Rare-Class Signal Suppression in Long-Tailed Multi-Expert Fine-Tuning

Zhiheng Gong[1] **YuHeng Yang**[2] **Pengkun Wang**[1,3]* **Yang Wang**[1,3,4]*
[1]University of Science and Technology of China [2]Shaoguan University
[3]Suzhou Institute for Advanced Research, USTC
[4]Anhui Provincial Key Laboratory of High Performance Computing

Abstract

Multi-expert ensemble frameworks have achieved strong performance on long-tailed visual recognition by training heterogeneous expert heads on a shared backbone, but their gradient coordination strategies are inherited from balanced multi-task learning, where resolving directional conflict is the canonical challenge. We show that head-class bias acts sequentially across three pipeline interfaces, namely loss geometry, gradient coordination, and inference fusion, forming a suppression cascade in which rare-class gradient energy can be erased before it influences model parameters, a failure mode that direction-focused analysis cannot detect. A controlled study reveals that direction-modifying and simple-aggregation solvers are statistically equivalent once rare-class source signal is amplified, while energy-balancing objectives erase this amplification entirely. To break this cascade, we propose GALE (**G**radient-energy **A**llocation for **L**ong-tail **E**nsembles), a framework that applies gradient-energy allocation at every cascade stage, amplifying tail-class source signal through frequency-calibrated loss geometry, preserving it through a non-penalizing coordination rule, and routing predictions through class-specific expert competence. On CIFAR-100-LT (ρ=100), GALE achieves **59.2%** overall accuracy and **46.1%** few-class accuracy across four seeds (σ=0.23%), surpassing the prior best by **+6.2%**, with consistent improvements on ImageNet-LT and iNaturalist 2018. Code is available at Supplementary Material.

1 Introduction

Multi-expert ensemble frameworks have emerged as a leading paradigm for long-tailed visual recognition, training heterogeneous expert heads on a shared backbone to cover different class-frequency regimes and fusing their predictions at inference [35, 15, 42]. Their gradient coordination strategies, however, inherit from balanced multi-task learning the assumption that the operative bottleneck between expert gradients is *directional conflict*, not scalar magnitude asymmetry [38, 24]. In the tail-sensitive fine-tuning stage, this assumption does not hold. When the backbone is frozen and expert heads differ primarily in loss geometry, gradient signals entering coordination are structurally asymmetric in *magnitude*, driven by the class-frequency gap in each expert's loss rather than by angular misalignment. We term this the **coordination bottleneck misdiagnosis**.

The misdiagnosis has a concrete consequence. The three pipeline interfaces, loss geometry, gradient coordination, and inference fusion, are arranged in series, so a head-class bias that is left uncorrected at one interface is passed intact to the next. Head-class procedural bias therefore acts sequentially across the pipeline, forming a **suppression cascade** in which rare-class gradient energy generated at the source can be erased before it influences model parameters. Prior work has adapted multi-objective

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

solvers to the long-tailed setting [42, 41, 20, 45], showing that decoupled Stage-II fine-tuning with heterogeneous experts meaningfully improves tail-class performance. These methods, however, leave the coordination objective family unexamined, applying task-agnostic rules regardless of whether a given solver is compatible with tail-class energy preservation. Whether the choice of coordination objective matters, and in which direction, has not been tested under controlled conditions.

To quantify this risk, we conduct a controlled solver study in which the source signal and inference rule are fixed and only the coordination objective varies. The results reveal a clean family separation. Projection-based direction modification and simple aggregation [38] are statistically indistinguishable, confirming that pairwise angular projection is not the operative bottleneck once a frequency-calibrated source signal is present. Conflict-averse and global-objective solvers, including CAGrad, MGDA, Nash-MTL, and FAMO [24, 11, 27, 23], by contrast, compress rare-sensitive effective update energy by approximately three orders of magnitude regardless of source-side amplification, translating to a 12–24% collapse in few-class accuracy. This structural incompatibility is the central diagnostic finding of this work.

To address this cascade, we propose GALE (**G**radient-energy **A**llocation for **L**ong-tail **E**nsembles), a cascade-correction framework built on a single principle of replacing the head-favoring default at each pipeline stage with a frequency-proportional gradient-energy allocation. The framework is driven by three components. TAILSPEC constructs a frequency-calibrated source expert whose loss geometry reallocates gradient energy toward rare classes, generating the signal that downstream stages must preserve. GES provides the non-destructive coordination floor that keeps this signal intact. The family study confirms this is the minimal sufficient design, as energy-penalizing solvers such as Nash-MTL [27] erase the amplified signal entirely even when the source expert is frequency-calibrated, collapsing few-class accuracy by 24.3%. COMPARB completes the chain at inference by routing each prediction through the expert with the highest class-specific competence, preventing confidence-biased fusion from discarding well-preserved tail representations. Module decomposition confirms that source amplification drives the majority of the few-class gain (+9.2%), with coordination and routing providing conditional corrections that prevent this gain from being erased downstream.

Our contributions are as follows:

- **A diagnostic finding on solver-family incompatibility in long-tailed multi-expert learning.** We provide the first systematic evidence that global energy-balancing and equilibrium-based coordination objectives are structurally incompatible with tail-class optimization. Under controlled conditions, these solvers compress rare-sensitive effective update energy by orders of magnitude regardless of source-side amplification, a failure mode distinct from directional gradient conflict and undetectable by direction-focused analysis.
- **GALE: a cascade-correction framework driven by gradient-energy allocation.** We introduce a unified framework grounded in a single corrective principle applied at three interfaces, namely source amplification, coordination floor, and inference routing, that jointly prevent rare-class gradient energy from being suppressed along the pipeline. Module decomposition confirms that each interface addresses a distinct failure mode, with source amplification providing the prerequisite signal that coordination and routing then preserve and deliver.
- **State-of-the-art performance with four-seed reproducibility.** On CIFAR-100-LT (ρ=100), GALE achieves **59.2%** overall and **46.1%** few-class accuracy (σ=0.23%), surpassing the prior best by **+6.2%**, with consistent gains on ImageNet-LT and iNaturalist 2018.

2 Related Work

Long-tailed recognition. The literature has progressed from data-level interventions, including over-sampling [7] and class-balanced re-weighting [10, 29], to loss-geometry modifications including focal re-weighting [21], margin adjustment [2], Balanced Softmax [28], logit adjustment [26], and causal deconfounding of momentum bias [32]. Label-distribution disentanglement [14] further addresses changes in class priors, while a unified generalization analysis [36] examines re-weighting and logit adjustment under class imbalance. Decoupled training [16, 44] separates feature learning from classifier refinement, and weight balancing [1], distribution alignment [39], and calibration-aware mixing [43] extend this paradigm from complementary angles. Label over-smoothing [31] and geometric regularization [34] further refine the optimization target and feature space respectively.

Related data-efficient methods address limited supervision. Distribution disagreement with Lorentzian focal representations [5] improves active sampling in hyperbolic space, while hyperspherical energy minimization [4] provides a geometric basis for representative sample selection. Mentored learning [3] incorporates teacher-guided feedback to improve learner generalization and label efficiency, while a survey [6] systematizes approaches to low-resource learning. Unlike GALE, these methods do not directly examine frequency-dependent bias in multi-expert optimization. Multi-expert methods [35, 15] assign heterogeneous expert heads to different frequency regimes, with balanced contrastive objectives [46, 17] further improving representation quality. PaCo [8] and GPaCo [9] extend this direction through parametric contrastive learning, incorporating learnable class centers to address imbalanced representation learning. These studies address complementary aspects of class imbalance or data scarcity, while GALE identifies that head-class biases can compound into a suppression cascade requiring a unified corrective principle.

Gradient coordination and multi-objective optimization. Building on multi-task learning as multi-objective optimization [30], PCGrad [38] mitigates directional conflict via pairwise gradient projection, while CAGrad [24], MGDA [11], Nash-MTL [27], and FAMO [23] apply objective-level coordination from conflict-averse regularization to energy-balancing and equilibrium objectives. Pareto-based methods [22] seek Pareto-optimal trade-offs, and all of these strategies remain task-agnostic with respect to class-frequency imbalance [12]. TS-MOF [41] adapts this paradigm to heterogeneous-expert long-tailed learning, yet leaves objective-family compatibility unexamined. Recent work frames head-tail trade-offs as Pareto problems [20, 45], but either adopts conflict-averse strategies without examining energy-level compatibility [45] or formulates long-tailed learning as gradient-based MOO without inspecting whether the chosen objective family preserves tail-sensitive update energy [20]. GALE diagnoses coordination as one stage of a suppression cascade, finding through a controlled family study that global energy-penalizing objectives are structurally misaligned with tail-sensitive effective-update preservation, a distinction invisible to both direction-focused and task-agnostic design.

Inference fusion and expert routing. Ensemble fusion strategies span static averaging [44], logit calibration [26], and class-conditional routing [42], and a consistent finding is that naive confidence averaging under-utilizes tail-specialist predictions due to the systematic over-confidence of deep models on frequent classes [19]. Prior routing methods select experts by global weighting or input-feature similarity, which inherits the confidence bias of head-dominant models and discards well-preserved tail representations at the final stage. Test-agnostic approaches further address uncertainty in test label distributions: SADE [40] combines diverse experts through self-supervised aggregation, while DirMixE [37] models hierarchical label-distribution variations to improve robustness under unknown test priors. GALE treats inference routing as the final stage of the suppression cascade rather than an independent post-processing step, where competence-based routing by class-specific accuracy estimates completes the frequency-proportional correction that source amplification initiates.

3 Methodology

3.1 Problem Setup

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a long-tailed training set with C classes and imbalance ratio $\rho = N_{\max}/N_{\min}$. We adopt two-stage decoupled training [16]: Stage I trains backbone $\mathcal{F}_\theta$ with cross-entropy; Stage II freezes the convolutional encoder and fine-tunes a lightweight shared residual adapter A_ϕ inserted after global pooling, together with K heterogeneous expert heads $\{h_{\psi_k}\}_{k=1}^{K}$. Let $\Phi = \phi \cup \{\psi_k\}_{k=1}^{K}$ denote the trainable Stage-II parameters. Expert gradients $\mathbf{g}_k = \nabla_\Phi \mathcal{L}_k$ must be combined into a single update, and this coordination is the focus of GALE.

Under severe imbalance, the expert gradients are structurally asymmetric in magnitude before any coordination occurs, driven by the class-frequency gap in each expert's loss geometry rather than by angular conflict between objectives. Existing solvers apply identical coordination rules regardless of this asymmetry.

Definition 1 (Procedural Symmetry)**.** A coordination rule satisfies *procedural symmetry* if it applies identical operations to all expert gradients regardless of their relative magnitude.

Standard solvers including PCGrad [38], CAGrad [24], MGDA [11], Nash-MTL [27], and FAMO [23] satisfy Definition 1. Note that breaking procedural symmetry alone does not guarantee benefit: Nash-MTL and MGDA also assign non-uniform weights to expert gradients, yet their energy-penalizing objectives impose a *frequency-inversely-proportional* asymmetry that suppresses low-energy tail signals. GALE instead breaks procedural symmetry through *frequency-proportional gradient-energy allocation*: TAILSPEC reallocates energy toward rare-class gradients at the source, GES reallocates energy across expert gradients during coordination, and COMPARB routes predictions to the expert with the highest class-specific competence at inference.

3.2 Tail-Specialist Constructor (TAILSPEC)

TAILSPEC is the source-interface gradient-energy allocator of GALE: a frequency-calibrated expert head whose loss geometry reallocates gradient energy toward rare classes, ensuring that the tail-class signal entering coordination is already elevated relative to frequency-agnostic experts, providing the signal that downstream stages must preserve.

Frequency-dependent logit calibration. Given per-class counts $\{n_c\}$, let $\bar{n}_c = (n_c - n_{\min})/(n_{\max} - n_{\min} + \epsilon)$. Each class receives a frequency-calibrated margin and scale:

$$m_c = m_{\max}(1 - \bar{n}_c), \qquad s_c = 1 + (s_{\max}{-}1)\,\frac{\log(n_c{+}1) - \min_j \log(n_j{+}1)}{\max_j \log(n_j{+}1) - \min_j \log(n_j{+}1) + \epsilon}, \tag{1}$$

where m_c uses linear normalization to assign larger margins to rarer classes, and s_c uses logarithmic normalization to compress dynamic range. For sample i with logits $\mathbf{z}_i$ and label y_i, the calibrated logit is $\tilde{z}_{i,c} = s_c z_{i,c} - m_c \cdot \mathbb{I}[y_i{=}c]$.

Inverse-frequency class weighting. Class weights are computed as $w_c \propto 1/(n_c + \epsilon)$, scaled and capped at $w_{\max}{=}2.0$, assigning proportionally larger gradient contributions to tail classes. A per-sample boost $a_i = 1 + \delta_w$ is applied to samples from the least-frequent quartile ($\bar{n}_c \leq 0.25$), where δ_w transitions from $0.15w$ to $0.25w$ after epoch r (the reweight epoch), and is controlled by the weight strength parameter w.

Unified objective.

$$\mathcal{L}_{\text{TAILSPEC}} = -\frac{1}{N}\sum_{i=1}^{N} a_i\, w_{y_i}\, \log \frac{\exp(\tilde{z}_{i,y_i}/\tau)}{\sum_c \exp(\tilde{z}_{i,c}/\tau)}, \tag{2}$$

where τ is a fixed temperature.

3.3 Gradient Energy Scaler (GES)

GES is the coordination-interface gradient-energy allocator of GALE: it reallocates inter-expert gradient magnitudes in proportion to task importance, without modifying gradient directions. Its primary role within the cascade is not to generate accuracy gains in isolation but to serve as the minimal sufficient coordination floor that prevents global energy-penalizing objectives from erasing the rare-class source signal that TAILSPEC constructs at the loss interface.

Importance estimation. For expert t with gradient $\mathbf{g}_t$ and current loss $\mathcal{L}_t$, a loss-normalized signal score measures gradient strength relative to optimization progress:

$$q_t = \frac{\|\mathbf{g}_t\|_2}{\mathcal{L}_t + \epsilon}. \tag{3}$$

A conflict diagnostic $c_t = \frac{1}{T-1}\sum_{j\neq t} \max(0, -\cos(\mathbf{g}_t, \mathbf{g}_j))$ captures pairwise directional interference. Raw importance fuses both via max-pooling:

$$u_t = \max\left(\frac{q_t}{\max_j q_j + \epsilon},\ \beta \cdot \frac{c_t}{\max_j c_j + \epsilon}\right). \tag{4}$$

The raw importance is then shifted and mean-normalized to obtain a stable scaling basis:

$$\tilde{u}_t = \frac{(u_t - \min_j u_j) + \epsilon}{\frac{1}{T}\sum_j (u_j - \min_j u_j) + \epsilon}. \tag{5}$$

Magnitude reallocation. Scaling factors clip the normalized importance to $[\gamma_{\min}, \gamma_{\max}]$:

$$\gamma_t = \mathrm{clip}(\tilde{u}_t,\ \gamma_{\min},\ \gamma_{\max})\,, \tag{6}$$

with experts whose importance exceeds $\max(0.40,\ \mu_{\tilde{u}} + k\sigma_{\tilde{u}})$ receiving additional boosting capped at γ_{abs}. The coordinated gradient aggregates the rescaled expert gradients:

$$\mathbf{g}_{\textsc{Ges}} = \frac{1}{T}\sum_{t=1}^{T} \gamma_t\, \mathbf{g}_t. \tag{7}$$

Remark 1 (Projection-free design). GES deliberately omits pairwise projection. The design hypothesis is that once inter-expert gradient energies are balanced through Eqs. (6)–(7), residual directional conflicts are negligible and $\mathcal{O}(T^2)$ projection adds no benefit. Empirical validation in §4.4 confirms this: projection variants reproduce identical results, and direction-modifying solvers are statistically equivalent to pure magnitude allocation, establishing $\mathcal{O}(T)$ importance-proportional reallocation as the minimal sufficient coordination floor.

TAILSPEC *and* GES *calibrate gradient magnitude through training; at inference, confidence-biased fusion can still suppress well-preserved tail representations in favor of head-dominant experts.*

3.4 Competence-based Arbiter (COMPARB)

Standard confidence-weighted fusion inherits the over-confidence of deep models on frequent classes [19], discarding tail-specialist predictions at the final decision stage. COMPARB completes the cascade correction by routing each prediction through the expert with the highest class-specific competence, measured by observed validation accuracy rather than prediction confidence.

Competence weight estimation. Per-class validation accuracy $\mathcal{A}_{\mathrm{val}}(t, c)$ is converted into per-expert competence weights via a difficulty-aware softmax:

$$W_{t,c} = \frac{\exp\big(\mathcal{A}_{\mathrm{val}}(t,c)^{\alpha} \cdot b_c\big)}{\sum_j \exp\big(\mathcal{A}_{\mathrm{val}}(j,c)^{\alpha} \cdot b_c\big)}, \tag{8}$$

where α sharpens competence differences and $b_c = (1 + 0.6\, d_c)(1 + 0.25\, r_c)$ upweights difficult classes ($d_c = 1 - \bar{\mathcal{A}}_c$) with sufficient validation support ($r_c = \mathbb{I}[n_c^{\mathrm{val}} \geq 50]$).

Consensus fusion. When all experts agree on the predicted class $\hat{y}$, predictions are fused by competence-weighted averaging:

$$\mathbf{p}_{\mathrm{out}} = \sum_{t=1}^{K} W_{t,\hat{y}} \cdot \mathrm{softmax}(\mathbf{z}_t). \tag{9}$$

Disagreement arbitration. When experts disagree, calibrated confidence is scored with an asymmetric temperature ($\tau = 0.55$ for high-support classes, $\tau = 1.0$ for tail classes) to reduce head-class over-confidence, and the prediction of the most confident expert is selected:

$$t^{\star} = \arg\max_t\ p_{t,\hat{y}_t}, \qquad \hat{y} = \hat{y}_{t^{\star}}. \tag{10}$$

Remark 2 (Validation-set calibration). $W_{t,c}$ is computed from validation-set labels, following the standard model-selection protocol in which a held-out set calibrates but does not train the model. No labels are accessed at test time.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on three benchmarks of increasing scale. **CIFAR-100-LT** [2] provides long-tailed variants of CIFAR-100 with imbalance ratios $\rho \in \{100, 50, 10\}$, following the evaluation protocol of [25], we report per-split accuracy for Many (>100 training samples), Medium (20–100), and Few (<20) classes; at ρ=100, the most frequent class holds 500 samples while the rarest holds 5. **ImageNet-LT** [25] contains 115.8K training images across 1,000 classes (ρ=256). **iNaturalist 2018** [33] provides 437.5K images across 8,142 species with natural long-tail distribution.

Table 1: **Top-1 accuracy (%) on CIFAR-100-LT.** GALE results: mean $\pm$ std over four seeds. All baselines were reproduced under a unified training and evaluation protocol. Few-class std is 0.23% before rounding.

	$\rho = 100$				$\rho = 50$				$\rho = 10$			
Method	OA	Many	Med.	Few	OA	Many	Med.	Few	OA	Many	Med.	Few
CE [13]	38.1	65.6	36.2	8.2	43.8	63.9	36.2	15.2	56.5	63.2	40.3	–
CE-DRW [2]	41.4	63.4	41.2	15.7	45.0	60.6	39.0	22.9	58.2	62.5	48.6	–
LDAM-DRW [2]	43.2	62.8	42.6	21.1	47.2	63.0	41.2	25.1	57.5	62.7	46.1	–
BS [28]	42.8	59.6	42.3	23.7	47.9	60.3	41.3	34.3	58.1	61.5	50.6	–
RIDE [35]	47.5	65.7	48.6	25.0	52.2	65.7	47.7	31.8	61.1	66.4	49.4	–
BCL [46]	44.2	63.1	42.9	23.9	49.1	61.6	43.1	34.3	58.9	62.2	51.8	–
KPS [18]	42.2	41.9	39.5	48.7	50.5	51.6	49.5	52.4	59.5	61.7	58.7	–
SHIKE [15]	46.9	66.0	39.0	12.0	49.5	67.0	43.0	23.0	59.0	66.0	45.0	–
LOS [31]	53.9	70.3	52.3	36.6	58.0	72.4	51.4	40.2	69.0	71.9	62.3	–
TS-MOF [41]	56.8	79.0	49.2	39.9	60.2	75.2	50.7	47.5	70.8	74.7	62.0	–
GALE **(Ours)**	$\mathbf{59.2}_{\pm.2}$	$\mathbf{77.9}_{\pm.2}$	$\mathbf{51.8}_{\pm.7}$	$\mathbf{46.1}_{\pm.2}$	$\mathbf{64.6}_{\pm.4}$	$\mathbf{76.3}_{\pm.2}$	$\mathbf{54.9}_{\pm.4}$	$\mathbf{60.1}_{\pm.9}$	$\mathbf{74.4}_{\pm.2}$	$\mathbf{75.9}_{\pm.1}$	$\mathbf{71.2}_{\pm.8}$	–

Baselines. We compare against four paradigms: *(i)* one-stage: CE, CE-DRW, LDAM-DRW [2], BS [28]; *(ii)* decoupled: cRT [16], LOS [31]; *(iii)* ensemble: RIDE [35], BCL [46], KPS [18], SHIKE [15]; *(iv)* MOO-based: TS-MOF [41], the most directly comparable baseline sharing the same two-stage paradigm and expert pool structure.

Implementation. Stage I trains ResNet-34 (CIFAR-100-LT) or ResNet-50 (ImageNet-LT, iNaturalist) with SGD and cosine annealing for 200 epochs. Stage II freezes the convolutional encoder and fine-tunes a shared residual feature adapter together with three expert heads (TAILSPEC, BCL, LOS) for 20 epochs. Key hyperparameters: TAILSPEC temperature τ=0.1; GES conflict weight β=0.65; COMPARB sharpness α=1.3. Full hyperparameter configuration is provided in Appendix B. All CIFAR-100-LT results are mean $\pm$ std over four independent seeds.

4.2 Main Results

CIFAR-100-LT (Table 1). At ρ=100, GALE achieves **59.2%** overall and **46.1%** few-class accuracy (σ=0.23%), improving over TS-MOF by $+2.4\%$ overall and $+6.2\%$ few-class. Tail-class gains are concentrated in the few-class split while many-class accuracy remains competitive. Performance transfers consistently across $\rho \in \{50, 10\}$ without retuning.

ImageNet-LT and iNaturalist 2018 (Table 2). On ImageNet-LT, GALE achieves **58.4%** overall ($+2.1\%$ over TS-MOF) with the largest gain in the few-class split ($+4.4\%$). On iNaturalist 2018, GALE reaches **74.6%** overall ($+2.3\%$), with consistent gains across medium and few splits. The tail-concentrated gain pattern replicates across all three benchmarks.

Table 2: **Top-1 accuracy (%) on ImageNet-LT and iNaturalist 2018.**

	ImageNet-LT				iNaturalist 2018			
Method	OA	Many	Med.	Few	OA	Many	Med.	Few
CE [13]	41.6	64.0	33.8	5.8	61.0	**73.9**	63.5	55.5
CE-DRW [2]	50.1	61.7	47.3	28.8	67.0	68.2	67.3	66.4
LDAM-DRW [2]	49.8	60.4	46.9	30.7	66.1	–	–	–
BS [28]	51.0	60.9	48.8	32.1	67.3	65.7	67.4	67.5
cRT [16]	47.3	58.8	44.0	26.1	65.2	69.0	66.0	63.2
KPS [18]	52.3	59.7	49.2	35.9	69.6	68.1	69.5	70.2
RIDE [35]	53.6	64.9	50.4	34.4	70.6	69.5	71.0	70.4
BCL [46]	55.6	65.3	53.5	36.3	70.9	69.5	70.9	71.3
LOS [31]	54.4	63.2	50.7	42.3	70.8	69.2	70.7	71.3
TS-MOF [41]	56.3	**65.9**	52.5	42.6	72.3	72.8	73.6	70.5
GALE (**Ours**)	**58.4**	65.8	**55.5**	**47.0**	**74.6**	72.0	**74.7**	**75.1**

4.3 Cascade Correction: Module Decomposition

Figure 1 and Tables 3–4 decompose the calibration chain on CIFAR-100-LT (ρ=100).

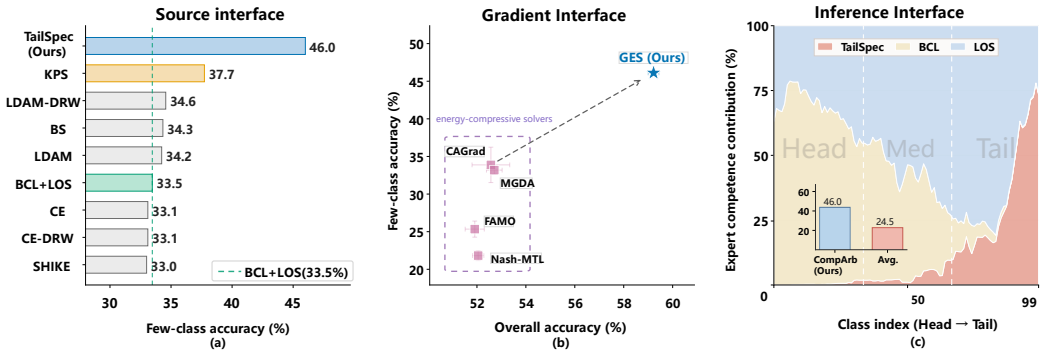


Figure 1: **Cascade correction evidence on CIFAR-100-LT** (ρ=100). **(a)** TAILSPEC leads nine expert-pool configurations monotonically by frequency-calibration strength. **(b)** Non-penalizing solvers cluster in the high-Few regime, energy-penalizing objectives collapse to the low-Few regime. **(c)** Competence routing shifts toward TAILSPEC for rare classes, uniform averaging collapses Few-Class accuracy by 21.5%.

Table 4: **Ablation Analysis on CIFAR-100-LT** ($\rho = 100$)**.** We conduct a comprehensive ablation by progressively enabling TAILSPEC, GES, and COMPARB. (✓: Enabled, –: Disabled).

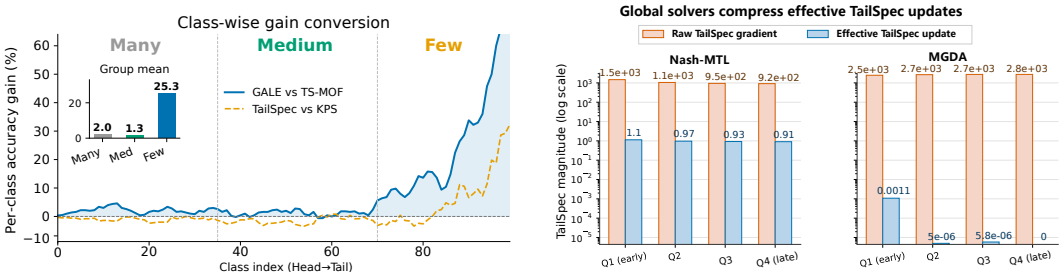
TS	GES	CA	Many	Med.	Few	OA
–	–	–	$77.5_{\pm.1}$	$50.0_{\pm.5}$	$34.9_{\pm.4}$	$55.1_{\pm.2}$
✓	–	–	$77.4_{\pm.1}$	$49.9_{\pm.4}$	$44.1_{\pm.7}$	$57.8_{\pm.2}$
–	✓	–	$77.7_{\pm.2}$	$50.2_{\pm.3}$	$35.6_{\pm.5}$	$55.4_{\pm.2}$
–	–	✓	$78.2_{\pm.1}$	$51.8_{\pm.5}$	$37.0_{\pm.2}$	$56.6_{\pm.1}$
✓	✓	–	$77.3_{\pm.1}$	$49.2_{\pm.3}$	$44.5_{\pm.4}$	$57.6_{\pm.2}$
✓	–	✓	$78.0_{\pm.1}$	$52.4_{\pm.2}$	$45.4_{\pm.6}$	$59.3_{\pm.1}$
–	✓	✓	$78.4_{\pm.1}$	$51.9_{\pm.3}$	$38.3_{\pm.5}$	$57.1_{\pm.2}$
✓	✓	✓	$77.9_{\pm.2}$	$51.8_{\pm.7}$	$\mathbf{46.1}_{\pm.2}$	$59.2_{\pm.2}$

Source interface (Figure 1a). Under fixed GES and COMPARB, TAILSPEC achieves **46.0%** few-class accuracy, outperforming the next-best configuration KPS by 8.3% and all frequency-agnostic alternatives by more than 10%. Performance ranks with frequency-calibration strength: TAILSPEC > KPS > LDAM-DRW > LDAM > CE, confirming that the gain is specific to frequency-calibrated loss geometry and not to ensemble size.

Table 3: **Controlled solver comparison on CIFAR-100-LT** (ρ=100), 4 seeds. All strategies are evaluated within the GALE framework: each combines TAILSPEC's frequency-amplified source signal with the tested coordination rule and COMPARB inference. Only the gradient coordination rule varies.

Solver	OA	Many	Med.	Few
Non-penalizing				
GES (Ours)	$\mathbf{59.2}_{\pm.2}$	$77.9_{\pm.2}$	$51.8_{\pm.7}$	$\mathbf{46.1}_{\pm.2}$
PCGrad [38]	$59.3_{\pm.1}$	$78.0_{\pm.1}$	$52.0_{\pm.4}$	$45.9_{\pm.8}$
RD-PCGrad	$59.3_{\pm.1}$	$78.0_{\pm.1}$	$52.4_{\pm.2}$	$45.4_{\pm.6}$
raw-merge-mean	$59.3_{\pm.1}$	$78.0_{\pm.1}$	$52.4_{\pm.4}$	$45.5_{\pm.5}$
Global-objective solvers (structurally incompatible)				
CAGrad [24]	$52.6_{\pm.8}$	$76.3_{\pm.5}$	$44.9_{\pm.5}$	$33.9_{\pm2.4}$
Nash-MTL [27]	$52.1_{\pm.2}$	$78.3_{\pm.3}$	$51.7_{\pm.7}$	$21.8_{\pm.7}$
MGDA [11]	$52.7_{\pm.3}$	$71.4_{\pm.4}$	$50.7_{\pm.3}$	$33.2_{\pm.5}$
FAMO [23]	$51.9_{\pm.4}$	$75.4_{\pm.2}$	$51.2_{\pm.5}$	$25.3_{\pm1.1}$

Gradient interface (Figure 1b, Table 3). All coordination strategies share TAILSPEC's frequency-amplified source signal and COMPARB inference; only the gradient coordination rule varies. The results reveal a family separation. All non-penalizing strategies are statistically indistinguishable: GES, PCGrad, RD-PCGrad, and raw-merge-mean, with PCGrad adding pairwise projection yet yielding no improvement over simple averaging, confirming that angular conflict is not the operative bottleneck once a frequency-calibrated source signal is present. Global multi-objective solvers that impose energy-balancing or equilibrium constraints, by contrast, collapse few-class accuracy by 12–24%, erasing the source gain entirely. The mechanism is analyzed in §4.4.



(a) Per-class gain decomposition (b) Effective update compression

Figure 2: **Mechanistic evidence on CIFAR-100-LT** (ρ=100). **(a)** Per-class gain over TS-MOF baseline: source construction (TAILSPEC vs. KPS, orange) and full GALE chain (blue); gain concentrates in the few-class regime. **(b)** Raw TAILSPEC gradient norm vs. effective update under Nash-MTL and MGDA; the raw signal remains large while the effective update collapses by orders of magnitude.

Inference interface (Figure 1c). Class-specific competence shifts from LOS toward TAILSPEC as class frequency decreases, consistent with Eq. (8). Replacing COMPARB with uniform averaging collapses few-class accuracy by 21.5%, confirming that confidence-biased fusion discards rare-class representations preserved by the upstream chain.

Chain dependency. Table 4 shows that source amplification accounts for the majority of the total few-class gain (+9.2% from TAILSPEC alone), establishing loss-interface gradient-energy reallocation as the primary cascade contribution. GES contributes modestly in isolation; its necessity is revealed not by standalone gain but by the counterfactual in Table 3: replacing GES with Nash-MTL within the same TAILSPEC+COMPARB framework collapses few-class accuracy from 46.1% to 21.8%, a 24.3% drop that erases the source gain entirely.

The statistical equivalence between GES, direction-modifying solvers, and simple averaging confirms that once gradient energy is frequency-calibrated at the source, the coordination interface requires only a non-destructive allocation floor. what matters is the absence of global energy penalization, not the specific form of the non-penalizing rule. Although GES reduces overall accuracy by 0.1% relative to TAILSPEC+ COMPARB alone, it improves few-class accuracy by 0.7% (three standard deviations), a trade-off that reflects the framework's goal of tail-class correction rather than overall accuracy maximization.

4.4 Further Analysis

Module component analysis (Table 5). Table 5 identifies the operative mechanism within each module by ablating individual components (seed 42, all other modules at full configuration).

TAILSPEC *components.* Removing frequency-calibrated scale (s_c=1) drops few-class accuracy by 11.5%, the largest single-component drop in the study, isolating per-class logit scale as the primary source-amplification mechanism. Replacing class weights with uniform weighting drops few-class by 5.2%, showing that inverse-frequency class weighting provides gradient-energy reallocation beyond logit calibration alone. Removing margin (m_c=0) or tail reweighting causes only 1.0–1.2% drops, indicating their secondary contributions.

Table 5: **Module component analysis on CIFAR-100-LT** (ρ=100, seed 42). Each row removes or replaces one component; all other modules at full configuration.

Module	Variant	OA	Few
GALE		**59.2**	**46.0**
TAILSPEC	No freq. scale (s_c=1)	56.1	34.5
	Uniform class weight	58.0	40.8
	No margin (m_c=0)	58.7	44.8
	No tail reweight	58.9	45.0
GES	No boost	59.0	43.9
	Conflict-only importance	58.8	45.3
	+ projection	59.2	46.0
COMPARB	Confidence-only (λ=0)	58.6	46.0
	Entropy routing (λ=0.5)	57.3	41.7
	Frozen val weights	58.6	46.0
	No arbitration (avg.)	46.4	24.5
	No arbitration (weighted)	44.1	21.8

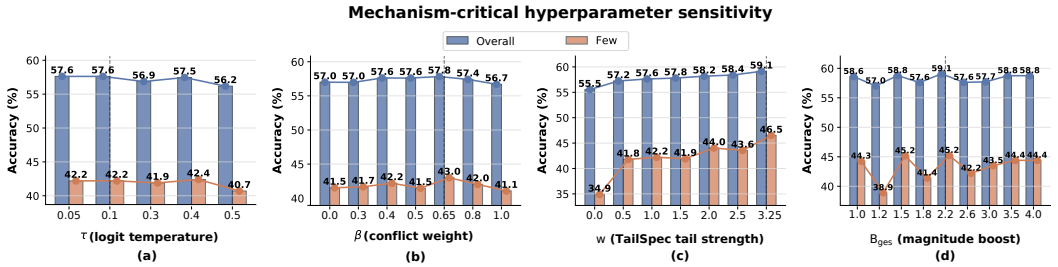


Figure 3: **Mechanism-critical hyperparameter sensitivity.** Each panel varies one control while keeping the remaining configuration fixed: source temperature τ, conflict weight β, TAILSPEC tail strength w, and the GES magnitude boost. Blue bars report overall accuracy and orange bars report Few-class accuracy.

GES *components.* Removing the importance boost drops few-class by 2.1%, confirming that tail-sensitive experts benefit from additional scaling above the adaptive threshold. Using conflict-only importance drops few-class by 0.7%, consistent with the loss-normalized signal being the primary importance diagnostic and conflict intensity providing complementary information via the max-pooling fusion of Eq. (4). The projection variant reproduces identical results, confirming that importance-proportional scaling is the operative mechanism and that $\mathcal{O}(T^2)$ projection provides no additional benefit.

COMPARB *components.* Confidence-only routing (λ=0) and frozen validation weights both match the full configuration on few-class accuracy (46.0%), establishing competence-weighted confidence as the operative routing mechanism. Entropy routing (λ=0.5) drops few-class by 4.3%, showing that rank-entropy tie-breaking actively interferes with competence routing. Removing arbitration entirely collapses few-class accuracy to 24.5% under average fusion and 21.8% under weighted fusion, confirming that competence-based routing is the essential final stage of the cascade.

Per-class gain decomposition (Figure 2a). Source construction (TAILSPEC vs. KPS, orange) generates substantial gains concentrated in the low-frequency regime; the full GALE chain (blue) amplifies this further. Averaged within each frequency band over the TS-MOF baseline, the per-class gain is modest for Many (+2.0%) and Medium (+1.3%) classes but rises to +25.3% for Few classes—a ratio that directly reflects the frequency-proportional treatment applied at each cascade stage and confirms that the gain is tail-directed rather than a head-class calibration artifact.

Effective update compression (Figure 2b). Figure 2b provides the mechanistic explanation for the 12–24% few-class collapse of energy-penalizing solvers in Table 3. In both cases the source signal is present but the coordination objective suppresses it before it reaches the parameters. MGDA seeks a minimum-norm convex combination of task gradients, while Nash-MTL solves a bargaining-equilibrium formulation for a joint update direction; in our residual-adapter Stage-II regime, both assign near-zero effective contribution to the low-energy tail-specialist channel, compressing the raw TAILSPEC gradient by approximately three orders of magnitude in the late training window. This structural incompatibility is not a consequence of missing source signal—the raw TAILSPEC gradient remains consistently large throughout training—but of the solvers' global energy criteria, which inherently penalize the low-energy tail-specialist signal that TAILSPEC deliberately elevated. Full diagnostics appear in Appendix C.

Hyperparameter sensitivity (Figure 3). The source-side strength parameter w has the largest effect on few-class accuracy, consistent with source construction providing the dominant rare-class evidence in the cascade. GES-related controls (β, boost) affect the operating point but do not by themselves explain the full gain, confirming that GES functions as a coordination floor within the cascade rather than a standalone optimizer. The framework is stable within broad ranges of all three mechanism-critical controls, requiring no per-dataset retuning in the reported experiments.

5 Conclusion

We identified the coordination bottleneck misdiagnosis as a systemic failure in Stage-II multi-expert fine-tuning: existing solvers treat directional conflict as the operative bottleneck, while head-class procedural bias compounds across the loss, gradient, and inference interfaces into a suppression cascade. This finding suggests that the critical failure mode in long-tailed expert coordination is not primarily angular disagreement among gradients, but the attrition of rare-class signal magnitude before it can produce an effective parameter update. It further reframes expert coordination as a problem of rare-signal preservation rather than symmetric conflict resolution. We proposed GALE, a gradient-energy allocation framework that corrects the suppression cascade at three successive interfaces. Source amplification generates the dominant rare-class signal, a non-destructive coordination floor preserves it, and competence routing completes the correction at inference. Together, these stages form a signal-fidelity chain in which rare-class evidence must be generated, protected, and delivered rather than merely averaged across experts. A controlled solver study reveals that global energy-penalizing objectives remain structurally incompatible with tail-class optimization even when the source signal is amplified, compressing rare-sensitive effective update energy by orders of magnitude. On CIFAR-100-LT (ρ=100), GALE achieves **59.2%** overall and **46.1%** few-class accuracy across four seeds (σ=0.23%), with consistent improvements on ImageNet-LT and iNaturalist 2018.

Acknowledgements.

The authors gratefully acknowledge the support from the National Natural Science Foundation of China (NSFC) under Grant Nos. 62402472, and 12227901. This work was also supported by the Natural Science Foundation of Jiangsu Province (No. BK20240461), the Jiangsu Provincial Young Science and Technology Talent Support Program (No. JSTJ-2026-622), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No. YSBR-005), and the Academic Leaders Cultivation Program at USTC. The AI-driven experiments, simulations, and model training were performed on the robotic AI-Scientist platform of the Chinese Academy of Sciences.

Competing Interests.

The authors declare no competing financial interests related to this work.

References

[1] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6907, 2022.

[2] 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.

[3] Xiaofeng Cao, Yaming Guo, Heng Tao Shen, Ivor W Tsang, and James T Kwok. Mentored learning: Improving generalization and convergence of student learner. *Journal of Machine Learning Research*, 25(325):1–45, 2024.

[4] Xiaofeng Cao, Weiyang Liu, and Ivor W Tsang. Data-efficient learning via minimizing hyperspherical energy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13422–13437, 2023.

[5] Xiaofeng Cao and Ivor W Tsang. Distribution disagreement via lorentzian focal representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6872–6889, 2021.

[6] Xiaofeng Cao, Mingwei Xu, Xin Yu, Jiangchao Yao, Wei Ye, Shengjun Huang, Minling Zhang, Ivor Tsang, Yew-Soon Ong, James T Kwok, et al. Analytical survey of learning with low-resource data: From analysis to investigation. *ACM Computing Surveys*, 58(6):1–47, 2025.

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

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

[9] 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.

[10] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[11] Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus. Mathématique*, 350(5-6):313–318, 2012.

[12] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021.

[13] 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*, pages 770–778, 2016.

[14] 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)*, pages 6622–6632. IEEE, 2021.

[15] Yan Jin, Mengke Li, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23695–23704, 2023.

[16] 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.

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[18] 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.

[19] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6929–6938, 2022.

[20] Weiqi Li, Fan Lyu, Fanhua Shang, Liang Wan, and Wei Feng. Long-tailed learning as multi-objective optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3190–3198, 2024.

[21] 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*, pages 2980–2988, 2017.

[22] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019.

[23] Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. Famo: Fast adaptive multitask optimization. *Advances in Neural Information Processing Systems*, 36:57226–57243, 2023.

[24] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.

[25] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

[26] 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.

[27] Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*, 2022.

[28] 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.

[29] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR, 2018.

[30] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

[31] 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 *The Thirteenth International Conference on Learning Representations*, 2025.

[32] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in neural information processing systems*, 33:1513–1524, 2020.

[33] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[34] Jinping Wang, Zhiqiang Gao, and Zhiwu Xie. Space alignment matters: The missing piece for inducing neural collapse in long-tailed learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 26302–26309, 2026.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021.

[40] 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.

[41] 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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[42] 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*, 2024.

[43] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021.

[44] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020.

[45] Zhipeng Zhou, Liu Liu, Peilin Zhao, and Wei Gong. Pareto deep long-tailed recognition: A conflict-averse solution. In *The Twelfth International Conference on Learning Representations*, 2023.

[46] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6908–6917, 2022.

Appendix

This appendix contains the material needed to audit the central diagnosis of the paper: in residual-adapter Stage-II long-tailed multi-expert fine-tuning, rare-class evidence can be generated at the loss interface but still be suppressed before it reaches the final prediction. The appendix therefore follows the same claim chain as the main text: source-side rare-signal construction, non-destructive gradient-energy coordination, and competence-aware inference routing. We do not repeat the main benchmark tables, the full controlled solver comparison, or the interface-decomposition results already reported in the main paper. Instead, we provide the formal scope, reproducibility details, and supplementary diagnostics needed to verify the suppression-cascade interpretation and the solver-family separation.

- **Appendix A: Formalizing the Suppression Cascade.** We define the residual-adapter Stage-II expert-objective regime, formalize head-class procedural bias and rare-signal retention, and separate the two coordination failures that motivate GALE: local direction correction is non-amplifying, whereas global energy-penalizing objectives can compress rare-sensitive effective update energy even when the raw TAILSPEC source gradient remains large.
- **Appendix B: Implementation, Protocol, and Complexity.** We specify the executable GALE path, the Stage-II trainable parameter set, the four-seed reporting convention, the compact mechanism-level configuration, validation-set competence calibration, diagnostic logging convention, computational resources, and overhead.
- **Appendix C: Supplementary Empirical Diagnostics.** We report diagnostics not duplicated in the main text, including source-interface transfer, solver-weighted raw-versus-effective rare-signal compression, late-window solver metric decompositions, solver-channel heatmaps. These figures are used for mechanism auditing, not as additional benchmark claims.
- **Appendix D: Scope, Limitations, and Broader Impacts.** We state the boundary of the claims, including residual-adapter Stage-II fine-tuning, heterogeneous-expert dependence, validation-set competence calibration, single-expert diagnostics, closed-set recognition, and potential societal implications of improved long-tailed recognition.

A Formalizing the Suppression Cascade

This section formalizes the claim structure used in the main paper. The goal is not to provide a universal theory of multi-task optimization, but to bound the empirical claim made in the reported setting: residual-adapter Stage-II long-tailed multi-expert fine-tuning. In this setting, rare-class evidence must pass through three interfaces. It must be generated by the source loss, preserved by gradient coordination, and used by the inference rule. A failure at any interface can reduce the final rare-class signal, and multiple failures compound multiplicatively.

The formalization separates two coordination phenomena that should not be conflated. Local direction correction can remove an angularly conflicting component, but it cannot generate missing rare-class gradient energy. Global energy-penalizing objectives are different: they can actively compress rare-sensitive effective update components even when the raw rare-sensitive source gradient remains present before coordination. This distinction is the formal counterpart of the main paper's solver-family separation: direction-modifying and simple aggregation rules remain close under the controlled setting, whereas global energy-penalizing objectives can collapse few-class accuracy.

A.1 Residual-Adapter Stage-II Expert-Objective Regime

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a long-tailed training set over C classes. Stage I trains a backbone on $\mathcal{D}$. In Stage II, we decompose the model into a frozen convolutional encoder, a shared residual feature adapter, and T heterogeneous expert heads:

$$f_t(x) = h_{\psi_t}(A_\phi(E_{\theta_{\text{enc}}}(x))), \qquad t \in \{1, \dots, T\}. \tag{11}$$

Here $E_{\theta_{\text{enc}}}$ denotes the frozen convolutional encoder, A_ϕ denotes the shared trainable residual adapter inserted after global pooling, and h_{ψ_t} denotes the classifier head of expert t.

The trainable Stage-II parameter set is

$$\Phi = \phi \cup \{\psi_t\}_{t=1}^{T}. \tag{12}$$

Each expert head induces a loss

$$\{\mathcal{L}_t\}_{t=1}^{T}, \tag{13}$$

and the corresponding expert-objective gradient is

$$g_t = \nabla_\Phi \mathcal{L}_t. \tag{14}$$

A coordination rule $\mathcal{C}$ maps expert gradients into a single update,

$$g_{\mathcal{C}} = \mathcal{C}(g_1, \ldots, g_T). \tag{15}$$

This is not a strict head-only protocol, because the shared residual adapter belongs to Φ. The comparison remains controlled because all solver families are evaluated with the same trainable parameter set Φ, the same expert losses, and the same downstream inference rule. When $T = 1$, the coordination interface disappears; any scalar rescaling of the single gradient is equivalent to changing the effective learning rate rather than reallocating energy across experts. This is why single-expert runs are used only as source-construction diagnostics rather than as tests of inter-expert gradient coordination (Appendix D).

A.2 Head-Class Procedural Bias

Definition 2 (Head-class procedural bias)**.** In a long-tailed multi-expert pipeline, an operation has *head-class procedural bias* if, under class-frequency imbalance, it assigns systematically larger effective contribution to head-dominant evidence than to rare-sensitive evidence without an explicit correction for frequency-proportional signal strength.

This definition is procedure-level rather than module-specific. A source loss can be biased by producing weak rare-class gradients. A coordination rule can be biased by assigning little effective update energy to rare-sensitive components. An inference rule can be biased by following over-confident head-dominant predictions. These biases are sequential: each interface receives a signal already shaped by the preceding interface.

Let r denote a rare-sensitive expert and h a head-dominant expert. One observable symptom of source-interface rare-signal deficiency is a small pre-coordination gradient share for the rare-sensitive expert,

$$\pi_t = \frac{\|g_t\|_2}{\sum_{j=1}^{T} \|g_j\|_2 + \epsilon}, \tag{16}$$

with

$$\pi_r \ll \pi_h. \tag{17}$$

This condition is not a complete definition of rare-class evidence. It is an optimization-budget proxy: when rare-sensitive evidence enters coordination with weak magnitude, downstream coordination and inference have little signal to preserve. This diagnostic share π_t is separate from the loss-normalized importance score used by GES in the main method,

$$q_t = \frac{\|g_t\|_2}{\mathcal{L}_t + \epsilon}, \tag{18}$$

which measures gradient strength relative to the current expert loss rather than the share of total gradient norm.

A.3 Rare-Signal Retention and Cascade Multiplication

Let S_0 denote the rare-class evidence available before the three-interface correction chain. Let S_1, S_2, and S_3 denote the rare-class evidence retained after source construction, gradient coordination, and inference arbitration, respectively. Here S_m is a conceptual scalar proxy for retained rare-class

evidence, not an additional objective optimized by the model. Define the interface-wise retention factor

$$\rho_m = \frac{S_m}{S_{m-1} + \epsilon}, \qquad m \in \{1, 2, 3\}. \tag{19}$$

The final retained rare-class signal is

$$S_3 = S_0 \prod_{m=1}^{3} \rho_m. \tag{20}$$

Definition 3 (Suppression cascade)**.** A suppression cascade occurs when one or more interfaces impose $\rho_m < 1$ on rare-class evidence, so that the final retained signal S_3 is reduced after source construction, gradient coordination, or inference arbitration. Multiple such reductions compound multiplicatively.

This definition makes the chain dependency explicit. If the source interface does not generate rare-class evidence, the coordination interface has little rare-sensitive signal to preserve. If source evidence is generated but the coordination objective compresses its effective update component, the gain is erased before it reaches the parameters. If both source construction and coordination are corrected but inference follows confidence-biased averaging, the preserved rare-sensitive expert can still be under-used at prediction time.

The controlled experiments provide an observable bridge to this conceptual cascade model. Since the latent signal variables S_m are not directly observable, we use Few-class accuracy only as an accuracy-level proxy for retained rare-class evidence, rather than as a direct measurement of S_m.

On CIFAR-100-LT ($\rho = 100$), the uncorrected baseline obtains 34.92% Few-class accuracy. Activating TAILSPEC alone raises the Few-class result to 44.10%, corresponding to an observable source-side proxy gain of

$$\widehat{\rho}_1^{\text{TAILSPEC}} = \frac{44.10}{34.92} = 1.26. \tag{21}$$

Equivalently, this is a $+9.18\%$ Few-class gain at the source interface. This empirical bridge supports the main-paper claim that source construction contributes the dominant rare-class evidence in the reported configuration.

However, source-side amplification is not sufficient. In the controlled solver comparison, all solvers use the same expert pool and downstream inference rule, while only the coordination objective is changed. Under this controlled setting, the full GALE path reaches 46.11% Few-class accuracy, whereas replacing GES with Nash-MTL yields 21.82%. The corresponding coordination-stage collapse proxy is

$$\widehat{\kappa}_{\text{coord}}^{\text{Nash}} = \frac{21.82}{46.11} = 0.47, \tag{22}$$

with an absolute Few-class drop of

$$\Delta_{\text{Few}}^{\text{Nash}} = 21.82 - 46.11 = -24.29\ \%. \tag{23}$$

Thus, the Nash-MTL failure is not explained by the absence of rare-class source evidence alone. Under the same source and inference interfaces, an energy-penalizing coordination objective can still erase the effective rare-sensitive update. This is the empirical counterpart of the suppression-cascade model: rare-class evidence must be generated at the source, preserved through coordination, and finally used by the inference rule.

A.4 Direction-Only Correction Is Non-Amplifying

We next formalize the limited role of local direction correction. Consider two expert-objective gradients g_r and g_h, where r denotes a rare-sensitive expert and h denotes a head-dominant expert. A PCGrad-style projection modifies g_r only when $\langle g_r, g_h \rangle < 0$:

$$\tilde{g}_r = g_r - \frac{\langle g_r, g_h \rangle}{\|g_h\|_2^2} g_h. \tag{24}$$

This removes a conflicting component of g_r, but it does not introduce additional rare-class energy.

Proposition 1 (Projection is non-amplifying)**.** For any nonzero vector u and any projection-only update of the form

$$\tilde{g} = g - \frac{\langle g, u \rangle}{\|u\|_2^2} u, \tag{25}$$

the projected vector satisfies

$$\|\tilde{g}\|_2 \leq \|g\|_2. \tag{26}$$

Thus, projection cannot by itself amplify a rare-sensitive signal.

Proof. Let

$$p = \frac{\langle g, u \rangle}{\|u\|_2^2} u \tag{27}$$

be the orthogonal projection of g onto the span of u. Then $\tilde{g} = g - p$ is orthogonal to p. By the Pythagorean theorem,

$$\|g\|_2^2 = \|\tilde{g}\|_2^2 + \|p\|_2^2. \tag{28}$$

Since $\|p\|_2^2 \geq 0$, we have $\|\tilde{g}\|_2^2 \leq \|g\|_2^2$, and therefore $\|\tilde{g}\|_2 \leq \|g\|_2$, as claimed.

Proposition 1 does not imply that projection is useless in all multi-task learning problems. It supports a narrower statement: projection is not a signal-generation or signal-amplification operation. Therefore, if the operative bottleneck is missing rare-class source evidence or insufficient rare-sensitive effective update energy, direction-only correction should not be expected to create the missing signal. The controlled solver comparison in the main paper tests the stronger empirical claim that, under the reported residual-adapter Stage-II setting, projection-style solvers remain close to simple aggregation once the source signal has already been frequency-calibrated.

A.5 Structural Risk of Global Energy-Penalizing Objectives

Global multi-objective solvers differ from local projection. Many construct an update by optimizing a global objective such as minimum norm, constrained descent, or task-weighted energy balancing. Such objectives can be misaligned with rare-sensitive optimization because they may prefer low-energy updates over rare-specific components that are necessary for few-class recognition.

We illustrate this risk with a minimal example. Let a head-dominant gradient and a rare-sensitive gradient be

$$g_h = (1, 0), \qquad g_r = (1, \delta), \qquad 0 < \delta \ll 1. \tag{29}$$

The two gradients have approximately comparable magnitudes because

$$\|g_h\|_2 = 1, \qquad \|g_r\|_2 = \sqrt{1 + \delta^2} \approx 1. \tag{30}$$

However, g_r contains a rare-specific component in the second coordinate. Consider a convex combination

$$g(\alpha) = \alpha g_r + (1 - \alpha) g_h = (1, \alpha\delta), \qquad \alpha \in [0, 1]. \tag{31}$$

A minimum-norm energy objective minimizes

$$\|g(\alpha)\|_2^2 = 1 + \alpha^2 \delta^2, \tag{32}$$

whose minimizer is $\alpha^\star = 0$. The resulting update is

$$g(\alpha^\star) = g_h, \tag{33}$$

and the rare-specific component is eliminated:

$$\big[g(\alpha^\star)\big]_2 = 0. \tag{34}$$

This example is not intended to model every global solver exactly. It shows the structural risk: a task-agnostic energy objective can compress rare-sensitive effective update energy because it has no term that rewards its frequency-proportional value. This is distinct from local direction correction. Local projection is non-amplifying, whereas global energy penalization can actively prefer low-energy updates that remove tail-sensitive update components.

The empirical evidence in the main paper and Appendix C tests this structural risk under the actual GALE expert pool. Nash-MTL and MGDA are used as clean diagnostics because their global

objectives make the energy-compression failure directly observable: the raw TAILSPEC gradient remains large, while the effective TAILSPEC update is compressed by orders of magnitude. This is why the claim is not that every non-GES solver fails, nor that GES is a universal optimizer. The claim is that global energy-penalizing objectives can be structurally incompatible with preserving rare-sensitive effective updates in this residual-adapter Stage-II long-tailed multi-expert regime.

A.6 Formal Scope

The formal statements above have the same scope as the main paper. We do not claim that direction correction is irrelevant in all multi-task learning settings, nor that every global multi-objective solver fails in every long-tailed problem. The claim is restricted to residual-adapter Stage-II multi-expert fine-tuning for long-tailed recognition, where the convolutional encoder is frozen and the trainable parameters are the shared residual adapter and heterogeneous expert heads.

The formal conclusion is therefore not that GES is a universal optimizer. It is that the reported pipeline contains a suppression cascade. Source construction must generate rare-class evidence; coordination must avoid compressing its effective update contribution; and inference must route the preserved evidence to the final prediction. This is why GALE is formulated as a three-interface cascade-correction framework rather than as a standalone gradient solver.

B Implementation, Protocol, and Complexity

This section specifies the executable GALE path used for all reported results. Appendix A defines the formal claim scope; here we specify the reproducibility-relevant protocol: datasets, reporting convention, residual-adapter Stage-II training, the reported projection-free GES path, validation-set calibration, diagnostic logging, computational resources, and overhead. We avoid repeating the full method derivations and benchmark tables already given in the main paper.

B.1 Datasets, Metrics, and Reporting Convention

We evaluate on CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018 using the same long-tailed recognition splits and metrics as in the main paper. CIFAR-100-LT is constructed by exponentially downsampling CIFAR-100 according to imbalance ratio ρ; the primary setting is $\rho = 100$, where the most frequent class contains 500 training samples and the rarest class contains 5 samples.

We report Top-1 overall accuracy (OA) and group-wise accuracy over Many, Medium, and Few classes following the standard class-frequency partition used by the evaluation scripts. For the primary CIFAR-100-LT ($\rho = 100$) result, we report mean and standard deviation over four independent seeds:

$$\mathcal{S} = \{42, 123, 456, 789\}. \tag{35}$$

For each seed, we select the Stage-II checkpoint according to the validation protocol and report the corresponding test accuracy. The reported value is mean$\pm$std using the sample standard deviation. Standard deviations in the main tables are rounded to one decimal place; the reported few-class standard deviation for the full GALE configuration is 0.23% before rounding.

Single-seed diagnostics, one-at-a-time sweeps, and branch-removal experiments are used only for mechanism analysis. They are explicitly marked as diagnostics and are not mixed with the four-seed benchmark result.

B.2 Reported Residual-Adapter Stage-II Protocol

Stage I. Stage I trains the backbone on the imbalanced training set. We use ResNet-34 for CIFAR-100-LT and ResNet-50 for ImageNet-LT and iNaturalist 2018. The backbone is trained with SGD and cosine annealing following the corresponding long-tailed recognition protocol.

Stage II. Stage II follows a residual-adapter expert fine-tuning protocol. The convolutional encoder is frozen up to the pooled feature representation, while a shared residual feature adapter A_ϕ and three heterogeneous classifier heads are fine-tuned for 20 epochs:

$$\{h_t\}_{t=1}^{T} = \{\text{TAILSPEC}, \text{BCL}, \text{LOS}\}, \qquad T = 3. \tag{36}$$

For an input x, expert t predicts through

$$f_t(x) = h_{\psi_t}(A_\phi(E_{\theta_{\text{enc}}}(x))), \tag{37}$$

where $E_{\theta_{\text{enc}}}$ denotes the frozen encoder, A_ϕ denotes the shared trainable residual adapter, and h_{ψ_t} denotes the classifier head of expert t.

The trainable Stage-II parameter set is

$$\Phi = \phi \cup \{\psi_t\}_{t=1}^{T}. \tag{38}$$

All controlled solver comparisons use the same trainable parameter set Φ, the same expert pool, and the same downstream inference rule; only the gradient coordination rule is changed.

For a mini-batch (x, y), each expert induces a loss ℓ_t and a Stage-II expert-objective gradient

$$g_t = \nabla_\Phi \ell_t. \tag{39}$$

The reported GES update uses the same scaled mean aggregation rule as Eq. 7 in the main paper:

$$g_{\text{GES}} = \frac{1}{T} \sum_{t=1}^{T} \gamma_t g_t. \tag{40}$$

The trainable Stage-II parameters are then updated by

$$\Phi \leftarrow \Phi - \eta g_{\text{GES}}. \tag{41}$$

When only a single expert objective is used, the coordination interface disappears and GES degenerates to scalar gradient scaling. Such single-expert runs are used only as source-construction diagnostics, not as multi-expert coordination experiments.

B.3 Executable GALE Path

The reported GALE path contains three interfaces.

Source interface. TAILSPEC constructs the rare-sensitive source expert using class-frequency calibrated logit geometry, inverse-frequency weighting, and source temperature. Its role is to generate rare-class evidence before gradient coordination. TAILSPEC is not treated as a standalone replacement for the full GALE chain.

Gradient interface. GES estimates expert importance from a loss-normalized gradient signal and a conflict diagnostic. The conflict term is used only as an importance feature. The reported update is scaled averaging over the shared Stage-II trainable parameter set Φ. GES path.

Inference interface. After Stage-II training, COMPARB estimates a class-wise expert competence table from the validation split and keeps this table fixed during test-time prediction. Test-time prediction uses model outputs and the fixed competence table; no test label is used for routing or probability fusion.

Compact procedure.

1. Train the backbone in Stage I.
2. Freeze the convolutional encoder up to the pooled feature representation.
3. Fine-tune the shared residual adapter with $\{\text{TAILSPEC}, \text{BCL}, \text{LOS}\}$ heads.
4. For each Stage-II mini-batch, compute expert losses and expert-objective gradients over $\Phi = \phi \cup \{\psi_t\}_{t=1}^{T}$.
5. Apply GES to obtain the scaled mean update.
6. Update Φ, i.e., the shared residual adapter and expert heads.
7. After Stage-II training converges, estimate validation-set class-wise competence from the final checkpoint and fix the competence table for test-time prediction.

Table 6: **Key mechanism controls for the reported GALE path.** Only the controls most directly tied to source construction, gradient-energy allocation, and validation-calibrated arbitration are shown. Secondary implementation constants are deferred to the released code and sensitivity diagnostics.

Module	Control	Value	Role
TAILSPEC	τ	0.1	Source temperature for frequency-calibrated evidence construction.
TAILSPEC	w	3.25	Strength of tail-sensitive source evidence.
GES	β	0.65	Conflict contribution in expert-importance fusion.
GES	λ_{boost}	2.2	Boost for high-importance experts (`-pc_boost`).
COMPARB	α	1.3	Sharpness of class-wise competence estimation.
COMPARB	τ_{fuse}	1.0	Temperature for competence-based fusion.

B.4 Mechanism-Level Configuration

The main paper exposes only the mechanism-critical controls needed to define the reported path: source temperature in TAILSPEC, conflict weighting in GES, and competence sharpness in COMPARB. Table 6 reports a compact view of these controls. We list at most two mechanism-facing controls per interface and avoid inactive compatibility arguments.

The reported GES update is scaled averaging over the trainable Stage-II parameter set Φ. Conflict statistics enter the importance estimate through β. One-at-a-time sweeps are used only for implementation sanity checks and fixed-constant selection. They are not used as benchmark evidence and are not mixed with the four-seed main result.

B.5 Diagnostic Quantities

The supplementary figures in Appendix C audit the mechanism behind the controlled solver comparison rather than introducing additional benchmark claims. We define the diagnostic quantities used in these figures here to keep the reporting convention consistent across source, coordination, and late-window analyses.

Let e index logged Stage-II epochs or mini-batch summaries, and let W_{late} denote the last 20% of the logged Stage-II trajectory. Epoch-quartile figures partition the same trajectory into four consecutive windows, denoted Q1–Q4.

For solver s, the raw TAILSPEC source magnitude is measured before coordination as

$$M_{\text{raw}}^{(s)}(e) = \left\| g_{\text{TAILSPEC}}^{(s)}(e) \right\|_2 . \tag{42}$$

The solver-weighted effective TAILSPEC update magnitude is measured after coordination and is denoted by

$$M_{\text{eff}}^{(s)}(e). \tag{43}$$

Thus, M_{raw} asks whether the source interface produced a rare-sensitive gradient, whereas M_{eff} asks whether the coordination rule preserved that signal in the final update.

The late-window compression ratio is

$$R_s = \frac{\text{mean}_{e \in W_{\text{late}}} M_{\text{raw}}^{(s)}(e)}{\text{mean}_{e \in W_{\text{late}}} M_{\text{eff}}^{(s)}(e) + \epsilon}. \tag{44}$$

Larger R_s indicates stronger post-coordination compression of the TAILSPEC signal. Ratios below 1 indicate that the effective TAILSPEC update magnitude exceeds the raw late-window magnitude. These diagnostics are used only to interpret solver behavior within the reported Stage-II setting and should not be read as four-seed benchmark statistics.

Table 7: **Additional complexity of GALE.** Here D denotes the trainable Stage-II parameter dimension, including the shared residual adapter and expert heads. The pairwise GES term is a conflict diagnostic, not a projection step.

Component	Training overhead	Inference overhead
TAILSPEC	$\mathcal{O}(BC)$	none beyond expert logits
GES scaling	$\mathcal{O}(TD)$	not used
GES conflict diagnostic	$\mathcal{O}(T^2D)$	not used
COMPARB table estimation	$\mathcal{O}(TC)$ on validation split	$\mathcal{O}(TC)$ or $\mathcal{O}(T)$

B.6 Validation-Set Calibration

COMPARB estimates class-wise expert competence from the validation split after Stage-II training. Validation labels are used only for calibration and model selection. They are not used to update the shared residual adapter, expert heads, or any other Stage-II trainable parameters. No test labels are used to route test samples or compute final predictions. During test evaluation, the competence table is fixed before predictions are made.

This calibration follows the standard model-selection setting rather than test-time label routing. Datasets with very small validation splits, noisy validation labels, or validation-test distribution shift may require cross-validation or confidence-regularized competence estimation.

B.7 Computational Resources

All experiments reported in this paper were conducted using a compute node equipped with four NVIDIA V100 GPUs. For the CIFAR-100-LT four-seed experiments, independent runs were scheduled across the four GPUs; each run used one V100 unless otherwise stated. Diagnostic logging runs, one-at-a-time hyperparameter sweeps, and controlled solver comparisons used the same hardware pool. We report the GPU resources for reproducibility checklist compliance; the method does not require specialized hardware beyond standard GPU training for the corresponding backbone.

B.8 Complexity

Let T be the number of experts, D the number of trainable Stage-II parameters in Φ, B the mini-batch size, and C the number of classes. In the reported residual-adapter Stage-II protocol, D includes the shared residual adapter and the expert-head parameters.

The additional training-time cost of TAILSPEC is $\mathcal{O}(BC)$ for class-wise logit calibration and weighting. GES costs $\mathcal{O}(TD)$ for expert-gradient norms and scaling, and $\mathcal{O}(T^2D)$ when the conflict diagnostic is computed. The $\mathcal{O}(T^2D)$ term is only the pairwise cosine diagnostic used to form the importance feature c_t; it is distinct from the $\mathcal{O}(T^2)$ residual projection branch that the reported GES path deliberately omits. COMPARB stores a competence table of size $\mathcal{O}(TC)$. At inference, competence-weighted probability fusion costs $\mathcal{O}(TC)$, while predicted-class arbitration after expert top predictions are available costs $\mathcal{O}(T)$.

C Supplementary Empirical Diagnostics

This section reports supplementary diagnostics that are not duplicated in the main paper. The goal is not to add more benchmark evidence, but to audit the mechanisms and practical boundary behind the suppression-cascade interpretation: (i) whether the source interface produces transferable rare-class evidence, (ii) whether coordination-stage failures arise from solver-weighted compression rather than absence of the raw rare-sensitive source gradient, (iii) whether the mechanism-critical controls operate in stable regimes, and (iv) whether the additional computation is concentrated in Stage-II training rather than in deployed inference.

Table 8: **Source-interface transfer diagnostic on CIFAR-100-LT ($\rho = 100$).** Each block fixes one co-expert family and LOS, then compares the KPS reference with alternative source constructions under the same residual-adapter Stage-II protocol and downstream inference rule. These are diagnostic runs rather than four-seed benchmark comparisons. TAILSPEC consistently improves Few-class accuracy across co-expert families, indicating transferable rare-class source evidence.

Co-expert family	Source construction	Many	Medium	Few	OA
CE	KPS reference	77.77	52.60	39.13	57.37
	BCL	78.31	52.23	33.10	55.62
	SHIKE	78.97	52.43	33.33	55.99
	TAILSPEC replacing KPS	77.91	52.26	**46.03**	**59.37**
	TAILSPEC position control	78.34	51.86	44.30	58.86
LDAM-DRW	KPS reference	75.51	55.57	39.03	57.59
	BCL	78.89	55.17	34.57	57.29
	SHIKE	78.83	54.46	34.50	57.00
	TAILSPEC replacing KPS	74.57	54.11	**46.00**	**58.84**
	TAILSPEC position control	76.86	54.37	42.10	58.56
Balanced Softmax	KPS reference	76.89	55.00	39.57	58.03
	BCL	78.20	54.11	34.33	56.61
	SHIKE	78.80	54.49	34.20	56.91
	TAILSPEC replacing KPS	76.57	53.89	**46.90**	**59.73**
	TAILSPEC position control	77.29	53.37	45.50	59.38
RIDE	KPS reference	78.31	51.03	41.17	57.62
	BCL	79.23	52.09	32.83	55.81
	SHIKE	78.86	52.14	33.57	55.92
	TAILSPEC replacing KPS	78.26	49.91	**46.33**	**58.76**
	TAILSPEC position control	78.26	50.17	45.73	58.67
Average gain of TAILSPEC replacing KPS		–	–	**+6.59**	**+1.52**
Average gain of TAILSPEC position control		–	–	**+4.68**	**+1.22**

C.1 Source-Interface Transfer Diagnostic

Source-transfer interpretation. Table 8 tests whether the source-interface gain of TAILSPEC transfers beyond the canonical co-expert configuration used in the main text. Replacing KPS with TAILSPEC improves Few-class accuracy across all four co-expert families, with an average Few-class gain of $+6.59\%$ and an average OA gain of $+1.52\%$. The position-control variant also improves over the KPS reference, showing that the gain is not tied to a single expert ordering. In contrast, BCL and SHIKE replacements reduce Few-class accuracy in most blocks. This supports TAILSPEC as a transferable source-interface correction rather than a gain from merely changing the expert-pool size.

The $+8.3\%$ advantage of TAILSPEC over KPS reported in the main text is measured in the fixed BCL+LOS co-expert setting (46.0% versus 37.7% Few-class accuracy). The $+6.59\%$ gain in Table 8 is instead averaged over four different co-expert families. These two numbers therefore have different denominators and should not be compared as repeated estimates of the same quantity: the main-text $+8.3\%$ measures the canonical source-interface comparison, whereas this appendix table tests whether the TAILSPEC source advantage transfers across co-expert choices.

C.2 Solver-Weighted Rare-Signal Compression

Late-window compression ratio. To summarize how much of the raw TAILSPEC signal survives solver coordination, we compute the late-window compression ratio using the same diagnostic convention as Eq. 44. For completeness, the quantity visualized in Figure 4 is

$$R_s = \frac{\mathrm{mean}_{e \in W_{\mathrm{late}}} \left\| g^{(s)}_{\mathrm{TAILSPEC}}(e) \right\|_2}{\mathrm{mean}_{e \in W_{\mathrm{late}}} \left\| \tilde{g}^{(s)}_{\mathrm{TAILSPEC}}(e) \right\|_2 + \epsilon}. \tag{45}$$

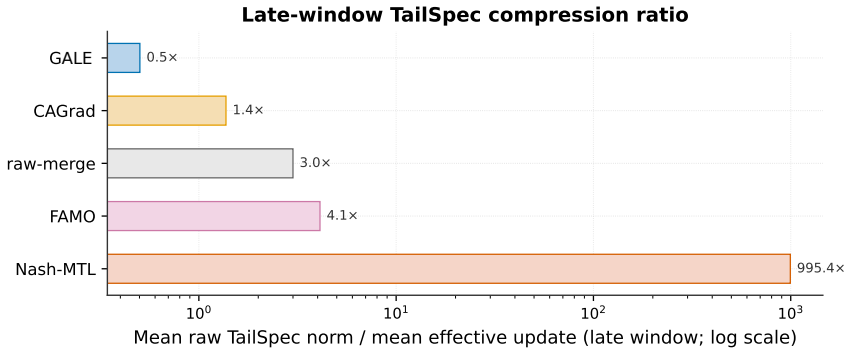


Figure 4: **Late-window TAILSPEC compression ratio.** Ratios larger than 1 indicate that the raw TAILSPEC gradient is compressed after solver coordination. GES is the only method with a ratio below 1, indicating that its effective TAILSPEC contribution exceeds the raw late-window magnitude. In contrast, Nash-MTL compresses the effective TAILSPEC contribution by nearly three orders of magnitude. These ratios are representative mechanism diagnostics and are not four-seed benchmark statistics.

Here W_{late} denotes the last 20% of the logged Stage-II trajectory, and $\tilde{g}^{(s)}_{\text{TAILSPEC}}$ denotes the solver-weighted effective TAILSPEC contribution under solver s. Larger values indicate stronger compression of the raw rare-sensitive gradient into the effective update.

Compression interpretation. Nash-MTL exhibits a compression ratio of $995.4\times$, indicating that the TAILSPEC raw gradient remains large before coordination but contributes little solver-weighted effective update magnitude in the late training window. FAMO and raw-merge show moderate compression, CAGrad shows mild compression, and GES avoids late-window down-weighting. This supports the interpretation that the collapse of global energy-penalizing solvers is not a source-interface failure alone; the coordination objective can convert a present rare-sensitive source signal into a weak effective update.

The moderate compression under raw-merge reflects its uniform $1/T$ weighting of all experts, which reduces the absolute TAILSPEC magnitude but does not selectively suppress the rare-sensitive update component. This explains why raw-merge can achieve statistically comparable Few-class accuracy to GES despite its higher compression ratio. In contrast, the Nash-MTL ratio reflects selective solver-weighted compression of the rare-sensitive effective update, which is consistent with its severe Few-class collapse in the controlled solver comparison.

C.3 Quartile-Level Effective Update Diagnostics

The scalar compression ratio summarizes only the late training window. We therefore inspect raw and effective TAILSPEC magnitudes across training quartiles in Figure 5. This separates two cases that would otherwise be conflated: a solver may receive a large raw rare-sensitive gradient but still suppress its solver-weighted effective contribution.

Quartile-level interpretation. Figure 5 shows that GES maintains effective TAILSPEC updates comparable to or larger than the raw TAILSPEC magnitude in later training quartiles. Raw-merge and CAGrad reduce the effective contribution more noticeably, while FAMO retains high absolute effective magnitude but under a substantially larger raw-gradient scale. These patterns explain why compression should be interpreted through both raw and effective channels rather than raw gradient magnitude alone.

Together with the main-text Nash-MTL and MGDA quartile panels, Figure 5 shows that solver failures do not have a single visual form. Nash-MTL and MGDA provide the cleanest examples of global energy-penalizing collapse; FAMO and CAGrad show weaker or trajectory-dependent compression; raw-merge shows uniform averaging rather than selective global energy minimization. This is why the main claim is a solver-family separation, not a claim that every non-GES coordination rule fails by the same mechanism.

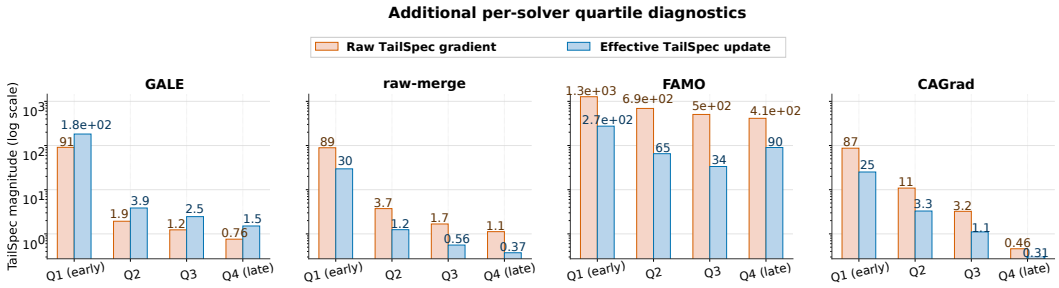


Figure 5: **Additional per-solver quartile diagnostics.** Bars compare raw TAILSPEC gradient magnitude and the solver-weighted effective TAILSPEC update across training quartiles for GES, raw-merge, FAMO, and CAGrad. Nash-MTL and MGDA quartile diagnostics are shown in the main text (Figure 2b); this figure focuses on the remaining four solvers.

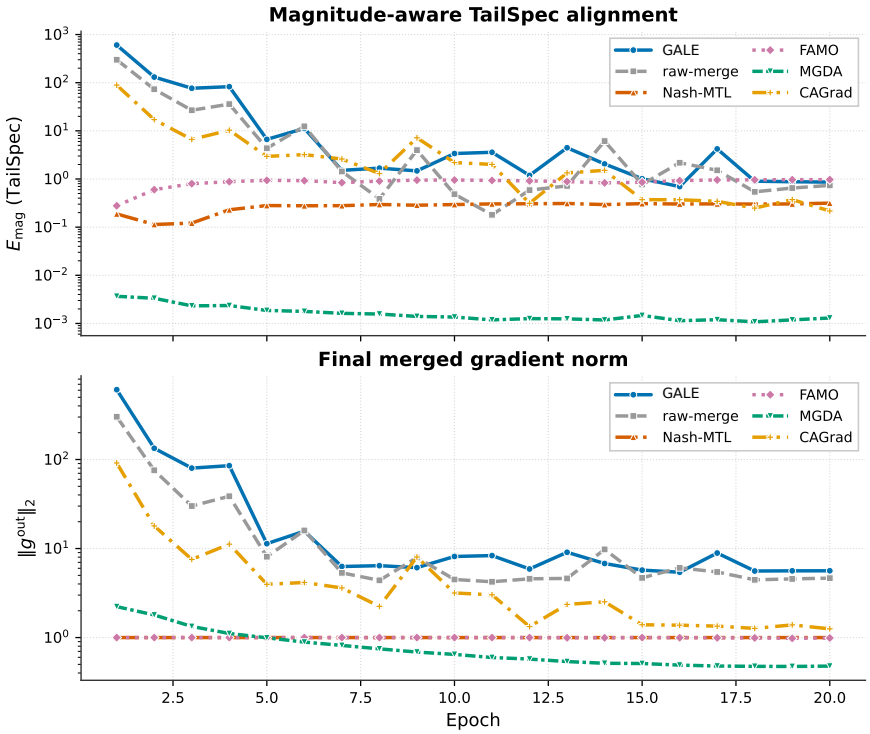


Figure 6: **TAILSPEC coordination diagnostics.** The upper panel tracks magnitude-aware TAILSPEC alignment over Stage-II training, while the lower panel tracks the final merged gradient norm after solver coordination. Together, the two panels show whether a solver preserves rare-sensitive TAILSPEC-aligned update energy and whether the coordinated update retains sufficient overall magnitude. The diagnostic illustrates that solvers differ not only in final accuracy but also in how much rare-sensitive update energy survives after coordination.

C.4 Coordination Dynamics and Late-Window Alignment

We next examine solver dynamics over epochs and summarize late-window alignment across solver–expert pairs. These diagnostics are intended to verify the mechanism behind the solver comparison, not to introduce additional benchmark claims. Figure 6 tracks TAILSPEC-related coordination dynamics during Stage-II training, while Figure 7 summarizes late-window solver–expert alignment.

Coordination-dynamics interpretation. Figure 6 shows that solver differences are not reducible to final accuracy alone: the retained TAILSPEC-related coordination signal varies substantially across solver families during Stage-II training. This supports the paper's claim that the coordination interface

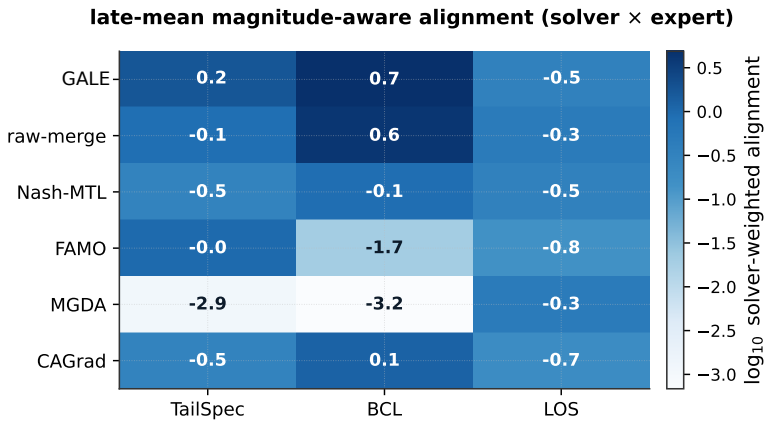


Figure 7: **Late-window magnitude-aware alignment by solver and expert.** Each cell reports the late-window $\log_{10}$ solver-weighted alignment for one solver–expert pair. The heatmap highlights how different solver families preserve or suppress expert-specific rare-sensitive update channels.

Table 9: **GES scale stability across four seeds.** Values report mean expert scales over logged Stage-II batches. The non-uniform but stable allocation supports GES as a consistent magnitude-allocation rule, while the main performance gain remains primarily attributed to source construction.

Expert	Mean GES scale
TAILSPEC	1.392 ± 0.004
LOS	1.822 ± 0.004
BCL	1.200 ± 0.000

must be evaluated through effective rare-sensitive update preservation, not only through aggregate accuracy.

Figure 7 further shows that the solver failure is not merely a change in overall gradient norm. The low TAILSPEC/BCL alignment under MGDA and the weak rare-sensitive alignment under Nash-MTL indicate that global objectives can suppress the channels needed for tail-side correction. GES, in contrast, keeps the late-window TAILSPEC and BCL channels active while avoiding a direction-projection objective.

Together, these diagnostics complement the four-seed solver comparison in Table 3. The table establishes the accuracy-level family separation, whereas Figures 6 and 7 provide mechanism-level evidence that the harmful solvers suppress rare-sensitive effective update channels rather than simply changing final prediction behavior.

C.5 Channel-Level Diagnostic Boundary

Figure 8 summarizes which late-window channel-level quantities are available for each solver. This audit is included only to delimit the scope of the compression analysis: channel-level compression claims are made only for solvers with compatible effective-update diagnostics.

Diagnostic-boundary interpretation. Figure 8 prevents a mechanism-level over-claim. The compression analysis is supported for solvers with available effective-update channels. For PCGrad and RD-PCGrad, the paper's conclusion is therefore restricted to the controlled accuracy comparison in Table 3, where projection-style coordination remains close to simple aggregation. The channel-level compression mechanism should not be extrapolated to those solvers from this diagnostic.

C.6 GES Scale Stability

Finally, we verify that GES does not produce arbitrary seed-specific scaling. Table 9 reports the mean expert scale over logged Stage-II batches and four seeds.

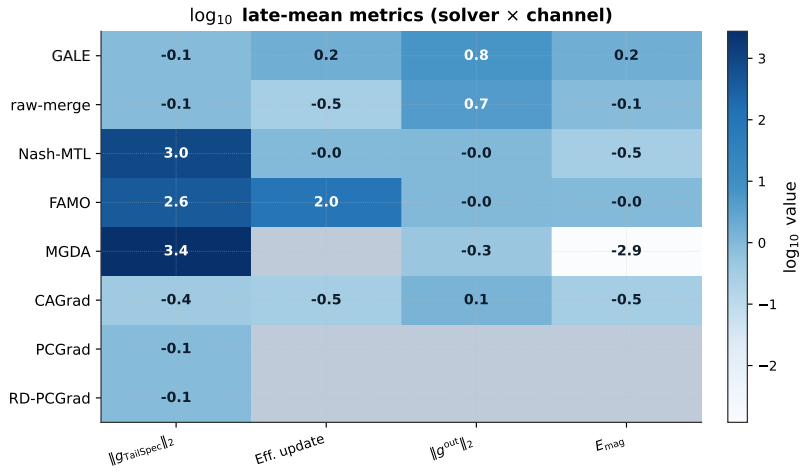


Figure 8: **Late-window solver-channel diagnostic boundary.** Each cell reports a late-window $\log_{10}$ value for one solver and one channel-level quantity. Blank cells indicate unavailable or non-positive quantities and should not be interpreted as measured solver behavior. The figure is used only to delimit the scope of the raw-versus-effective compression analysis.

Table 10: **Measured computation overhead on CIFAR-100-LT.** All runs use seed 42 and complete the full 20-epoch Stage-II schedule without internal early stopping. Forward-only profiler estimates are identical across methods at 2.32 GFLOPs/img, so the table focuses on wall-clock training cost, inference latency, and peak memory. Runtime statistics are diagnostic measurements and are not used as benchmark accuracy evidence.

Method	Train time	Epoch (s)	Train × CE	Train × TS-MOF	Infer. ms/img	Peak mem. GB
CE	0:02:54.95	8.75	1.00×	–	0.151	3.37
LOS	0:02:55.50	8.77	1.00×	–	0.168	3.37
SHIKE	0:03:05.24	9.26	1.06×	–	0.154	3.37
RIDE	0:03:12.39	9.62	1.10×	–	0.142	3.38
TS-MOF	0:05:42.07	17.10	1.96×	1.00×	0.147	3.40
GALE	0:09:12.85	27.64	3.16×	1.62×	0.159	3.40

Scale-stability interpretation. GES applies a stable non-uniform allocation across the expert pool. This supports its role as a lightweight magnitude-allocation rule within the cascade. It should not be read as evidence that GES is the primary source of the accuracy gain; the source-interface results in Appendix C.1 show that TAILSPEC provides the dominant rare-class evidence.

BCL's scale of 1.200 with zero variance indicates that it consistently reaches the lower allocation floor across seeds and logged training batches. This is consistent with the intended minimum-guaranteed allocation behavior: low-importance experts are not eliminated, but the adaptive boost is reserved for more important experts. This supports the interpretation of GES as a non-destructive energy-allocation rule rather than an energy-minimizing solver.

C.7 Measured Computation Overhead

The preceding diagnostics audit the mechanism of the suppression cascade. We additionally report measured runtime statistics to separate mechanism evidence from practical overhead. This diagnostic answers a different question: whether the additional accuracy of GALE comes with substantially larger deployed inference cost or memory footprint.

Runtime-overhead interpretation. Table 10 shows that the additional cost of GALE is primarily a Stage-II training cost. Relative to CE, GALE increases the full-run wall-clock time from 0:02:54.95 to 0:09:12.85, corresponding to 3.16× training time. Relative to the multi-expert TS-MOF baseline, the increase is 1.62×. In contrast, inference latency remains in the same range as TS-MOF (0.159 versus 0.147 ms/img), and peak memory is unchanged at 3.40GB. Thus, the overhead comes mainly

from Stage-II multi-objective backward/coordination and calibration paths rather than from a larger deployed forward model.

D Scope and Limitations

This section states the boundary of the claims. Implementation details, reporting convention, validation calibration, diagnostic logging, and computational resources are specified in Appendix B; we do not repeat them here.

D.1 Claim Scope

Residual-adapter Stage-II regime. GALE is designed for a decoupled two-stage pipeline. Stage I trains the backbone. Stage II freezes the convolutional encoder up to the pooled feature representation and fine-tunes a shared residual feature adapter together with heterogeneous expert heads. Thus, the reported setting is neither strict head-only frozen-backbone fine-tuning nor fully end-to-end long-tailed representation learning. The supported claim is restricted to residual-adapter Stage-II multi-expert fine-tuning, where the optimization problem is concentrated in the shared adapter and expert-objective heads.

Not a universal solver claim. GALE should not be interpreted as proposing GES as a standalone optimizer that must dominate every merging, projection, or multi-objective optimization variant. The empirical claim is narrower: under the controlled Stage-II expert-objective interface, local direction correction remains close to simple aggregation, whereas global energy-penalizing objectives are structurally misaligned with tail-sensitive effective update preservation in this setting, collapsing Few-class accuracy by 12–24 percentage points in the controlled solver comparison.

Dependence on heterogeneous experts. The method assumes that the expert pool contains complementary long-tailed inductive biases. If all heads are trained with nearly identical losses and converge to similar predictions, the source and arbitration interfaces become less meaningful, and the setting becomes closer to ordinary model averaging than to cascade correction.

Single-expert diagnostics. When Stage II uses only one expert objective, the coordination interface disappears. In that case, GES degenerates to scalar gradient scaling and should not be interpreted as inter-expert energy allocation. Single-expert runs are therefore used only for source-construction diagnosis, not for evaluating gradient coordination.

Closed-set assumption. The current formulation targets closed-set long-tailed recognition. It does not explicitly address open-set recognition, label noise, incremental class arrival, or test-time distribution shift. Since COMPARB relies on class-wise competence, unseen or severely shifted classes would require additional uncertainty handling.

D.2 Limitations

Limitations. The main limitation of GALE is that its claims are tied to the reported residual-adapter Stage-II multi-expert setting and to closed-set long-tailed recognition. The framework assumes that the expert pool contains useful and complementary rare-class evidence; if the source interface fails to generate such evidence, GES cannot recover it through gradient scaling alone. COMPARB also depends on a representative validation split for estimating class-wise expert competence, so very small, noisy, or shifted validation sets may require additional smoothing or cross-validation. Finally, the diagnostic quantities in Appendix C are mechanism audits for the reported solver implementations rather than a universal causal decomposition of all long-tailed optimization failures.

D.3 Broader Impacts

GALE is a method for long-tailed visual recognition. Its potential positive impact is to improve recognition of rare visual categories when such categories are important but under-represented in training data, for example in scientific, ecological, industrial, or specialized inspection settings. However, improved recognition of rare categories should not be interpreted as a general fairness

guarantee: rare classes are not necessarily socially under-represented groups, and deployment fairness depends on dataset construction, label quality, application context, and downstream decision rules. As with other recognition technologies, the method could also be misused in surveillance or other high-stakes settings where improved identification of infrequent categories may increase monitoring capability. Deployment in sensitive domains should therefore require application-specific risk assessment, validation under the target distribution, and human or institutional oversight. The paper does not introduce new personal data, biometric datasets, or deployment systems; its empirical claims remain limited to benchmark long-tailed recognition.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We list our contributions in detail at the end of the Introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the relevant assumptions and proofs in Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide code links in the abstract and implementation details in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the source of the dataset and code in Appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe the relevant details in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report performance metrics such as average accuracy on multiple datasets.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We mention the computational resources used for the experiments in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and understood the code of ethics; and have done our best to conform.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our experiments are conducted on publicly available data.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We are not shipping our code with any source code or binary files from any other existing libraries, so there are no concerns over getting permission or including a license. We did cite open-sourced libraries, e.g. PyTorch, in our paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release our code base with included readme files. We do not ship any source code or binary files from any other existing libraries.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our work does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.