EQUIVALENT PREDICTORS, DIFFERENT UPDATES: CANONICAL-COORDINATE STRATEGY FUSION FOR LONG-TAILED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-Tailed Recognition (LTR) increasingly relies on heterogeneous learning strategies with complementary strengths across class-frequency regions. Gradient-based strategy fusion coordinates these strategies through their full-parameter gradients. We identify a non-invariance: functionally equivalent parameterizations of an expert-private classifier can preserve the predictor, loss, and shared-parameter gradient while producing different coordinated updates. The chosen private coordinates therefore become an unintended factor in how LTR strategies are coordinated, a phenomenon we call *coordinate leakage*. We propose **Canonical-Coordinate Strategy Fusion (CCSF)**, which transports the affected gradients to a common reference coordinate, applies the original fusion rule and optimizer there, and maps the resulting update back to parameter storage. Controlled interventions across multiple full-gradient coordinators reveal the predicted coordinate sensitivity, trace its effect on expert complementarity across frequency regions, and show that CCSF numerically recovers the corresponding reference trajectories. Experiments evaluate the complete configuration on CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018. Code is provided in the supplementary material.

1 INTRODUCTION

Long-tailed distributions are prevalent in real-world visual data and remain challenging for deep models, which tend to favor data-rich head classes while underperforming on data-scarce tail classes (Liu et al., 2019; Cao et al., 2019). Existing LTR methods address this imbalance through reweighting and logit adjustment (Ren et al., 2020; Menon et al., 2020), decoupled representation learning and classifier adjustment (Kang et al., 2019; Zhang et al., 2021), contrastive learning (Cui et al., 2021; Zhu et al., 2022), and multi-expert learning (Wang et al., 2020; Aimar et al., 2023; Zhao et al., 2023). These families address different aspects of the imbalance, and individual strategies are not uniformly strong across class-frequency regions. Their complementary behavior motivates coordinating heterogeneous LTR strategies.

Recent LTR strategy-fusion methods exploit this complementarity by jointly optimizing heterogeneous objectives within a shared model (Zhao et al., 2026). Related gradient-based multi-task and multi-objective methods coordinate objectives by weighting or projecting task gradients according to their magnitudes, directions, and pairwise relations (Sener & Koltun, 2018; Chen et al., 2018; Yu et al., 2020; Liu et al., 2021; Navon et al., 2022), making the resulting coordination dependent on task-gradient geometry. In heterogeneous LTR fusion, these full-parameter gradients contain components for both shared parameters and expert-private classifiers. The gradient geometry used for coordination can therefore depend on how an expert-private classifier is parameterized, even when the predictor and learning objective are unchanged.

We identify a non-invariance of full-gradient strategy fusion under function-preserving changes of private coordinates. Consider the private classifier of an LTR strategy, with effective weight W and an equivalent stored parameterization $U = aW$, where $a > 0$. Evaluating the classifier through $W = U/a$ preserves its logits, loss, and shared-parameter gradient, while the private gradient rescales as $g_U = g_W/a$. Since full-gradient fusion depends on gradient norms, cosine similarities, and projection terms, this private-coordinate rescaling alters the geometry used to

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

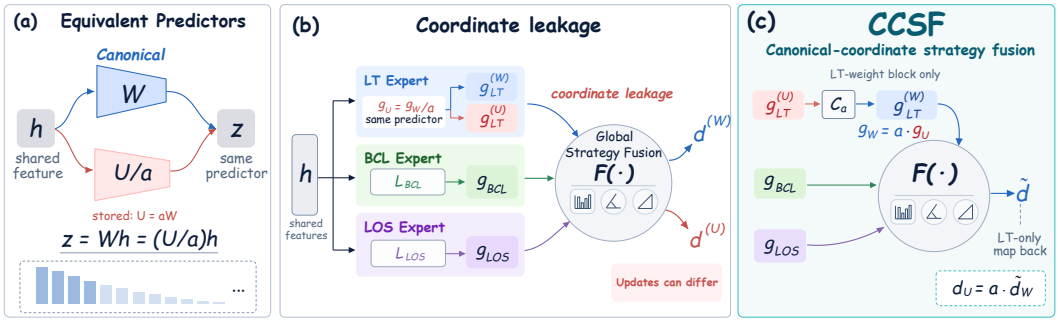


Figure 1: Conceptual overview of coordinate leakage and CCSF. (a) Two LT parameterizations, W and $U = aW$, represent the same predictor when the stored parameter is evaluated through $W = U/a$. (b) The equivalent predictor exposes different LT-private gradient coordinates to full-gradient fusion, which can change the geometry read by the same fusion rule and produce different coordinated updates. (c) CCSF canonicalizes only the affected LT-private gradient before fusion, applies the base fusion rule unchanged, and maps the corresponding LT update back to the stored coordinate.

coordinate heterogeneous LTR objectives. Functionally equivalent predictors can therefore produce different coordinated updates. We term this phenomenon *coordinate leakage*.

Coordinate leakage motivates coordinating equivalent private representations in a common reference coordinate. We therefore propose **Canonical-Coordinate Strategy Fusion (CCSF)**, which separates parameter storage from strategy coordination. As illustrated in Figure 1, CCSF expresses the affected private gradients in the reference coordinate, applies the unchanged fusion rule and optimizer there, and transports the resulting increment back to storage. This removes storage-coordinate dependence without changing the represented predictor, fusion rule, or optimizer.

Controlled studies across multiple full-gradient coordinators show the predicted coordinate sensitivity: functionally equivalent private representations produce different coordinated training behavior, while CCSF numerically recovers the corresponding reference trajectories. Beyond scalar rescaling, non-uniform and factor-basis conversions preserve the predictor while exposing the same coordination dependence. Expert-level analysis further shows that a stronger standalone LT expert can become less complementary after the coordinate change within the fused system. The complete configuration is evaluated on CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018.

We summarize our contributions as follows:

❶ **Coordinate leakage in LTR strategy fusion:** we show that functionally equivalent private parameterizations can preserve the predictor, loss, and shared-parameter gradient yet change coordinated updates. We identify this unintended dependence on private coordinates as *coordinate leakage*.

❷ **Canonical-Coordinate Strategy Fusion:** we propose CCSF, which executes the original fusion rule and optimizer in a common reference coordinate and transports the resulting increment back to storage, removing this coordinate dependence without changing the represented predictor.

❸ **Mechanism and LTR evidence:** experiments across full-gradient coordinators show coordinate-sensitive fusion and recovery of reference trajectories. Expert-level analysis shows coordinate leakage changes complementarity across frequency regions, with results on CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018.

2 Related Work

► **Long-Tailed Recognition and Multi-Expert Learning.** Long-tailed recognition methods span loss re-balancing and classifier adjustment (Cui et al., 2019; Cao et al., 2019; Ren et al., 2020; Menon et al., 2020; Zhang et al., 2021), decoupled and contrastive representation learning (Kang et al., 2019; Cui et al., 2021; Zhu et al., 2022), and multi-expert modeling. Multi-expert methods such as RIDE, SHIKE, BalPoE, SADE, and MDCS learn complementary or distribution-aware experts specialized across class-frequency regions (Wang et al., 2020; Jin et al., 2023; Aimar et al., 2023;

Zhang et al., 2022; Zhao et al., 2023). TS-MOF further coordinates heterogeneous LTR strategies through multi-objective Stage 2 fine-tuning (Zhao et al., 2026). These methods focus on learning complementary experts or coordinating their objectives. CCSF studies a different property of strategy fusion: how function-preserving reparameterizations of an expert-private classifier can alter the full-gradient coordination of the same LTR strategies.

► **Gradient-Based Multi-Task and Multi-Objective Optimization.** Multi-task and multi-objective methods coordinate competing objectives by modifying their relative weights or joint update directions. GradNorm adapts task weights from relative training rates (Chen et al., 2018), MGDA seeks a common descent direction (Sener & Koltun, 2018), PCGrad projects conflicting gradients (Yu et al., 2020), CAGrad optimizes a conflict-averse direction (Liu et al., 2021), and Nash-MTL formulates coordination as a bargaining problem (Navon et al., 2022). Aligned-MTL further aligns task gradients through an independent-component formulation (Senushkin et al., 2023). Recent ConsMTL uses task-specific parameters to shape gradients propagated to shared parameters and incorporates this interaction into conflict mitigation (Qin et al., 2025). These methods determine how task gradients should be coordinated. CCSF studies a complementary issue at the coordinator input: functionally equivalent private parameterizations can preserve the effective predictor and shared-parameter gradient while presenting different full-gradient geometry to the same coordinator.

► **Parameterization and Optimization Geometry.** Parameterization is known to shape optimization geometry. Batch normalization and weight normalization alter effective optimization through normalization and reparameterization (Ioffe & Szegedy, 2015; Salimans & Kingma, 2016), Path-SGD constructs rescaling-aware updates (Neyshabur et al., 2015), and function-preserving transformations can change quantities such as sharpness (Dinh et al., 2017); parameter scale can likewise alter effective training dynamics (Van Laarhoven, 2017). Natural-gradient methods use the Fisher geometry to construct parameterization-aware updates (Amari, 1998), with later analysis clarifying their approximate parameterization-invariance properties (Martens, 2020); K-FAC provides a scalable natural-gradient approximation for neural networks (Martens & Grosse, 2015). These works study how parameterization or optimization geometry affects model-level updates. CCSF instead studies how expert-private coordinates enter cross-objective full-gradient coordination: it expresses the affected private gradients in a common reference coordinate before applying the unchanged fusion rule and optimizer.

3 METHODOLOGY

Figure 2 summarizes CCSF within Stage 2 optimization: affected private gradients are expressed in a common reference coordinate, the unchanged fusion rule and optimizer execute there, and the resulting parameter increment is transported back to storage; shared and unaffected private blocks pass through unchanged.

3.1 HETEROGENEOUS FULL-GRADIENT STRATEGY FUSION

We instantiate CCSF in the two-stage heterogeneous strategy-fusion setting of prior work (Zhao et al., 2026). Stage 1 learns the visual representation, frozen during Stage 2, while a shared decoder and expert-specific heads optimize the adaptive long-tailed (LT), BCL (Zhu et al., 2022), and LOS-style (Sun et al., 2025) objectives. The stateful full-gradient operator in Appendix B.4 serves as the base coordinator; CCSF changes only its coordinate interface, and the argument applies to coordinators that jointly read shared and private gradients.

Let the trainable Stage 2 parameters be partitioned as $\theta = (\theta_s, \theta_{\mathrm{LT}}, \theta_{\mathrm{BCL}}, \theta_{\mathrm{LOS}})$, where θ_s denotes the shared decoder and the remaining blocks denote expert-specific classifiers. For each objective $t \in \mathcal{T} = \{\mathrm{LT}, \mathrm{BCL}, \mathrm{LOS}\}$, the *full task gradient* is $g_t = \nabla_\theta \mathcal{L}_t(\theta)$. A stateful fusion rule maps these full task gradients and the current fusion state to a fused direction and updated state,

$$(d, \xi^+) = \mathcal{F}\left(g_{\mathrm{LT}}, g_{\mathrm{BCL}}, g_{\mathrm{LOS}}; \xi\right), \tag{1}$$

where ξ denotes the current fusion state. Because $\mathcal{F}$ operates on the full task gradients, its weighting, similarity, projection, and fusion-state updates can depend on both shared and private gradient components. The coordinate representation of a private classifier can therefore enter the geometry used for fusion.

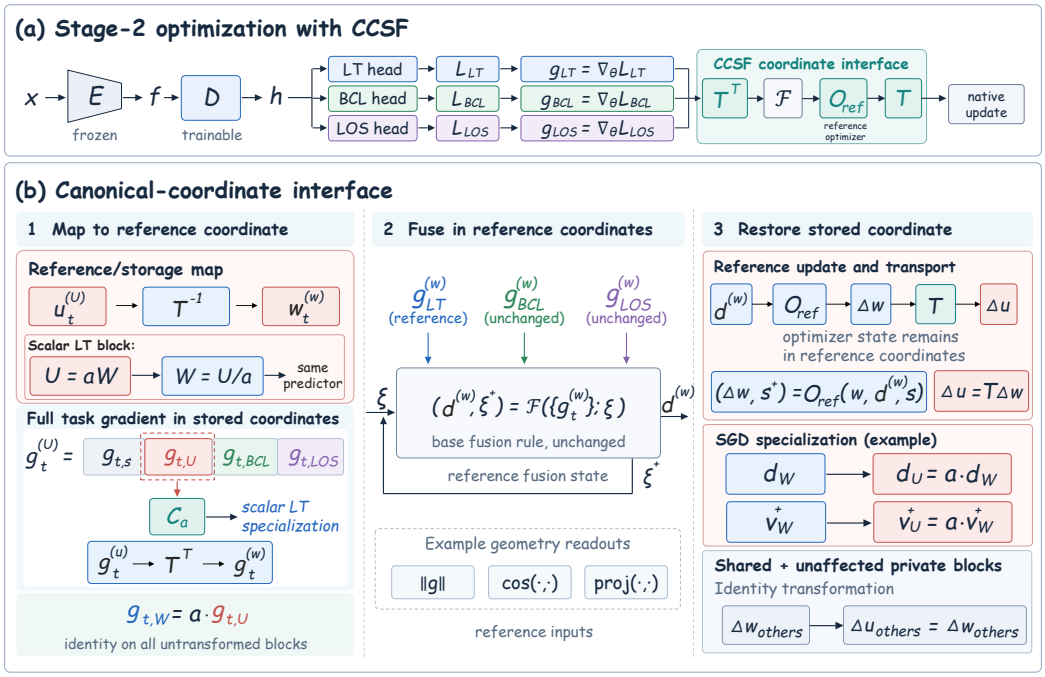


Figure 2: CCSF within Stage 2 optimization. (a) Full task gradients from the LT, BCL, and LOS heads enter the CCSF coordinate interface, where the affected private gradients are expressed in a common reference coordinate before applying the unchanged fusion rule and optimizer. (b) Only the reparameterized LT-private block is transformed; shared and unaffected private blocks pass through unchanged, and the resulting LT parameter increment is transported back to the stored U coordinate.

3.2 Coordinate Leakage in Full-Gradient Fusion

Observation 1 (Coordinate Leakage). *A function-preserving reparameterization of an expert-private classifier can preserve the predictor, loss, and gradients on all untransformed parameters while changing the private-gradient coordinates seen by full-gradient fusion. When the fusion rule depends on this coordinate-sensitive geometry, functionally equivalent parameterizations need not induce the same coordinated output.*

Functional equivalence. Let $W \in \mathbb{R}^{C\times d}$ denote the effective LT classifier weight and let U denote its stored parameterization. For a scale $a > 0$,

$$U = aW, \qquad W = U/a, \qquad z_U(h) = (U/a)h = Wh = z_W(h). \tag{2}$$

Thus, the two parameterizations define the same logits and loss at the same effective model state, and the raw gradients on all untransformed blocks remain identical. By the chain rule, the LT-weight gradient transforms as

$$g_U = \nabla_U \mathcal{L}_{\mathrm{LT}} = \frac{1}{a}\nabla_W \mathcal{L}_{\mathrm{LT}} = \frac{1}{a} g_W. \tag{3}$$

Hence, functional equivalence preserves the predictor, loss, and all untransformed raw gradients while presenting a different private-gradient coordinate to full-gradient fusion.

Full-gradient geometry. For the LT objective, write $g_{\mathrm{LT}}^{(W)} = (g_{\mathrm{LT},s}, g_W)$ in the reference coordinate and $g_{\mathrm{LT}}^{(U)} = (g_{\mathrm{LT},s}, a^{-1} g_W)$ in the stored coordinate, suppressing unaffected private blocks. For another task gradient g_j with disjoint private support, the inner product and LT-gradient norms satisfy

$$\begin{cases} \left\langle g_{\mathrm{LT}}^{(U)}, g_j \right\rangle = \langle g_{\mathrm{LT},s}, g_{j,s} \rangle = \left\langle g_{\mathrm{LT}}^{(W)}, g_j \right\rangle, \\ \left\| g_{\mathrm{LT}}^{(U)} \right\|_2^2 = \|g_{\mathrm{LT},s}\|_2^2 + a^{-2} \|g_W\|_2^2, \qquad \left\| g_{\mathrm{LT}}^{(W)} \right\|_2^2 = \|g_{\mathrm{LT},s}\|_2^2 + \|g_W\|_2^2. \end{cases} \tag{4}$$

Thus, the private rescaling leaves the cross-objective inner product unchanged but changes the full LT-gradient norm.

Proposition 1 (Private Rescaling Changes Fusion-Relevant Gradient Geometry). *Assume $a \neq 1$, $g_W \neq 0$, and $\langle g_{\mathrm{LT},s}, g_{j,s} \rangle \neq 0$. Then the private rescaling preserves the cross-objective inner product but changes the cosine similarity and the norm-normalized projection coefficient.*

Specifically, the cosine similarity satisfies

$$\frac{\cos\left(g_{\mathrm{LT}}^{(U)}, g_j\right)}{\cos\left(g_{\mathrm{LT}}^{(W)}, g_j\right)} = \frac{\left\|g_{\mathrm{LT}}^{(W)}\right\|_2}{\left\|g_{\mathrm{LT}}^{(U)}\right\|_2} \neq 1, \tag{5}$$

Likewise, the standard projection coefficient satisfies

$$\beta_{j\leftarrow\mathrm{LT}}^{(U)} = \frac{\left\langle g_j, g_{\mathrm{LT}}^{(U)} \right\rangle}{\left\|g_{\mathrm{LT}}^{(U)}\right\|_2^2} \neq \frac{\left\langle g_j, g_{\mathrm{LT}}^{(W)} \right\rangle}{\left\|g_{\mathrm{LT}}^{(W)}\right\|_2^2} = \beta_{j\leftarrow\mathrm{LT}}^{(W)}. \tag{6}$$

Thus, a coordinate change confined to one private block can alter fusion-relevant gradient geometry even though the cross-objective inner product is unchanged.

Proposition 2 (Conditional Orthogonal Equivariance). *Let $G = [g_1, \ldots, g_m]$ and $(d, \xi^+) = \mathcal{F}(G; \xi)$. Suppose the coefficients, ordering, and task-level state transition of $\mathcal{F}$ depend only on the Gram matrix $G^\top G$ and coordinate-compatible state. If Q is block orthogonal and identity on shared coordinates, with coordinate-compatible masking and coupled solver, tie-breaking, and random choices, then in exact arithmetic*

$$(QG)^\top (QG) = G^\top G, \qquad \mathcal{F}(QG; \xi) = (Qd, \xi^+). \tag{7}$$

Thus, Gram-based coordination is equivariant to such transformations, while non-orthogonal or coordinate-wise statistics can remain coordinate sensitive. Fixed-coefficient fusion provides the corresponding representation-independent boundary. Stateful coordinators can additionally retain coordinate dependence after the current shared fused direction coincides; Appendix B.2 gives the proof, full conditions, and hidden-state construction.

3.3 Canonical-Coordinate Strategy Fusion

The non-invariance above motivates separating parameter storage from the coordinate used for strategy coordination. We first describe CCSF for the scalar reparameterization in Eq. 2, using the predictor-facing weight W as the reference coordinate while retaining U as the stored representation; the general invertible case is given below.

❶ Canonicalization before fusion. For the scalar case, we define a blockwise map $\mathcal{C}_a$ that multiplies only the reparameterized LT-weight coordinates by a and acts as the identity elsewhere. Applied to each full task gradient,

$$\widetilde{g}_t = \mathcal{C}_a\left(g_t^{(U)}\right), \qquad t \in \mathcal{T}, \qquad \widetilde{g}_{\mathrm{LT}} = g_{\mathrm{LT}}^{(W)}. \tag{8}$$

The fusion rule therefore receives the affected LT-weight block in the reference coordinate, while all shared and unaffected private coordinates remain unchanged.

❷ Fusion in the reference coordinate. CCSF applies the *base fusion rule unchanged* to these reference-coordinate gradients,

$$(\widetilde{d}, \widetilde{\xi}^+) = \mathcal{F}\left(\widetilde{g}_{\mathrm{LT}}, \widetilde{g}_{\mathrm{BCL}}, \widetilde{g}_{\mathrm{LOS}}; \widetilde{\xi}\right). \tag{9}$$

The fused direction and fusion state are computed from the same reference-coordinate gradients, and the base optimizer and its state remain in this coordinate. Step 3 transports only the resulting parameter increment back to storage.

❸ Transport to the stored coordinate. Let ΔW denote the LT-weight increment produced by the reference-coordinate optimizer. For $U = aW$, CCSF applies

$$\Delta U = a\Delta W, \qquad U^+ = U + \Delta U = a(W + \Delta W) = aW^+. \tag{10}$$

All shared and unaffected private increments are applied without transformation. Thus, the stored parameter follows the same effective update as the reference-coordinate execution.

General private-coordinate maps. The scalar construction extends to a fixed, known, invertible map $u = Tw$, with T acting as the identity on shared and unaffected blocks. Parameters are evaluated through $w = T^{-1}u$, and

$$g_t^{(w)} = T^\top g_t^{(u)}, \qquad \Delta u = T\Delta w. \tag{11}$$

The first map expresses native gradients in the reference coordinate, while the second transports reference-coordinate parameter increments back to storage.

Proposition 3 (Reference-Consistent Execution). *Let $\mathcal{S}$ collect all retained states relevant to subsequent updates. Suppose $u_0 = Tw_0$, the corresponding states are initialized identically, objectives are evaluated through $w = T^{-1}u$, and both executions use the same data and stochastic choices. If CCSF applies Eq. 11 at every step while leaving the coordinator and optimizer unchanged, then in exact arithmetic*

$$T^{-1}u_k = w_k, \qquad \mathcal{S}_k^{\mathrm{CCSF}} = \mathcal{S}_k^{\mathrm{ref}} \quad \text{for every } k. \tag{12}$$

Thus, CCSF preserves the chosen reference execution across equivalent private storage representations. The scalar case is $T = aI$ on the transformed private block; optimizer-specific state evolution and factorized-coordinate extensions are given in Appendix B.3.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate CCSF on three standard long-tailed recognition benchmarks: CIFAR-100-LT (Cao et al., 2019), ImageNet-LT (Liu et al., 2019), and iNaturalist 2018 (Van Horn et al., 2018). CIFAR-100-LT follows the standard long-tailed construction with imbalance ratios (IRs) of 10, 50, and 100. ImageNet-LT and iNaturalist 2018 extend the evaluation to larger-scale long-tailed settings with 1,000 and 8,142 classes, respectively. Detailed dataset statistics and Stage 2 budgets are summarized in Table 7.

Evaluation Metrics. We report Overall Top-1 accuracy (%) together with Many/Medium/Few accuracies following the standard LTR protocol (Liu et al., 2019). Unique accuracy is used only for expert-complementarity analysis.

Comparison Baselines. We compare CCSF with representative long-tailed recognition methods spanning the main learning paradigms in recent LTR studies. For CIFAR-100-LT, we use a unified ResNet-34 comparison protocol across all methods. For ImageNet-LT and iNaturalist 2018, we follow the reported comparison protocol and baseline setting (Zhao et al., 2026). The compared methods are reported in Tables 1 and 2.

Implementation Details. All Stage 2 variants use the same frozen Stage 1 representation. We use ResNet-34 for CIFAR-100-LT and ResNet-50 for ImageNet-LT and iNaturalist 2018 (He et al., 2016). During Stage 2, the shared decoder and three classifier heads are optimized with SGD using a batch size of 64, learning rate 5×10^{-4}, momentum 0.90, and zero weight decay. CIFAR-100-LT uses LOS smoothing 0.98 for 20 cosine-annealed epochs; ImageNet-LT and iNaturalist 2018 use smoothing 0.99 for up to 20 epochs. CCSF uses $a = 2$, and the mechanism analyses report multi-seed results. Additional architecture, optimization, and evaluation details are provided in Appendix A.

4.2 Main Results

CIFAR-100-LT. As shown in Table 1, the complete CCSF configuration achieves the best Overall Top-1 accuracy across all three imbalance ratios. Relative to the reported strategy-fusion baseline (Zhao et al., 2026), CCSF is higher by 2.5%, 3.6%, and 1.2% at IR=10, 50, and 100, respectively. At IR=100, Few-class accuracy reaches 41.5%. The frequency-group response is non-uniform, motivating analysis of how coordinate choice redistributes expert contribution across the long tail. Section 4.3 further analyzes this mechanism under matched Canonical–Transformed–CCSF controls.

ImageNet-LT and iNaturalist 2018. Table 2 extends the complete CCSF configuration to two large-scale long-tailed benchmarks. CCSF reaches 56.5% on ImageNet-LT and 72.4% on iNaturalist

Table 1: Top-1 accuracy (%) on CIFAR-100-LT. Overall accuracy is reported at IR=10, 50, and 100, with the Many/Medium/Few breakdown reported at IR=100. Best results are in **bold**, second-best results are underlined; shaded CCSF results are three-seed means.

Method	Overall Top-1			IR=100		
	IR=10	IR=50	IR=100	Many	Medium	Few
CE (He et al., 2016)	67.1	52.7	47.0	77.1	45.9	13.1
CE-DRW (Cao et al., 2019)	68.1	55.1	49.2	75.3	49.1	18.9
LDAM-DRW (Cao et al., 2019)	63.4	50.4	42.8	67.3	41.4	15.9
Balanced Softmax (Ren et al., 2020)	68.3	56.7	51.3	76.4	50.9	22.5
DBM (Son et al., 2025)	65.3	50.9	46.0	75.8	44.5	12.8
KPS (Li et al., 2022)	65.8	54.1	50.0	72.8	49.5	23.9
BCL (Zhu et al., 2022)	68.5	55.7	51.2	75.6	51.6	22.3
Focal-SAM (Li et al., 2025)	68.7	56.7	51.0	76.2	56.0	15.7
BalPoE (Aimar et al., 2023)	72.4	60.5	55.9	77.7	**57.1**	29.2
ProCo (Du et al., 2024)	72.2	60.7	55.1	**79.3**	53.2	29.1
PRL (Zhao et al., 2024)	71.8	60.3	55.1	75.9	56.5	29.3
LOS (Sun et al., 2025)	68.8	57.4	53.2	72.5	48.7	35.8
SAMN (Zhang et al., 2026)	66.1	54.0	49.8	70.9	51.1	23.7
SHIKE (Jin et al., 2023)	65.1	51.2	42.2	70.8	40.8	10.6
ICL (Chen et al., 2025)	66.1	53.4	49.0	73.3	49.9	19.7
SADE (Zhang et al., 2022)	69.4	57.9	52.8	73.4	56.2	24.7
RIDE (Wang et al., 2020)	68.8	55.5	50.5	75.3	51.6	20.4
TS-MOF (Zhao et al., 2026)	70.8	60.2	56.8	79.0	49.2	39.9
CCSF	**73.3**	**63.8**	**58.0**	77.7	52.5	**41.5**

Table 2: Top-1 accuracy (%) on ImageNet-LT and iNaturalist 2018. Baseline results follow the reported comparison protocol (Zhao et al., 2026); CCSF reports multi-seed means with fixed Stage 1 models.

Method	ImageNet-LT				iNaturalist 2018			
	Many	Medium	Few	All	Many	Medium	Few	All
CE (He et al., 2016)	64.0	33.8	5.8	41.6	73.9	63.5	55.5	61.0
CE-DRW (Cao et al., 2019)	61.7	47.3	28.8	50.1	68.2	67.3	66.4	67.0
cRT (Kang et al., 2019)	58.8	44.0	26.1	47.3	69.0	66.0	63.2	65.2
LDAM-DRW (Cao et al., 2019)	60.4	46.9	30.7	49.8	–	–	–	66.1
BS (Ren et al., 2020)	60.9	48.8	32.1	51.0	65.7	67.4	67.5	67.3
KPS (Li et al., 2022)	59.7	49.2	35.9	52.3	68.1	69.5	70.2	69.6
RIDE (3 experts) (Wang et al., 2020)	64.9	50.4	34.4	53.6	69.5	71.0	70.4	70.6
BCL (Zhu et al., 2022)	65.3	**53.5**	36.3	55.6	69.5	70.9	71.3	70.9
LOS (Sun et al., 2025)	63.2	50.7	42.3	54.4	69.2	70.7	71.3	70.8
TS-MOF (Zhao et al., 2026)	65.9	52.5	**42.6**	56.3	72.8	**73.6**	70.5	72.3
CCSF	**68.0**	51.8	40.7	**56.5**	**76.2**	69.9	**74.5**	**72.4**

2018, slightly above the reported strategy-fusion baseline on both datasets. The frequency-group effects are non-uniform: on ImageNet-LT, Many-class accuracy is higher while Medium and Few are lower than TS-MOF; on iNaturalist 2018, Many and Few are higher while Medium is lower. Multi-seed coordinate controls show that CCSF matches the Canonical result on both large-scale benchmarks; matched-state analyses are reported in Appendix C.4.

4.3 In-depth Analysis

We test whether functionally equivalent private parameterizations alter full-gradient strategy fusion and whether CCSF restores the corresponding reference execution, using CIFAR-100-LT at IR=100 unless otherwise stated. At matched effective states, Transformed changes the shared fused direction in 91 of 288 evaluations (31.6%) while preserving logits, losses, and all untransformed raw

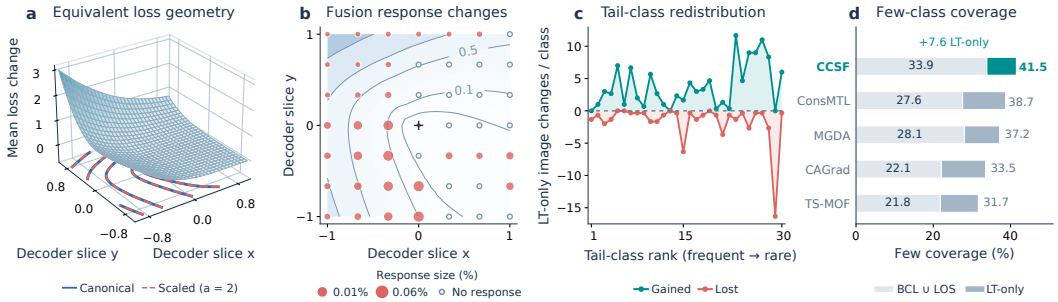


Figure 3: Mechanism and LTR evidence for coordinate leakage: (a) functional equivalence, (b) fusion non-invariance, (c) tail-class redistribution, and (d) Few-class coverage.

gradients. We then trace this local mismatch through end-to-end restoration and its class-frequency consequences.

► **Coordinate leakage extends across full-gradient fusion rules.** Under scalar rescaling, PCGrad changes 40 of 72 shared directions and MGDA 69 of 72, while fixed equal weighting remains unchanged (Table 3). Diagonal maps preserve the same qualitative separation. Orthogonal mixing follows Proposition 2 for Gram-based readers, whereas the residual response of the stateful base rule isolates coordinate-wise statistics as an additional pathway.

Table 3: Coordinate sensitivity across coordinators (72 matched pairs per map; cells report count and rate).

Coordinator	Family	Scalar	Diagonal	Orthogonal
Equal weighting	Fixed	0 (0.0%)	0 (0.0%)	0 (0.0%)
PCGrad (full)	Proj.	40 (55.6%)	30 (41.7%)	0 (0.0%)
MGDA	Descent	69 (95.8%)	69 (95.8%)	0 (0.0%)
CAGrad (full)	Conflict	72 (100.0%)	72 (100.0%)	0 (0.0%)
Aligned-MTL	Align.	72 (100.0%)	72 (100.0%)	0 (0.0%)
Base stateful	Stateful	25 (34.7%)	8 (11.1%)	8 (11.1%)

► **Functional equivalence does not imply fusion invariance.** At IR=10, Canonical reaches 73.11% whereas Transformed reaches 72.16%, creating a 0.95-point gap despite the same effective predictor; CCSF returns the result to 73.11% (Table 4). The same reference recovery appears at IR=50 and IR=100, and it remains under a non-uniform diagonal map and a differently structured stateful coordinator. Figure 3(a)–(b) provides the fixed-state counterpart: matched logits, losses, and untransformed raw gradients can still yield different coordinated updates. The agreement between local response and end-to-end restoration links the observed trajectory change directly to the reference-consistency property in Proposition 3.

Table 4: Controlled restoration across fusion rules and IRs.

Fusion rule	Setting	Canonical	Transformed	CCSF
Base scalar	IR=10	73.11 ± 0.40	72.16 ± 0.28	**73.11 ± 0.40**
Base scalar	IR=50	63.46 ± 0.31	63.22 ± 0.12	**63.46 ± 0.31**
Base scalar	IR=100	58.01 ± 0.30	57.71 ± 0.21	**58.01 ± 0.30**
Non-uniform diagonal	IR=100	58.01 ± 0.30	58.00 ± 0.30	**58.01 ± 0.30**
Alternative stateful	IR=100	57.65 ± 0.17	57.40 ± 0.17	**57.65 ± 0.17**

► **Coordinate leakage is not reducible to ordinary private-step correction.** Matching the LT-private step removes the direct storage-scale mismatch, but 41 of 216 shared-direction responses and 180 of 216 fusion-state deviations remain under private correction; CCSF reduces both to zero (Table 5). The much larger state residual is consequential because later fusion decisions read statistics written from the stored-coordinate geometry, so a corrected current step can still inherit a different coordination history. In other words, endpoint agreement on the LT branch does not imply that the coordinator has seen the same problem. CCSF instead canonicalizes both the geometry consumed by the coordinator and the state retained for subsequent updates. Appendix C.2 gives the matched controls.

Table 5: Residual coordination after private correction.

Diagnostic	Private correction	CCSF
Shared direction	41 (19.0%)	**0 (0.0%)**
Fusion state	180 (83.3%)	**0 (0.0%)**
Geometry input	Stored	**Reference**
State update	Stored	**Reference**

These controls distinguish reference-consistent coordination from ordinary update scaling and show why private-step correction alone is insufficient.

Figure 4 localizes the system-level advantage: CCSF gains +7.1, +6.2, and +6.1 points in the three rarest deciles, while the common end remains near zero. The alternatives show weaker or less tail-concentrated profiles, placing the gain where heterogeneous strategies are expected to add complementary coverage.

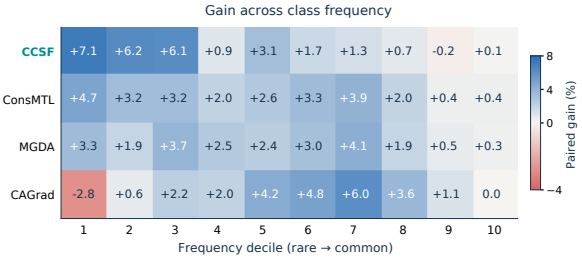


Figure 4: Frequency-localized gains over the controlled baseline.

► **Coordinate leakage reshapes expert complementarity despite a stronger LT expert.** Transformed raises standalone LT Top-1 by 9.00 points (34.34% to 43.34%), yet Few LT-unique support falls by 1.72 points and fused Few accuracy by 1.98 points (Table 6). CCSF restores the reference contribution pattern. Together with Figure 4, this inversion shows that branch accuracy can increase while heterogeneous usefulness decreases.

Table 6: Matched LT-expert complementarity at common epoch 2 (%).

	Fused-system accuracy				LT expert		
Condition	Overall	Many	Medium	Few	Top-1	Unique	Few unique
CCSF	**58.01**	77.69	52.49	**41.50**	34.34	**2.92**	**7.62**
Transformed	57.70	77.98	53.02	39.52	**43.34**	2.65	5.90
CCSF - Trans.	+0.31	-0.29	-0.53	**+1.98**	-9.00	+0.27	**+1.72**

► **Coordinate leakage is stateful and persists beyond the current update.** State-only replay changes coordinator outputs even when the current gradients are held fixed, showing that a private-coordinate intervention can remain in retained fusion state. Under a common 16-step canonical suffix, those state differences change the shared fused direction in 20 of 144 subsequent evaluations, so the effect can reappear after current inputs return to the reference coordinate. Mid-training scalar and diagonal switches provide the trajectory counterpart: the effective predictor is preserved at the switch, yet naive continuation leaves the reference path. Factor-basis rebasing reproduces the same behavior for a non-scalar representation change, while CCSF keeps model, optimizer, and fusion state aligned with the reference execution. Detailed controls are reported in Appendices C.1 and C.5.

Together, these analyses characterize coordinate leakage as an invariance problem of heterogeneous full-gradient coordination with a direct LTR consequence: function-preserving private coordinates can alter shared updates, retained state, and rare-class expert contribution. CCSF restores the reference coordination by evaluating fusion and retained state in a common reference representation. This distinction matters for LTR because a stronger standalone expert can still contribute less complementary tail coverage after coordinate leakage changes its fused trajectory.

5 Conclusion

We identified *coordinate leakage* in heterogeneous LTR strategy fusion: function-preserving private-coordinate changes can preserve the predictor yet alter the geometry consumed by full-gradient coordination and thereby change its updates and trajectory. CCSF removes this dependence by executing the unchanged fusion rule and optimizer in a reference coordinate and transporting updates back to storage. The resulting analyses reveal an LTR-specific consequence: expert value under heterogeneous fusion depends not only on standalone accuracy, but also on which examples the expert contributes jointly with other strategies, and this complementarity can be distorted by arbitrary private coordinates. Results on CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018 support coordinate consistency as an interface principle for heterogeneous LTR fusion. By separating parameter storage from strategy coordination, CCSF makes interactions among functionally equivalent experts depend on their predictive role rather than incidental private representation, while preserving the original fusion mechanism itself.

AI USE STATEMENT

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

REPRODUCIBILITY STATEMENT

Reproducibility details are provided throughout the main paper and appendices. Section 4.1 and Appendix A specify the datasets, Stage 2 architecture, optimization settings, and evaluation protocol. Appendix B provides the complete theoretical derivations and Algorithms 1–2, while Appendix C reports additional coordinate controls, multi-seed analyses, and continuation experiments. Anonymous source code is included in the supplementary material.

REFERENCES

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

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

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

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

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

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

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

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

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

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

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

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

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

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

Sicong Li, Qianqian Xu, Zhiyong Yang, Zitai Wang, Linchao Zhang, Xiaochun Cao, and Qingming Huang. Focal-sam: Focal sharpness-aware minimization for long-tailed classification. *arXiv preprint arXiv:2505.01660*, 2025.

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.

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

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

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.

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.

Behnam Neyshabur, Russ R Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *Advances in neural information processing systems*, 28, 2015.

Xiaohan Qin, Xiaoxing Wang, and Junchi Yan. Towards consistent multi-task learning: Unlocking the potential of task-specific parameters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10067–10076. IEEE, 2025.

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.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

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

Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20083–20093. IEEE, 2023.

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

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

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8769–8778. IEEE, 2018.

Twan Van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

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.

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

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

Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2361–2370. IEEE, 2021.

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

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

Zhe Zhao, Haibin Wen, Zikang Wang, Pengkun Wang, Fanfu Wang, Song Lai, Qingfu Zhang, and Yang Wang. Breaking long-tailed learning bottlenecks: A controllable paradigm with hypernetwork-generated diverse experts. *Advances in Neural Information Processing Systems*, 37:7493–7520, 2024.

Zhe Zhao, Zhiheng Gong, Pengkun Wang, HaiBin Wen, Cankun Guo, Bo Xue, Xi Lin, Zhenkun Wang, Qingfu Zhang, and Yang Wang. Ts-mof: Two-stage multi-objective fine-tuning for long-tailed recognition. *Advances in Neural Information Processing Systems*, 38:5224–5252, 2026.

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

A Experimental Details

This section complements Section 4.1 with the dataset statistics, Stage 2 architecture, optimization settings, and evaluation details used in our experiments.

A.1 CIFAR-100-LT Setting and Stage 2 Architecture

CIFAR-100-LT follows the standard exponential long-tailed construction with imbalance ratios of 10, 50, and 100. The corresponding dataset statistics are summarized in Table 7. Stage 1 uses ResNet-34, whose representation is frozen during Stage 2. The trainable Stage 2 module contains a shared refinement block with a 64-dimensional bottleneck between 512-dimensional feature layers, together with three expert-specific classifier heads. The three heads optimize the adaptive LT objective, BCL, and the LOS-style objective, respectively.

Adaptive LT objective. The LT branch uses the following adaptive objective, whose formulation, hyperparameters, and state updates are fixed across all matched coordinate conditions. Let n_c denote the number of training samples in class c, let B be the batch size, and let $\delta = 10^{-9}$. With class prior $\pi_c = (n_c + \delta)/(\sum_j n_j + C\delta)$, the adjusted logit for sample i and class c is

$$b_{ic} = z_{ic} + \log \max(\pi_c, \delta) - \mathbf{1}[c = y_i] m_{y_i}. \tag{13}$$

The epoch-dependent scale is $s_e = 30\left(1 + \frac{1}{2}\operatorname{clip}(e/80, 0, 1)\right)$, and the sample weight is

$$w_{i,e} = \begin{cases} 1, & e < 16, \\ n_{y_i} \operatorname{clip}(1 - A_{y_i}, 0, 1)^2, & e \geq 16. \end{cases} \tag{14}$$

The LT objective is

$$\mathcal{L}_{\mathrm{LT}} = \frac{1}{B} \sum_{i=1}^{B} \left[w_{i,e} \operatorname{CE}\left(\frac{s_e}{\tau_{y_i}} b_i, y_i \right) + \lambda_{y_i} \left\| h_i - W_{y_i} \right\|_2^2 \right], \tag{15}$$

where h_i is the shared Stage 2 feature and W_{y_i} is the corresponding LT-classifier row in the effective coordinate.

After computing the current-batch loss, the class-wise accuracy estimate is

$$\widehat{A}_c = \frac{\sum_i \mathbf{1}[y_i = c]\, \mathbf{1}[\arg\max_j b_{ij} = c]}{\sum_i \mathbf{1}[y_i = c] + \delta}. \tag{16}$$

The adaptive states are then updated as

$$\begin{aligned} A_c^+ &= 0.90 A_c + 0.10 \widehat{A}_c, \\ m_c^+ &= 0.90 m_c + 0.10 \operatorname{clip}\big(0.50(1 - A_c^+), 0, 0.50\big), \\ \lambda_c^+ &= 0.98 \lambda_c + 0.02(1 - A_c^+), \\ \tau_c^+ &= \operatorname{clip}\left(\tau_c \left[1 - 0.25 \operatorname{clip}\left(C^{-1} \sum_j A_j^+ - A_c^+, 0, 1 \right) \right], 0.50, 3 \right). \end{aligned} \tag{17}$$

The states are initialized by $A_c = 0$, $\lambda_c = 0$, $\tau_c = 1.10$, and $m_c = \operatorname{clip}(0.50[1 - n_c/(n_{\max} + \delta)], 0, 0.50)$. Classes absent from a batch use $\widehat{A}_c = 0$ and still update their states; the current-batch loss uses the pre-update states.

A.2 Optimization and Evaluation Details

Optimization. Stage 2 uses SGD with learning rate 5×10^{-4}, momentum 0.90, zero weight decay, and batch size 64. CIFAR-100-LT is optimized for 20 epochs with cosine learning-rate scheduling and LOS smoothing 0.98. ImageNet-LT and iNaturalist 2018 are optimized for up to 20 epochs with LOS smoothing 0.99. The scalar coordinate experiments use $a = 2$.

Fusion configuration. The gradient-fusion module uses $(\alpha_n, \alpha_c, \alpha_e, \alpha_v) = (0.82, 0.19, 0.60, 0.08)$, importance momentum 0.69, scale power 1.70, scale range $[0.40, 3.00]$, attenuation 0.81, and sigmoid

Table 7: Dataset statistics and Stage 2 training settings used in our experiments.

Dataset	IR	Classes	Train	Val	Backbone	Max. Stage 2 epochs
CIFAR-100-LT	10	100	19,064	10,000	ResNet-34	20
CIFAR-100-LT	50	100	12,286	10,000	ResNet-34	20
CIFAR-100-LT	100	100	10,566	10,000	ResNet-34	20
ImageNet-LT	–	1,000	115,846	20,000	ResNet-50	20
iNaturalist 2018	–	8,142	437,513	24,426	ResNet-50	20

multiplier 8.00. For each matched coordinate comparison, the coordinator and its coefficients are fixed across the compared representations.

Evaluation protocol. Stage 2 validation and prediction aggregation follow the same protocol as TSMOF (Zhao et al., 2026). Fixed-epoch analyses are identified explicitly. The multi-seed mechanism analysis uses Stage 2 seeds 0, 42, and 200. During inference, the LT classifier uses the effective weight $W = U/a$.

B Theoretical and Algorithmic Details

This section provides the formal analysis and algorithmic specification supporting Sections 3.2 and 3.3. We formalize the private-coordinate transformation and its effect on full-gradient fusion, prove the conditional invariance boundary and state-mediated coordinate dependence, derive reference-consistent execution for fixed invertible private maps, and then give the scalar Stage 2 specification together with its computational analysis.

B.1 Coordinate Transformation and Functional Equivalence

Let the reference Stage 2 parameter vector be

$$w = (\theta_s, W, \theta_{\mathrm{BCL}}, \theta_{\mathrm{LOS}}), \tag{18}$$

and define the corresponding stored coordinate by

$$u = T_a w, \qquad T_a = \mathrm{diag}(I_s, aI_{\mathrm{LT}}, I_{\mathrm{BCL}}, I_{\mathrm{LOS}}), \qquad a > 0. \tag{19}$$

The model represented by u is evaluated through $w = T_a^{-1}u$. Thus, T_a changes only the LT-private storage coordinate, while the shared, BCL, and LOS blocks remain in their reference coordinates. Every parameter-dependent term in the objective is evaluated through the effective reference parameters, including the adaptive LT center term in Eq. 15; matched coordinate comparisons also hold the corresponding loss states fixed.

Lemma 1 (Gradient transformation under private scaling)**.** *For any differentiable Stage 2 objective $\mathcal{L}(w)$ evaluated through $w = T_a^{-1}u$, the stored-coordinate gradient satisfies*

$$\nabla_u \widetilde{\mathcal{L}} = T_a^{-\top} \nabla_w \mathcal{L}, \qquad \widetilde{\mathcal{L}}(u) = \mathcal{L}(T_a^{-1}u). \tag{20}$$

Consequently, the shared and unaffected private gradients are unchanged, while the LT-private block satisfies

$$g_U = a^{-1} g_W. \tag{21}$$

Proof. Since $w = T_a^{-1}u$, the multivariate chain rule gives

$$\nabla_u \widetilde{\mathcal{L}} = \left(\frac{\partial w}{\partial u}\right)^{\top} \nabla_w \mathcal{L} = T_a^{-\top} \nabla_w \mathcal{L}. \tag{22}$$

The identity blocks of $T_a^{-\top}$ leave the shared, BCL, and LOS gradients unchanged, while its LT-private block is $a^{-1}I_{\mathrm{LT}}$.

Lemma 2 (Functional equivalence under private scaling)**.** *Let $u = T_a w$ and evaluate all parameter-dependent terms through $w = T_a^{-1}u$. In particular, the LT classifier is stored as $U = aW$ and evaluated through $W = U/a$. At corresponding effective parameters and loss states, the reference and stored implementations have identical logits and objective values, and identical raw gradients on every untransformed parameter block.*

Proof. For every feature vector h,

$$\frac{1}{a}Uh = Wh. \tag{23}$$

All remaining parameter-dependent terms read the same effective parameters $w = T_a^{-1}u$, so the complete objective agrees at the corresponding state. Lemma 1 then gives equality on every untransformed block. The equivalence therefore concerns the complete objective, not only the logits.

B.2 Gradient Geometry and Conditional Invariance

We first retain the scalar geometry underlying Proposition 1 and then characterize the corresponding invariance boundary for fixed coefficients, orthogonal maps, and stateful coordinate-wise statistics.

Restatement of Proposition 1 (Private Rescaling Changes Fusion-Relevant Gradient Geometry). Let

$$g_{\mathrm{LT}}^{(w)} = (g_s, g_p, 0, 0), \qquad g_j^{(w)} = (h_s, 0, h_p, 0), \tag{24}$$

where $g_p \neq 0$ is the LT-private component and the two experts have disjoint private support. Under the coordinate map $u = T_a w$,

$$g_{\mathrm{LT}}^{(u)} = (g_s, a^{-1}g_p, 0, 0), \qquad g_j^{(u)} = g_j^{(w)}. \tag{25}$$

Hence,

$$\left\| g_{\mathrm{LT}}^{(u)} \right\|_2^2 = \|g_s\|_2^2 + a^{-2}\|g_p\|_2^2, \tag{26}$$

$$\left\langle g_{\mathrm{LT}}^{(u)}, g_j^{(u)} \right\rangle = \langle g_s, h_s \rangle = \left\langle g_{\mathrm{LT}}^{(w)}, g_j^{(w)} \right\rangle. \tag{27}$$

For $a \neq 1$, $g_p \neq 0$, and $\langle g_s, h_s \rangle \neq 0$, the preserved numerator and changed norm imply

$$\frac{\cos\left(g_{\mathrm{LT}}^{(u)}, g_j^{(u)}\right)}{\cos\left(g_{\mathrm{LT}}^{(w)}, g_j^{(w)}\right)} = \frac{\left\| g_{\mathrm{LT}}^{(w)} \right\|_2}{\left\| g_{\mathrm{LT}}^{(u)} \right\|_2} \neq 1, \tag{28}$$

and the corresponding projection coefficient satisfies

$$\beta_{j\leftarrow\mathrm{LT}}^{(u)} = \frac{\left\langle g_j^{(u)}, g_{\mathrm{LT}}^{(u)} \right\rangle}{\left\| g_{\mathrm{LT}}^{(u)} \right\|_2^2} \neq \frac{\left\langle g_j^{(w)}, g_{\mathrm{LT}}^{(w)} \right\rangle}{\left\| g_{\mathrm{LT}}^{(w)} \right\|_2^2} = \beta_{j\leftarrow\mathrm{LT}}^{(w)}. \tag{29}$$

Proof of Proposition 1. Lemma 1 gives Eq. 25. Disjoint private support removes the private cross term, yielding Eq. 27, while the rescaled private component gives Eq. 26. Substituting these relations into the cosine similarity and projection definitions yields Eqs. 28 and 29.

Proposition 1 identifies the scalar mechanism behind coordinate leakage: full-gradient fusion rules that depend on gradient norms, normalized similarities, or projection coefficients can receive different geometric inputs from functionally equivalent private parameterizations.

Fixed-coefficient boundary. For a fixed invertible private map T that acts as the identity on shared coordinates, let P_s select the shared block. Since $P_s T^{-\top} = P_s$, fixed fusion coefficients $\{\alpha_t\}$ satisfy

$$P_s \sum_t \alpha_t T^{-\top} g_t = \sum_t \alpha_t P_s g_t. \tag{30}$$

Thus, fixed coefficients preserve the shared fused output under such a private reparameterization. The full fused direction transforms as $T^{-\top}d$, which generally differs from the parameter-space increment transport Td; shared-output invariance alone therefore does not imply ordinary private-step equivalence.

Proof of Proposition 2. Let Q be block orthogonal on the transformed private coordinates and the identity on shared coordinates, and let $G = [g_1, \ldots, g_m]$. Orthogonality gives

$$(QG)^\top (QG) = G^\top G. \tag{31}$$

Hence any weights, task ordering, projection coefficients, and task-level state transitions that depend only on the Gram matrix receive identical task-level inputs in the two representations. Under the stated coordinate-compatible masks and coupled solver, tie-breaking, and stochastic choices, both executions therefore follow the same branch. Linear combinations commute with Q, and norm-normalized projections are preserved because both their dot-product numerators and squared-norm denominators are unchanged. Induction over the fusion operations yields the transformed output Qd and the same updated task-level state. Because Q is the identity on the shared block, the shared fused direction is unchanged.

For a non-orthogonal map, the corresponding Gram matrix contains $T^{-1}T^{-\top}$ and can change. Proposition 2 therefore gives a sufficient equivariance boundary, while observable responses under non-orthogonal maps remain state dependent because clipping, inactive projections, or other piecewise behavior can suppress the current response.

Stateful coordinate dependence. Coordinate-wise state lies outside the Gram-only equivariance conditions and can retain coordinate dependence even when the current shared output is unchanged. For a nonzero private gradient p, let $H(p)$ denote the entropy of $|p_i|/\|p\|_1$ with $0 \log 0 = 0$, and let σ be the logistic function. Consider the stateful rule

$$\xi^+ = \beta\xi + (1-\beta)H(p), \qquad d_s = \sigma(\xi^+) g_{1,s} + \left[1 - \sigma(\xi^+)\right] g_{2,s}, \qquad 0 < \beta < 1. \tag{32}$$

An orthogonal map can send $p = (1, 0)$ to $p' = (1, 1)/\sqrt{2}$, preserving Gram geometry while changing H from 0 to $\log 2$. Starting from $\xi = 0$ with $g_{1,s} = g_{2,s} = 1$, both executions produce the same shared output $d_s = 1$, but their updated states become 0 and $(1-\beta)\log 2$. On the next call, identical reference inputs with shared gradients $(1, -1)$ and $H(p) = 0$ produce outputs 0 and

$$2\sigma(\beta(1-\beta)\log 2) - 1 > 0, \tag{33}$$

respectively. This construction isolates a state-mediated route by which coordinate dependence can reappear under common subsequent inputs; Appendix C.1 provides the corresponding state-only and continuation controls.

B.3 Reference-Consistent Execution

We next prove Proposition 3 for the complete state relevant to subsequent updates, and then give the SGD-momentum, adaptive-optimizer, and factor-coordinate specializations.

Restatement of Proposition 3 (Reference-Consistent Execution). Let $u = Tw$ for a fixed known invertible map T, and let

$$\mathcal{S} = (\xi, s_{\text{opt}}, s_{\text{loss}}, s_{\text{rng}}, \ldots) \tag{34}$$

collect every retained state that can affect the next update, expressed in the reference execution. Here ξ is the fusion state, s_{opt} is optimizer state, and s_{loss} includes the adaptive LT states in Appendix A.1. Suppose $u_0 = Tw_0$, the corresponding states are initialized consistently, all objective terms are evaluated through $w = T^{-1}u$, and both executions use coupled data and stochastic choices. At each step, CCSF forms

$$g_{k,t}^{(w)} = T^\top g_{k,t}^{(u)}, \qquad (d_t^{(w)}, \xi_{t+1}) = \mathcal{F}\Big(\{g_{k,t}^{(w)}\}; \xi_t\Big), \tag{35}$$

applies the reference optimizer

$$(\Delta w_t, s_{\text{opt},t+1}) = \mathcal{O}\Big(w_t, d_t^{(w)}, s_{\text{opt},t}\Big), \qquad u_{t+1} = u_t + T\Delta w_t, \tag{36}$$

and updates the remaining coupled states identically. Then, in exact arithmetic,

$$u_t = Tw_t, \qquad \mathcal{S}_t^{\text{CCSF}} = \mathcal{S}_t^{\text{ref}} \tag{37}$$

for every Stage 2 update t.

Proof of Proposition 3. Assume the correspondence holds at step t. Evaluating the model and every parameter-dependent objective term through $w = T^{-1}u$ gives identical predictions, objective values, and adaptive loss-state transitions. By the chain rule, $g_{k,t}^{(w)} = T^\top g_{k,t}^{(u)}$, so the unchanged coordinator receives the same reference gradients and fusion state and therefore returns the same reference direction and updated fusion state. The reference optimizer then receives the same parameters, direction, and optimizer state and returns the same increment Δw_t and next optimizer state. Consequently,

$$u_{t+1} = u_t + T\Delta w_t = T(w_t + \Delta w_t) = Tw_{t+1}. \tag{38}$$

Coupled data, stochastic choices, and auxiliary-state transitions preserve the remaining components of $\mathcal{S}$. The corresponding initialization establishes the induction.

Scalar specialization with SGD momentum. In the reference-state realization, SGD with momentum maintains $v^{(w)}$ in the reference coordinate,

$$v_{t+1}^{(w)} = \mu v_t^{(w)} + d_t^{(w)}, \qquad \Delta w_t = -\eta_t v_{t+1}^{(w)}. \tag{39}$$

For a fixed linear map, an equivalent native-storage realization exists because $v_t^{(u)} = Tv_t^{(w)}$ and $d_t^{(u)} = Td_t^{(w)}$ imply

$$v_{t+1}^{(u)} = \mu T v_t^{(w)} + Td_t^{(w)} = Tv_{t+1}^{(w)}, \qquad u_{t+1} = T(w_t - \eta_t v_{t+1}^{(w)}). \tag{40}$$

For the scalar LT transformation $U = aW$, this reduces to

$$g_W = ag_U, \qquad d_U = ad_W, \qquad v_U = av_W, \tag{41}$$

and therefore

$$\frac{U_t}{a} = W_t, \qquad \frac{v_{U,t}}{a} = v_{W,t}. \tag{42}$$

Algorithm 1 uses this equivalent scalar native-storage specialization.

Extension to general linear coordinates. For

$$u = Tw, \qquad w = T^{-1}u, \tag{43}$$

the chain rule gives

$$g_u = T^{-\top} g_w, \qquad g_w = T^\top g_u, \tag{44}$$

while parameter increments transport as $\Delta u = T\Delta w$. A block-diagonal map $T = \mathrm{diag}(I_s, T_1, \ldots, T_K)$ applies the same construction independently to multiple private blocks.

Reference-state AdamW. The execution in Proposition 3 extends directly to AdamW when its optimizer state is retained in the reference coordinate. The first and second moments, bias-correction counts, and decoupled weight decay are updated from the reference gradient and reference parameters. With bias-corrected moments $\widehat{m}_t$ and $\widehat{v}_t$, the reference increment is

$$\Delta w_t = -\eta_t \left[\frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon} + \lambda w_t \right], \tag{45}$$

followed by $\Delta u_t = T\Delta w_t$. For a dense non-orthogonal map, ordinary native-coordinate AdamW generally defines a different execution because its elementwise second-moment state is not obtained by a simple linear transport. CCSF therefore preserves the chosen reference AdamW execution by retaining those states in reference coordinates.

Factorized private coordinates. Let $A \in \mathbb{R}^{c\times r}$ and $B \in \mathbb{R}^{r\times d}$ with effective classifier $W = AB$. For a fixed invertible $Q \in \mathbb{R}^{r\times r}$,

$$A' = AQ, \qquad B' = Q^{-1}B \tag{46}$$

preserves W. Under column-major vectorization,

$$u_p = T_Q w_p, \qquad w_p = \begin{bmatrix} \mathrm{vec}\, A \\ \mathrm{vec}\, B \end{bmatrix}, \qquad T_Q = \mathrm{diag}\big(Q^\top \otimes I_c,\, I_d \otimes Q^{-1}\big). \tag{47}$$

The reference execution is defined on the factors (A, B) themselves. Their gradients transform by $T_Q^\top$ and their increments by T_Q, while factor-dependent objective terms are evaluated in the reference factors. This gives a fixed linear private-coordinate instance of Proposition 3 without replacing the many-to-one factorization map $(A, B) \mapsto AB$ by an artificial inverse.

B.4 CCSF Stage 2 Training and Gradient Fusion

Algorithm 1 gives the scalar Stage 2 procedure used in the main LTR experiments and shows where canonical-coordinate fusion enters the heterogeneous expert optimization loop. The general linear interface follows Proposition 3, while the scalar fusion operation is detailed in Algorithm 2.

Algorithm 1 CCSF Stage 2 Training

Require: Training set $\mathcal{D}_{\text{train}}$, validation set $\mathcal{D}_{\text{val}}$, frozen encoder E, Stage 2 epochs E_2
Require: Trainable decoder D, expert heads $H_{\text{LT}}, H_{\text{BCL}}, H_{\text{LOS}}$, losses $\mathcal{L}_{\text{LT}}, \mathcal{L}_{\text{BCL}}, \mathcal{L}_{\text{LOS}}$, coordinate scale a
Ensure: Fine-tuned Stage 2 model
Initialize D, H_{LT}, H_{BCL}, and H_{LOS}
Let W denote the effective LT classifier weight and store $U \leftarrow aW$
Initialize SGD momentum v, fusion state ξ, and the adaptive loss states in Appendix A.1
for $e = 1$ to E_2 **do**
  for each mini-batch $(X, Y) \sim \mathcal{D}_{\text{train}}$ **do**
    $F \leftarrow E(X)$ {Frozen Stage 1 representation}
    $Z \leftarrow D(F)$ {Shared Stage 2 representation}
    $W \leftarrow U/a$ {Effective LT coordinate}
    $P_{\text{LT}} \leftarrow H_{\text{LT}}(Z; W)$
    $P_{\text{BCL}} \leftarrow H_{\text{BCL}}(Z)$, $P_{\text{LOS}} \leftarrow H_{\text{LOS}}(Z)$
    $L_{\text{LT}} \leftarrow \mathcal{L}_{\text{LT}}(P_{\text{LT}}, Y)$
    $L_{\text{BCL}} \leftarrow \mathcal{L}_{\text{BCL}}(P_{\text{BCL}}, Y)$, $L_{\text{LOS}} \leftarrow \mathcal{L}_{\text{LOS}}(P_{\text{LOS}}, Y)$
    for $t \in \{\text{LT}, \text{BCL}, \text{LOS}\}$ **do**
      $g_t \leftarrow \nabla_\theta L_t$ and record the gradient-existence mask h_t
    end for
    Update the adaptive LT states (A, m, λ, τ) using Eqs. 16–17; the current gradients use the pre-update states
    $(d^{(U)}, \xi) \leftarrow \textsc{CanonicalGradientFusion}(\{g_t, h_t\}, a, \xi)$
    $v \leftarrow \mu v + d^{(U)}$
    $\theta \leftarrow \theta - \eta_e v$
  end for
  Update the learning rate by cosine annealing
  Perform Stage 2 validation following the baseline protocol using the effective LT weight $W = U/a$
end for
return Fine-tuned Stage 2 model

Algorithm 2 specifies the scalar LT coordinate interface inserted around the base full-gradient fusion rule. For each packed task gradient $g_t \in \mathbb{R}^D$, it expresses the LT-private block in the reference coordinate, performs all fusion statistics and conflict projections there, and maps the fused LT direction back to storage; Proposition 3 gives the corresponding general-T form.

Fusion-statistic definitions. Let $G^{(0)}$ denote the reference-coordinate gradient matrix before task scaling and projection, and let $\epsilon = 10^{-8}$. The coordinate-wise sample variance and its cross-task normalization are

$$\mu_t = \frac{1}{D}\sum_{k=1}^{D} G_{tk}^{(0)},$$
$$v_t = \frac{1}{D-1}\sum_{k=1}^{D}\left(G_{tk}^{(0)} - \mu_t\right)^2, \tag{48}$$
$$V_t = \begin{cases} 0, & v_{\max} - v_{\min} < \epsilon, \\ \dfrac{v_t - v_{\min}}{v_{\max} - v_{\min} + \epsilon}, & \text{otherwise}, \end{cases}$$

where $v_{\min} = \min_j v_j$ and $v_{\max} = \max_j v_j$.

Algorithm 2 Canonical-Coordinate Gradient Fusion

Require: Task gradients $\{g_t\}_{t=1}^{T}$, gradient-existence masks $\{h_t\}_{t=1}^{T}$, LT-private coordinates $\mathcal{I}_{\mathrm{LT}}$, coordinate scale a, fusion state ξ, stability constant $\epsilon = 10^{-8}$

Ensure: Fused direction $d^{(U)}$ in the stored coordinate and updated fusion state ξ

{**Canonicalize before fusion**}
Copy $\widetilde{g}_t \leftarrow g_t$ for every task t
for $t = 1$ to T **do**
  $\widetilde{g}_t[\mathcal{I}_{\mathrm{LT}}] \leftarrow a\, g_t[\mathcal{I}_{\mathrm{LT}}]$ {Reference LT-private coordinate}
end for
Stack $\{\widetilde{g}_t\}$ into $G^{(0)} \in \mathbb{R}^{T\times D}$
{**Apply the base fusion rule**}
$r_t \leftarrow \max(\|G_t^{(0)}\|_2, \epsilon), \quad \nu_t \leftarrow \max(r_t^{-1}, \epsilon)$
$c_{tj} \leftarrow (G_t^{(0)})^{\top} G_j^{(0)} / \max(r_t r_j, \epsilon)$
$q_t \leftarrow \sum_{j\neq t} \max(-c_{tj}, 0)/(T-1+\epsilon)$ {Conflict score}
$p_{tk} \leftarrow (|G_{tk}^{(0)}| + \epsilon)/(\sum_l |G_{tl}^{(0)}| + D\epsilon)$
$H_t \leftarrow -\sum_k p_{tk} \log(p_{tk}+\epsilon)/(\max(\log D, \epsilon) + \epsilon)$ {Gradient entropy}
Compute normalized coordinate-wise gradient variance V_t as defined below
$I_t \leftarrow (\nu_t+\epsilon)^{\alpha_n}(q_t+\epsilon)^{\alpha_c}(H_t+\epsilon)^{\alpha_e}(1+\alpha_v V_t)$
Compute percentile-enriched importance I'_t, update EMA state M_t^{+}, and obtain normalized importance $\bar{I}_t$
$s_t \leftarrow \mathrm{clip}_{[s_{\min}, s_{\max}]}\big[(\bar{I}_t/(\mathrm{mean}(\bar{I})+\epsilon))^{\gamma}\big]$
Set $G_t \leftarrow s_t G_t^{(0)}$, order tasks by $\bar{I}_t$, and cache $B_t \leftarrow \max(\|G_t\|_2^2, \epsilon)$
for each task i in the ordered list **do**
  for each higher-priority task j already processed **do**
    if $G_i^{\top} G_j < 0$ **then**
      $\lambda_{ij} \leftarrow \mathrm{clip}_{[0,1]}\big(1 - \rho\,\sigma(\kappa \max(\bar{I}_j - \bar{I}_i, 0))\big)$
      $G_i \leftarrow G_i - \lambda_{ij} \dfrac{G_i^{\top} G_j}{B_j} G_j$
    end if
  end for
end for
Aggregate the projected gradients with the original existence masks $\{h_t\}$ using the masked mean defined below to obtain the reference fused direction $d^{(W)}$
{**Map the fused direction back**}
Copy $d^{(U)} \leftarrow d^{(W)}$
$d^{(U)}[\mathcal{I}_{\mathrm{LT}}] \leftarrow a\, d^{(W)}[\mathcal{I}_{\mathrm{LT}}]$ {Stored LT-private coordinate}
return $d^{(U)}$ and updated fusion state ξ

For the current raw task-importance vector I, define

$$R = \mathrm{argsort}(\mathrm{argsort}(I)), \qquad P_t = \frac{R_t + 1}{T+1}, \qquad I'_t = I_t(1 + P_t^2). \tag{49}$$

Here $R_t \in \{0, \ldots, T-1\}$ is the ascending ordinal rank across the current tasks. The rank uses only the current tasks and does not aggregate across earlier steps. Ties follow the deterministic `argsort` order used by the fusion rule; tied ranks are not averaged.

The stored importance state M and normalized importance $\bar{I}$ are updated as

$$M_t^{+} = \beta M_t + (1-\beta) I'_t, \qquad J_t = M_t^{+} - \min_j M_j^{+} + \epsilon, \qquad \bar{I}_t = \frac{J_t}{T^{-1}\sum_j J_j + \epsilon}, \tag{50}$$

with $\beta = 0.69$ and $M_t = I'_t$ on the first call. The fusion state stores M^{+}; scaling, ordering, and projection attenuation use $\bar{I}$.

Let $\widehat{G}$ denote the scaled and projected gradients. The masked mean used by the fusion rule is

$$d_k^{(W)} = \frac{\sum_{t=1}^{T} h_{tk} \widehat{G}_{tk}}{\sum_{t=1}^{T} h_{tk} + \epsilon}. \tag{51}$$

The mask records gradient existence before fusion: a numerically zero entry in an existing gradient tensor has $h_{tk} = 1$, whereas a missing parameter gradient has $h_{tk} = 0$ and is zero-filled before fusion. Statistics are computed on the full zero-padded gradients. Masks are applied after scaling and projection, using the original existence masks in both numerator and denominator; coordinates missing from all tasks therefore receive zero update.

All task statistics, adaptive weighting, task ordering, and conflict projection are computed from reference-coordinate full gradients. The base fusion rule itself is unchanged; CCSF modifies only the coordinate interface through which the LT-private gradient enters fusion and the corresponding fused direction returns to the stored parameterization.

B.5 Computational Complexity

For m objectives and p transformed private parameters, scalar or diagonal gradient transport requires $O(mp)$ operations and increment transport requires $O(p)$. These operations introduce no trainable parameters and do not change the dimensionality of the gradients processed by the base fusion rule. Table 8 reports the matched Stage 2 profile using cached features for the full reference execution, scalar CCSF, and the shared-only control.

Table 8: Matched cached-Stage 2 computational cost in FP32 with batch size 64. Time is mean ± sample SD over six order-balanced repetitions; each repetition uses 10 warm-up and 50 timed steps. Stage 1, data loading, validation, and I/O are excluded.

Method	Time per step (ms)	Peak allocated (MiB)
Full reference	36.00 ± 1.29	90.69
Scalar CCSF	35.77 ± 1.07	90.89
Shared-only control	37.24 ± 1.29	80.54

The scalar CCSF profile is comparable to the full reference execution in both time and peak allocated memory; the observed timing difference is smaller than repeat-to-repeat variation. An explicit dense $p \times p$ map costs $O(mp^2)$ for gradient transport and $O(p^2)$ for increment transport and map storage, while block-diagonal maps sum the corresponding block costs. For the factor gauge in Eq. 47, structured multiplication by Q costs $O(mr^2(c+d))$ for gradients and $O(r^2(c+d))$ for increments with $O(r^2)$ map storage, without materializing T_Q. The measured profile therefore characterizes the scalar interface used in the main LTR experiments; dense and factorized maps follow the analytical costs above.

C Additional Experimental Analysis

This section complements Section 4.3 with fixed-state evidence on coordinate leakage, multi-seed reference recovery, coordination-level controls, class-frequency localization, large-scale validation, and coordinate-consistent continuation beyond scalar rescaling.

C.1 Fixed-State Coordinate Leakage and Stateful Propagation

Fixed-state interventions show that coordinate leakage can arise before optimization trajectories diverge. At $a = 2$, the scaled representation changes the fused shared direction in 91 of 288 matched evaluations (31.60%) while preserving the effective logits, losses, and shared raw gradients. Table 9(a) further shows that either the current gradient representation or the accumulated fusion state can independently induce a response under both tested stateful fusion rules.

The alternative stateful control is a ranking-based coordinator that combines inverse norm, conflict, conflict strength, uniqueness, and variance with an eight-step importance window, window-mean scaling, and ordered conflict-gated projection.

Table 9: Fixed-state coordinate dependence on CIFAR-100-LT with IR=100. Panel (a) separates current-gradient and fusion-state effects; panel (b) localizes responses when clipped task scales and task ordering are unchanged.

(a) Current representation and fusion state

Fusion rule	Current only	State only	Both
Base	12/36	**14/36**	15/36
Alternative stateful	10/36	**20/36**	20/36

(b) Projection localization

Quantity	Changed / evaluated
Shared fused direction	31/226
Active projection set	0/31
Projection attenuation	**31/31**
Projection denominator	29/31
Cross-task dot term	0/71

Table 9(b) shows that the shared fused direction can change with fixed active projections, clipped scales, and task ordering, localizing leakage to continuous projection geometry. In an additional control using the KPS+BCL+LOS baseline configuration (Zhao et al., 2026), BCL-private reparameterization changes 44 of 144 shared directions, and CCSF restores the corresponding reference direction.

The coordinate-sensitivity matrix in Table 3 evaluates the same reference interface across scalar, non-uniform diagonal, and orthogonal maps; the maximum CCSF relative L_2 restoration residual is 6.89×10^{-13}.

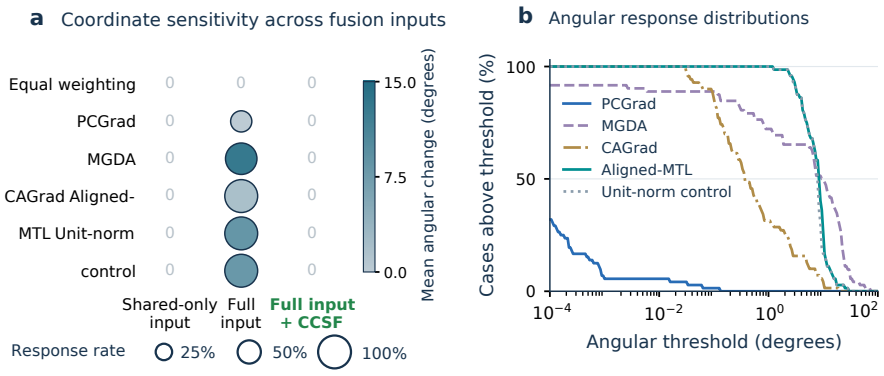


Figure 5: **Cross-coordinator coordinate sensitivity.** Shared-only inputs remain invariant, whereas full-gradient private coordinates induce method-dependent angular changes that disappear after CCSF canonicalization.

Figure 5 applies the same private-coordinate intervention across representative coordinators. Shared-only inputs produce no response, whereas full-gradient inputs alter several geometry-dependent methods; fixed equal weighting remains invariant. Canonicalizing the affected private block with CCSF removes the observed responses, while the angular distributions show their method-dependent magnitudes. This visualization complements Table 3 by separating the coordinate interface from the underlying fusion rule.

Propagation under a common suffix. State-only replay shows that retained fusion state can differ even when the current gradients are held fixed. To test whether this difference can affect later coordination, we take nine real Stage 2 states and continue each with the same 16-step canonical suffix. Across the resulting 144 subsequent evaluations, 20 change the shared fused direction. The effect is therefore not restricted to the intervention step: a representation-induced state difference can remain latent and reappear under common subsequent inputs.

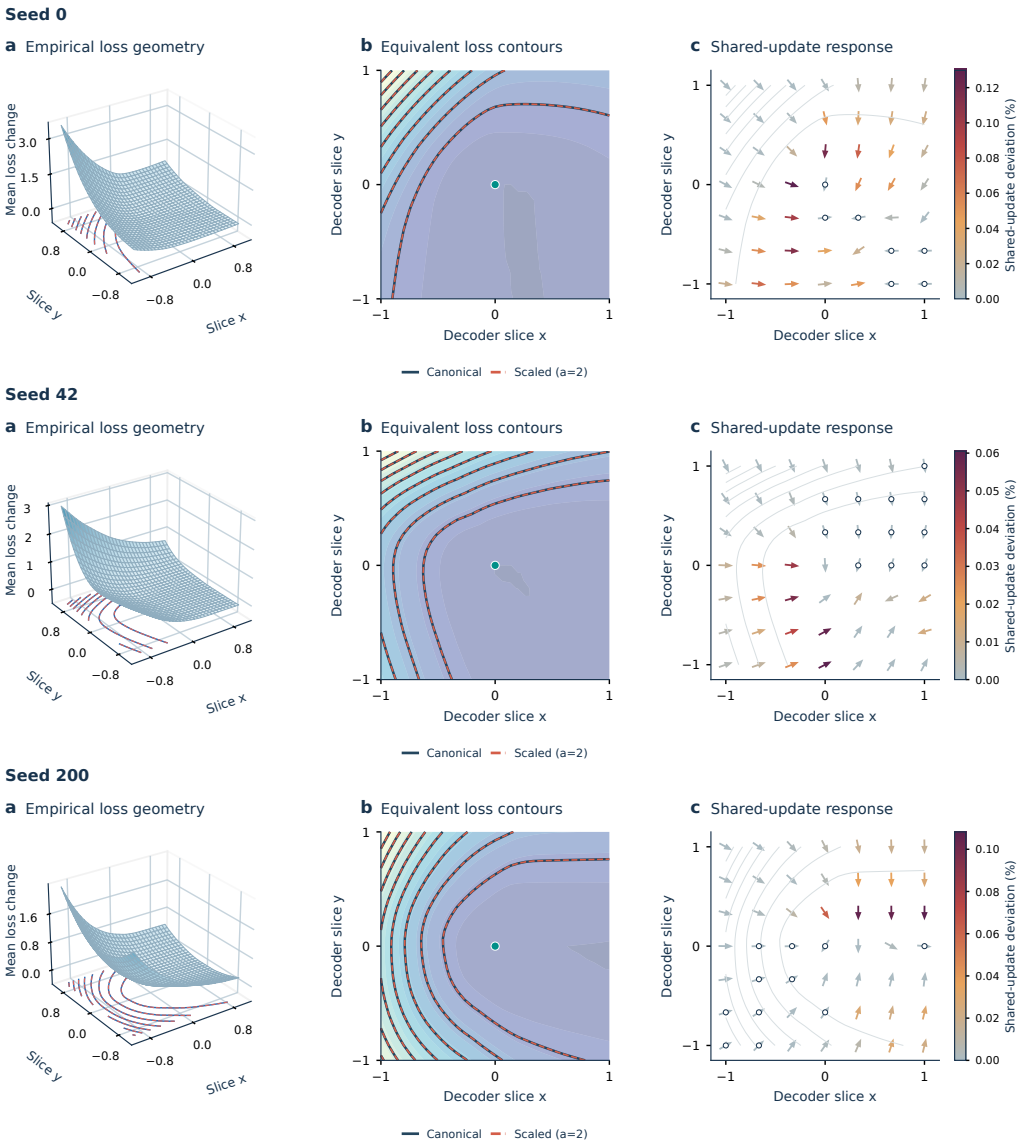


Figure 6: **Multi-seed fixed-state replication of coordinate leakage.** Fixed-batch analyses for Stage 2 seeds 0, 42, and 200 on CIFAR-100-LT with IR=100. For each seed, the figure shows the empirical mean-loss surface, the reference and scaled ($a = 2.00$) effective-loss contours at matched effective states, and the corresponding shared-update response field. Overlapping loss contours illustrate the matched effective-loss geometry, while the response fields show that full-gradient fusion can remain coordinate sensitive.

C.2 Coordination and Expert Complementarity

The private-step control isolates ordinary update scaling from coordination-level dependence. Across 216 matched scalar controls, private-step/output correction still leaves 41 shared-fusion responses and 180 fusion-state deviations, whereas CCSF removes both, matching the reference interface reported in Table 5.

The downstream consequence is concentrated in the transformed LT expert's rare-class contribution. At common epoch 2, CCSF recovers 1.98% Few-class fused accuracy relative to Transformed, with 1.83% supplied by LT-specific predictions. The same matched comparison increases LT unique accuracy from 2.65% to 2.92% and Few LT unique accuracy from 5.90% to 7.62%, even though

standalone LT Top-1 moves in the opposite direction. Table 6 reports this complementarity contrast in the main text.

Table 10: Matched prediction-source attribution at common epoch 2 on CIFAR-100-LT with IR=100. Values are CCSF minus Transformed, averaged over three Stage 2 seeds, in percentage points. LT, BCL, and LOS provide a post-hoc attribution of changed correct predictions to the corresponding experts.

Group	Δ Acc.	LT	BCL	LOS
Overall	+0.31	+0.29	+0.00	+0.02
Many	-0.29	-0.29	+0.00	+0.00
Medium	-0.53	-0.46	+0.00	-0.08
Few	**+1.98**	**+1.83**	+0.00	+0.15

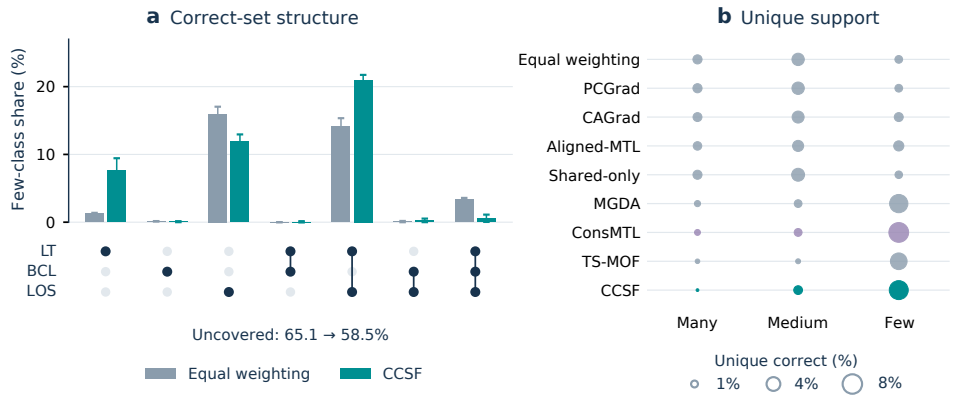


Figure 7: **Few-class correct-set structure and unique support.** Left: correct-set overlap for equal weighting and CCSF on Few classes, where the uncovered share decreases from 65.1% to 58.5%. Right: unique-correct support across Many/Medium/Few groups for the tested coordination configurations; marker area encodes the unique-correct percentage.

Figure 7 provides a complementary system-level view of expert coverage. CCSF reduces the uncovered Few-class share and places more unique-correct support in the Few region. The matched coordinate-specific comparison remains Table 6 and Table 10, which isolate the CCSF–Transformed difference.

Class-frequency localization. Figure 4 reports the system-level frequency profile relative to the controlled baseline. The CCSF gains are largest in the rarest deciles and become small near the common end, whereas the alternative coordinators show weaker or differently distributed profiles. The matched CCSF–Transformed comparison in Table 6 provides the coordinate-specific counterpart, where the fused-system recovery is concentrated in Few classes and accompanied by a larger LT-specific unique contribution.

C.3 Controls Across Fusion Rules and Seeds

For the system-level frequency comparisons in Figure 4, the reference is the controlled TS-MOF reproduction using KPS/BCL/LOS with its original decoder, PLA, PCGrad, and aggregation protocol; differences are paired by seed using the corresponding Stage 2 result for each method.

Five-seed scalar control. With a fixed Stage 1 model and data split, the scalar control was extended to Stage 2 seeds $0, 1, 2, 42, 200$. Table 11 shows identical Canonical and CCSF means, while the paired CCSF–Scaled Overall accuracy difference is positive for all five seeds.

The paired Overall accuracy difference is positive for all five tested Stage 2 seeds, with values of $+0.07\%$, $+0.08\%$, $+0.47\%$, $+0.75\%$, and $+0.09\%$ and a mean of $+0.29\%$; the mean Few-class

Table 11: Five-seed scalar coordinate control on CIFAR-100-LT at IR=100. These are five Stage 2 seeds with fixed Stage 1 and data split.

Condition	Many	Medium	Few	Overall
Canonical	77.93	52.51	40.99	57.95 ± 0.23
Scaled	78.33	53.04	38.94	57.66 ± 0.19
CCSF	**77.93**	**52.51**	**40.99**	**57.95 ± 0.23**

difference is $+2.05\%$. Table 4 also reports scalar restoration across IR=10/50/100 and an alternative stateful IR=100 control. A separate Stage 1 initialization control gives 62.25 ± 0.10, 62.18 ± 0.14, and 62.25 ± 0.10 Overall accuracy for Canonical, Scaled, and CCSF, respectively.

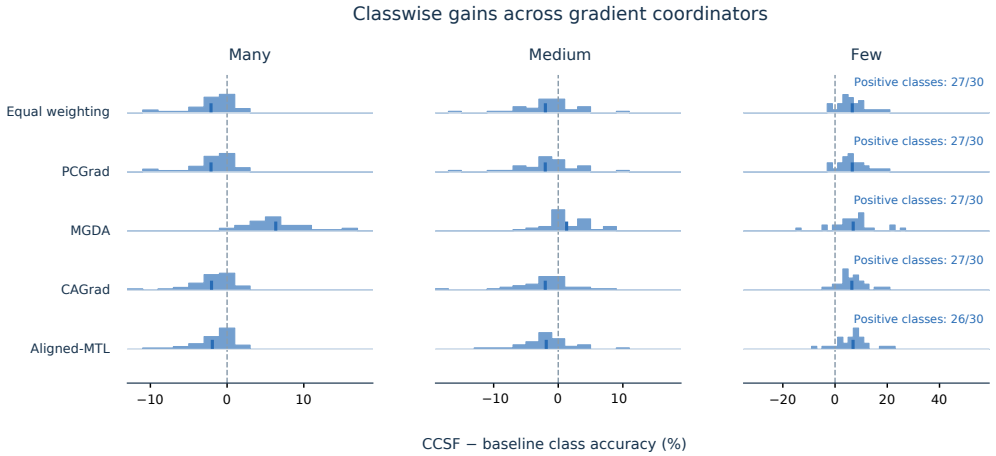


Figure 8: **Classwise gains across gradient coordinators.** Per-class CCSF-minus-baseline Top-1 differences on CIFAR-100-LT with IR=100. Rows compare CCSF with equal weighting, PCGrad, MGDA, CAGrad, and Aligned-MTL; columns show Many, Medium, and Few classes. The Few-column annotations report the number of classes with positive differences.

Figure 8 resolves the tail-localized pattern at individual-class granularity. CCSF yields positive differences for 26–27 of the 30 Few classes across the tested coordinators, while the Many- and Medium-class distributions remain mixed. The tail advantage is therefore distributed across most Few classes rather than being driven by a small subset.

C.4 Large-Scale Coordinate Controls

The large-scale controls extend reference-consistent execution to ImageNet-LT and iNaturalist 2018 using fixed Stage 1 models, three Stage 2 seeds, and LOS smoothing 0.99. Canonical and CCSF both obtain $56.53 \pm 0.17\%$ Overall accuracy on ImageNet-LT and $72.35 \pm 0.27\%$ on iNaturalist 2018 under the paired execution. These controls extend reference-consistent coordination beyond CIFAR-100-LT.

C.5 Beyond Scalar Rescaling and Coordinate-Consistent Continuation

Full-trajectory restoration beyond uniform scaling. Full-trajectory restoration is not restricted to the scalar map $U = aW$. Under the non-uniform diagonal transformation $U = WD$, where the 512 feature dimensions cycle through scales $0.50, 1.00, 2.00, 4.00$, Canonical, Transformed, and CCSF obtain 58.01 ± 0.30, 58.00 ± 0.30, and 58.01 ± 0.30 Overall accuracy, respectively, over three Stage 2 seeds. Across the 21 recorded epoch states for each seed, the CCSF effective model, mapped optimizer state, loss, and fusion state remain numerically aligned with the Canonical run, while the Transformed parameter trajectory reaches a maximum relative L_2 separation of 2.67×10^{-3}. This provides trajectory-level evidence beyond uniform scaling even when endpoint accuracy is nearly unchanged.

Mid-training storage-coordinate switch. To test continuation from an already trained state, we switch a saved Stage 2 model to functionally equivalent scalar or diagonal storage coordinates at step 831. Immediately before continuation, the effective weight, logits, mapped momentum, shared raw gradients, and fixed-probe loss match the reference. The naive switches subsequently depart from the reference continuation, whereas CCSF numerically preserves the Canonical trajectory for both scalar and diagonal maps.

Reference-state AdamW. We additionally evaluate reference-state AdamW with decay 0.01. Its first and second moments, update count, and decay remain in the reference coordinate before increment transport. Canonical and CCSF trajectories and optimizer states remain numerically aligned. Native-coordinate AdamW follows a distinct execution because its coordinate-wise moments evolve in storage coordinates.

Factor-basis representation conversion. For a full-column-rank left factor, write its thin QR decomposition as $A = Q_A R$ and set $S = R^{-1}$. This defines a function-preserving gauge rebase of $W = AB$ through $A' = AS$, $B' = S^{-1}B$, with fixed S after the switch. The rebase changes the stored left-factor conditioning from $\mathrm{cond}_2(A) = 2.18$ to 1.00, providing a concrete non-scalar basis conversion. We evaluate early, middle, and late Stage 2 switch points over three seeds; Table 12 summarizes the continuation and fusion-state responses.

Table 12: QR factor-basis continuation across training stages. Overall/Few (%) are mean ± sample SD over three Stage 2 seeds. CCSF deltas are relative to Reference; the final column reports naive/CCSF state responses.

Switch	Reference Overall/Few	Δ CCSF Overall/Few	State response
Early	55.28 ± 0.17 / 27.48 ± 0.36	**0.00 / 0.00**	3/3; **0/3**
Middle	55.28 ± 0.17 / 27.48 ± 0.36	**0.00 / 0.00**	3/3; **0/3**
Late	55.28 ± 0.17 / 27.48 ± 0.36	**0.00 / 0.00**	3/3; **0/3**

Across early, middle, and late switches, CCSF keeps the post-switch model, optimizer, and fusion-state trajectories numerically aligned with the reference execution. At the middle switch, the naive fusion state changes in 3/3 cases while CCSF removes the state response, extending reference-consistent coordination to a concrete non-scalar factor-basis conversion.

D Scope and Future Work

The analysis covers fixed, known, invertible private-coordinate maps, including scalar and non-uniform diagonal rescaling, orthogonal mixing, and factor-coordinate rebasing, with reference-consistent execution analyzed for SGD with momentum and reference-state AdamW. Extending CCSF to settings where the reference representation or coordinate map must be inferred, or where the transformation evolves nonlinearly during training, provides a natural direction for future work.