expert set S is*

$$G(e \mid S) = \Pr(\mathcal{C}_e \cap \mathcal{C}_S^c) = B(S \cup \{e\}) - B(S). \tag{1}$$

Coverage interpretation. Equation 1 measures the increase in correctness coverage obtained by adding e to the current expert set. Here, "conditional" refers to dependence on the reference set S, rather than normalization by the size of the residual pool. In contrast, standalone accuracy $A(e) = \Pr(\mathcal{C}_e)$ measures all correct predictions made by the candidate, including those already covered by S.

Correctness overlap. Equivalently, $G(e \mid S) = A(e) - \Pr(\mathcal{C}_e \cap \mathcal{C}_S)$. The second term is precisely the correctness overlap ignored by standalone evaluation. Fixing $A(e)$ therefore fixes the size of the candidate's correctness set but not its overlap with existing coverage. Two candidates with identical standalone accuracy can consequently provide different amounts of new correctness, as illustrated in Figure 1(a).

Residual recovery. Conditional gain measures the absolute probability mass of newly recovered samples. When reference sets leave residual pools of different sizes, we additionally report *residual recovery*. For $B(S) < 1$,

$$R(e \mid S) = \Pr(\mathcal{C}_e \mid \mathcal{C}_S^c) = \frac{G(e \mid S)}{1 - B(S)}. \tag{2}$$

Thus, $G(e \mid S)$ measures the absolute increase in coverage, whereas $R(e \mid S)$ measures the recovered fraction of the residual pool; for fixed S, they induce the same candidate ordering.

► **Set dependence.** Conditional gain is a property of the pair (e, S) rather than an intrinsic score of e. Let $a = A(e)$ and $b = B(S)$. Two direct properties follow.

$$\textit{Gain bounds:} \qquad \max(0, a - b) \le G(e \mid S) \le \min(a, 1 - b). \tag{3}$$

$$\textit{Reference-set monotonicity:} \qquad S \subseteq T \Longrightarrow G(e \mid T) \le G(e \mid S). \tag{4}$$

Equation 3 bounds the amount of correctness that can lie outside the current coverage. Equation 4 follows because enlarging the reference set can only shrink the residual region available to the same candidate.

The same candidate can therefore have different conditional gain under different reference sets despite unchanged standalone predictions. Accordingly, our controlled correspondence analyses use a fixed reference set, enabling candidate gains to be compared against the same existing coverage. The coverage function $B(S)$ is also monotone submodular, with conditional gain as its marginal contribution; complete derivations are provided in Appendix B.1.

Class-wise gain decomposition. Let $\pi_c = \Pr(Y = c)$ be the evaluation prior, $d_c(S) = \Pr(\mathcal{C}_S^c \mid Y = c)$ the class-wise residual rate, and $a_c(e) = \Pr(\mathcal{C}_e \mid Y = c)$ the candidate's class-wise accuracy. Define

$$G_{\mathrm{ind}}(e \mid S) = \sum_c \pi_c d_c(S) a_c(e), \qquad H(e; S) = G(e \mid S) - G_{\mathrm{ind}}(e \mid S). \tag{5}$$

Here, G_{ind} is the class-conditional independence reference and H captures the remaining within-class dependence. For two correspondence assignments under the same reference, $\Delta G = \Delta G_{\mathrm{ind}} + \Delta H$, separating class-wise accuracy allocation from additional within-class residual-overlap change.

3.2 Frequency-Structured Auxiliary Learning (FSAL)

► **Design principle.** Conditional gain measures how much new correctness an auxiliary expert contributes; FSAL addresses where it emerges by using one class-to-frequency assignment to orient the Stage-II auxiliary scores. The complete configuration carries the same assignment into deferred weighting. Within the TS-MOF two-stage framework (Zhao et al., 2026), FSAL modifies only the Stage-II auxiliary objective, preserving the frozen Stage-I representation and expert budget.

❶ Class-frequency correspondence. Let n_c denote the number of training samples of class c. We represent its empirical frequency using the class prior p_c and a normalized log-frequency coordinate r_c,

$$p_c = \frac{n_c}{\sum_j n_j}, \qquad r_c = \frac{\log n_c - \min_j \log n_j}{\max_j \log n_j - \min_j \log n_j + \epsilon}, \qquad \Phi_\sigma(c) = \big(p_{\sigma(c)}, r_{\sigma(c)}\big). \tag{6}$$

Here, $r_c \in [0, 1]$, with smaller values corresponding to rarer classes, and σ is a bijection over class indices that assigns frequency information to class identities. The structural object in FSAL is this shared assignment Φ_σ, rather than either frequency statistic in isolation. The identity mapping $\sigma(c) = c$ gives the natural class-frequency correspondence; alternative mappings reassign the same frequency statistics across class identities while preserving their multiset. The bijective assignment and multiset-preservation property are formalized in Appendix C.1.

Role of the two statistics. The prior p_c retains the empirical class-probability scale and is used for prior-aware score adjustment and, in the complete configuration, deferred sample weighting, whereas the normalized coordinate r_c provides a smooth frequency axis for class-dependent temperature and margin functions. For compactness, define

$$p_c^\sigma = p_{\sigma(c)}, \qquad r_c^\sigma = r_{\sigma(c)}. \tag{7}$$

❷ Correspondence-indexed score transformation. FSAL first uses the shared correspondence to transform the auxiliary training scores. Define

$$\tau(r) = \tau_{\min} + \tau_{\text{span}} r, \qquad q(p) = \log(p + \epsilon), \qquad m(r) = m_{\max}(1 - r)^\gamma + m_{\text{floor}}, \tag{8}$$

where $\tau_{\min}$ and τ_{span} determine the temperature profile, while $m_{\max}$, m_{floor}, and γ determine the margin profile.

Shared correspondence. A single correspondence Φ_σ jointly indexes all frequency-dependent terms in the auxiliary score transformation.

For sample i with auxiliary logits $\mathbf{z}_i$ and target y_i, the correspondence-indexed score of class c is

$$\widetilde{z}_{ic} = s \left[\frac{z_{ic}}{\tau(r_c^\sigma)} + \lambda q(p_c^\sigma) - \mathbf{1}[c = y_i]\, m(r_{y_i}^\sigma) \right], \tag{9}$$

where λ controls the strength of the prior adjustment and $s > 0$ is a global score scale.

Transformation effects. The temperature $\tau(r_c^\sigma)$ controls the class-wise logit scale along the frequency axis, the prior term $\lambda q(p_c^\sigma)$ introduces a frequency-dependent score offset, and the target margin $m(r_{y_i}^\sigma)$ increases the separation required for rarer targets. Because all three terms inherit the same Φ_σ, this shared correspondence jointly determines their class-wise orientation.

❸ Deferred weighting in the complete configuration. The same correspondence also controls the sample-level weighting used during Stage II. For Stage-II epoch t, define

$$w_c^\sigma = \frac{(p_c^\sigma + \epsilon)^{-1}}{\frac{1}{C}\sum_{j=1}^{C}(p_j^\sigma + \epsilon)^{-1}}, \qquad \bar{w}_c^{(t,\sigma)} = \begin{cases} 1, & t < t_{\text{rw}}, \\ w_c^\sigma, & t \geq t_{\text{rw}}, \end{cases} \tag{10}$$

where C is the number of classes, t_{rw} denotes the epoch at which deferred reweighting is activated, and the inverse-prior weights are normalized to have unit mean across classes.

Deferred emphasis. Weights are uniform before t_{rw} and follow w_c^σ thereafter. Applied to the transformed scores in Equation 9, they yield

$$\mathcal{L}_{\text{CF}} = \frac{\sum_{i \in \mathcal{B}} \bar{w}_{y_i}^{(t,\sigma)}\, \ell_{\text{CE}}(\widetilde{\mathbf{z}}_i, y_i)}{\sum_{i \in \mathcal{B}} \bar{w}_{y_i}^{(t,\sigma)}}. \tag{11}$$

Equation 9 defines the core correspondence-indexed score transformation; Equations 10–11 specify the deferred weighting used by the complete configuration. Both use the same class-frequency correspondence Φ_σ. The complete Stage-II procedure is given in Appendix C.2.

► **Relation to existing objectives and expert integration.** FSAL's structural contribution is the shared correspondence that gives prior-aware adjustment (Ren et al., 2020; Menon et al., 2020), class-dependent score shaping (Cao et al., 2019), and complete-configuration deferred weighting (Cui et al., 2019) a common class-wise orientation. Inference follows TS-MOF (Zhao et al., 2026) with the same expert budget and prediction procedure.

4 EXPERIMENTS

We evaluate FSAL on three standard long-tailed recognition benchmarks. We first report recognition performance and then analyze expert value, correspondence strength, target-prior dependence, representation robustness, and transfer.

Table 1: **Top-1 accuracy (%) on CIFAR-100-LT.** Overall accuracy is reported at IR=10, 50, and 100, with Many/Medium/Few at IR=100. Baselines follow the unified ResNet-34 protocol; bold/underline denote best/second-best.

Method	Overall accuracy			IR=100		
	IR=10	IR=50	IR=100	Many	Medium	Few
CE (He et al., 2016)	67.1	52.7	47.0	77.1	45.9	13.1
CE-DRW (Cao et al., 2019)	68.1	55.1	49.2	75.3	49.1	18.9
LDAM-DRW (Cao et al., 2019)	63.4	50.4	42.8	67.3	41.4	15.9
Balanced Softmax (Ren et al., 2020)	68.3	56.7	51.3	76.4	50.9	22.5
LADE (Hong et al., 2021)	67.9	56.0	51.1	73.7	47.7	28.7
DBM (Son et al., 2025)	65.3	50.9	46.0	75.8	44.5	12.8
KPS (Li et al., 2022)	65.8	54.1	50.0	72.8	49.5	23.9
PaCo (Cui et al., 2021)	71.4	<u>60.8</u>	54.6	77.6	55.4	26.7
BCL (Zhu et al., 2022)	68.5	55.7	51.2	75.6	51.6	22.3
ProCo (Du et al., 2024)	72.2	60.7	55.1	**79.3**	53.2	29.1
Focal-SAM (Li et al., 2025)	68.7	56.7	51.0	76.2	56.0	15.7
LOS (Sun et al., 2025)	68.8	57.4	53.2	72.5	48.7	35.8
SAMN (Zhang et al., 2026)	66.1	54.0	49.8	70.9	51.1	23.7
Expert-based methods						
BalPoE (Aimar et al., 2023)	<u>72.4</u>	60.5	55.9	77.7	**57.1**	29.2
SHIKE (Jin et al., 2023)	65.1	51.2	42.2	70.8	40.8	10.6
ICL (Chen et al., 2025)	66.1	53.4	49.0	73.3	49.9	19.7
SADE (Zhang et al., 2022)	69.4	57.9	52.8	73.4	56.2	24.7
PRL (Zhao et al., 2024b)	71.8	60.3	55.1	75.9	<u>56.5</u>	29.3
RIDE (Wang et al., 2020)	68.8	55.5	50.5	75.3	51.6	20.4
TS-MOF (Zhao et al., 2026)	70.8	60.2	<u>56.8</u>	<u>79.0</u>	49.2	<u>39.9</u>
FSAL	**73.5**	**64.1**	**59.5**	78.1	54.0	**44.1**

4.1 EXPERIMENTAL SETUP

► **Datasets.** We evaluate FSAL on CIFAR-100-LT (Krizhevsky et al., 2009; Cao et al., 2019), ImageNet-LT (Liu et al., 2019), and iNaturalist 2018 (Van Horn et al., 2018). CIFAR-100-LT serves as the primary benchmark for recognition and correspondence analysis and is evaluated at imbalance ratios $\mathrm{IR} \in \{10, 50, 100\}$, where IR denotes the ratio between the largest and smallest training-class counts. ImageNet-LT extends the analysis to a larger long-tailed setting, while iNaturalist 2018 evaluates transfer under naturally occurring class imbalance. Detailed dataset definitions, class partitions, and evaluation settings are provided in <u>Appendix A.1</u>.

► **Backbones and baselines.** We use ResNet-34 for CIFAR-100-LT and ResNet-50 for ImageNet-LT and iNaturalist 2018 (He et al., 2016). We compare FSAL with representative methods spanning loss adjustment, representation learning, multi-expert learning, and two-stage adaptation; Tables 1 and 2 report the main comparisons. For the standard two-stage setting, Stage-I representations are frozen during Stage II, where the residual decoder and three expert heads are optimized for 20 epochs with SGD. Unless otherwise stated, repeated analyses use three Stage-II seeds with a fixed Stage-I

representation and report mean and sample SD; independent Stage-I analyses are identified separately. Full training and evaluation settings and implementation details are provided in Appendix A.

► **Metrics.** Following standard LTR evaluation, we report Overall Top-1 accuracy together with Many, Medium, and Few accuracies, where classes contain more than 100, 20–100, and fewer than 20 training samples, respectively. For expert-level analysis, we additionally report conditional gain and residual recovery from Eqs. 1 and 2.

To characterize where newly recovered correctness is concentrated along the class-frequency spectrum, we define rare-class enrichment as

$$E_{\mathrm{rare}}(e \mid S) = \frac{\Pr(D_{\mathrm{rare}} \mid \mathcal{C}_e \cap \mathcal{C}_S^c)}{\Pr(D_{\mathrm{rare}} \mid \mathcal{C}_S^c)}, \tag{12}$$

where D_{rare} contains the rarest three frequency deciles (D1–D3). The denominator accounts for the rare-class share already present in the residual pool, so the statistic reflects where newly recovered correctness is concentrated rather than how many rare samples remain uncovered. Values above $1\times$ indicate that these classes account for a larger share of newly recovered samples than of the residual pool itself. This decile-based definition is distinct from the threshold-based Few split.

4.2 Benchmark Results

► **CIFAR-100-LT.** As shown in Table 1, FSAL attains higher Overall Top-1 than TS-MOF across all three imbalance ratios, with the largest difference at IR=50 (64.1% vs. 60.2%). Among all compared methods, FSAL achieves the highest Overall Top-1 accuracy across all three imbalance ratios. At IR=100, FSAL reaches 54.0% Medium and 44.1% Few accuracy, compared with 49.2% and 39.9% for TS-MOF, while maintaining comparable Many-class accuracy. The largest class-group differences relative to TS-MOF occur in Medium and Few classes. Full class-group breakdowns are provided in Appendix D.3.

► **ImageNet-LT and iNaturalist 2018.** Table 2 reports the highest Overall and Few accuracies among the compared methods for FSAL on both large-scale benchmarks. On iNaturalist 2018, FSAL reaches 74.7% Few accuracy, compared with the strongest reference value of 71.3%. These results support transfer beyond CIFAR-100-LT. Section 4.3 further examines correspondence-dependent recovery on ImageNet-LT under controlled intervention; full large-scale class-group results are provided in Appendix D.4.

Table 2: **Top-1 accuracy (%) on ImageNet-LT and iNaturalist 2018.** Reference baselines follow Zhao et al. (2026); bold/underline denote best/second-best.

Method	ImageNet-LT		iNaturalist 2018	
	Overall	Few	Overall	Few
CE (He et al., 2016)	41.6	5.8	61.0	55.5
CE-DRW (Cao et al., 2019)	50.1	28.8	67.0	66.4
cRT (Kang et al., 2019)	47.3	26.1	65.2	63.2
LDAM-DRW (Cao et al., 2019)	49.8	30.7	66.1	–
BS (Ren et al., 2020)	51.0	32.1	67.3	67.5
KPS (Li et al., 2022)	52.3	35.9	69.6	70.2
RIDE (Wang et al., 2020)	53.6	34.4	70.6	70.4
BCL (Zhu et al., 2022)	55.6	36.3	70.9	71.3
LOS (Sun et al., 2025)	54.4	42.3	70.8	71.3
TS-MOF (Zhao et al., 2026)	56.3	42.6	72.3	70.5
FSAL	**56.8**	**43.4**	**73.1**	**74.7**

4.3 Analysis and Ablations

► **Expert value and residual localization.** Under fixed common references, 17.46–26.59% of eligible cross-method orderings reverse across imbalance ratios, showing that standalone accuracy and conditional gain can order candidates differently against the same existing coverage. Figure 2(a) further shows that prediction disagreement is informative but does not determine expert complementarity; similar disagreement can correspond to markedly different pairwise utility.

Figure 2(b,c) shows that residual specialization is frequency-localized: alignment concentrates recovery in rare deciles, reversal shifts it toward more frequent classes, and permutation lies between. Here E_d normalizes each decile's recovered-sample share by its residual-pool share, complementing the D1–D3 aggregate E_{rare} in Eq. 12. Thus, complementarity also differs in where new correctness emerges. Together, these diagnostics separate the magnitude of complementary correctness from its

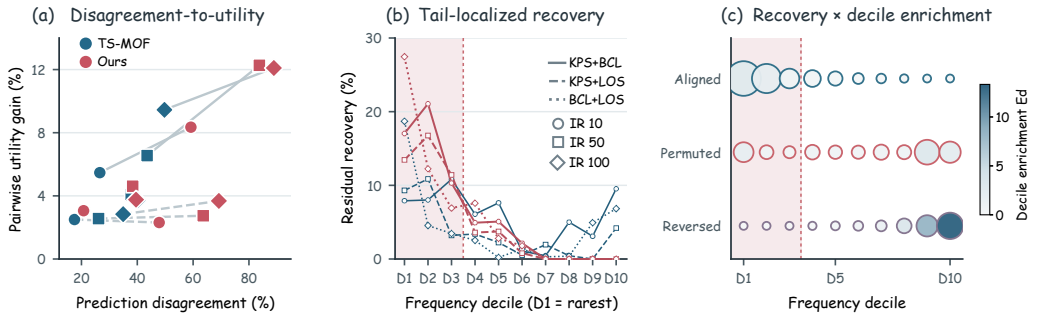


Figure 2: **Expert complementarity beyond predictive disagreement.** (a) Pairwise utility gain versus prediction disagreement across expert pairs and imbalance ratios. (b) Residual recovery by frequency decile (D1 = rarest; D1–D3 = rare region). (c) Recovery and decile enrichment E_d under aligned, permuted, and reversed class-frequency correspondence; bubble size denotes recovery and color denotes E_d.

frequency allocation across the long-tailed spectrum. This frequency-resolved view motivates testing whether the class-to-frequency assignment itself shapes where residual recovery emerges.

► **Class-frequency correspondence shapes residual specialization.** To isolate class-frequency correspondence, we keep the representation, decoder, and reference experts fixed and vary only the class-frequency assignment used by the auxiliary objective. Table 3 shows an accuracy–gain inversion: reversing correspondence raises standalone accuracy from 22.85% to 36.92%, yet reduces conditional gain from 7.78% to 0.86% and Few recovery from 26.88% to 0.01%, with permutation in between. Because the reference predictions are fixed, the comparison uses the same residual pool, so correspondence changes both the amount and frequency location of newly recovered correctness relative to existing coverage.

Table 3: **Class-frequency correspondence at IR=100.** Three-seed mean ± SD; fixed-reference coverage 47.57 ± 0.24%; percentages except E_{rare}; bold marks row maxima.

Metric	Aligned	Permuted	Reversed
Stand. acc.	22.85 ± 1.38	31.86 ± 1.85	**36.92** ± 1.30
Cond. gain	**7.78** ± 0.20	3.33 ± 0.58	0.86 ± 0.04
Residual recov.	**14.83** ± 0.31	6.36 ± 1.12	1.65 ± 0.08
Few recov.	**26.88** ± 1.90	6.49 ± 1.15	0.01 ± 0.02
E_{rare} (×)	**1.81** ± 0.10	1.02 ± 0.11	0.01 ± 0.01

Figure 3 shows the score-level counterpart of Table 3. Margin profiles remain close on Many residuals but separate progressively on Medium and Few classes: alignment yields larger true-class margins, reversal smaller margins, and permutation lies between. The same ordering persists across five correspondence levels (Appendix D.2).

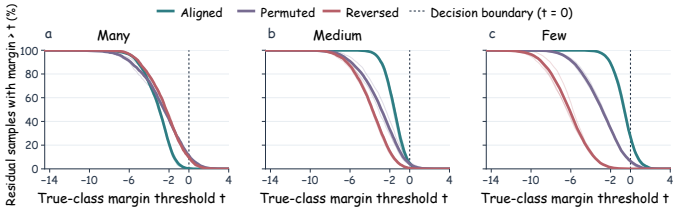


Figure 3: **Correspondence reshapes residual margins.** Share of residuals above true-class margin t; dashed line marks $t = 0$.

Taken together, Table 3 and Figure 3 show that correspondence shapes both the amount and location of newly recovered correctness. A class-wise decomposition further shows that class-wise accuracy allocation does not fully explain the aligned–reversed gain shift: $\Delta G = 3.33 \pm 0.28\%$, with $\Delta G_{\text{ind}} = 2.84 \pm 0.46\%$ and $\Delta H = 0.48 \pm 0.18\%$. The same gain ordering recurs across related frequency-aware objectives.

► **Correspondence strength and evaluation dependence.** Having established correspondence-driven residual specialization, we next test whether the effect varies continuously with assignment strength and persists across independently trained Stage-I representations. We combine a five-level

intervention with matched aligned–reversed comparisons across three representations; cross-objective recurrence is reported in Appendix C.3.

Table 4 shows a graded effect: conditional gain rises from 0.81% to 7.70% and Few recovery from 0.00% to 26.78% as ρ moves from reversed to aligned. Across three Stage-I representations, the aligned–reversed gap remains 6.71–6.88 points in gain and 25.02–26.20 points in Few recovery, showing dose-responsive and representation-stable specialization.

Table 4: **Correspondence strength and representation robustness at IR=100.** Five-level sweep and cross-Stage-I Aligned-minus-Reversed effects.

(a) ρ	Overall	Gain	Recovery	Few recov.
-1.000	47.90	0.81	1.54	0.00
-0.436	49.21	2.27	4.34	2.60
0.000	50.48	3.44	6.58	7.37
+0.431	52.41	5.55	10.61	12.28
+1.000	**54.52**	**7.70**	**14.72**	**26.78**
(b) Stage-I seed		Δ Overall	Δ Gain	Δ Few recov.
0		6.57	6.86	26.20
42		6.68	6.88	25.88
200		6.36	6.71	25.02

Because conditional expert value also depends on the evaluation distribution, we next vary the target prior while keeping predictions fixed; this changes only the class-wise weighting of recovered correctness.

Table 5: **Target-prior dependence of conditional gain (%) at IR=100.** F/B denote Forward-/Backward-LT priors.

Correspondence ρ	Forward-LT					Uni.	Backward-LT				
	F50	F25	F10	F5	F2	U	B2	B5	B10	B25	B50
$+1.00$	1.2	1.7	2.8	3.9	**5.9**	**7.7**	**9.8**	**12.7**	**15.0**	**17.8**	**19.8**
$+0.50$	**2.8**	**3.1**	**3.5**	**4.0**	4.8	5.6	6.4	7.5	8.3	9.3	10.0
0.00	2.0	2.2	2.4	2.6	3.1	3.4	3.9	4.5	5.0	5.7	6.2
-0.50	2.2	2.2	2.2	2.2	2.3	2.3	2.3	2.4	2.4	2.5	2.6
-1.00	1.6	1.5	1.3	1.2	1.0	0.8	0.7	0.5	0.4	0.3	0.2

Table 5 shows that conditional expert value varies with the evaluation prior. Full alignment leads under the uniform and Backward-LT priors, whereas the $+0.50$ mapping leads under F50. Thus, correspondence shapes where correctness is recovered, while the evaluation distribution changes how that recovery is valued.

► **Robustness and transfer.** On ImageNet-LT, reversal reduces Few recovery from 18.96% to 0.00% while increasing Many recovery from 0.16% to 5.74%. The same pattern persists at lower imbalance ratios, across independent Stage-I representations, and throughout Stage II; ΔH remains positive with ResNet-18, while ImageNet-LT reproduces the rare-to-frequent shift. Across these settings, residual recovery is consistently redistributed along the class-frequency spectrum.

5 Conclusion

This paper established a set-conditioned view of expert value in multi-expert long-tailed recognition. Conditional expert gain measures correctness beyond existing coverage, while our analyses show that standalone accuracy and prediction disagreement are incomplete proxies for complementarity. Controlled correspondence interventions identify class-frequency correspondence as a key factor shaping residual specialization: reversal raises standalone accuracy while sharply reducing conditional gain and shifting recovery away from rare classes. A class-wise decomposition further shows that class-wise accuracy allocation does not fully explain this shift, leaving an additional within-class residual-overlap contribution. Graded interventions further show that residual specialization changes progressively with correspondence strength and persists across imbalance levels, representations, and ImageNet-LT. Guided by this insight, FSAL structures auxiliary-expert training through shared correspondence and attains competitive recognition results across CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018, linking complementary expert design to existing coverage and class-frequency correspondence.

AI USE STATEMENT

Generative AI tools were used for language editing and polishing. All scientific content, experimental results, mathematical claims, and final manuscript content were checked and approved by the authors.

REPRODUCIBILITY STATEMENT

The main paper specifies the model formulation, evaluation metrics, and principal experimental settings. Appendix A provides the complete dataset definitions, training configuration, evaluation protocols, and statistical settings, while Appendix C details the FSAL objective, Stage-II procedure, objective analyses, correspondence evaluations, and sensitivity studies. The accompanying code and supplementary material provide the corresponding implementation and additional experimental details needed to reproduce the reported results.

REFERENCES

Emanuel Sanchez Aimar, Arvi Jonnarth, Michael Felsberg, and Marco Kuhlmann. Balanced product of calibrated experts for long-tailed recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19967–19977. IEEE, 2023.

Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6887–6897. IEEE, 2022.

Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information fusion*, 6(1):5–20, 2005.

Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 112–121. IEEE, 2021.

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

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.

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

Jiequan Cui, Zhisheng Zhong, Zhuotao Tian, Shu Liu, Bei Yu, and Jiaya Jia. Generalized parametric contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 7463–7474, 2023.

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

Yuan Dong, Zhe Zhao, Liheng Yu, Di Wu, and Pengkun Wang. Guide: Gated uncertainty-informed disentangled experts for long-tailed recognition. In *International Conference on Learning Representations*, volume 2026, pp. 26004–26034, 2026.

Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9): 5890–5904, 2024.

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

Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6622–6632. IEEE, 2021.

Yan Jin, Mengke Li, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23695–23704. IEEE, 2023.

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

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

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Mengke Li, Yiu-Ming Cheung, and Zhikai Hu. Key point sensitive loss for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4812–4825, 2022.

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. *arXiv preprint arXiv:2505.01660*, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

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

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

Haolin Pan, Yong Guo, Mianjie Yu, and Jian Chen. Enhanced long-tailed recognition with contrastive cutmix augmentation. *IEEE Transactions on Image Processing*, 33:4215–4230, 2024.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

Benedek Rozemberczki and Rik Sarkar. The shapley value of classifiers in ensemble games. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 1558–1567, 2021.

Lixiang Ru, Xin Guo, Lei Yu, Yingying Zhang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yansheng Li, and Ming Yang. Parameter-efficient complementary expert learning for long-tailed visual recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5393–5402, 2024.

Minseok Son, Inyong Koo, Jinyoung Park, and Changick Kim. Difficulty-aware balancing margin loss for long-tailed recognition. In *Proceedings of the aaai conference on artificial intelligence*, volume 39, pp. 20522–20530, 2025.

Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan. Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In *International Conference on Learning Representations*, volume 2025, pp. 26681–26697, 2025.

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 *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8769–8778. IEEE, 2018.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. *Advances in Neural Information Processing Systems*, 36:48417–48430, 2023.

Zhiyong Yang, Qianqian Xu, Zitai Wang, Sicong Li, Boyu Han, Shilong Bao, Xiaochun Cao, and Qingming Huang. Harnessing hierarchical label distribution variations in test agnostic long-tail recognition. In *Forty-first International Conference on Machine Learning*, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Shuo Zhang, Chenqi Li, and Tingting Zhu. Why not hyperparameter-friendly optimisation? a monotonic adaptive norm rescaling approach for long-tailed recognition. *arXiv preprint arXiv:2606.02526*, 2026.

Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in neural information processing systems*, 35:34077–34090, 2022.

Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu. Mdcs: More diverse experts with consistency self-distillation for long-tailed recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11563–11574. IEEE, 2023.

Zhe Zhao, Pengkun Wang, HaiBin Wen, Wei Xu, Song Lai, Qingfu Zhang, and Yang Wang. Two fists, one heart: Multi-objective optimization based strategy fusion for long-tailed learning. In *Forty-first International Conference on Machine Learning*, 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. *Advances in Neural Information Processing Systems*, 37:7493–7520, 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. *Advances in Neural Information Processing Systems*, 38:5224–5252, 2026.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6898–6907. IEEE, 2022.

Appendix

Beyond Standalone Accuracy: Class-Frequency Correspondence Shapes Conditional Expert Gain in Long-Tailed Recognition

The Appendix is organized as follows:

- **Section A: Experimental Settings.** We provide the dataset definitions, evaluation settings, Stage-II implementation, and auxiliary evaluation protocols used in our experiments.
  - Section A.1: Datasets and Evaluation
  - Section A.2: Implementation Details
  - Section A.3: Evaluation Protocols
- **Section B: Expert Value, Theory, and Decomposition.** We derive the properties of conditional expert gain, its submodular coverage structure, relation to coalition-based valuation, class-wise gain decomposition, and the distinction between correctness coverage and aggregation-based prediction.
  - Section B.1: Direct Properties, Submodularity, and Coalition Valuation
  - Section B.2: Class-Wise Gain Decomposition
- **Section C: FSAL Objective, Training, and Analysis.** We present the correspondence assignment, Stage-II training procedure, cross-objective correspondence analysis, decomposition controls, and local sensitivity analysis.
  - Section C.1: Correspondence Assignment
  - Section C.2: FSAL Stage-II Training Procedure
  - Section C.3: Correspondence Generality and Decomposition
  - Section C.4: Hyperparameter Sensitivity
- **Section D: Extended Empirical Analysis.** We examine expert-value diagnostics, residual localization, correspondence strength, evaluation priors, training imbalance, representation robustness, training progress, and large-scale transfer.
  - Section D.1: Expert Value and Residual Localization
  - Section D.2: Correspondence Strength and Residual Geometry
  - Section D.3: Robustness Across Evaluation and Training Conditions
  - Section D.4: Large-Scale Transfer

A Experimental Settings

This section provides additional details for the experimental setup in Section 4.1, including the datasets, evaluation settings, Stage-II implementation, and evaluation protocols used by the auxiliary analyses.

A.1 Datasets and Evaluation

We evaluate FSAL on CIFAR-100-LT (Krizhevsky et al., 2009; Cao et al., 2019), ImageNet-LT (Liu et al., 2019), and iNaturalist 2018 (Van Horn et al., 2018). Table 6 summarizes the benchmark configurations.

Table 6: **Dataset and evaluation settings.** CIFAR-100-LT is evaluated at three imbalance ratios, while ImageNet-LT and iNaturalist 2018 use their standard long-tailed splits.

Dataset	Classes	Backbone	Setting
CIFAR-100-LT	100	ResNet-34	IR $\in \{10, 50, 100\}$
ImageNet-LT	1,000	ResNet-50	Standard LT split
iNaturalist 2018	8,142	ResNet-50	Natural LT split

For CIFAR-100-LT, the imbalance ratio is defined as

$$\mathrm{IR} = \frac{\max_c n_c}{\min_c n_c}, \tag{13}$$

where n_c denotes the number of training samples in class c.

Following standard LTR evaluation, classes are divided into Many, Medium, and Few groups according to their training sample counts:

$$\mathcal{C}_{\mathrm{Many}} = \{c : n_c > 100\}, \qquad \mathcal{C}_{\mathrm{Medium}} = \{c : 20 \le n_c \le 100\}, \qquad \mathcal{C}_{\mathrm{Few}} = \{c : n_c < 20\}. \tag{14}$$

CIFAR-100-LT at IR=10 contains no Few classes under this definition, so Few-class accuracy is not reported. For frequency-resolved analyses, we additionally rank classes by n_c and divide them into ten deciles, with D1 denoting the rarest decile. D1–D3 define the rare region used by E_{rare} and are distinct from the threshold-based Few group.

For ImageNet-LT and iNaturalist 2018, the complete large-scale tables use Head/Medium/Tail terminology, while the compact main-text discussion refers to the corresponding high- and low-frequency groups as Many/Few.

Set-conditioned and target-prior analyses are applied post hoc to fixed prediction outputs after training.

A.2 Implementation Details

Table 7 summarizes the Stage-II implementation used in our experiments.

Table 7: **Stage-II implementation details.**

Setting	Configuration
Stage-I representation	Frozen
Trainable Stage-II modules	Residual decoder (evaluation mode, trainable parameters) and three expert heads
Optimizer	SGD, momentum 0.9, weight decay 0
Stage-II epochs	20
Batch size / workers	64 / 16
Learning rate	5×10^{-4}
LR schedule	CosineAnnealingLR
FSAL default parameters	$\lambda = 1.5$, $\tau_{\min} = 0.8$, $\tau_{\mathrm{span}} = 0.5$, $m_{\max} = 0.6$, $m_{\mathrm{floor}} = 0$, $\gamma = 1$, $s = 34$
Deferred reweighting	Activated from epoch 17 through epoch 20
LOS label smoothing	0.98
Gradient coordination	Single PCGrad-based coordination step following TS-MOF (Yu et al., 2020; Zhao et al., 2026)
Stage-II validation	Follows the TS-MOF protocol (Zhao et al., 2026)
Inference protocol	Same three-expert prediction protocol as TS-MOF (Zhao et al., 2026)

Unless otherwise stated, repeated analyses use three Stage-II seeds with a fixed Stage-I representation and report mean and sample standard deviation; analyses across independently trained Stage-I representations are identified separately. Training uses the training split, while Stage-II validation and prediction aggregation follow the TS-MOF protocol.

Frozen-base correspondence setting. For the principal correspondence intervention, we fix the Stage-I representation, residual decoder, and two reference experts (BCL and LOS), and optimize only the auxiliary classifier head. Writing the Stage-II parameter blocks as θ_{dec}, θ_1, θ_2, and θ_{aux}, the trainable parameter sets are

$$\Theta_{\mathrm{std}} = \{\theta_{\mathrm{dec}}, \theta_1, \theta_2, \theta_{\mathrm{aux}}\}, \qquad \Theta_{\mathrm{frozen}} = \{\theta_{\mathrm{aux}}\}. \tag{15}$$

Across correspondence conditions, the representation, decoder, and reference predictions remain fixed; only the class-frequency assignment within the auxiliary objective changes.

TS-MOF comparison. For direct comparison with TS-MOF, we keep the Stage-I representation, Stage-II architecture, expert count, and inference protocol fixed, and change only the auxiliary training objective.

A.3 Evaluation Protocols

Reference protocols. For auxiliary-expert analyses, we distinguish own-reference and fixed-reference evaluations. Own-reference analyses pair each candidate with the BCL+LOS reference experts from the same Stage-II seed. Fixed-reference analyses evaluate all candidates against the aligned FSAL BCL+LOS predictions from the same Stage-II seed, so conditional gains are measured against identical existing coverage. The principal correspondence intervention in Section 4.3 uses the fixed-reference setting.

Statistical unit. Three-seed auxiliary analyses use Stage-II seeds 42, 2024, and 3407 with a shared frozen Stage-I representation within each dataset/backbone setting. Reported sample standard deviation is computed across these Stage-II seeds. Experiments that retrain Stage I are identified explicitly and summarized separately from Stage-II seed variation.

Prediction outputs. Set-conditioned metrics are computed from each head's Top-1 predictions. Auxiliary analyses use raw head scores without test-time augmentation unless a different prediction rule is stated explicitly. Standard benchmark reporting follows the inherited TS-MOF inference protocol.

B Expert Value: Theory and Decomposition

This section develops the set-conditioned expert-value formulation introduced in Section 3.1. We derive the direct properties and submodular coverage structure of conditional expert gain, relate it to coalition-based valuation, distinguish correctness coverage from aggregation-based prediction, and derive the class-wise gain decomposition used in the correspondence analysis.

B.1 Direct Properties, Submodularity, and Coalition Valuation

For a candidate expert e, let $\mathcal{C}_e = \{(x, y) : e(x) = y\}$ denote its correctness set and $\mathcal{C}_S = \bigcup_{s \in S} \mathcal{C}_s$ the correctness covered by a reference set S. Decomposing $\mathcal{C}_e$ into correctness inside and outside $\mathcal{C}_S$ gives

$$\begin{aligned} G(e \mid S) &= A(e) - \Pr(\mathcal{C}_e \cap \mathcal{C}_S) \\ &= B(S \cup \{e\}) - B(S). \end{aligned} \tag{16}$$

Equation 16 makes the role of overlap explicit. Standalone accuracy determines the total correctness mass of e, but not how much of that correctness is already covered by S.

Let $a = A(e)$ and $b = B(S)$. The overlap satisfies

$$\max(0, a + b - 1) \leq \Pr(\mathcal{C}_e \cap \mathcal{C}_S) \leq \min(a, b).$$

Substituting these bounds into Eq. 16 yields

$$\max(0, a - b) \leq G(e \mid S) \leq \min(a, 1 - b). \tag{17}$$

Thus, standalone accuracy constrains conditional expert gain without determining it.

Set dependence follows directly from inclusion. If $S \subseteq T$, then $\mathcal{C}_S \subseteq \mathcal{C}_T$ and hence $\mathcal{C}_T^c \subseteq \mathcal{C}_S^c$, giving $G(e \mid T) \leq G(e \mid S)$. For $T = S \cup \{s\}$,

$$G(e \mid S \cup \{s\}) = G(e \mid S) - \Pr(\mathcal{C}_e \cap \mathcal{C}_s \cap \mathcal{C}_S^c)\,. \tag{18}$$

The final term is exactly the correctness of e that becomes newly covered when s is added to the reference set.

Submodular coverage. The coverage function $B(S) = \Pr(\mathcal{C}_S)$ is normalized, $B(\varnothing) = 0$, and monotone under set inclusion. Moreover, for $S \subseteq T$ and $e \notin T$,

$$\begin{aligned} B(S \cup \{e\}) - B(S) &= \Pr(\mathcal{C}_e \cap \mathcal{C}_S^c) \\ &\geq \Pr(\mathcal{C}_e \cap \mathcal{C}_T^c) \\ &= B(T \cup \{e\}) - B(T). \end{aligned} \tag{19}$$

Hence B is a monotone submodular set function and conditional expert gain is exactly its marginal contribution.

Relation to coalition-based valuation. Let $\mathcal{E}$ be a finite expert universe and define the correctness-coverage game

$$v(S) = B(S) = \Pr\left(\bigcup_{s\in S} \mathcal{C}_s\right). \tag{20}$$

Conditional expert gain is the marginal contribution to a specified coalition,

$$G(e \mid S) = v(S \cup \{e\}) - v(S). \tag{21}$$

A Shapley-style classifier valuation (Rozemberczki & Sarkar, 2021), in contrast, averages the same marginal contribution over all possible coalitions:

$$\phi_e = \sum_{S\subseteq\mathcal{E}\setminus\{e\}} \frac{|S|!(|\mathcal{E}| - |S| - 1)!}{|\mathcal{E}|!} \left[v(S \cup \{e\}) - v(S)\right]. \tag{22}$$

The key distinction is the reference coalition. Conditional expert gain evaluates one specified expert set, whereas the Shapley value averages marginal contributions over possible coalitions.

Coverage and aggregation. Conditional expert gain measures correctness coverage across an expert set, whereas an aggregation rule produces a single prediction. Available complementary correctness need not translate one-to-one into aggregate accuracy: a correct expert prediction need not determine the aggregate output, and aggregation can recover the correct class even when no individual expert selects it as Top-1. Realized prediction accuracy therefore depends on the aggregation rule in addition to the available correctness coverage.

B.2 Class-Wise Gain Decomposition

Let $\pi_c = \Pr(Y = c)$ denote the evaluation prior and define the class-wise residual rate

$$d_c(S) = \Pr(\mathcal{C}_S^c \mid Y = c). \tag{23}$$

For classes with $d_c(S) > 0$, let

$$u_c(e; S) = \Pr(\mathcal{C}_e \mid \mathcal{C}_S^c, Y = c) \tag{24}$$

denote the candidate's correctness on the residual samples of class c. Conditioning incremental correctness on the class label gives

$$G(e \mid S) = \sum_c \pi_c d_c(S) u_c(e; S). \tag{25}$$

If $d_c(S) = 0$, class c contributes zero to Eq. 25, so the corresponding conditional recovery need not be evaluated.

Let

$$a_c(e) = \Pr(\mathcal{C}_e \mid Y = c) \tag{26}$$

be the candidate's class-wise accuracy. If candidate correctness and reference residual membership were conditionally independent given the class label, then $u_c(e; S) = a_c(e)$. This defines the class-conditional independence reference

$$G_{\text{ind}}(e \mid S) = \sum_c \pi_c d_c(S) a_c(e). \tag{27}$$

The remaining within-class contribution is

$$\begin{aligned} H(e; S) &= G(e \mid S) - G_{\text{ind}}(e \mid S) \\ &= \sum_c \pi_c d_c(S) \big[u_c(e; S) - a_c(e)\big]. \end{aligned} \tag{28}$$

Equivalently, H is the class-prior-weighted conditional covariance between the candidate-correctness indicator and the reference-residual indicator. Positive H indicates that candidate correctness is

enriched on reference residuals relative to class-conditional independence, whereas negative H indicates depletion.

For two correspondence assignments evaluated against the same reference set, the gain difference therefore satisfies

$$\Delta G = \Delta G_{\mathrm{ind}} + \Delta H. \tag{29}$$

The first term captures the part explained by changes in class-wise accuracy allocation under the fixed residual prevalence $d_c(S)$; the second captures the remaining change in within-class dependence between candidate correctness and residual membership. We use H as a diagnostic of within-class dependence; it can reflect shared sample difficulty or other within-class error structure not determined by class-wise accuracy alone.

Both $d_c(S)$ and $u_c(e; S)$ depend on the reference set because changing S changes the identity and composition of residual samples within each class. By contrast, in the target-prior analysis the predictions and within-class residual structure are fixed while only the class weights π_c are changed. The empirical class-accuracy-preserving and cross-setting decompositions are reported with the correspondence controls in Appendix C.3.

C FSAL Objective, Training, and Analysis

This section details the class-frequency correspondence used by FSAL, presents the Stage-II training procedure, and reports cross-objective correspondence, decomposition, and sensitivity results supporting the main experiments.

C.1 Correspondence Assignment

Let σ denote a bijection over the C class indices. The correspondence assigned to class c is

$$\Phi_\sigma(c) = \big(p_{\sigma(c)}, r_{\sigma(c)}\big), \tag{30}$$

with $\sigma(c) = c$ defining the aligned correspondence. Alternative mappings reassign the same frequency statistics across class identities. Because σ is bijective,

$$\{p_{\sigma(c)}\}_{c=1}^{C} = \{p_c\}_{c=1}^{C}, \qquad \{r_{\sigma(c)}\}_{c=1}^{C} = \{r_c\}_{c=1}^{C}. \tag{31}$$

The intervention changes only the class-to-frequency assignment while preserving the underlying frequency-statistic multiset. Aligned, permuted, and reversed correspondence therefore use the same prior and frequency coordinates but assign them to different class identities.

C.2 FSAL Stage-II Training Procedure

Algorithm 1 summarizes the core correspondence-indexed score transformation and the deferred weighting used by the complete configuration during Stage-II optimization. Stage I follows the underlying two-stage framework, while FSAL structures the auxiliary objective in Stage II through the shared correspondence defined above.

Algorithm 1 FSAL Stage-II Training

Require: Training set $\mathcal{D}_{\text{train}}$, validation set $\mathcal{D}_{\text{val}}$, frozen representation E, decoder D, reference heads $\{H_k\}_{k=1}^{2}$, auxiliary head H_{aux}, class counts $\{n_c\}$, correspondence σ
Ensure: Stage-II parameters
Compute p_c, r_c, $\Phi_\sigma(c)$, and normalized inverse-prior weights w_c^σ
for Stage-II epoch $t = 1, \ldots, E_{S2}$ **do**
  for each minibatch $\mathcal{B}$ **do**
    Compute frozen features $h_i = E(x_i)$ and decoded features $u_i = D(h_i)$ with D in evaluation mode
    Compute logits of the two reference heads and auxiliary head
    Apply the correspondence-indexed score transformation
    Apply deferred weighting for the complete configuration and compute $\mathcal{L}_{\text{CF}}$
    Compute the two reference-head objectives
    Update D, the reference heads, and H_{aux} using gradient coordination
  end for
  Perform Stage-II validation following the baseline protocol
end for
return Stage-II parameters

C.3 Correspondence Generality and Decomposition

Cross-objective analysis tests whether the aligned–neutral–reversed ordering persists beyond one auxiliary-loss formulation. We then use fixed-reference decomposition and cross-backbone controls to test whether the residual-overlap signature recurs beyond the primary ResNet-34 setting.

Correspondence across objectives. Table 8 evaluates aligned, neutral, and reversed correspondence for FSAL and KPS. Each objective uses its own jointly trained BCL+LOS reference experts, so the relevant comparison is the correspondence ordering within each objective rather than the absolute gain across objectives.

Table 8: **Joint Stage-II correspondence across objectives.** Mean ± SD over three Stage-II seeds on balanced validation at IR=100 (%). Candidates use their own jointly trained BCL+LOS references; KPS uses its fixed scale-30 configuration. Bold marks the largest mean within each objective and metric.

Objective	Mapping	A	G	R	Few recov.
FSAL	Aligned	21.42 ± 1.07	**4.55 ± 0.43**	**10.06 ± 0.83**	**18.97 ± 2.08**
	Neutral	30.80 ± 0.64	1.44 ± 0.05	3.23 ± 0.10	2.12 ± 0.28
	Reversed	**34.25 ± 0.28**	0.48 ± 0.01	1.07 ± 0.03	0.00 ± 0.00
KPS	Aligned	43.00 ± 2.16	**2.29 ± 0.34**	**5.13 ± 0.75**	**9.34 ± 2.58**
	Neutral	41.96 ± 1.44	1.70 ± 0.14	3.81 ± 0.31	4.26 ± 0.58
	Reversed	**43.56 ± 0.68**	0.82 ± 0.12	1.84 ± 0.28	0.65 ± 0.13

Both objectives retain the aligned–neutral–reversed ordering in conditional gain, while standalone accuracy does not follow the same ordering.

Class-accuracy-preserving residual overlap. For the fixed-reference decomposition, candidate correctness is jointly permuted within each class while the reference residual mask remains fixed. This preserves class-wise candidate accuracy and cross-candidate relationships while removing within-class alignment between candidate correctness and residual membership. We use 5,000 paired permutations per Stage-II seed and evaluation split. With N_c samples, K_c candidate-correct samples, and M_c reference residuals in class c, the expected recovered count is $M_c K_c / N_c$, matching G_{ind} in Eq. 27.

Cross-backbone decomposition. Table 9 summarizes the aligned–reversed decomposition on CIFAR-100-LT with ResNet-34 and ResNet-18. The ResNet-34 row corresponds to the decomposition reported in the main analysis.

Table 9: **Residual-overlap decomposition across evaluated settings.** Global aligned–reversed gain decomposition; ΔG and ΔG_{ind} are three-seed means, while ΔH is mean ± sample SD. Values are percentages.

Setting	Contrast	Mean ΔG	Mean ΔG_{ind}	ΔH
CIFAR / ResNet-34	Aligned–Reversed	3.33	2.84	0.482 ± 0.177
CIFAR / ResNet-18	Aligned–Reversed	5.43	4.58	0.849 ± 0.231

Global ΔH remains positive with ResNet-18, supporting recurrence beyond the primary ResNet-34 analysis. Across the evaluated backbones, the decomposition therefore identifies residual structure that is not determined by class-wise accuracy allocation alone.

C.4 Hyperparameter Sensitivity

We vary one auxiliary hyperparameter at a time around the reference configuration $\lambda = 1.5$, $\tau_{\text{span}} = 0.5$, $m_{\max} = 0.6$, and $s = 34$, while keeping the remaining FSAL defaults in Appendix A.2 fixed. Each setting uses one Stage-II seed under the standard Stage-II protocol.

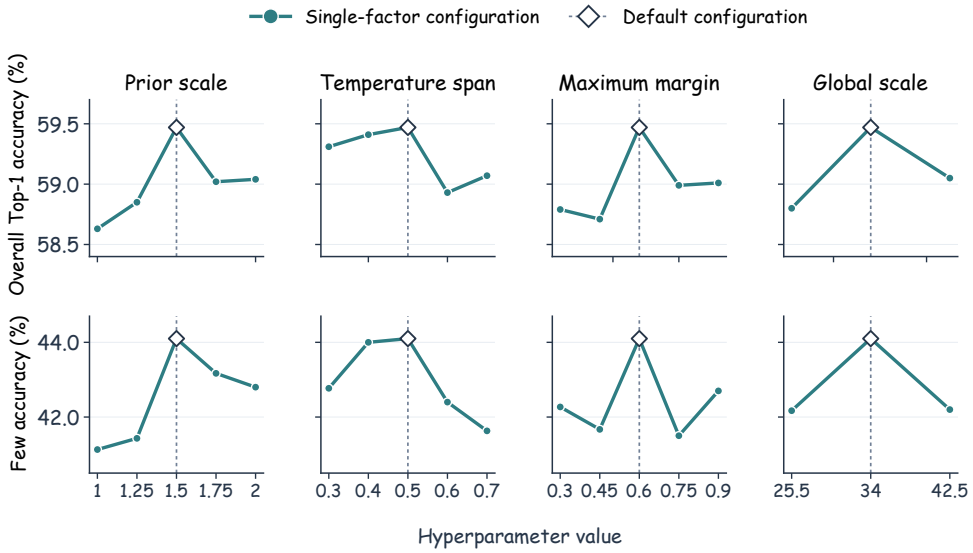


Figure 4: **Hyperparameter sensitivity at IR=100.** Overall Top-1 and Few accuracy are shown in the upper and lower rows, respectively; the reference configuration is marked in each column.

Across the displayed configurations, Overall Top-1 accuracy ranges from 58.63% to 59.47%, while Few accuracy ranges from 41.13% to 44.10%. The reference configuration attains the highest Overall Top-1 and Few accuracy among the evaluated settings.

D Extended Empirical Analysis

This section extends the main analysis with expert-value diagnostics, reference-set dependence, frequency-resolved residual localization, correspondence strength, evaluation-prior dependence, training-imbalance robustness, representation and optimization controls, and large-scale transfer.

D.1 Expert Value and Residual Localization

For the disagreement analysis in Figure 2(a), we define pairwise utility gain as $\Pr(\mathcal{C}_{e_i} \cup \mathcal{C}_{e_j}) - \max\{A(e_i), A(e_j)\}$. Across 81 within-method expert pairs, pairwise utility gain and prediction disagreement have Spearman correlation 0.6911. The partial Spearman correlation controlling for

pair accuracy, $\max\{A(e_i), A(e_j)\}$, is 0.7546. Disagreement is informative about complementarity but does not determine pairwise utility.

Conditional gain also varies with the reference set. In Figure 5, the same candidate is evaluated against BCL, LOS, and BCL+LOS reference sets under aligned, neutral, and reversed correspondence. In the figure, D denotes candidate correctness already covered by the reference set.

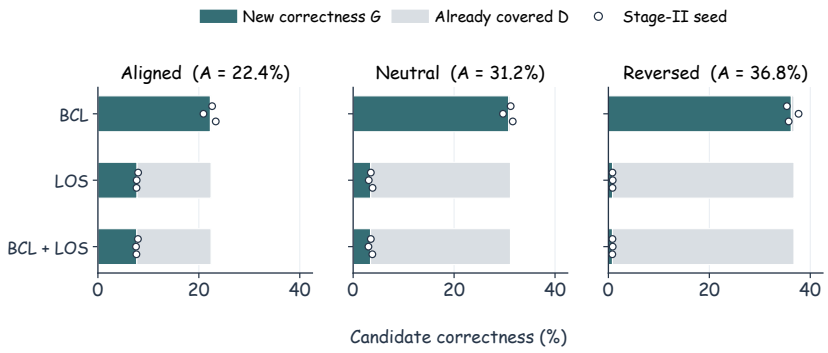


Figure 5: **Reference-set dependence of candidate correctness.** Columns fix the Aligned, Neutral, or Reversed candidate, while rows use BCL, LOS, or BCL+LOS as the reference set. Candidate correctness is partitioned into new correctness G and already-covered correctness D; circles denote Stage-II seed values of G. Within each column, standalone accuracy is fixed while its partition relative to the reference set changes.

Within each panel, standalone accuracy is unchanged across the three reference sets, while the new correctness attributed to the candidate changes with existing coverage. Across the correspondence conditions, higher standalone accuracy can therefore coincide with lower conditional gain relative to the larger reference set.

D.2 Correspondence Strength and Residual Geometry

We vary correspondence continuously. Here ρ denotes the achieved rank correlation between the natural and assigned class-frequency orders. Table 10 reveals a graded pattern: conditional gain, residual recovery, and Few recovery increase consistently with ρ.

Table 10: **Graded correspondence on CIFAR-100-LT at IR=100.** Balanced-validation mean ± SD over three Stage-II seeds; values are percentages except ρ; bold marks the largest mean in each column.

ρ	Overall	Gain	Recovery	Few recovery
-1.000	47.90 ± 0.15	0.81 ± 0.03	1.54 ± 0.07	0.00 ± 0.00
-0.436	49.21 ± 0.15	2.27 ± 0.10	4.34 ± 0.18	2.60 ± 0.46
0.000	50.48 ± 0.21	3.44 ± 0.32	6.58 ± 0.60	7.37 ± 1.02
0.431	52.41 ± 0.18	5.55 ± 0.47	10.61 ± 0.87	12.28 ± 1.20
1.000	**54.52 ± 0.16**	**7.70 ± 0.16**	**14.72 ± 0.25**	**26.78 ± 1.60**

Figure 6 provides the corresponding score-space view. As correspondence moves from aligned toward reversed, residual true-class margins and recovery progressively shift away from low-frequency classes. The ordered margin profiles provide complementary geometric evidence for the same redistribution.

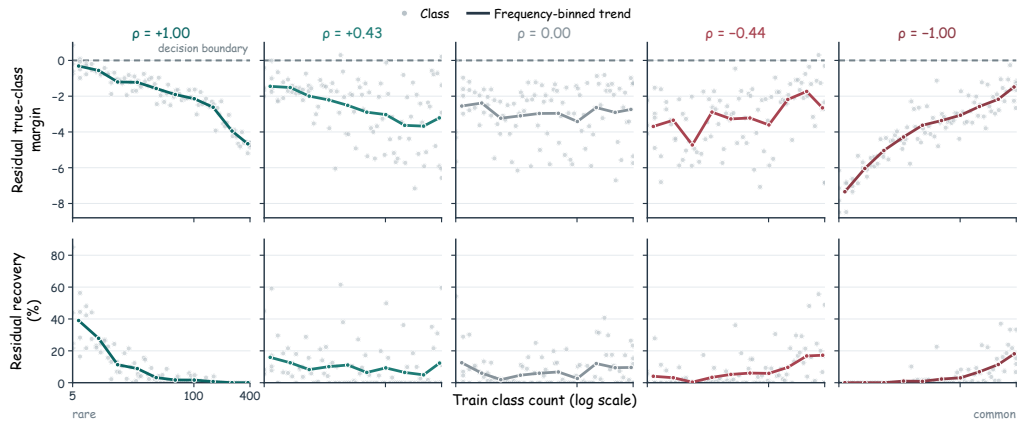


Figure 6: **Residual-margin and recovery profiles across five correspondence levels at IR=100.** Upper panels show residual true-class margins; lower panels show residual recovery along the training-frequency axis.

D.3 Robustness Across Evaluation and Training Conditions

We first examine how conditional expert value changes with the evaluation distribution. Let π_c denote a normalized target-class prior and define the class-conditional contribution g_c as

$$g_c = \Pr(\mathcal{C}_e \cap \mathcal{C}_S^c \mid y = c), \qquad G_\pi(e \mid S) = \sum_{c=1}^{C} \pi_c g_c. \tag{32}$$

Because predictions are fixed, varying π changes only the class-wise weighting of recovered correctness. This formulation underlies the fixed-prediction sweep in Table 5, where the preferred correspondence changes with the target prior. Conditional expert value therefore depends jointly on class-frequency correspondence and the evaluation distribution. Intermediate correspondence labels in Table 5 are nominal, while the achieved rank correlations are reported in Table 10.

We next examine training imbalance. Figure 7 shows that the frequency location of conditional gain follows the same ordered pattern across imbalance ratios. Aligned correspondence accumulates a larger share of gain toward the rare end of the class-frequency ranking, whereas reversal shifts useful corrections toward more frequent classes; the near-zero correspondence condition remains intermediate. At IR=10, conditional gain is 6.01% under alignment and 1.19% under reversal. At IR=50, the corresponding values are 7.49% and 0.95%. The same ordering is retained at IR=100. Thus, increasing imbalance changes the magnitude of conditional gain without changing the correspondence-dependent location of residual specialization.



Figure 7: **Recovery location across training imbalance.** All three imbalance ratios show Aligned, Neutral, and Reversed correspondence, with Neutral denoting the near-zero correspondence condition. Upper panels show class-level gain contributions; lower panels show cumulative gain shares, with classes ordered from rare to common.

Cross-Stage-I robustness is reported in main-text Table 4. Table 11 complements it with Stage-II persistence: the correspondence separation is already pronounced at epoch 2 and remains through epoch 20.

Table 11: **Stage-II persistence of correspondence-dependent specialization.** Three-seed means at epochs 2 and 20; values are percentages.

Metric	Aligned	Permuted	Reversed
Gain	7.31 / 7.30	3.05 / 3.21	0.81 / 0.79
Recovery	13.80 / 13.79	5.75 / 6.07	1.54 / 1.49
Few recov.	25.15 / 25.48	5.58 / 6.24	0.03 / 0.00

Table 12 provides the threshold-based Many/Medium/Few breakdown across imbalance ratios, complementing the class-rank analysis in Figure 7; IR=10 contains no Few classes under the standard threshold.

Table 12: **Complete CIFAR-100-LT results across imbalance ratios.** Many/Medium/Few/Overall Top-1 accuracy (%). IR=10 contains no Few classes. The table extends Table 1 with additional long-tailed recognition baselines. Best results are in bold and second-best results are underlined.

Method	IR=10				IR=50				IR=100			
	Many	Med.	Few	Overall	Many	Med.	Few	Overall	Many	Med.	Few	Overall
CE (He et al., 2016)	73.6	52.5	–	67.1	75.3	44.6	19.6	52.7	77.1	45.9	13.1	47.0
CE-DRW (Cao et al., 2019)	72.9	57.3	–	68.1	74.7	47.7	26.9	55.1	75.3	49.1	18.9	49.2
LDAM-DRW (Cao et al., 2019)	69.3	50.2	–	63.4	69.3	43.9	22.3	50.4	67.3	41.4	15.9	42.8
Balanced Softmax (Ren et al., 2020)	72.8	58.2	–	68.3	75.4	49.4	30.6	56.7	76.4	50.9	22.5	51.3
DBM (Son et al., 2025)	71.9	50.6	–	65.3	73.2	42.6	18.9	50.9	75.8	44.5	12.8	46.0
KPS (Li et al., 2022)	70.4	55.6	–	65.8	71.3	47.6	29.6	54.1	72.8	49.5	23.9	50.0
PaCo (Cui et al., 2021)	75.0	63.3	–	71.4	75.7	55.4	39.4	60.8	77.6	55.4	26.7	54.6
BCL (Zhu et al., 2022)	74.0	56.3	–	68.5	75.5	47.8	28.7	55.7	75.6	51.6	22.3	51.2
ProCo (Du et al., 2024)	**77.2**	60.9	–	72.2	**77.9**	53.3	38.3	60.7	**79.3**	53.2	29.1	55.1
Focal-SAM (Li et al., 2025)	73.3	58.5	–	68.7	74.4	53.0	24.7	56.7	76.2	56.0	15.7	51.0
LOS (Sun et al., 2025)	71.4	62.9	–	68.8	69.8	50.7	44.3	57.4	72.5	48.7	35.8	53.2
LADE (Hong et al., 2021)	72.3	58.0	–	67.9	73.1	46.9	38.0	56.0	73.7	47.7	28.7	51.1
SAMN (Zhang et al., 2026)	69.4	58.6	–	66.1	69.9	48.9	29.5	54.0	70.9	51.1	23.7	49.8
Expert-based methods												
SHIKE (Jin et al., 2023)	70.5	53.3	–	65.1	72.5	43.4	20.6	51.2	70.8	40.8	10.6	42.2
ICL (Chen et al., 2025)	70.4	56.7	–	66.1	71.1	47.2	27.1	53.4	73.3	49.9	19.7	49.0
SADE (Zhang et al., 2022)	72.5	62.3	–	69.4	73.0	53.0	34.4	57.9	73.4	56.2	24.7	52.8
RIDE (Wang et al., 2020)	74.0	57.3	–	68.8	72.7	49.7	29.7	55.5	75.3	51.6	20.4	50.5
BalPoE (Aimar et al., 2023)	75.9	64.6	–	72.4	75.9	55.4	37.0	60.5	77.7	**57.1**	29.2	55.9
PRL (Zhao et al., 2024b)	75.2	64.5	–	71.8	76.5	54.5	36.9	60.3	75.9	56.5	29.3	55.1
TS-MOF (Zhao et al., 2026)	74.7	62.0	–	70.8	75.2	50.7	47.5	60.2	79.0	49.2	39.9	56.8
FSAL	76.8	**66.1**	–	**73.5**	76.5	**56.9**	**52.1**	**64.1**	78.1	54.0	**44.1**	**59.5**

The complete comparison shows that FSAL attains the highest Overall Top-1 accuracy across all three imbalance ratios and the strongest Few-class accuracy at IR=50 and IR=100.

D.4 Large-Scale Transfer

The correspondence-dependent redistribution also appears on ImageNet-LT (Liu et al., 2019). We apply the same aligned, permuted, and reversed intervention used in the main CIFAR analysis, then report the complete large-scale benchmark results to complement the controlled correspondence view.

Table 13 shows the same rare-to-frequent redistribution on ImageNet-LT: alignment retains more low-frequency recovery, whereas reversal shifts recovery toward frequent classes.

Table 13: **Correspondence intervention on ImageNet-LT (Liu et al., 2019).** Values are percentages except enrichment; bold marks row maxima.

Metric	Aligned	Permuted	Reversed
Overall Top-1	**56.8**	55.2	55.0
Few acc.	**43.4**	30.6	30.2
Overall recov.	**6.4**	2.7	2.4
Many recov.	0.16	4.8	**5.74**
Few recov.	**18.96**	0.5	0.00
E_{rare} ($\times$)	**2.2**	0.3	0.1

Table 14 reports the complete class-group breakdown for ImageNet-LT (Liu et al., 2019) and iNaturalist 2018 (Van Horn et al., 2018).

Table 14: **Complete large-scale benchmark results.** Head/Medium/Tail/Overall Top-1 accuracy (%) on ImageNet-LT (Liu et al., 2019) and iNaturalist 2018 (Van Horn et al., 2018). Reference baseline values follow Zhao et al. (2026). Best results are in bold and second-best results are underlined.

Method	ImageNet-LT				iNaturalist 2018			
	Head	Med.	Tail	Overall	Head	Med.	Tail	Overall
CE (He et al., 2016)	64.0	33.8	5.8	41.6	<u>73.9</u>	63.5	55.5	61.0
CE-DRW (Cao et al., 2019)	61.7	47.3	28.8	50.1	68.2	67.3	66.4	67.0
cRT (Kang et al., 2019)	58.8	44.0	26.1	47.3	69.0	66.0	63.2	65.2
LDAM-DRW (Cao et al., 2019)	60.4	46.9	30.7	49.8	–	–	–	66.1
BS (Ren et al., 2020)	60.9	48.8	32.1	51.0	65.7	67.4	67.5	67.3
KPS (Li et al., 2022)	59.7	49.2	35.9	52.3	68.1	69.5	70.2	69.6
RIDE (Wang et al., 2020)	64.9	50.4	34.4	53.6	69.5	<u>71.0</u>	70.4	70.6
BCL (Zhu et al., 2022)	65.3	**53.5**	36.3	55.6	69.5	70.9	<u>71.3</u>	70.9
LOS (Sun et al., 2025)	63.2	50.7	42.3	54.4	69.2	70.7	<u>71.3</u>	70.8
TS-MOF (Zhao et al., 2026)	<u>65.9</u>	<u>52.5</u>	<u>42.6</u>	<u>56.3</u>	72.8	**73.6**	70.5	<u>72.3</u>
FSAL	**67.4**	52.0	**43.4**	**56.8**	**77.2**	70.9	**74.7**	**73.1**

The controlled ImageNet-LT analysis shows the same correspondence-dependent redistribution, while Table 14 reports the corresponding benchmark transfer results on ImageNet-LT and iNaturalist 2018.