FALOTS: ORTHOGONAL TARGET-AWARE SAM FOR FEDERATED ACTIVE LEARNING UNDER GLOBAL CLASS IMBALANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated active learning (FAL) reduces annotation costs by selectively labeling data distributed across clients. However, under global class imbalance, the class proportions in the aggregate labeled data may differ substantially from those desired for evaluation (target prior), limiting the global model's performance. Existing FAL methods mainly improve sample acquisition, leaving this mismatch insufficiently addressed during local training. Sharpness-aware minimization (SAM) offers a promising foundation for local training by seeking parameters whose loss remains low under nearby perturbations. Yet, its empirical-gradient perturbation does not explicitly account for the mismatch between global labeled and target class priors. To address this limitation, we propose `FalOTS`, a target-aware SAM optimizer for local training within FAL loop. At each acquisition phase, `FalOTS` constructs class weights from the ratio between the target prior and current global labeled prior. It computes a target-aware gradient on each labeled mini-batch and adds its normalized orthogonal residual to SAM perturbation. This correction preserves SAM empirical-gradient component while adding a complementary direction. The subsequent descent step uses the original unweighted loss. Our analysis characterizes the correction's geometric properties and provides sufficient conditions for a lower target risk after one update than SAM. `FalOTS` improves global model accuracy under limited labeling budgets, outperforming FAL baselines using SGD- and SAM-based optimizers across diverse benchmarks and settings[1].

1 INTRODUCTION

In many federated learning (FL) applications, labeled data is scarce (McMahan et al., 2017). Federated active learning (FAL) tackles this by allowing clients to select examples worth annotating using acquisition criteria (Beluch et al., 2018; Ash et al., 2020) from active learning (AL) and to train a shared model on a small fraction of labeled data without centralizing the data (Cao et al., 2023; Kim et al., 2023a). Because labeled data is limited and grows through acquisition phases, FAL relies on both example selection and model training strategies. Yet, prior FAL studies mainly improve acquisition through the choice of query signals and models. In contrast, we focus on local training to enhance generalization rather than acquisition. To this end, sharpness-aware minimization (SAM) (Foret et al., 2021) offers a useful local training objective by seeking parameters within low-loss neighborhoods to improve model generalization. This motivates us to adapt SAM to the evolving, globally imbalanced labeled sets in FAL, a direction that remains largely unexplored.

SAM studies in FL have focused on a challenge: clients have different data distributions, so a locally computed perturbation may not reflect the aggregate training objective (Qu et al., 2022; Fan et al., 2024; Lu et al., 2026). Yet, a second obstacle remains in FAL: even when all clients are conceptually combined, the overall labeled dataset can exhibit class imbalance relative to the distribution on which the model is evaluated (Shen et al., 2021). This global prior mismatch is distinct from client heterogeneity. Client heterogeneity concerns differences among local distributions, while global class imbalance concerns class frequencies in their aggregate. Making clients similar does not resolve this mismatch. Moreover, active acquisition changes the aggregate labeled prior across phases; thus, the distribution used for local training can change even when the target prior remains fixed.

[1] We will make our code available upon publication.

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.

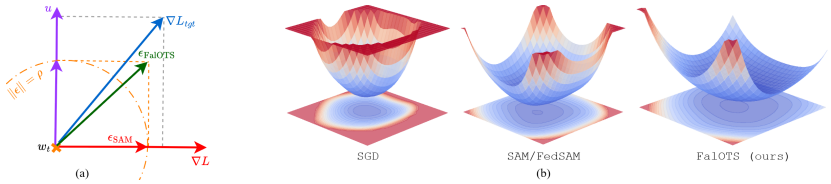


Figure 1: a) For a nonzero empirical gradient ∇L, SAM perturbation ϵ_{SAM} lies along that gradient, which can be substantially mismatched with the target-weighted gradient ∇L_{tgt}. When ∇L_{tgt} has a nonzero orthogonal component, its normalized direction u cannot be reached by changing ρ. `FalOTS` retains ϵ_{SAM} and adds this complementary direction: $\epsilon_{\texttt{FalOTS}} = \epsilon_{\text{SAM}} + \rho u$. b) In the global loss landscapes on CIFAR-10, `FalOTS` reaches the flattest minimum compared to others.

This mismatch matters for SAM because its practical perturbation follows the empirical training gradient ∇L, as illustrated in Fig. 1a. Consequently, frequent classes dominate the loss gradient, causing local SAM perturbations to align with dominant classes rather than the target class mixture. Increasing the perturbation radius scales this empirical perturbation, but cannot recover a target-gradient component. The research question is how to incorporate the current global prior mismatch into local perturbations while retaining SAM empirical-gradient component.

In this paper, we propose `FalOTS` (`F`ederated `a`ctive `l`earning with `O`rthogonal `T`arget-aware `S`AM), a local optimizer for FAL under global class imbalance. At each acquisition phase, the server aggregates labeled class counts and broadcasts weights based on the desired target prior. Each client uses these weights to compute a target-aware gradient on the same mini-batch as its empirical gradient. We retain SAM perturbation and add only the normalized orthogonal residual of the target-weighted gradient. The correction is first-order neutral to the empirical loss. When class priors align, the correction vanishes and `FalOTS` reduces to SAM. By integrating global class information into local optimization, `FalOTS` requires no changes to FAL acquisition rules or server aggregation algorithms. It consistently outperforms FAL baselines using SGD- and SAM-based optimizers, even under nearly independent and identically distributed (i.i.d.) client partitions where global class imbalance persists. Moreover, `FalOTS` yields noticeably global flatter loss surfaces (Fig. 1b).

Our main contributions are as follows:

1. We distinguish global class imbalance from client heterogeneity and address it as a key FAL challenge. We also identify a limitation of empirical SAM perturbations: merely changing their perturbation radius cannot recover an orthogonal target-aware direction (Sec. 2).
2. We propose `FalOTS` to address the global class imbalance problem in FAL by adding an orthogonal, target-aware correction to SAM perturbation while preserving its first-order training-loss increase (Sec. 3). We further isolate its first-order effect relative to SAM and characterize when the one-step target-risk comparison is favorable (Sec. 4).
3. We conduct experiments across eight benchmarks spanning both image and text modalities under severe global class imbalance to show the superior performance of our `FalOTS` algorithm for FAL and its ability to reduce global sharpness (Sec. 5).

2 Revisiting SAM-Based Algorithms

In this section, we formally describe the federated active learning (FAL) setup, review SAM and its variants that alter its perturbation, and outline our observation and motivation.

2.1 Preliminaries

Federated Active Learning (FAL). Consider a system with K clients, each holding a data pool $\mathcal{P}_k$ and a labeled subset $\mathcal{L}_k \subseteq \mathcal{P}_k$, with $n_k = |\mathcal{L}_k| > 0$ labeled examples. Clients collaboratively train a shared model $\theta \in \mathbb{R}^P$ without centralizing raw data using the aggregate empirical objective in Eq. 1. FAL alternates between training and acquisition. Each training phase comprises T communication rounds on fixed labeled sets: in each round, the server distributes the current global model, clients perform local updates, and the server aggregates their updates. At an acquisition step, clients select examples from the unlabeled remainder $\mathcal{U}_k = \mathcal{P}_k \setminus \mathcal{L}_k$ for annotation, expanding the labeled sets for the next training phase. Communication rounds update the model, while acquisition steps

Table 1: Overview of SAM-based algorithms. *Client*: explicit correction of local-global perturbation mismatch; *Class*: class-dependent treatment; *Target*: explicit use of a specified target class prior. θ^{old} is global model of previous round; L_{tail} is loss on tail classes; other notations are in Appx. B

Category	Method	Perturbation direction	Client	Class	Target
Federated	`FedSAM` (Qu et al., 2022; Caldarola et al., 2022)	$\rho \nabla L_k / \|\nabla L_k\|$ (local)	×	×	×
	`FedLESAM` (Fan et al., 2024)	$\rho(\theta^{\text{old}} - \theta) / \|\theta^{\text{old}} - \theta\|$	✓	×	×
	`FedGMT` (Li et al., 2025b)	global-trajectory guidance (EMA)	✓	×	×
	`FedScar` (Lu et al., 2026)	historical geometric-bias correction	✓	×	×
Class-aware	`ImbSAM` (Zhou et al., 2023)	$\rho \nabla L_{\text{tail}} / \|\nabla L_{\text{tail}}\|$	×	✓	×
	`Focal-SAM` (Li et al., 2025a)	class-weighted sharpness	×	✓	×
FAL	`FalOTS` (ours)	$\rho \hat{g} + \rho u,\ u \perp \hat{g}$	×	✓	✓

expand the training data. We index training phases by the labeled fraction $b = \sum_k |\mathcal{L}_k| / \sum_k |\mathcal{P}_k|$.

$$L(\theta) = \sum_{k=1}^{K} q_k L_k(\theta), \qquad q_k = n_k / \sum_j n_j, \qquad L_k(\theta) = \frac{1}{n_k} \sum_{(x,y)\in\mathcal{L}_k} \ell(\theta; x, y). \tag{1}$$

Class Priors. A class prior is a distribution over C labels. Let π_k be the client k's labeled set prior, π_{pool} the prior of the full pooled data $\bigcup_k \mathcal{P}_k$, and π_{tgt} the desired evaluation prior. The aggregate labeled prior is $\pi_{\text{emp}}(y) = \sum_k q_k \pi_k(y)$. These serve different roles: π_{pool} defines the fixed available data, π_{emp} defines the labels acquired so far, and π_{tgt} specifies the target class mixture. We use π_{pool} for the experimental data distribution; computing `FalOTS` weights needs only current labeled class counts and π_{tgt}, without access to labels of unqueried examples. The empirical loss L averages over the labeled data, while the target surrogate risk is $L_{\text{tgt}}(\theta) = \mathbb{E}_{(x,y)\sim P_{\text{tgt}}} \ell(\theta; x, y)$, where P_{tgt} specifies both target prior and class-conditionals. A uniform π_{tgt} assigns equal importance to each class. Class-balanced accuracy equals ordinary accuracy when test class frequencies are uniform. The client partition controls heterogeneity, acquisition changes π_{emp}, and the evaluation objective defines π_{tgt}. Thus, nearly identical client priors can coexist with a global labeled-target mismatch.

Sharpness-aware Minimization (SAM). SAM (Foret et al., 2021) is motivated by minimizing the neighborhood objective $\mathcal{M}_\rho(\theta) = \max_{\|e\|\le\rho} L(\theta + e)$. Its practical one-step approximation is:

$$\widetilde{L}^{\text{SAM}}(\theta) = L(\theta + \epsilon_{\text{SAM}}), \qquad \epsilon_{\text{SAM}} = \rho \widehat{g}, \quad \widehat{g} = \nabla L(\theta) / \|\nabla L(\theta)\| \quad (\nabla L(\theta) \neq 0). \tag{2}$$

We set $\widehat{g} = 0$ when $\nabla L(\theta) = 0$. The perturbation maximizes the linearized loss over the radius-ρ ball; it need not maximize the nonlinear loss. For a locally Lipschitz gradient,

$$S(\theta) \triangleq L(\theta + \epsilon_{\text{SAM}}) - L(\theta) = \rho \|\nabla L(\theta)\| + O(\rho^2). \tag{3}$$

Here, S is the loss increase along the chosen perturbation. The practical descent direction is $\nabla L(\theta + \epsilon_{\text{SAM}})$, with the perturbation held fixed (stop-gradient); it is not the total derivative of $\widetilde{L}^{\text{SAM}}(\theta)$ through $\epsilon_{\text{SAM}}(\theta)$. SAM requires one extra forward and backward pass compared with SGD.

2.2 Limitations of Empirical SAM Perturbations

Current refinements of SAM perturbation can be divided into two groups (Tab. 1). 1) The *federated family* addresses local-global differences in optimization. `FedSAM` (Qu et al., 2022) constructs a perturbation from each client's empirical loss. Since local flatness need not reflect global flatness, `FedLESAM` (Fan et al., 2024) estimates a global perturbation from previously received global models. `FedGMT` (Li et al., 2025b) uses global model trajectory information, while `FedScar` (Lu et al., 2026) maintains a historical geometric-bias correction. These methods address client heterogeneity but do not explicitly specify the desired evaluation class (target) prior. 2) The *class-aware family* changes how classes contribute to sharpness-aware training. `ImbSAM` (Zhou et al., 2023) constructs perturbations from a tail-class loss, while `Focal-SAM` (Li et al., 2025a) assigns different penalties to class-wise sharpness. Our focus is the evolving global labeled-target prior mismatch in FAL, which is not addressed in the above SAM-based algorithms, by using an explicit target prior while keeping SAM empirical-gradient component. In AL, SAM is adopted only in `SAAL` (Kim et al., 2023b) that uses sharpness for an acquisition score; thus, we also include it in the comparison.

Motivation. To see why global labeled-target prior mismatch matters, recall that SAM constructs its perturbation from the empirical gradient $\nabla L(\theta)$, which determines the leading loss increase in

Eq. 3. Yet, the target risk L_{tgt} reflects the desired class mixture π_{tgt}. Even under matched class-conditionals, a mismatch between the aggregate labeled prior π_{emp} and π_{tgt} can cause the empirical and target-risk gradients to differ in direction. To identify component that SAM perturbations cannot capture, we decompose the target-risk gradient into components parallel and orthogonal to $\widehat{g}$:

$$\nabla L_{\mathrm{tgt}}(\theta) = \langle \nabla L_{\mathrm{tgt}}, \widehat{g} \rangle \, \widehat{g} + v, \qquad v = \nabla L_{\mathrm{tgt}} - \langle \nabla L_{\mathrm{tgt}}, \widehat{g} \rangle \, \widehat{g}, \qquad v \perp \widehat{g}, \tag{4}$$

The first term ($\langle \nabla L_{\mathrm{tgt}}, \widehat{g} \rangle \, \widehat{g}$) is collinear with SAM perturbation, while the second term (v) is orthogonal to it. Adjusting perturbation radius ρ only rescales SAM perturbation along its original direction, making it impossible to recover the orthogonal residual v. Conversely, reweighting or replacing the perturbation direction introduces an orthogonal component but destroys the original empirical-gradient SAM component. This motivates our formulation in Sec. 3: use the current global labeled prior to form a target-aware gradient, then add its orthogonal residual to SAM perturbation.

3 METHODOLOGY

Motivated by Sec. 2, we present `FalOTS` as a local training optimizer within FAL loop. After initial labeling and each acquisition step, the server aggregates labeled class counts and broadcasts the resulting class weights. Clients use these fixed weights throughout the training phase to construct a target-aware gradient (Sec. 3.1) and retain its component orthogonal to their empirical SAM direction (Sec. 3.2) before assembling the final perturbation (3.3). The correction adapts to the newly acquired labels while leaving the acquisition rule and aggregation formula unchanged.

3.1 TARGET-AWARE GRADIENT ESTIMATION FROM CLASS PRIORS

Assumption 1 (Label shift with support overlap)**.** Let P_{emp} and P_{tgt} be fixed joint distributions of (x, y), with $y \in \{1, \ldots, C\}$ and class priors π_{emp} and π_{tgt}. For every y with $\pi_{\mathrm{tgt}}(y) > 0$,

$$\pi_{\mathrm{emp}}(y) > 0, \qquad P_{\mathrm{emp}}(\cdot \mid y) = P_{\mathrm{tgt}}(\cdot \mid y).$$

The priors may be equal or different and do not depend on θ. Class-wise subsampling that is independent of x conditional on y maintains the class-conditional distributions at the population level. Active acquisition can create within-class selection bias. Therefore, Assumption 1 is an idealization, not a result of FAL. P_{emp} is the aggregate labeled-data law, and we do not assume that finite-sample class-conditionals equal deployment class-conditionals. We then define the fixed class weights as:

$$w(y) = \frac{\pi_{\mathrm{tgt}}(y)}{\pi_{\mathrm{emp}}(y)} \quad \text{if } \pi_{\mathrm{emp}}(y) > 0, \qquad w(y) = 0 \quad \text{otherwise.} \tag{5}$$

The second convention concerns classes absent from both supports under Assumption 1.

Proposition 1 (Target-gradient identity)**.** *Under Assumption 1, suppose $\ell(\theta; x, y)$ is differentiable in a neighborhood of θ, its gradient is dominated there by an integrable envelope under each relevant source and target class-conditional, and the corresponding risks are finite. Then,*

$$\mathbb{E}_{\mathrm{emp}}[w(y) \nabla_\theta \ell(\theta; x, y)] = \nabla L_{\mathrm{tgt}}(\theta). \tag{6}$$

This identity applies to the unnormalized weighted gradient with exact population ratios. For a shared model parameter, aggregating client expectations weighted by $q_k = n_k / \sum_i n_i$ recovers the source expectation when $P_{\mathrm{emp}} = \sum_k q_k P_k$. Furthermore, projecting and normalizing client gradients does not commute with aggregation. In practice, the server computes the aggregate labeled histogram once per acquisition phase and broadcasts the corresponding class ratios. For a mini-batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^n$ with a normalization factor $S_w = \sum_i w(y_i) > 0$, we define:

$$L_w(\theta) = \frac{\sum_i w(y_i) \ell(\theta; x_i, y_i)}{S_w}, \qquad g_{\mathrm{tgt}} = \nabla_\theta L_w(\theta). \tag{7}$$

The unweighted objective on the same batch is $L_{\mathcal{B}}(\theta) = n^{-1} \sum_i \ell(\theta; x_i, y_i)$. Self-normalization rescales unnormalized weighted gradient by a positive scalar, leaving the normalized orthogonal residual directionally invariant in exact arithmetic, though this property does not imply finite-batch unbiasedness. If $S_w = 0$, we omit the target correction. Moreover, the population identity in Proposition 1 relies on label shift and support overlap assumptions. If a target class is absent globally, its gradient cannot be recovered from the available labels. Thus, g_{tgt} serves as a target-aware reweighted gradient whose exact population interpretation depends on these constraints, while its subsequent geometric properties require only a well-defined, finite correction.

3.2 ORTHOGONAL TARGET-AWARE CORRECTION

For the following geometric statements, L is the objective used to compute the empirical direction: the aggregate objective for aggregate analysis, L_k for exact local analysis, or $L_{\mathcal{B}}$ for a realized mini-batch. Simply replacing the empirical gradient with the target-weighted gradient g_{tgt} from Sec. 3.1 would alter SAM ascent direction and its empirical-gradient component. Thus, we extract only information in g_{tgt} that is absent in the empirical SAM direction. We define:

$$g_{\mathrm{emp}} = \nabla L(\theta), \qquad \widehat{g} = \begin{cases} g_{\mathrm{emp}} / \left\| g_{\mathrm{emp}} \right\|, & g_{\mathrm{emp}} \neq 0, \\ 0, & g_{\mathrm{emp}} = 0, \end{cases} \qquad v = g_{\mathrm{tgt}} - \langle g_{\mathrm{tgt}}, \widehat{g} \rangle \, \widehat{g}. \tag{8}$$

For $v \neq 0$, we define $u = v / \|v\|$; otherwise we set $u = 0$. When $g_{\mathrm{emp}} = 0$, this convention yields $v = g_{\mathrm{tgt}}$, rendering every direction first-order neutral to L at that stationary point. By construction, $u \perp g_{\mathrm{emp}}$. We call $\widehat{g}$ the sharpness direction (matching the standard SAM perturbation path) and u the target-residual direction (which is orthogonal to $\widehat{g}$). Normalizing v decouples direction from magnitude, ensuring that the intensity of the target-aware correction is governed explicitly by the perturbation budget rather than the variable scale of the importance-weighted gradient. This normalization also renders u invariant to any positive rescaling of g_{tgt}, including the self-normalization used in Eq. 7. The following proposition characterizes this vector as a constrained ascent direction.

Proposition 2 (Optimal target-aware direction under first-order neutrality)**.** *At a fixed θ, if $v \neq 0$, then u is the unique solution of*

$$u = \arg\max_{d \in \mathbb{R}^P} \left\{ \langle g_{\mathrm{tgt}}, d \rangle : \|d\| = 1, \ \langle g_{\mathrm{emp}}, d \rangle = 0 \right\}. \tag{9}$$

Moreover, $\langle g_{\mathrm{emp}}, u \rangle = 0$. Hence, if L has a locally Lipschitz gradient,

$$|L(\theta + tu) - L(\theta)| \leq \frac{M_L}{2} t^2 \tag{10}$$

for sufficiently small $|t|$, where M_L is a Lipschitz constant of ∇L on the segment in question. If $v = 0$, we choose $u = 0$ by convention; no uniqueness of a unit maximizer is asserted.

Here, u maximizes the first-order increase of the target-weighted loss among all unit directions that are first-order neutral to the empirical loss. This first-order ascent optimality concerns the target-aware reweighted mini-batch loss; it does not guarantee target-risk reduction after descent step. On a mini-batch, neutrality holds as regards $L_{\mathcal{B}}$, not necessarily full local or aggregate empirical risk.

3.3 ASSEMBLING THE PERTURBATION

We now have two components: the sharpness direction $\widehat{g}$ and the target-residual direction u. u is orthogonal to ∇L, ensuring it does not change the training loss to first order. We use a single scalar $\rho \geq 0$ to allocate an equal perturbation budget to each active direction:

$$\epsilon_{\texttt{FalOTS}} = \rho \widehat{g} + \rho u, \qquad \theta^{+} = \theta - \eta \nabla L(\theta + \epsilon_{\texttt{FalOTS}}), \tag{11}$$

with $u = 0$ when the orthogonal residual vanishes. The subsequent descent step evaluates the gradient at the displaced parameter vector while holding the perturbation fixed. Assigning the same ρ to both directions keeps SAM's interpretation of the radius along $\widehat{g}$. The equal-budget choice also yields a straightforward geometric characterization. The total perturbation norm increases to $\rho\sqrt{2}$ when both $g_{\mathrm{emp}} \neq 0$ and $u \neq 0$. We examine the role of this larger norm through radius-scaling controls in Fig. 6c. Proposition 3 formalizes these exact geometric properties.

Proposition 3 (Preservation and reduction)**.** *Under the conventions above, for $\rho \geq 0$:*

1. *$\langle \nabla L, \epsilon_{\texttt{FalOTS}} \rangle = \rho \|\nabla L\|$, equal to SAM linearized empirical-loss increase with budget ρ.*
2. *If $\pi_{\mathrm{tgt}} = \pi_{\mathrm{emp}}$ and both gradients use the same objective and examples with exact ratios in Eq. 5, then $L_w = L$, $v = 0$, $u = 0$, and the update equals SAM with the same zero-gradient convention.*
3. *$\|\epsilon_{\texttt{FalOTS}}\|^2 = \rho^2(\|\widehat{g}\|^2 + \|u\|^2) \leq 2\rho^2$. If $\nabla L \neq 0$, the norm is $\rho\sqrt{2}$ when $u \neq 0$ and ρ otherwise. If $\nabla L = 0$, the norm is $\rho \|u\|$.*

These equalities concern exact normalization. Numerical safeguards must preserve the unit projector and explicitly handle zero residuals; in particular, adding a smoothing constant $\delta > 0$ to the denominator ($\|g_{\mathrm{emp}}\| + \delta$) in Eq. 8 generally destroys exact orthogonality.

4 First-Order Analysis

We analyze how the additional perturbation modifies a local update before federated aggregation. We compare one plain gradient-descent step from the same parameter θ, using the same differentiable objective L, learning rate $\eta > 0$, and component budget ρ. The directions $\widehat{g}$ and u are computed at θ and fixed during descent. For a single client or mini-batch, L can be replaced throughout by its corresponding objective. Suppose L is C^2 on a convex neighborhood of θ and its Hessian is M_H-Lipschitz in operator norm there. Let $H = \nabla^2 L(\theta)$, for sufficiently small ρ we have:

$$\nabla L(\theta + \epsilon_{\texttt{FalOTS}}) = \nabla L(\theta + \rho\widehat{g} + \rho u) = \nabla L(\theta) + \rho H\widehat{g} + \rho H u + O(\rho^2), \tag{12}$$

$$\nabla L(\theta + \epsilon_{\text{SAM}}) = \nabla L(\theta + \rho\widehat{g}) = \nabla L(\theta) + \rho H\widehat{g} + O(\rho^2). \tag{13}$$

Proposition 4 (Isolated first-order correction)**.** *Under the preceding regularity assumptions, let $\theta^+_{\texttt{FalOTS}}$ and θ^+_{SAM} be the two updates defined above. Then,*

$$\theta^+_{\texttt{FalOTS}} - \theta^+_{\text{SAM}} = -\eta\rho H u + r_\rho, \qquad \|r_\rho\| \le \tfrac{3}{2} M_H \eta \rho^2. \tag{14}$$

Thus, the term $-\eta\rho H\widehat{g}$ is preserved to first order in ρ. Although $u \perp \nabla L$, Hu need not be orthogonal to ∇L or to $H\widehat{g}$; orthogonality remains a structural property of the perturbation construction itself. To evaluate this one-step comparison on the target surrogate risk, let $r = \nabla L_{\text{tgt}}(\theta)$. Suppose ∇L_{tgt} is M_{tgt}-Lipschitz on a convex neighborhood of θ containing both updates and their connecting segment. Let B_L bound $\|\nabla L\|$ and M_L bound $\left\|\nabla^2 L\right\|_{\text{op}}$ on a fixed neighborhood containing the perturbation segments and the segment joining the two perturbed points. Then, we have:

$$\begin{aligned} L_{\text{tgt}}(\theta^+_{\texttt{FalOTS}}) - L_{\text{tgt}}(\theta^+_{\text{SAM}}) &= -\eta\rho\,\langle r, Hu\rangle + \mathcal{E}_{\eta,\rho}, \\ |\mathcal{E}_{\eta,\rho}| &\le \tfrac{3}{2} M_H \|r\|\,\eta\rho^2 + M_{\text{tgt}} B_L M_L \eta^2 \rho. \end{aligned} \tag{15}$$

Consequently, the leading comparison is governed by

$$A_\perp(u) = \langle r, Hu\rangle = \langle Hr, u\rangle\,. \tag{16}$$

A sufficient condition for a strictly smaller target risk than SAM after this step is $A_\perp(u) > \frac{3}{2} M_H \|r\|\,\rho + M_{\text{tgt}} B_L M_L \eta$. Thus, $A_\perp(u) > 0$ yields an improvement for sufficiently small positive η and ρ.This is a relative one-step comparative guarantee, not an assertion of absolute target-risk reduction. Also, gradient alignment alone is insufficient to satisfy this curvature condition (Appx. E).

For the aggregate objective, this describes a hypothetical aggregate step. For one exact local step per client followed by averaging, the leading correction is instead $-\eta\rho \sum_k q_k H_k u_k$, with $H_k = \nabla^2 L_k(\theta)$; it is not generally $-\eta\rho H u$ for an aggregate residual u. The corresponding leading target-risk comparison depends on $\langle r, \sum_k q_k H_k u_k\rangle$, highlighting the roles of local curvature and aggregation explicitly. Appx. D separately treats multiple local steps and same-batch stochastic updates in a fixed acquisition phase, under smoothness and bounded client dissimilarity assumptions.

5 Experiments

Datasets. Following recent work (Zong & Huang, 2026), we evaluate `FalOTS` on both natural and medical image classification benchmarks. The natural image datasets include FMNIST (Xiao et al., 2017), CIFAR-10/100 (Krizhevsky et al., 2009) and Tiny ImageNet (Le et al., 2015). We use OctMNIST and DermaMNIST from MedMNIST (Yang et al., 2023) as medical benchmarks. We further evaluate `FalOTS` on AGNews and DBpedia-14 datasets (Zhang et al., 2015) for text classification using a pretrained DistilBERT Sanh et al. (2019) model.

Baselines. Following (Zong & Huang, 2026), we compare `FalOTS` against diverse query strategies: 1) Random sampling; 2) Uncertainty-based methods: `Entropy` (Holub et al., 2008), `Margin` (Balcan et al., 2007); 3) Diversity-based method: `Coreset` (Sener & Savarese, 2017); 4) Hybrid approaches: `BADGE` (Ash et al., 2020), `ALFA-Mix` (Parvaneh et al., 2022) that utilize uncertainty and diversity cues; 5) FAL methods: `KAFAL` (Cao et al., 2023), `LoGo` (Kim et al., 2023a), `IFAL` (Zong et al., 2025), `FairFAL` (Zong & Huang, 2026). We then compare with SAM-based algorithms in long-tailed settings: `ImbSAM` (Zhou et al., 2023) and `Focal-SAM` (Li et al., 2025a). We also examine AL approach: `SAAL` (Kim et al., 2023b) and FL methods: `FedSAM` (Qu et al., 2022), `FedLESAM` (Fan et al., 2024), `FedGMT` (Li et al., 2025b), and `FedScar` (Lu et al., 2026).

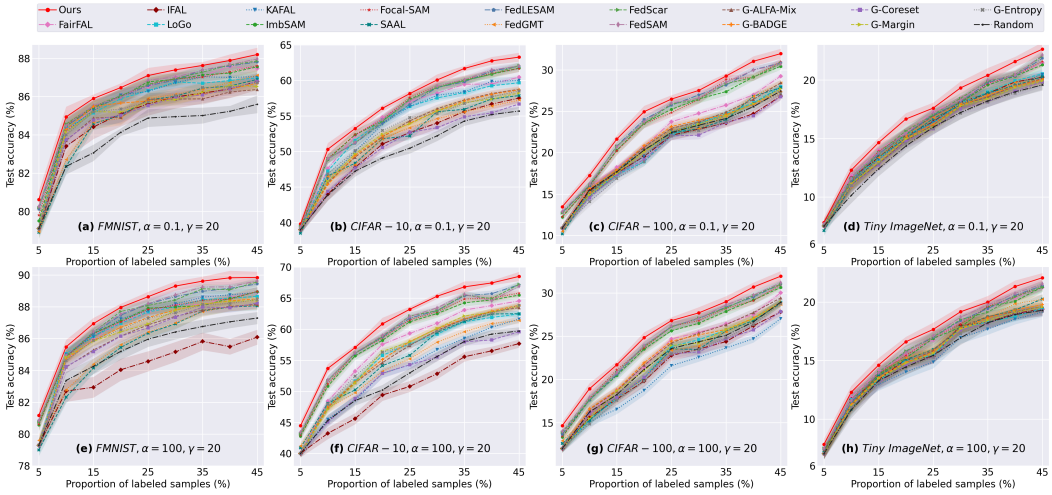


Figure 2: Test accuracy vs. proportion of labeled samples on FMNIST, CIFAR-10/100 and Tiny ImageNet under global imbalance ($\gamma = 20$) and two heterogeneity levels ($\alpha \in \{0.1, 100\}$). Results are averaged over five seeds. G denotes the global query-selector model. Local query-selector model performance is in Fig. 8 (Appx. F.5). Numerical results appear in Tabs. 4– 5 in Appx. F.5.2.

Implementation Details. The main experiments are conducted using the standard `FedAvg` (McMahan et al., 2017) framework to ensure consistency with recent works (Kim et al., 2023a; Zong et al., 2025; Zong & Huang, 2026). We also evaluate with `FedProx` (Li et al., 2020a) and `SCAFFOLD` (Karimireddy et al., 2020) as ablations in Appx. I.1. We use FairFAL's query strategy as our baseline; training with SGD instead of `FalOTS` algorithm achieves identical performance to `FairFAL`. All SAM-based methods use the same FAL framework for a fair comparison. We also evaluate with recent FAL query strategies from `KAFAL`, `LoGo` and `IFAL` as ablations in Appx. I.2. Main experiments are run on five random seeds. More details are in Appx. F.

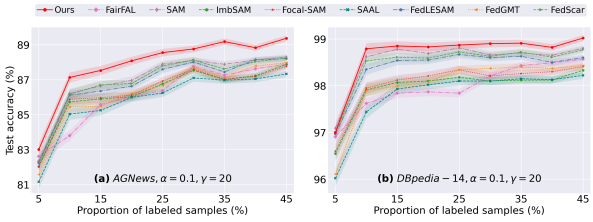


Figure 3: Test accuracy vs. proportion of labeled samples on AGNews and DBpedia-14 with $\gamma = 20$, $\alpha = 0.1$. Numerical results appear in Tab. 6 in Appx. F.5.2.

5.1 Performance Comparison

General Performance. For image classification, Fig. 2 and Tabs. 4–5 (Appx. F.5.2) compare test accuracy across labeled budgets on FMNIST, CIFAR-10/100, and Tiny ImageNet, under severe global class imbalance ($\gamma = 20$) and both heterogeneous ($\alpha = 0.1$) and nearly i.i.d. ($\alpha = 100$) client partitions. `FalOTS` achieves superior performance over other methods. On CIFAR-100 with extreme heterogeneity ($\alpha = 0.1$), `FalOTS` achieves **2.71%** and **1.22%** improvement compared to `FairFAL` and `FedSAM` at data ratio 45%. FMNIST benchmark is nearly saturated, but we still observe consistent gains with `FalOTS` compared to other methods. Tab. 2 presents results from two naturally imbalanced medical datasets, OctMNIST and DermaMNIST. `FalOTS` also consistently achieves the highest accuracy across these datasets. These findings highlight `FalOTS`'s robustness in addressing highly imbalanced real-world clinical scenarios. For text classification, Fig. 3 shows the performance comparison of the FairFAL baseline and SAM-based approaches on AGNews and DBpedia-14 datasets. Since performance on these two datasets is saturated, we only evaluate under severe global class imbalance ($\gamma = 20$) and client heterogeneity ($\alpha = 0.1$). `FalOTS` also achieves superior performance compared to other methods, further validating its effectiveness.

Computation and Communication. SAM/`FedSAM` requires two gradient evaluations per local update: one to construct the perturbation and one at the perturbed point. `FalOTS` adds a target-weighted gradient evaluation at the unperturbed point. Its projection and normalization require $O(P)$ vector operations for P model parameters. In addition to the communication required by

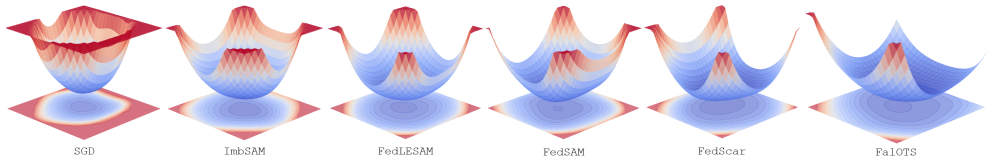


Figure 5: Visualization of the global loss surface on CIFAR-10 under $(\gamma, \alpha) = (20, 0.1)$

Table 2: Final-round test accuracy at 45% of labeled-data ratio comparison on OctMNIST and DermaMNIST under both client heterogeneity $\alpha = 0.1, 100$.

α	Dataset	Random	KAFAL	LoGo	IFAL	FairFAL	ImbSAM	Focal-SAM	SAAL	FedLESAM	FedGMT	FedScar	FedSAM	FalOTS
0.1	OctMNIST	68.30	70.40	70.00	68.40	70.76	71.37	71.52	69.18	71.93	69.87	72.01	72.23	**72.86**
	DermaMNIST	72.32	73.27	73.62	72.97	73.53	73.11	73.24	72.86	73.47	72.93	73.50	73.41	**74.59**
100	OctMNIST	68.30	70.40	70.00	68.40	74.32	74.64	75.02	73.19	75.27	73.81	76.47	76.46	**76.88**
	DermaMNIST	72.32	73.27	73.62	72.97	75.12	75.18	75.96	74.88	76.62	75.13	77.04	77.36	**77.42**

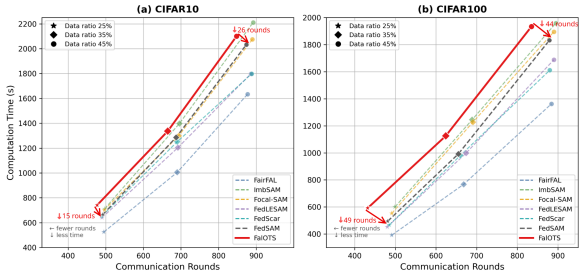


Figure 4: Efficiency trade-off on CIFAR benchmarks.

the underlying federated training protocol, each client sends a C-dimensional labeled class histogram and receives C class weights once per acquisition phase. This adds $O(KC)$ scalar communication per phase across K clients, compared with $O(KP)$ per full-participation model-exchange round. Fig. 4 compares communication rounds and computation time to reach fixed accuracy thresholds: $56\%, 59\%, 60\%$ on CIFAR-10 and $23\%, 25\%, 29\%$ on CIFAR-100 at labeled fractions $25\%, 35\%, 45\%$. `FairFAL` with SGD has the lowest run time, whereas `FalOTS` requires the fewest communication rounds. Relative to `FedSAM`, the reduction reaches 49 rounds on CIFAR-100 with added compute. In FAL, reducing rounds is crucial because of network latency, limited bandwidth, stragglers, and the cost of uploading and downloading models. `FalOTS` also achieves superior results across data ratios compared to other methods (Fig. 2). So, the current trade-off between computation and communication is reasonable.

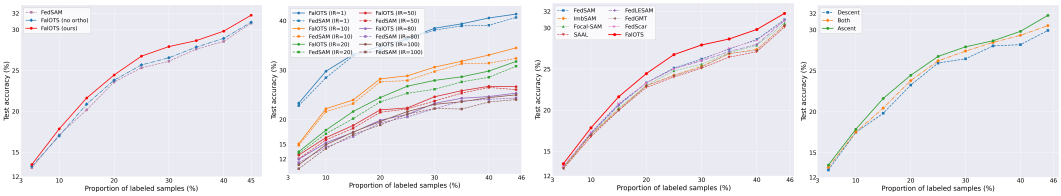
Visualization of Global Loss Surface. Fig. 5 shows the global loss surface visualization on CIFAR-10. All representative SAM-based algorithms exhibit a much flatter loss landscape than SGD within the same FAL framework. Moreover, `FalOTS` further reduces global loss-surface sharpness, showing improved generalization and complementing the accuracy results.

5.2 Ablation Study

Impact of orthogonal target-aware correction. Fig. 6a examines whether to retain the target-aware reweighted gradient component aligned with SAM empirical direction. The variant retaining that component improves over `FedSAM`, but using only the orthogonal correction gives larger gains. This indicates that preserving the component aligned with SAM direction may amplify pre-existing noise within SAM perturbation, particularly in severe global class imbalance. By ignoring this overlapping component, `FalOTS` offers a more complementary optimization signal.

Impact of global class imbalance factor γ. We vary $\gamma \in \{1, 10, 20, 50, 80, 100\}$ to assess its sensitivity. Fig. 6b shows gains over `FedSAM` at every tested ratio, with the largest gains at $\gamma \in \{20, 50\}$. As γ increases, benchmarks are increasingly challenging, resulting in smaller yet still consistent gains over `FedSAM`. We also examine the balanced case ($\gamma = 1$), showing that `FalOTS` still improves over `FedSAM`, though the gains are less pronounced than those in imbalanced conditions (e.g., $\gamma = 20$, 50). A balanced pool ($\gamma = 1$) does not imply that the labeled prior π_{emp} equals π_{tgt}. Thus, this does not contradict the exact matched-prior reduction in Proposition 3.

Matched-radius comparison. By Proposition 3, the total perturbation norm of `FalOTS` is $\rho\sqrt{2}$ when both directions are nonzero. To examine whether simply enlarging the perturbation explains its gains, we increase each SAM-based baseline's selected radius from ρ to $\rho\sqrt{2}$, where ρ is the optimal



(a) Impact of orthogonality (b) Impact of γ (c) Impact of matched ρ (d) Impact of target scope

Figure 6: Ablation studies for our proposed FalOTS on CIFAR-100 dataset

hyperparameter for each method specified in Appx. F.6. Fig. 6c shows that these enlarged-radius baselines do not close the gap to FalOTS. This supports the importance of perturbation direction beyond a simple radius increase. Since the starting radius is selected separately for each method, this is a radius-scaling control rather than necessarily an exact equal-norm comparison across methods.

Impact of target scope. We apply the orthogonal target-aware correction during ascent, descent, or both stages (Fig. 6d). Ascent-only correction gives the best results; applying it during descent, alone or together with ascent, gives lower accuracy. These results support using target-aware information to construct the perturbation while retaining the unweighted empirical objective for descent.

6 Related Work

Federated active learning (FAL). FAL integrates AL into FL to facilitate label-efficient model training without centralizing raw examples. Early studies (Ahmed et al., 2020; Ahn et al., 2024; Wu et al., 2022) directly applied centralized active learning strategies using either a global or a local model. However, these single-model approaches tend to degrade quickly under significant client heterogeneity. Recent methods (Cao et al., 2023; Chen et al., 2024; Kim et al., 2023a; Zong et al., 2025; Zong & Huang, 2026) have therefore adopted hybrid designs that combine global and local signals. For instance, LoGo (Kim et al., 2023a) performs local clustering followed by a global uncertainty scoring process, while KAFAL (Cao et al., 2023) and IFAL (Zong et al., 2025) leverage prediction discrepancies between global and local models to guide data acquisition. More recently, FairFAL (Zong & Huang, 2026) enables adaptive selection between global and local query models by inferring local-global divergence through lightweight prediction discrepancy. These methods primarily improve acquisition through the choice of query signals and models. Our contribution is complementary: FalOTS targets local sharpness-aware training to account for the aggregate labeled-target prior mismatch, enhancing the model generalization under global class imbalance.

Sharpness-Aware Optimization in FL. SAM (Foret et al., 2021) focuses on finding parameters for which the empirical loss remains low within a local neighborhood. FedSAM (Qu et al., 2022; Caldarola et al., 2022) and related methods, including FedGamma (Dai et al., 2023), FedLESAM (Fan et al., 2024), FedSMOO (Sun et al., 2023a), FedGMT (Li et al., 2025b) and FedScar (Lu et al., 2026), adapt SAM to heterogeneous clients in FL. These approaches primarily address the mismatch between local and global optimization and the quality of perturbation estimation in a federated context. FalOTS instead addresses the aggregate class-prior mismatch encountered in FAL settings under severe global class imbalance, which can persist even under nearly i.i.d. client partitions. It preserves SAM empirical-gradient component and adds the steepest unit target-aware direction that is first-order neutral to the empirical loss, whenever the orthogonal residual is nonzero.

7 Conclusion

We study global class imbalance in FAL through the local training objective. We propose FalOTS, which uses the current global labeled prior to add an orthogonal, target-aware correction to SAM perturbation while retaining the unweighted descent objective. We establish the correction's geometric properties and provide conditions for a favorable one-step target-risk comparison with SAM. Experiments show improved global model accuracy over FAL baselines using SGD- and SAM-based optimizers, including under nearly i.i.d. client partitions. This supports accounting for the evolving labeled-target prior mismatch during local training, alongside acquisition design. Handling within-class acquisition bias and characterizing target-risk behavior across phases remain open directions.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and grammar correction, and we have reviewed all AI-assisted work. We have not used generative AI tools for generating the core scientific content of this study, including the formulation of research ideas, the derivation of scientific conclusions, or the interpretation of results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We acknowledge that all datasets used in this work are publicly available and comply with relevant licensing terms. No personally identifiable information (PII) or sensitive data was collected or used. We believe this technology holds significant potential for beneficial applications. We are committed to the responsible advancement of this field and encourage continued research to establish clear ethical guidelines for deployment.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the proposed method and its theoretical and experimental settings in the main paper and appendix. Sec. 3 details the target-aware gradient construction and orthogonal correction, with the client-side update procedure provided in Appx. G. Our implementation builds on the original SAM (Foret et al., 2021)[2] and FairFAL (Zong & Huang, 2026)[3] repositories. Appx. F describes the benchmark datasets, client partitioning, class-imbalance settings, annotation budgets, model architectures, training hyperparameters, baseline configurations, and computational environment. We also report the implementation details for the SAM-based algorithm baselines, including hyperparameter selection, in Appx. F.6. We report the main experimental results averaged over five random seeds and provide the numerical results underlying the main comparison figure in Appx F.5.2. For the theoretical results, Appx. E provides complete proofs of the propositions and the first-order target-risk comparison, while Appx. D states the assumptions and proofs for within-phase convergence under exact and stochastic local-gradient updates. We believe that these details enable practitioners to implement our ideas straightforwardly. We will release our code upon acceptance to support independent reproduction and further research.

REFERENCES

Lulwa Ahmed, Kashif Ahmad, Naina Said, Basheer Qolomany, Junaid Qadir, and Ala Al-Fuqaha. Active learning based federated learning for waste and natural disaster image classification. *IEEE access*, 8:208518–208531, 2020.

Jin-Hyun Ahn, Yeeun Ma, Seoyun Park, and Cheolwoo You. Federated active learning (f-al): an efficient annotation strategy for federated learning. *IEEE Access*, 12:39261–39269, 2024.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryghZJBKPS.

Wonho Bae, Junhyug Noh, and Danica J Sutherland. Generalized coverage for more robust low-budget active learning. In *European conference on computer vision*, pp. 318–334. Springer, 2024.

Sikai Bai, Shuaicheng Li, Weiming Zhuang, Jie Zhang, Kunlin Yang, Jun Hou, Shuai Yi, Shuai Zhang, and Junyu Gao. Combating data imbalances in federated semi-supervised learning with dual regulators. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 10989–10997, 2024.

[2] https://github.com/google-research/sam

[3] https://github.com/chenchenzong/FairFAL

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pp. 35–50. Springer, 2007.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

MA Benteng, FENG Yu, CHE Geng, et al. Federated adaptive reweighting for medical image classification [j]. *Pattern Recognition*, 144:109880, 2023.

Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving generalization in federated learning by seeking flat minima. In *European Conference on Computer Vision*, pp. 654–672. Springer, 2022.

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.

Yu-Tong Cao, Ye Shi, Baosheng Yu, Jingya Wang, and Dacheng Tao. Knowledge-aware federated active learning with non-iid data. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22222–22232. IEEE, 2023.

Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9578–9587. IEEE, 2021.

Jiayi Chen, Benteng Ma, Hengfei Cui, and Yong Xia. Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11439–11449. IEEE, 2024.

Zhen Chen, Meilu Zhu, Chen Yang, and Yixuan Yuan. Personalized retrogress-resilient framework for real-world medical federated learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 347–356. Springer, 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.

Rong Dai, Xun Yang, Yan Sun, Li Shen, Xinmei Tian, Meng Wang, and Yongdong Zhang. Fedgamma: Federated learning with global sharpness-aware minimization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):17479–17492, 2023.

Ziqing Fan, Shengchao Hu, Jiangchao Yao, Gang Niu, Ya Zhang, Masashi Sugiyama, and Yanfeng Wang. Locally estimated global perturbations are better than local perturbations for federated sharpness-aware minimization. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=6axTFAlzRV.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pp. 510–526. Springer, 2020.

Yongxin Guo, Xiaoying Tang, and Tao Lin. Fedbr: Improving federated learning on heterogeneous data via local learning bias reduction. In *International conference on machine learning*, pp. 12034–12054. PMLR, 2023.

Guy Hacohen and Daphna Weinshall. How to select which active learning strategy is best suited for your specific problem and budget. *Advances in Neural Information Processing Systems*, 36: 13395–13407, 2023.

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.

Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 1–8. IEEE, 2008.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. Semi-supervised active learning with temporal output discrepancy. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3427–3436. IEEE, 2021.

Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16312–16322, June 2023.

Yongzhe Jia, Xuyun Zhang, Hongsheng Hu, Kim-Kwang Raymond Choo, Lianyong Qi, Xiaolong Xu, Amin Beheshti, and Wanchun Dou. Dapperfl: domain adaptive federated learning with model fusion pruning for edge devices. *Advances in Neural Information Processing Systems*, 37:13099–13123, 2024.

Meirui Jiang, Holger R Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair federated medical image segmentation via client contribution estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16302–16311. IEEE, 2023.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

SangMook Kim, Sangmin Bae, Hwanjun Song, and Se-Young Yun. Re-thinking federated active learning based on inter-class diversity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3944–3953. IEEE, 2023a.

Yoon-Yeong Kim, Youngjae Cho, JoonHo Jang, Byeonghu Na, Yeongmin Kim, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Saal: sharpness-aware active learning. In *International Conference on Machine Learning*, pp. 16424–16440. PMLR, 2023b.

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

Natsumaro Kutsuna, Takumi Higaki, Sachihiro Matsunaga, Tomoshi Otsuki, Masayuki Yamaguchi, Hirofumi Fujii, and Seiichiro Hasezawa. Active learning framework with iterative clustering for bioimage classification. *Nature communications*, 3(1):1032, 2012.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Mingkun Li and Ishwar K Sethi. Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1251–1261, 2006.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10708–10717. IEEE, 2021.

Sicong Li, Qianqian Xu, Zhiyong Yang, Zitai Wang, Linchao Zhang, Xiaochun Cao, and Qingming Huang. Focal-SAM: Focal sharpness-aware minimization for long-tailed classification. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36624–36651. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/li25dk.html.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020b.

Yuhang Li, Tong Liu, Yangguang Cui, Ming Hu, and Xiaoqiang Li. One arrow, two hawks: Sharpness-aware minimization for federated learning via global model trajectory. In *Forty-second International Conference on Machine Learning*, 2025b.

Xiaoxiao Liang, Yiqun Lin, Huazhu Fu, Lei Zhu, and Xiaomeng Li. Rscfed: Random sampling consensus federated semi-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10144–10153. IEEE, 2022.

Jianfeng Lu, YuZhao Xiang, Yue Chen, Gang Li, Shuqin Cao, and Guanghui Wen. Fedscar: Correcting geometric bias for flatness-consistent federated learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=sI5pTdXuXc.

Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. *Advances in neural information processing systems*, 34:15434–15447, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

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.

Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 79, 2004.

Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton Van Den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12237–12246, 2022.

Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8485–8494. IEEE, 2021.

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International conference on machine learning*, pp. 18250–18280. PMLR, 2022.

Yasar Abbas Ur Rehman, Yan Gao, Pedro Porto Buarque De Gusmao, Mina Alibeigi, Jiajun Shen, and Nicholas D Lane. L-dawa: Layer-wise divergence aware weight aggregation in federated self-supervised visual representation learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16418–16427. IEEE, 2023.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro. An agnostic approach to federated learning with class imbalance. In *International Conference on Learning Representations*, 2021.

Donald Shenaj, Marco Toldo, Alberto Rigon, and Pietro Zanuttigh. Asynchronous federated continual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5055–5063. IEEE, 2023.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

Yan Sun, Li Shen, Shixiang Chen, Liang Ding, and Dacheng Tao. Dynamic regularized sharpness aware minimization in federated learning: Approaching global consistency and smooth landscape. In *International conference on machine learning*, pp. 32991–33013. PMLR, 2023a.

Yan Sun, Li Shen, Tiansheng Huang, Liang Ding, and Dacheng Tao. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=bZjxxYURKT.

Ruth Urner, Sharon Wulff, and Shai Ben-David. Plal: Cluster-based active learning. In *Conference on learning theory*, pp. 376–397. PMLR, 2013.

Fan Wang, Zhongyi Han, Zhiyan Zhang, Rundong He, and Yilong Yin. Mhpl: Minimum happy points learning for active source free domain adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20008–20018. IEEE, 2023.

Jiacheng Wang, Yueming Jin, and Liansheng Wang. Personalizing federated medical image segmentation via local calibration. In *European Conference on Computer Vision*, pp. 456–472. Springer, 2022.

Thorben Werner, Johannes Burchert, Maximilian Stubbemann, and Lars Schmidt-Thieme. A cross-domain benchmark for active learning. *Advances in Neural Information Processing Systems*, 37: 62875–62911, 2024.

Xing Wu, Jie Pei, Cheng Chen, Yimin Zhu, Jianjia Wang, Quan Qian, Jian Zhang, Qun Sun, and Yike Guo. Federated active learning for multicenter collaborative disease diagnosis. *IEEE transactions on medical imaging*, 42(7):2068–2080, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xiao Xiao, Xinyue Yang, Yucheng Wang, Yilong Hui, Jianchuan Zhou, and Guoqiang Mao. Unsupervised domain adaptive vehicle re-identification: A federated learning scheme. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific data*, 10(1):41, 2023.

Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 93–102. IEEE, 2019.

Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International conference on machine learning*, pp. 12073–12086. PMLR, 2021.

Jiakang Yuan, Bo Zhang, Xiangchao Yan, Tao Chen, Botian Shi, Yikang Li, and Yu Qiao. Bi3d: Bi-domain active learning for cross-domain 3d object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15599–15608. IEEE, 2023.

Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3954–3963. IEEE, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Yixuan Zhou, Yi Qu, Xing Xu, and Hengtao Shen. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11311–11321. IEEE, 2023.

Chen-Chen Zong and Sheng-Jun Huang. Rethinking epistemic and aleatoric uncertainty for active open-set annotation: An energy-based approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10153–10162. IEEE, 2025.

Chen-Chen Zong and Sheng-Jun Huang. Federated active learning under extreme non-iid and global class imbalance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24534–24544, June 2026.

Chen-Chen Zong, Tong Jin, and Sheng-Jun Huang. Inconsistency-based federated active learning. In *IJCAI*, pp. 7300–7308, 2025.

Appendix

This appendix includes several supporting materials for the main paper, organized as follows:

Appendix Overview.

- **Extended Related Work** (Appendix A): reviews related advances in federated learning, active learning, and sharpness-aware optimization.
- **Comprehensive Notation Table** (Appendix B): summarizes the symbols and definitions used throughout the paper.
- **Visualization of Local Data Distributions** (Appendix C): illustrates how client heterogeneity and global class imbalance affect the distribution of examples across clients.
- **Within-Phase Convergence Analysis** (Appendix D): establishes empirical stationarity bounds under exact and stochastic local-gradient updates within a fixed acquisition phase.
- **Complete Proofs of the Propositions** (Appendix E): provides proofs of the target-gradient identity, perturbation properties, and first-order comparison with SAM.
- **Detailed Experimental Setups & Additional Results** (Appendix F): describes datasets, training settings, baseline configurations, and additional results supporting the main experiments.
- **`FalOTS` Algorithm** (Appendix G): presents the client-side procedure for constructing the perturbation and updating the model.
- **Verification of the Target-Aware Sharpness Motivation** (Appendix H): verify our observation that motivates our proposed orthogonal target-aware correction.
- **Additional Ablation Studies** (Appendix I): examines the effects of FL frameworks, acquisition strategies, model architectures, and training configurations.
- **Limitations and Discussion** (Appendix J): discusses limitations of the proposed method and open directions to enhance it.

A Extended Related Work

Federated learning (FL). FL allows multiple clients to collaboratively train a shared model without the need to exchange raw data. The foundational framework for FL was established by `FedAvg` (McMahan et al., 2017), but its performance significantly declines when faced with statistical heterogeneity, particularly with non-i.i.d. (independent and identically distributed) data. To address this issue, previous research has identified three main approaches. Regularization-based approaches (Li et al., 2020b; Karimireddy et al., 2020; Guo et al., 2023; Huang et al., 2023) aim to constrain local updates or feature representations in order to reduce client drift. Aggregation-based methods (Jiang et al., 2023; Benteng et al., 2023; Rehman et al., 2023; Zhang et al., 2023) adjust client contributions based on divergence or data quality. Personalized FL (Chen et al., 2021; Wang et al., 2022; Ye et al., 2023) further decouples global and client-specific components to improve adaptability across heterogeneous distributions. In addition to these standard approaches, FL has also been extended to semi-supervised (Liang et al., 2022; Bai et al., 2024), multi-task (Smith et al., 2017; Marfoq et al., 2021), continual (Yoon et al., 2021; Shenaj et al., 2023), and domain-adaptive (Jia et al., 2024; Xiao et al., 2025) learning, showcasing its versatility in real-world applications.

Active learning (AL). AL aims to reduce the costs of annotation by selectively querying the most informative samples from an unlabeled pool. Existing methods can be categorized into three main

groups. Uncertainty-based approaches (Li & Sethi, 2006; Balcan et al., 2007; Gal & Ghahramani, 2016; Yoo & Kweon, 2019; Parvaneh et al., 2022) focus on ambiguous samples that lie near the decision boundary. This uncertainty is typically measured using methods such as entropy (Holub et al., 2008) or margin (Balcan et al., 2007). Diversity-based strategies (Caramalau et al., 2021; Kutsuna et al., 2012; Nguyen & Smeulders, 2004; Sener & Savarese, 2017; Urner et al., 2013) such as Coreset (Sener & Savarese, 2017) concentrate on selecting representative points that cover the feature space effectively. Hybrid methods (Ash et al., 2020; Prabhu et al., 2021; Wang et al., 2023; Yuan et al., 2023; Zong & Huang, 2025) combine both uncertainty and diversity strategies. A well-known example of this is `BADGE` (Ash et al., 2020), which employs a two-stage process of uncertainty-diversity balanced sampling. Recent developments have expanded active learning into low-budget (Bae et al., 2024; Hacohen & Weinshall, 2023), semi-supervised (Gao et al., 2020; Huang et al., 2021), and cross-domain (Werner et al., 2024) scenarios.

Federated active learning (FAL). FAL integrates AL into FL to facilitate label-efficient model training while adhering to privacy constraints. Early studies (Ahmed et al., 2020; Ahn et al., 2024; Wu et al., 2022) directly applied centralized active learning strategies using either a global or a local model. However, these single-model approaches tend to degrade quickly when faced with significant client heterogeneity. Recent methods (Cao et al., 2023; Chen et al., 2024; Kim et al., 2023a; Zong et al., 2025; Zong & Huang, 2026) have therefore adopted hybrid designs that combine global and local signals. For instance, LoGo (Kim et al., 2023a) performs local clustering followed by a global uncertainty scoring process, while KAFAL (Cao et al., 2023) and IFAL (Zong et al., 2025) leverage prediction discrepancies between global and local models to guide data acquisition. More recently, FairFAL (Zong & Huang, 2026) enables adaptive selection between global and local query models by inferring local-global divergence through lightweight prediction discrepancies. These methods primarily improve acquisition through the choice of query signals and models. Our contribution is complementary: `FalOTS` targets local sharpness-aware training to account for the aggregate labeled-target prior mismatch, enhancing the model generalization under global class imbalance.

Sharpness-Aware Optimization in Federated Learning. SAM (Foret et al., 2021) focuses on finding parameters for which the empirical loss remains low within a local neighborhood. FedSAM (Qu et al., 2022; Caldarola et al., 2022) and related methods, including FedGamma (Dai et al., 2023), FedLESAM (Fan et al., 2024), FedSMOO (Sun et al., 2023a), FedGMT (Li et al., 2025b) and FedScar (Lu et al., 2026), adapt SAM for heterogeneous federated training. The primary motivation behind these approaches is to address the mismatch between local and global optimization, as well as the quality of perturbation estimation in a federated context. In contrast, `FalOTS` tackles a different issue in FAL setting under severe global class imbalance: even when the federation is nearly independent and identically distributed (i.i.d.), the overall training distribution can still exhibit global class imbalances when compared to a specified target distribution. `FalOTS` preserves SAM empirical-gradient component and adds the steepest unit target-aware direction that is first-order neutral to the empirical loss, whenever the orthogonal residual is nonzero

Long-tailed learning and class-aware SAM. Long-tailed recognition is often addressed through methods such as class re-weighting, margin modification, or logit adjustment (Cui et al., 2019; Cao et al., 2019; Menon et al., 2020). Approaches like ImbSAM (Zhou et al., 2023) and Focal-SAM (Li et al., 2025a) make SAM class-aware in various ways. However, our method is not merely another tail emphasis technique; it leverages an explicit train-target prior mismatch to estimate a target gradient. `FalOTS` introduces only the component of the gradient that is orthogonal to the empirical SAM direction, thereby preserving the original first-order component of SAM rather than replacing it.

B Comprehensive Notation Table

To enhance readability and ensure mathematical clarity, we summarize the commonly used symbols and their definitions in Tab. 3.

Table 3: Summary of frequently used notations.

Symbol	Description
Federated active learning setup	
K	Total number of clients
C	Number of classes
$\mathcal{P}_k$	Full data pool of client k, containing labeled and unlabeled examples
$\mathcal{L}_k$	Labeled subset of client k, $\mathcal{L}_k \subseteq \mathcal{P}_k$
$\mathcal{U}_k$	Unlabeled remainder, $\mathcal{U}_k = \mathcal{P}_k \setminus \mathcal{L}_k$
n_k	Number of labeled examples at client k, $n_k = \lvert\mathcal{L}_k\rvert$
q_k	Client weight in the aggregate empirical objective, $q_k = n_k / \sum_j n_j$
b	Labeled fraction, $b = \sum_k \lvert\mathcal{L}_k\rvert / \sum_k \lvert\mathcal{P}_k\rvert$; indexes the training phase
T, E	Communication rounds per training phase and local epochs per round, respectively
$\theta \in \mathbb{R}^P$	Model parameters; P is the parameter count
$L(\theta)$	Aggregate empirical loss on labeled data, $L(\theta) = \sum_k q_k L_k(\theta)$
$L_k(\theta)$	Local empirical loss of client k
η	Local learning rate
γ	Global class imbalance factor: largest-to-smallest class-count ratio in the pooled data
α	Dirichlet concentration controlling client heterogeneity; smaller values induce more heterogeneous partitions
Class priors and weighted objectives	
π_k	Class prior of client k's labeled set
π_{emp}	Aggregate labeled class prior, $\pi_{\text{emp}} = \sum_k q_k \pi_k$
π_{pool}	Class prior of the full pooled data across clients, including labeled and unlabeled examples
π_{tgt}	Desired target class prior; uniform for a class-balanced target
$L_{\text{tgt}}(\theta)$	Target surrogate risk, $\mathbb{E}_{(x,y)\sim P_{\text{tgt}}} \ell(\theta; x, y)$
$w(y)$	Class-prior ratio on the labeled support, defined in Eq. 5
$\mathcal{B}$, n	Labeled mini-batch and its size, $n = \lvert\mathcal{B}\rvert$
S_w	Total mini-batch weight, $S_w = \sum_{i=1}^{n} w(y_i)$
$L_w(\theta)$	Self-normalized weighted mini-batch loss, defined when $S_w > 0$
Perturbation geometry	
ρ	Perturbation budget per component
ϵ	Parameter perturbation; $\epsilon_{\text{SAM}} = \rho\widehat{g}$ and $\epsilon_{\texttt{FalOTS}} = \rho\widehat{g} + \rho u$
g_{emp}	Gradient of the empirical objective used in the perturbation step
g_{tgt}	Target-weighted gradient $\nabla L_w(\theta)$ when $S_w > 0$; zero otherwise
$\widehat{g}$	Normalized empirical direction, $g_{\text{emp}} / \lVert g_{\text{emp}}\rVert$ when $g_{\text{emp}} \neq 0$; zero otherwise
v	Orthogonal residual, $v = g_{\text{tgt}} - \langle g_{\text{tgt}}, \widehat{g}\rangle \widehat{g}$
u	Normalized residual, $u = v / \lVert v\rVert$ when $v \neq 0$; zero otherwise
Analysis	
$H(\theta)$	Hessian of the empirical objective under consideration
Λ	Smoothness constant (Lipschitz constant of the gradient)
J	Number of local gradient updates per client per communication round in the analysis
R	Number of communication rounds in the within-phase analysis
τ	Effective round step budget, $\tau = \eta J$
q_2	Squared-weight sum appearing in aggregation-noise bounds, $q_2 = \sum_k q_k^2$

C Visualization of Local Data Distributions

Fig. 7 illustrates the federated data landscape influenced by two key factors explored throughout this work: the Dirichlet concentration parameter α, which affects local client heterogeneity, and the imbalance ratio γ, which impacts global class imbalance. Each panel displays a 10×10 client-class matrix for CIFAR-10, where the area of the marker at position (i, c) corresponds to the number of class-c samples owned by client i.

At $\alpha = 100$, the distribution is nearly uniform, meaning every client observes all ten classes in equal proportions, and no client-class cell is left empty. Conversely, at $\alpha = 0.1$, the distribution concentrates on a limited number of classes per client. This results in roughly one-third of the cells

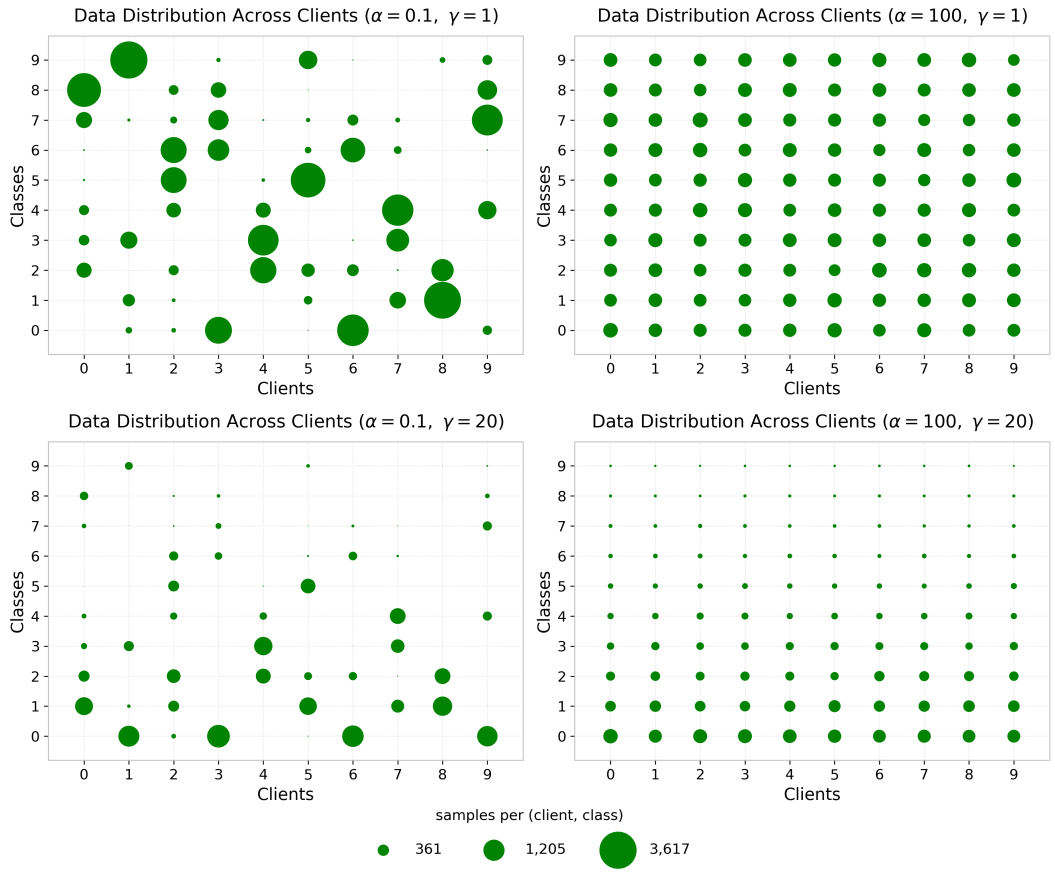


Figure 7: Visualization of data distributions on CIFAR-10 under the four (α, γ) in our experiments. These plots demonstrate how client heterogeneity and global class imbalance together influence the federated data landscape.

being empty. A single client may hold as many as $3,617$ samples of one class while having none from several others.

The rows in the matrix highlight the distinction between local class distribution and overall global imbalance. When $\gamma = 1$, class totals are uniform. However, at $\gamma = 20$, these totals decrease exponentially, causing every marker in the upper rows to shrink, regardless of the value of α. Therefore, these two factors are orthogonal yet compounding: α dictates how much each local objective deviates from the global one, while γ indicates the minimal contribution of the tail classes to either objective. The scenario of $(\alpha = 0.1, \gamma = 20)$ exemplifies a distribution where local supports are sparse amid an already tail-starved label distribution, leading to the most significant divergence between local and global loss geometry.

D Within-Phase Convergence Analysis

Why conduct a within-phase analysis? Federated active learning (FAL) generates a sequence of optimization problems that are indexed by the acquisition phase rather than solving a single, static objective. Within a given phase, the labeled sets $\{\mathcal{L}_k\}_{k=1}^K$ remain unchanged; consequently, the local objectives L_k, aggregation weights q_k, aggregate empirical loss L, and aggregate labeled prior π_{emp} also remain constant throughout the training process. When the next acquisition boundary is reached, the addition of newly labeled examples alters these quantities, prompting our protocol to reinitialize the model before optimizing the expanded labeled sets. Therefore, a single acquisition phase serves as the natural unit for optimization analysis. We establish stationarity within a fixed phase without making convergence claims across phase boundaries. Any cross-phase result would

necessitate additional assumptions that connect consecutive objectives, which are not required for the protocol being examined here.

Fixed-phase setting and scope. Condition on a single realized acquisition phase. The labeled sets, local objectives, aggregate prior, and aggregation weights remain fixed throughout optimization. We take $n_k > 0$, $q_k = n_k / \sum_i n_i$, and $L = \sum_k q_k L_k$, such that $q_k > 0$ and $\sum_k q_k = 1$.

Fix an initial parameter θ_0 with suboptimality gap $\Delta_0 = L(\theta_0) - L^\star < \infty$, expectations reflect subsequent optimization randomness. For random initializations, the stochastic statements hold with $\Delta_0 = \mathbb{E}[L(\theta_0)] - L^\star < \infty$, while the deterministic result holds conditionally for each realized instance. These guarantees do not compare performance across distinct acquired datasets or imply convergence across phase boundaries, regardless of whether subsequent phases are warm-started or reinitialized.

We analyze full client participation and $J \geq 1$ local gradient updates per client per round, for $R \geq 1$ rounds. The local learning rate $\eta > 0$ and component perturbation radius $\rho \geq 0$ remain constant within the analyzed trajectory. Let $\tau = \eta J$ be the effective round step budget. All local descent steps execute plain, unweighted gradient steps with stop-gradient perturbations. While our experimental setup incorporates weight decay, learning-rate schedule, and local epochs, where unequal dataset sizes may cause fixed local epochs to yield varying J across clients, the theoretical results examine this equal-step abstraction rather than guaranteeing convergence for every implementation variant.

Proof strategy and interpretation. We bound the perturbation, local drift, and gradient discrepancy before applying smooth descent and telescoping, following standard local-SGD/SAM analysis techniques such as `FedSAM` (Qu et al., 2022). All bounds are derived directly rather than imported from external SAM objectives or oracle assumptions.

The only method-specific input required is the pathwise perturbation bound $\|\epsilon\| \leq \rho\sqrt{2}$; neither a label shift assumption nor an unbiased target-gradient estimator is needed. The bounds establish empirical stationarity in a best-iterate or averaged sense, rather than monotone decrease, last-iterate convergence, target-risk convergence, or generalization. Because any perturbation satisfying this norm bound fulfills the proof conditions, these results establish local optimization stability rather than an explicit convergence speedup of the orthogonal direction over alternative bounded perturbations.

D.1 Assumptions and perturbation bound

Assumption 2 (Smoothness and lower boundedness)**.** Each $L_k : \mathbb{R}^P \to \mathbb{R}$ is differentiable, and there is a finite common constant $\Lambda > 0$ such that

$$\|\nabla L_k(\theta) - \nabla L_k(\theta')\| \leq \Lambda \|\theta - \theta'\|, \qquad \forall \theta, \theta'. \tag{17}$$

The aggregate empirical objective L is bounded below by $L(\theta) \geq L^\star > -\infty$ for all parameter configurations θ. In particular, L is also Λ-smooth.

Assumption 3 (Bounded client gradient dissimilarity)**.** There is a finite constant $\zeta \geq 0$ such that, for all θ,

$$\sum_{k=1}^{K} q_k \|\nabla L_k(\theta) - \nabla L(\theta)\|^2 \leq \zeta^2. \tag{18}$$

These conditions serve as formal assumptions on the analyzed objectives, rather than properties established by the experimental class-imbalance ratio or Dirichlet concentration. Specifically, the uniform dissimilarity bound acts as an additive heterogeneity assumption; class-prior mismatch alone neither guarantees it nor specifies the parameter ζ. Similarly, global smoothness represents a theoretical idealization rather than an automatic property of non-smooth neural network architectures. The theoretical proofs require these bounds to hold strictly across all iterates, comparison points, and connecting line segments evaluated in the analysis.

The following property is the only method-specific ingredient required by the convergence proofs.

Lemma 5 (Uniform perturbation bound)**.** *For every well-defined `FalOTS` perturbation, including the zero-gradient and zero-weight-sum conventions in Sec. 3,*

$$\left\|\epsilon_{k,j}^r\right\|^2 = \rho^2 \left(\left\|\widehat{g}_{k,j}^r\right\|^2 + \left\|u_{k,j}^r\right\|^2\right) \leq 2\rho^2. \tag{19}$$

The inequality holds pathwise even when the perturbation depends on the current mini-batch.

Proof. If the empirical gradient is nonzero, its normalized direction $\widehat{g}$ is a unit vector and perpendicular to u by construction ($u \perp \widehat{g}$). If the empirical gradient is zero, we set $\widehat{g} = 0$, causing the inner product $\langle \widehat{g}, u \rangle$ to vanish. In both scenarios, the norm constraints $\|\widehat{g}\|, \|u\| \in \{0, 1\}$ hold strictly. Expanding the squared norm of the combined perturbation $\|\rho\widehat{g} + \rho u\|^2$ yields:

$$\|\rho\widehat{g} + \rho u\|^2 = \rho^2 \|\widehat{g}\|^2 + 2\rho^2 \langle \widehat{g}, u \rangle + \rho^2 \|u\|^2 = \rho^2 \left(\|\widehat{g}\|^2 + \|u\|^2 \right)$$

Because $\|\widehat{g}\|^2 \leq 1$ and $\|u\|^2 \leq 1$, this identity directly implies the upper bound:

$$\|\rho\widehat{g} + \rho u\|^2 \leq 2\rho^2 \implies \|\epsilon\| \leq \rho\sqrt{2}$$

This pathwise norm bound holds strictly regardless of how the finite target-weighted gradient g_{tgt} was constructed or estimated. □

D.2 Exact local-gradient analysis

We first isolate the effect of deterministic optimization. In communication round r, all clients begin with θ_r and perform J exact local updates.

$$\begin{aligned} \theta_{k,0}^r &= \theta_r, \\ \theta_{k,j+1}^r &= \theta_{k,j}^r - \eta \nabla L_k\left(\theta_{k,j}^r + \epsilon_{k,j}^r\right), \\ \theta_{r+1} &= \sum_{k=1}^{K} q_k \theta_{k,J}^r. \end{aligned} \tag{20}$$

Recall that $\tau = \eta J$ denotes the effective local-step budget per communication round.

Theorem 6 (Exact-gradient within-phase stationarity)**.** *Under Assumptions 2 and 3, suppose all K clients participate and $0 < \tau \leq 1/(8\Lambda)$. Then the iterates in Eq. 20 satisfy*

$$\min_{0 \leq r < R} \|\nabla L(\theta_r)\|^2 \leq \frac{1}{R} \sum_{r=0}^{R-1} \|\nabla L(\theta_r)\|^2 \leq \frac{4\Delta_0}{\tau R} + 20\Lambda^2\rho^2 + 54\Lambda^2\tau^2\zeta^2. \tag{21}$$

The bound is uniform over the class-weight vectors used to construct the orthogonal directions.

For fixed ρ and τ, the bound establishes a limiting upper bound on the averaged squared gradient of order $O(\rho^2 + \tau^2\zeta^2)$, for a fixed smoothness parameter Λ. This guarantee applies to the averaged gradient magnitude across rounds rather than implying that every individual iterate enters or permanently resides within such a neighborhood. For a sequence of horizon-indexed runs on the same acquisition phase, choose $\tau_R = c_\tau R^{-1/2}$ and $\rho_R = c_\rho R^{-1/2}$, where $0 < c_\tau \leq 1/(8\Lambda)$ and $c_\rho \geq 0$ are fixed constants, and use $\eta_R = \tau_R / J$ throughout each run. Then,

$$\frac{\Delta_0}{\tau_R R} = O(R^{-1/2})$$

while the other terms are $O(R^{-1})$. Importantly, an upper bound $\tau_R = O(R^{-1/2})$ alone would not control the inverse step-size term. These horizon-indexed parameters describe a sequence of fixed-length optimization runs rather than a dynamic, within-run decaying step-size schedule.

D.2.1 Local drift

Throughout the deterministic proof, we define

$$g_r = \nabla L(\theta_r), \qquad h_{k,j}^r = \nabla L_k\left(\theta_{k,j}^r + \epsilon_{k,j}^r\right), \tag{22}$$

and we then define

$$D_{r,j} = \sum_{k=1}^{K} q_k \left\|\theta_{k,j}^r - \theta_r\right\|^2. \tag{23}$$

Lemma 7 (Local-model drift)**.** *Under Assumptions 2 and 3, the updates in Eq. 20, and* $0 < \tau = \eta J \leq 1/(8\Lambda)$*, for every round* r*,*

$$\max_{0\leq j\leq J} D_{r,j} \leq 6\tau^2 \left\|g_r\right\|^2 + 6\tau^2\zeta^2 + 12\tau^2\Lambda^2\rho^2. \tag{24}$$

Proof. By Assumption 2 and Lemma 5, we have:

$$\sum_k q_k \left\|h_{k,j}^r\right\|^2 \leq 3\sum_k q_k \left\|\nabla L_k(\theta_r)\right\|^2 + 3\Lambda^2 D_{r,j} + 6\Lambda^2\rho^2 \tag{25}$$

$$\leq 3\left\|g_r\right\|^2 + 3\zeta^2 + 3\Lambda^2 D_{r,j} + 6\Lambda^2\rho^2, \tag{26}$$

where the second line employs the weighted variance identity:

$$\sum_k q_k \left\|\nabla L_k(\theta_r)\right\|^2 = \left\|g_r\right\|^2 + \sum_k q_k \left\|\nabla L_k(\theta_r) - g_r\right\|^2. \tag{27}$$

Because $\theta_{k,j}^r - \theta_r = -\eta\sum_{s<j} h_{k,s}^r$, we can apply the Cauchy–Schwarz inequality to obtain:

$$D_{r,j} \leq \eta^2 j \sum_{s<j}\sum_k q_k \left\|h_{k,s}^r\right\|^2. \tag{28}$$

Let $D_r^{\max} = \max_{j\leq J} D_{r,j}$. Since $j \leq J$ and $\tau = \eta J$, Eqs. 26 and 28 lead to:

$$D_r^{\max} \leq \tau^2\Big(3\left\|g_r\right\|^2 + 3\zeta^2 + 6\Lambda^2\rho^2\Big) + 3\Lambda^2\tau^2 D_r^{\max}. \tag{29}$$

Under $\Lambda\tau \leq 1/8$, we have $1 - 3\Lambda^2\tau^2 \geq 61/64 > 1/2$, which confirms the validity of Eq. 24. □

D.2.2 Round-gradient approximation

We define the averaged perturbed local gradient as follows:

$$G_r = \frac{1}{J}\sum_{j=0}^{J-1}\sum_{k=1}^{K} q_k h_{k,j}^r, \qquad \delta_r = G_r - g_r. \tag{30}$$

Lemma 8 (Round-gradient approximation error)**.** *Under the hypotheses of Lemma 7,*

$$\left\|\delta_r\right\|^2 \leq 4\Lambda^2\rho^2 + 12\Lambda^2\tau^2\left\|g_r\right\|^2 + 12\Lambda^2\tau^2\zeta^2 + 24\Lambda^4\tau^2\rho^2. \tag{31}$$

Proof. Utilizing the fact that $\sum_k q_k \nabla L_k(\theta_r) = g_r$, we can decompose δ_r as follows:

$$\delta_r = \frac{1}{J}\sum_{j,k} q_k\left[\nabla L_k(\theta_{k,j}^r + \epsilon_{k,j}^r) - \nabla L_k(\theta_{k,j}^r)\right] + \frac{1}{J}\sum_{j,k} q_k\left[\nabla L_k(\theta_{k,j}^r) - \nabla L_k(\theta_r)\right]. \tag{32}$$

By applying Jensen's inequality alongside the smoothness properties and Lemma 5, we establish that the squared norms of the two averages are bounded by $2\Lambda^2\rho^2$ and $\Lambda^2 D_r^{\max}$, respectively. Consequently, we obtain:

$$\left\|\delta_r\right\|^2 \leq 4\Lambda^2\rho^2 + 2\Lambda^2 D_r^{\max}. \tag{33}$$

Substituting the results from Lemma 7 directly confirms the validity of Eq. 31. □

D.2.3 Proof of Theorem 6

Proof. By employing full participation and considering Eq. 20, we obtain:

$$\theta_{r+1} = \theta_r - \tau G_r = \theta_r - \tau(g_r + \delta_r). \tag{34}$$

Utilizing the Λ-smoothness condition, we derive:

$$L(\theta_{r+1}) \le L(\theta_r) - \tau \|g_r\|^2 - \tau \langle g_r, \delta_r \rangle + \frac{\Lambda \tau^2}{2} \|g_r + \delta_r\|^2 \tag{35}$$

$$\le L(\theta_r) - \tau \left(\frac{3}{4} - \Lambda \tau \right) \|g_r\|^2 + \tau(1 + \Lambda\tau) \|\delta_r\|^2, \tag{36}$$

where we employed the inequalities $-\langle a, b \rangle \le \frac{1}{4} \|a\|^2 + \|b\|^2$ and $\|a + b\|^2 \le 2 \|a\|^2 + 2 \|b\|^2$. The remaining coefficient of $-\tau \|g_r\|^2$ after substituting Lemma 8 is bounded below by

$$\frac{3}{4} - \Lambda\tau - 12(1 + \Lambda\tau)\Lambda^2\tau^2 \ge \frac{5}{8} - \frac{27}{128} = \frac{53}{128} \ge \frac{1}{4}.$$

Also $1 + \Lambda\tau \le 9/8$, giving

$$L(\theta_{r+1}) \le L(\theta_r) - \frac{\tau}{4} \|g_r\|^2 + \frac{9\tau}{8} \left(4\Lambda^2\rho^2 + 12\Lambda^2\tau^2\zeta^2 + 24\Lambda^4\tau^2\rho^2 \right). \tag{37}$$

Summing over $r = 0, \ldots, R-1$, applying $L(\theta_R) \ge L^\star$, and dividing by $\frac{\tau R}{4}$ yields:

$$\frac{1}{R} \sum_{r=0}^{R-1} \|g_r\|^2 \le \frac{4\Delta_0}{\tau R} + 18\Lambda^2\rho^2 + 54\Lambda^2\tau^2\zeta^2 + 108\Lambda^4\tau^2\rho^2. \tag{38}$$

Finally, since $108\Lambda^4\tau^2\rho^2 \le \frac{108}{64}\Lambda^2\rho^2$ under the condition $\Lambda\tau \le 1/8$, the perturbation terms are at most $20\Lambda^2\rho^2$. This completes the proof of Eq. 21. □

D.3 Stochastic mini-batch analysis with same-batch reuse

The update in Algorithm 1 constructs its perturbation and computes the descent gradient on the same mini-batch. As a result, the conditional unbiasedness of the perturbed stochastic gradient should not be presumed, as the perturbation is inherently data-dependent. Instead, we examine the same-batch update directly under a standard stochastic-gradient oracle at the unperturbed iterate, along with an almost-sure smoothness condition.

Let $\xi_{k,j}^r$ represent the mini-batch selected by client k at the local step (r, j), and define

$$g_k(\theta; \xi) = \nabla_\theta \ell_k(\theta; \xi) \tag{39}$$

as the corresponding mini-batch gradient.

Assumption 4 (Stochastic mini-batch oracle)**.** Let $\mathcal{F}_{r,j}$ denote the history before drawing the current mini-batches, including all current local parameters and the fixed acquisition-phase weights. The filtration includes all current draws and updates at the next local step; at a round boundary it also includes the deterministic aggregation and broadcast. Assume the gradient oracle is measurable. For every client k and every $\mathcal{F}_{r,j}$-measurable parameter θ, it holds that

$$\mathbb{E}\big[g_k(\theta; \xi_{k,j}^r) \mid \mathcal{F}_{r,j} \big] = \nabla L_k(\theta), \tag{40}$$

$$\mathbb{E}\Big[\big\| g_k(\theta; \xi_{k,j}^r) - \nabla L_k(\theta) \big\|^2 \mid \mathcal{F}_{r,j} \Big] \le \sigma^2. \tag{41}$$

Additionally, on a probability-one event the following holds simultaneously for all parameter pairs for each realized mini-batch ξ:

$$\|g_k(\theta; \xi) - g_k(\theta'; \xi)\| \le \Lambda \|\theta - \theta'\|, \qquad \forall \theta, \theta', \tag{42}$$

For $k \ne l$, define the centered noises at the current local parameters as $z_{k,j}^r$ in Eq. 44, and assume $\mathbb{E}[\big\langle z_{k,j}^r, z_{l,j}^r \big\rangle \mid \mathcal{F}_{r,j}] = 0$. Conditional independence of the client mini-batch draws is sufficient. Here $\sigma \ge 0$ is a finite bound on the variance of a mini-batch gradient, not of an individual-example gradient.

Eq. 42 presents a condition that is stronger than the smoothness of the expected local objective, yet it is sufficient for managing the data-dependent displacement arising from reusing the perturbation batch. The conditional unbiasedness statement represents the standard abstraction for mini-batches

with replacement; random reshuffling within an epoch generally does not satisfy conditional unbiasedness for the fixed full local objective and is not covered. The same oracle realization includes any stochastic augmentation or model randomness used for both gradient evaluations; reusing example indices alone need not ensure Eq. 42 applies to their difference.

The same-batch stochastic local update is expressed as follows:

$$\begin{aligned} \theta_{k,0}^r &= \theta_r, \\ \widehat{h}_{k,j}^r &= g_k\big(\theta_{k,j}^r + \epsilon_{k,j}^r; \xi_{k,j}^r\big), \\ \theta_{k,j+1}^r &= \theta_{k,j}^r - \eta \widehat{h}_{k,j}^r, \\ \theta_{r+1} &= \sum_{k=1}^K q_k \theta_{k,J}^r, \end{aligned} \tag{43}$$

where $\epsilon_{k,j}^r$ can be any measurable function of $(\theta_{k,j}^r, \xi_{k,j}^r)$ generated by `FalOTS`, and it consistently satisfies Lemma 5.

D.3.1 A SAME-BATCH DECOMPOSITION

Lemma 9 (Same-batch gradient decomposition)**.** *Under Assumption 4 and Lemma 5, define*

$$z_{k,j}^r = g_k(\theta_{k,j}^r; \xi_{k,j}^r) - \nabla L_k(\theta_{k,j}^r), \tag{44}$$

$$d_{k,j}^r = g_k(\theta_{k,j}^r + \epsilon_{k,j}^r; \xi_{k,j}^r) - g_k(\theta_{k,j}^r; \xi_{k,j}^r). \tag{45}$$

Then, we have the following decomposition:

$$\widehat{h}_{k,j}^r = \nabla L_k(\theta_{k,j}^r) + z_{k,j}^r + d_{k,j}^r, \tag{46}$$

with the properties:

$$\mathbb{E}[z_{k,j}^r \mid \mathcal{F}_{r,j}] = 0, \qquad \mathbb{E}[\left\|z_{k,j}^r\right\|^2 \mid \mathcal{F}_{r,j}] \le \sigma^2, \qquad \left\|d_{k,j}^r\right\|^2 \le 2\Lambda^2\rho^2. \tag{47}$$

Note that no independence assumption between $d_{k,j}^r$ and $z_{k,j}^r$ is required.

Proof. The decomposition is derived algebraically. The first two properties in Eq. 47 follow from Assumption 4. Regarding the last property, since the same realized mini-batch is involved on both sides of Eq. 45, we can apply Eq. 42 and Lemma 5 to obtain:

$$\left\|d_{k,j}^r\right\| \le \Lambda \left\|\epsilon_{k,j}^r\right\| \le \sqrt{2}\Lambda\rho. \tag{48}$$

□

D.3.2 CLIENT DISAGREEMENT

We define the virtual weighted average as follows:

$$\bar{\theta}_{r,j} = \sum_{k=1}^K q_k \theta_{k,j}^r, \qquad q_2 = \sum_{k=1}^K q_k^2, \tag{49}$$

ensuring that $\bar{\theta}_{r,0} = \theta_r$ and $\bar{\theta}_{r,J} = \theta_{r+1}$. Next, we introduce the pointwise client disagreement and its expectation, defined by

$$\mathcal{D}_{r,j} = \sum_{k=1}^K q_k \left\|\theta_{k,j}^r - \bar{\theta}_{r,j}\right\|^2, \qquad \bar{D}_{r,j} = \mathbb{E}[\mathcal{D}_{r,j}]. \tag{50}$$

Lemma 10 (Stochastic local disagreement)**.** *Under Assumptions 2, 3, and 4, for Eq. 43, if $\tau = \eta J \le 1/(8\Lambda)$, then for every round r,*

$$\max_{0 \le j \le J} \bar{D}_{r,j} \le \frac{64}{15}\tau^2(\zeta^2 + \sigma^2) + \frac{128}{15}\Lambda^2\tau^2\rho^2. \tag{51}$$

Proof. Let $\bar{h}_{r,j} = \sum_k q_k \widehat{h}^r_{k,j}$. The weighted variance is minimized at the weighted mean. By conditioning on $\mathcal{F}_{r,j}$ and selecting the comparison vector $\nabla L(\bar{\theta}_{r,j})$, we derive:

$$\sum_k q_k \mathbb{E}\left[\left\|\widehat{h}^r_{k,j} - \bar{h}_{r,j}\right\|^2 \mid \mathcal{F}_{r,j}\right] \le \sum_k q_k \mathbb{E}\left[\left\|\widehat{h}^r_{k,j} - \nabla L(\bar{\theta}_{r,j})\right\|^2 \mid \mathcal{F}_{r,j}\right] \tag{52}$$

$$\le 4\Lambda^2 \mathcal{D}_{r,j} + 4\zeta^2 + 8\Lambda^2\rho^2 + 4\sigma^2. \tag{53}$$

The last line follows from Lemma 9, the four-term inequality

$$\|a+b+c+d\|^2 \le 4(\|a\|^2 + \|b\|^2 + \|c\|^2 + \|d\|^2),$$

along with smoothness and Assumption 3 evaluated at $\bar{\theta}_{r,j}$.

Since all clients initiate the round from the same model, it follows that

$$\theta^r_{k,j} - \bar{\theta}_{r,j} = -\eta \sum_{s<j} \left(\widehat{h}^r_{k,s} - \bar{h}_{r,s}\right). \tag{54}$$

By applying Cauchy–Schwarz and taking the expectation in Eq. 53, we can infer, with the notation $\bar{D}^{\max}_r = \max_{j\le J} \bar{D}_{r,j}$, that

$$\bar{D}^{\max}_r \le \tau^2 \left(4\Lambda^2 \bar{D}^{\max}_r + 4\zeta^2 + 8\Lambda^2\rho^2 + 4\sigma^2\right). \tag{55}$$

Given that $\Lambda\tau \le 1/8$, it results in $1 - 4\Lambda^2\tau^2 \ge 15/16$, which leads to Eq. 51. □

D.3.3 Stochastic stationarity theorem

Theorem 11 (Same-batch stochastic within-phase stationarity)**.** *Under Assumptions 2, 3, and 4, suppose all clients participate and* $0 < \tau = \eta J \le 1/(8\Lambda)$*. Then the virtual averages generated by Eq. 43 satisfy*

$$\min_{\substack{0\le r<R\\ 0\le j<J}} \mathbb{E}\left[\left\|\nabla L(\bar{\theta}_{r,j})\right\|^2\right] \le \frac{1}{RJ}\sum_{r=0}^{R-1}\sum_{j=0}^{J-1} \mathbb{E}\left[\left\|\nabla L(\bar{\theta}_{r,j})\right\|^2\right]$$

$$\le \frac{2\Delta_0}{\eta RJ} + 10\Lambda^2\rho^2 + 20\Lambda^2\tau^2(\zeta^2+\sigma^2) + 4\Lambda\eta q_2\sigma^2. \tag{56}$$

Proof. Fix (r, j) and condition on $\mathcal{F}_{r,j}$. Let

$$g_{r,j} = \nabla L(\bar{\theta}_{r,j}), \tag{57}$$

$$a_{r,j} = \sum_k q_k \left[\nabla L_k(\theta^r_{k,j}) - \nabla L_k(\bar{\theta}_{r,j})\right], \tag{58}$$

$$p_{r,j} = \sum_k q_k d^r_{k,j}, \qquad \nu_{r,j} = \sum_k q_k z^r_{k,j}. \tag{59}$$

Then the virtual update is

$$\bar{\theta}_{r,j+1} = \bar{\theta}_{r,j} - \eta\left(g_{r,j} + a_{r,j} + p_{r,j} + \nu_{r,j}\right). \tag{60}$$

According to Jensen's inequality, along with Lemma 9 and the principle of conditional uncorrelatedness among clients,

$$\|a_{r,j}\|^2 \le \Lambda^2 \mathcal{D}_{r,j}, \qquad \|p_{r,j}\|^2 \le 2\Lambda^2\rho^2, \qquad \mathbb{E}[\nu_{r,j} \mid \mathcal{F}_{r,j}] = 0, \qquad \mathbb{E}[\|\nu_{r,j}\|^2 \mid \mathcal{F}_{r,j}] \le q_2\sigma^2. \tag{61}$$

The first inequality is pathwise, with the random disagreement $\mathcal{D}_{r,j}$; only after taking total expectations do we obtain $\mathbb{E}\|a_{r,j}\|^2 \le \Lambda^2 \bar{D}_{r,j}$. Also, $p_{r,j}$ and $\nu_{r,j}$ may be dependent; therefore, the proof presented below does not make an independence assumption to eliminate their cross term.

By utilizing the Λ-smoothness of L, we take the conditional expectation and apply the inequality $-\langle x, y\rangle \le \frac{1}{8}\|x\|^2 + 2\|y\|^2$ separately to $a_{r,j}$ and $p_{r,j}$, and using $\|x_1+x_2+x_3+x_4\|^2 \le$

$4\sum_{i=1}^{4}\|x_i\|^2$, we first obtain the coefficients $-\eta(3/4-2\Lambda\eta)$ for $\|g_{r,j}\|^2$, $\eta(2+2\Lambda\eta)$ for each displacement squared norm, and $2\Lambda\eta^2$ for the noise second moment. Thus

$$\mathbb{E}\big[L(\bar{\theta}_{r,j+1}) \mid \mathcal{F}_{r,j}\big] \le L(\bar{\theta}_{r,j}) - \frac{\eta}{2}\|g_{r,j}\|^2 + \frac{9\eta}{4}\Big(\|a_{r,j}\|^2 + \mathbb{E}[\|p_{r,j}\|^2 \mid \mathcal{F}_{r,j}]\Big) + 2\Lambda\eta^2 q_2\sigma^2, \tag{62}$$

where $\eta\Lambda \le \tau\Lambda \le 1/8$ was used in the last step. By substituting Eq. 61, taking the total expectation, and summing over all RJ local steps, we achieve a telescoping effect, since $\bar{\theta}_{r,J} = \bar{\theta}_{r+1,0}$. Therefore,

$$\frac{1}{RJ}\sum_{r,j}\mathbb{E}\Big[\|g_{r,j}\|^2\Big] \le \frac{2\Delta_0}{\eta RJ} + 9\Lambda^2\rho^2 + \frac{9}{2}\Lambda^2\max_{r,j}\bar{D}_{r,j} + 4\Lambda\eta q_2\sigma^2. \tag{63}$$

Using Lemma 10, we have

$$\frac{9}{2}\Lambda^2\max_{r,j}\bar{D}_{r,j} \le \frac{96}{5}\Lambda^2\tau^2(\zeta^2+\sigma^2) + \frac{192}{5}\Lambda^4\tau^2\rho^2. \tag{64}$$

Finally, $\Lambda^2\tau^2 \le 1/64$ implies $(192/5)\Lambda^4\tau^2\rho^2 \le (3/5)\Lambda^2\rho^2$. Rounding $9+3/5$ to 10 and $96/5$ to 20 gives Eq. 56. □

Rate and interpretation. For fixed J, take $\eta_R = c_\eta R^{-1/2}$ and $\rho_R = c_\rho R^{-1/2}$ in horizon-indexed runs, with $0 < c_\eta J \le 1/(8\Lambda)$ and $c_\rho \ge 0$. If the phase, initial gap, smoothness, and variance constants are fixed, the right-hand side is $O(R^{-1/2})$. This concerns squared-gradient stationarity, not a function-value or generalization rate. With fixed η, ρ, it is an averaged neighborhood bound. For equal client weights $q_2 = 1/K$; otherwise $q_2 = \sum_k q_k^2$. The factor q_2 reduces the centered aggregation-noise term, not the possibly biased same-batch displacement. The drift estimate is conservative and is not a claim of optimal dependence on J; for $J = 1$, the disagreement at every evaluated virtual iterate is exactly zero.

D.3.4 Stationarity of communicated iterates

The virtual iterates for $j > 0$ are analytical averages, not models actually communicated during a round. The following corollary makes the connection to the server models explicit.

Corollary 12 (Communicated-iterate stationarity)**.** *Under the hypotheses of Theorem 11,*

$$\min_{0\le r<R}\mathbb{E}\,\|\nabla L(\theta_r)\|^2 \le \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\,\|\nabla L(\theta_r)\|^2 \le \frac{5\Delta_0}{\tau R} + 24\Lambda^2\rho^2 + 56\Lambda^2\tau^2(\zeta^2+\sigma^2) + 10\Lambda\eta q_2\sigma^2. \tag{65}$$

Proof. Define $B_{r,j} = \sum_k q_k\mathbb{E}\left\|\theta_{k,j}^r - \theta_r\right\|^2$ and $B_r^{\max} = \max_{j\le J} B_{r,j}$. Decompose each $\widehat{h}_{k,j}^r$ into $\nabla L_k(\theta_r)$, $\nabla L_k(\theta_{k,j}^r) - \nabla L_k(\theta_r)$, $z_{k,j}^r$, and $d_{k,j}^r$. The four-term squared-norm inequality, the weighted variance identity, and Lemma 9 give

$$\sum_k q_k\mathbb{E}\left\|\widehat{h}_{k,j}^r\right\|^2 \le 4\mathbb{E}\,\|\nabla L(\theta_r)\|^2 + 4\zeta^2 + 4\Lambda^2 B_{r,j} + 4\sigma^2 + 8\Lambda^2\rho^2.$$

Since $\theta_{k,j}^r - \theta_r = -\eta\sum_{s<j}\widehat{h}_{k,s}^r$, Cauchy–Schwarz and $\eta j \le \tau$ imply

$$B_r^{\max} \le \frac{64}{15}\tau^2\big(\mathbb{E}\,\|\nabla L(\theta_r)\|^2 + \zeta^2 + \sigma^2 + 2\Lambda^2\rho^2\big),$$

where $1-4\Lambda^2\tau^2 \ge 15/16$ was used. Jensen's inequality gives $\mathbb{E}\left\|\bar{\theta}_{r,j} - \theta_r\right\|^2 \le B_r^{\max}$. Smoothness therefore yields, for every j,

$$\mathbb{E}\,\|\nabla L(\theta_r)\|^2 \le 2\mathbb{E}\left\|\nabla L(\bar{\theta}_{r,j})\right\|^2 + 2\Lambda^2 B_r^{\max}.$$

Absorbing the coefficient $(128/15)\Lambda^2\tau^2 \le 2/15$ and averaging over r, j gives

$$\frac{1}{R}\sum_r \mathbb{E}\,\|\nabla L(\theta_r)\|^2 \le \frac{30}{13RJ}\sum_{r,j}\mathbb{E}\,\big\|\nabla L(\bar\theta_{r,j})\big\|^2 + \frac{128}{13}\Lambda^2\tau^2(\zeta^2+\sigma^2+2\Lambda^2\rho^2).$$

Substituting Eq. 56 gives coefficients $60/13$ for $\Delta_0/(\tau R)$, at most $304/13$ for $\Lambda^2\rho^2$, $728/13 = 56$ for $\Lambda^2\tau^2(\zeta^2+\sigma^2)$, and $120/13$ for $\Lambda\eta q_2\sigma^2$. Rounding upward proves the claim. □

For $J = 1$, $\bar\theta_{r,0} = \theta_r$, so the sharper direct consequence of Eq. 63 is $R^{-1}\sum_r \mathbb{E}\,\|\nabla L(\theta_r)\|^2 \le 2\Delta_0/(\eta R) + 9\Lambda^2\rho^2 + 4\Lambda\eta q_2\sigma^2$, without a local-drift term. For fixed $J > 1$, the corollary has the same $O(R^{-1/2})$ horizon-indexed scaling specified above. Neither result requires continuity or differentiability of the map $\theta \mapsto u$: the possibly discontinuous normalization enters only through a bounded parameter displacement.

E Complete Proof of Propositions in the Main Paper

E.1 Proof of Proposition 1

Proof. Let $\mathcal{Y}_{\mathrm{tgt}} = \{y : \pi_{\mathrm{tgt}}(y) > 0\}$. All its elements have positive source mass by support overlap. Source classes outside this set receive zero weight. Thus

$$\begin{aligned}\mathbb{E}_{\mathrm{emp}}[w(y)\nabla_\theta\ell(\theta;x,y)] &= \sum_{y\in\mathcal{Y}_{\mathrm{tgt}}}\pi_{\mathrm{emp}}(y)\frac{\pi_{\mathrm{tgt}}(y)}{\pi_{\mathrm{emp}}(y)}\mathbb{E}_{P_{\mathrm{emp}}(\cdot|y)}[\nabla_\theta\ell(\theta;x,y)]\\ &= \sum_{y\in\mathcal{Y}_{\mathrm{tgt}}}\pi_{\mathrm{tgt}}(y)\mathbb{E}_{P_{\mathrm{tgt}}(\cdot|y)}[\nabla_\theta\ell(\theta;x,y)]\\ &= \nabla_\theta\sum_{y\in\mathcal{Y}_{\mathrm{tgt}}}\pi_{\mathrm{tgt}}(y)\mathbb{E}_{P_{\mathrm{tgt}}(\cdot|y)}[\ell(\theta;x,y)] = \nabla L_{\mathrm{tgt}}(\theta). \qquad (66)\end{aligned}$$

The second equality uses equality of the class-conditionals. The third uses the stated local integrable gradient envelopes to differentiate under each expectation; the sum over classes is finite and its weights are independent of θ. □

Population identity versus a finite labeled set. For a realized aggregate labeled dataset with $N_y > 0$ for every target class, let $N = \sum_y N_y$ and $\pi_{\mathrm{emp}}(y) = N_y/N$. Define the class loss by

$$\widehat{L}_y(\theta) = \frac{1}{N_y}\sum_{i:y_i=y}\ell(\theta;x_i,y_i).$$

Exact histogram weighting then gives the algebraic identity

$$\frac{1}{N}\sum_i w(y_i)\ell(\theta;x_i,y_i) = \sum_{y\in\mathcal{Y}_{\mathrm{tgt}}}\pi_{\mathrm{tgt}}(y)\widehat{L}_y(\theta).$$

The total weight is $\sum_i w(y_i) = N$, so full-dataset self-normalization gives the same expression. This is an empirical class-balanced or target-prior-weighted risk. Equality with the deployment population risk does not follow from histogram weighting.

Self-normalization and aggregation. On a fixed batch with $S_w > 0$, $g_{\mathrm{tgt}} = S_w^{-1}\sum_i w(y_i)\nabla_\theta\ell_i$. For any fixed empirical direction and any $a > 0$, the linear projector $P_\perp = I - \widehat{g}\widehat{g}^\top$ satisfies $P_\perp(ag_{\mathrm{tgt}}) = aP_\perp g_{\mathrm{tgt}}$. Its nonzero normalized residual is therefore unchanged, and a zero residual stays zero. This argument also holds when $\widehat{g} = 0$, since then $P_\perp = I$. It is a per-batch identity, not an unbiasedness statement for a random ratio. In particular, $\mathbb{E}[\sum_i w_i\nabla\ell_i/\sum_i w_i]$ generally differs from the ratio of the two expectations. Likewise, nonlinear normalization and projection of client gradients do not commute with client aggregation. The population identity is obtained by aggregating the unnormalized weighted expectations at a common θ.

Effect of within-class selection bias. Under support overlap and finite class-conditional gradient means, write $\mu_{\mathrm{emp},y} = \mathbb{E}_{P_{\mathrm{emp}}(\cdot|y)}\nabla\ell$ and $\mu_{\mathrm{tgt},y} = \mathbb{E}_{P_{\mathrm{tgt}}(\cdot|y)}\nabla\ell$. Without label shift, exact ratios give the bias identity

$$\mathbb{E}_{\mathrm{emp}}[w(y)\nabla\ell] - \nabla L_{\mathrm{tgt}} = \sum_{y\in\mathcal{Y}_{\mathrm{tgt}}} \pi_{\mathrm{tgt}}(y)(\mu_{\mathrm{emp},y} - \mu_{\mathrm{tgt},y}). \tag{67}$$

This follows by expanding both expectations class by class. Its norm is at most

$$\sum_{y\in\mathcal{Y}_{\mathrm{tgt}}} \pi_{\mathrm{tgt}}(y)\left\|\mu_{\mathrm{emp},y} - \mu_{\mathrm{tgt},y}\right\|,$$

which quantifies the remaining within-class acquisition bias when label shift does not hold. Under Assumption 1, the class-conditional gradient means agree for every target class, so the discrepancy vanishes and Proposition 1 follows. This population statement does not establish finite-batch unbiasedness or control direction error after normalizing a nearly vanishing residual.

E.2 Proof of Proposition 2

Proof. If $g_{\mathrm{emp}} \neq 0$, $\|\widehat{g}\| = 1$ and $P_\perp = I - \widehat{g}\widehat{g}^\top$ is the orthogonal projector onto $\{d : \langle g_{\mathrm{emp}}, d\rangle = 0\}$. If $g_{\mathrm{emp}} = 0$, the same formula gives $P_\perp = I$, the projector onto all of $\mathbb{R}^P$, which is the feasible neutral subspace in this case. In either case $v = P_\perp g_{\mathrm{tgt}}$, and when $v \neq 0$ the vector $u = v/\|v\|$ is feasible. For any feasible unit vector d,

$$\langle g_{\mathrm{tgt}}, d\rangle = \langle P_\perp g_{\mathrm{tgt}}, d\rangle = \langle v, d\rangle \leq \|v\|\,\|d\| = \|v\|.$$

The vector u attains the upper bound. Equality with the positive value $\|v\|$ requires d to be a positive scalar multiple of v; unit norm then forces $d = u$. This proves uniqueness for $v \neq 0$. If $v = 0$, every feasible unit vector has value zero; we select the zero correction instead. In dimension one with $g_{\mathrm{emp}} \neq 0$, the unit feasible set is empty, but then $v = 0$ and the proposition makes no unit-optimizer claim.

Finally, for sufficiently small $|t|$ with the segment in a region of M_L-Lipschitz gradient,

$$|L(\theta + tu) - L(\theta) - t\,\langle\nabla L(\theta), u\rangle| \leq \frac{M_L}{2}t^2\,\|u\|^2.$$

The linear term is zero and $\|u\| \leq 1$, proving the stated bound. The direction is held fixed at the base point for this expansion. □

E.3 Proof of Proposition 3

Proof. Part (i). If $g_{\mathrm{emp}} \neq 0$, then $\langle g_{\mathrm{emp}}, \widehat{g}\rangle = \|g_{\mathrm{emp}}\|$ and $\langle g_{\mathrm{emp}}, u\rangle = 0$. Consequently,

$$\langle\nabla L, \epsilon_{\texttt{FalOTS}}\rangle = \rho\,\langle g_{\mathrm{emp}}, \widehat{g}\rangle + \rho\,\langle g_{\mathrm{emp}}, u\rangle = \rho\,\|g_{\mathrm{emp}}\| = \langle\nabla L, \epsilon_{\mathrm{SAM}}\rangle.$$

If $g_{\mathrm{emp}} = 0$, both sides are zero, irrespective of u. This equality preserves the linearized increase, not necessarily the actual nonlinear loss at the perturbed points. For example, local M_L-smoothness bounds the difference between either loss increase and its common linear term by $(M_L/2)\,\|\epsilon\|^2$.

Part (ii). Under matched priors, every observed class has $w(y) = 1$. When both gradients use the same dataset or realized batch, Eq. 7 becomes its unweighted objective, and $g_{\mathrm{tgt}} = g_{\mathrm{emp}}$. If $g_{\mathrm{emp}} \neq 0$, then $v = g_{\mathrm{emp}} - \|g_{\mathrm{emp}}\|\,\widehat{g} = 0$; if $g_{\mathrm{emp}} = 0$, the conventions also give $v = g_{\mathrm{tgt}} = 0$. Thus $u = 0$ and both methods use $\epsilon = \rho\widehat{g}$, including the zero-gradient case. Their one-step updates coincide. Their trajectories coincide by induction only when initialization, optimizer states, oracle realizations, aggregation, and all subsequent choices are identical and the matched-prior condition continues to hold at every step. A balanced pool alone is insufficient to ensure this condition after acquisition.

Part (iii). Since $\langle\widehat{g}, u\rangle = 0$,

$$\|\rho\widehat{g} + \rho u\|^2 = \rho^2(\|\widehat{g}\|^2 + \|u\|^2) \leq 2\rho^2.$$

For nonzero empirical gradient, $\|\widehat{g}\| = 1$ and $\|u\| \in \{0, 1\}$, giving the two stated cases. For zero empirical gradient, $\widehat{g} = 0$, giving $\|\epsilon\| = \rho\,\|u\|$. □

E.4 Proof of Proposition 4

Proof. On the convex neighborhood in the proposition, let $H = \nabla^2 L(\theta)$. The integral Taylor formula gives, for a displacement e whose segment stays in this neighborhood,

$$\nabla L(\theta + e) = \nabla L(\theta) + He + R(e), \qquad \|R(e)\| \le \frac{M_H}{2} \|e\|^2 . \tag{68}$$

Indeed, $R(e) = \int_0^1 [\nabla^2 L(\theta + te) - H]e \, dt$, whose norm is bounded by $\int_0^1 M_H t \|e\|^2 \, dt$. Set $e_F = \rho\widehat{g} + \rho u$ and $e_S = \rho\widehat{g}$. The norm bounds $\|e_F\|^2 \le 2\rho^2$ and $\|e_S\|^2 \le \rho^2$ remain valid at zero gradients. Subtracting the two updates gives

$$\theta^+_{\texttt{FalOTS}} - \theta^+_{\mathrm{SAM}} = -\eta H(e_F - e_S) - \eta[R(e_F) - R(e_S)] = -\eta\rho H u + r_\rho,$$

where $\|r_\rho\| \le \eta(M_H/2)(2\rho^2 + \rho^2) = \frac{3}{2} M_H \eta \rho^2$. No derivative of the perturbation map is taken. If $u = 0$, the two updates are exactly equal, not merely equal up to a remainder. □

E.5 Derivation of Eq. 15

Write $x_F = \theta^+_{\texttt{FalOTS}}$, $x_S = \theta^+_{\mathrm{SAM}}$, and $d = x_F - x_S$. Both displacements from θ have norm at most ηB_L. The fundamental theorem of calculus along the segment joining x_S and x_F gives

$$L_{\mathrm{tgt}}(x_F) - L_{\mathrm{tgt}}(x_S) - \langle r, d\rangle = \int_0^1 \langle \nabla L_{\mathrm{tgt}}(x_S + td) - r, d\rangle \, dt.$$

Every point on this segment is within ηB_L of θ by convexity of the norm. Also, smoothness of L along the perturbation segment implies

$$\|d\| = \eta \|\nabla L(\theta + e_F) - \nabla L(\theta + e_S)\| \le \eta M_L \rho \|u\| \le \eta M_L \rho.$$

Thus the absolute integral remainder is at most $M_{\mathrm{tgt}} B_L M_L \eta^2 \rho$. Substituting $d = -\eta\rho H u + r_\rho$ and $|\langle r, r_\rho\rangle| \le \frac{3}{2} M_H \|r\| \eta\rho^2$ proves Eq. 15. The constants can be chosen on fixed compact neighborhoods for all sufficiently small η, ρ. Unlike separate $O(\eta^2)$ Taylor remainders, this comparison retains the factor ρ, so its sign implication is valid as both η and ρ tend to zero independently. The equality $\langle r, Hu\rangle = \langle Hr, u\rangle$ follows from the symmetry of the Hessian of a C^2 scalar objective.

Why positive curvature alone does not imply an improvement. At $\theta = 0$, consider two smooth quadratic objectives with the same positive-definite Hessian and gradients

$$H = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \qquad g_{\mathrm{emp}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad r = g_{\mathrm{tgt}} = \begin{pmatrix} -3 \\ 1 \end{pmatrix}.$$

These are realized by $L(z) = c_L + \langle g_{\mathrm{emp}}, z\rangle + \frac{1}{2} z^\top H z$ and $L_{\mathrm{tgt}}(z) = c_T + \langle r, z\rangle + \frac{1}{2} z^\top H z$; finite constants c_L, c_T can make both nonnegative. Here $\widehat{g} = (1, 0)^\top$, $u = (0, 1)^\top$, $\langle r, u\rangle = 1 > 0$, but $A_\perp(u) = \langle r, Hu\rangle = -1 < 0$. Thus the target-ascent optimality of u, even with an exact target gradient and a positive-definite H, does not guarantee a favorable target-risk comparison after the empirical descent step.

One local step followed by aggregation. If all clients start at the same θ, use one exact local step, and satisfy the Hessian condition with a common constant M_H, then applying Proposition 4 to each L_k and averaging gives

$$\theta^+_{F,\mathrm{fed}} - \theta^+_{S,\mathrm{fed}} = -\eta\rho \sum_k q_k H_k u_k + e, \qquad \|e\| \le \tfrac{3}{2} M_H \eta \rho^2.$$

The leading target-risk comparison involves $\langle r, \sum_k q_k H_k u_k\rangle$. An aggregate-gradient step instead involves $\langle r, Hu\rangle$. These expressions are generally different. The within-phase analysis treats multi-step federated dynamics separately and does not rely on equating these expressions.

F Detailed Experimental Setups & Additional Results

F.1 Environments

The code is implemented in PyTorch 2.9.1, and the overall code structure is based on the FairFAL (Zong & Huang, 2026) and SAM (Foret et al., 2021) official implementations. All experiments are conducted on a Linux (Ubuntu-20.04.6 LTS) server with NVIDIA GeForce RTX 4090 24GB GPUs.

F.2 Training Details

All experiments are conducted under the standard `FedAvg` framework (McMahan et al., 2017). We further conduct experiments with `FedProx` (Li et al., 2020a) and `SCAFFOLD` (Karimireddy et al., 2020), which were additionally examined in the ablation study in Appx. I.1. For the FL training pipeline, we set the number of FL rounds to 100, with each client performing 5 local epochs per round. We set the total number of clients to 10, while we conduct ablations for a varying number of clients in Appx. I.5. For the AL training pipeline, the FAL process involves 9 query cycles. In the first cycle, each client labels 5% of its local data randomly. In each subsequent cycle, another 5% is queried from the client's unlabeled data. We set the global imbalance ratio to $\gamma = 20$ for all experiments and conduct ablations for $\gamma \in \{1, 10, 20, 50, 80, 100\}$, as presented in Appx. 5.2. The data is distributed among 10 clients using a Dirichlet distribution, with $\alpha = 0.1$ for highly heterogeneous splits and $\alpha = 100$ for nearly homogeneous client distributions. We use FairFAL's query strategy as our baseline; training with SGD instead of `FalOTS` algorithm achieves identical performance to `FairFAL`. All SAM-based algorithms use the same FAL framework for a fair comparison. We also conduct experiments with recent FAL query strategies from `KAFAL`, `LoGo` and `IFAL` as ablations in Appx. I.2. We run all main experiments with five random seeds. The mean results and standard deviations are shown in Fig. 2.

For the image classification, we strictly follow recent studies (Kim et al., 2023a; Zong et al., 2025; Zong & Huang, 2026), which utilize a 4-layer CNN backbone and optimize it using Stochastic Gradient Descent (SGD) with a momentum of 0.9, a weight decay of 1×10^{-5}, and a batch size of 64. We further evaluate our method on more diverse backbone architectures, including MobileNet (Howard et al., 2017) and ResNet-18 (He et al., 2016), as in Appx. I.3. The learning rate is set to 0.01 and is reduced by a factor of 10 after round 75.

For text classification, we adopt the settings used in image classification, with the following distinctions. We utilize the pretrained `DistilBERT-base-uncased`[4] model as our text encoder, attaching a task-specific classification head with 4 output classes for AG News and 14 output classes for DBpedia-14. Each document is tokenized with the corresponding DistilBERT tokenizer, truncated or padded to a maximum sequence length of 128 tokens, and represented by its token IDs and attention mask. We utilize the complete official training and testing splits for both datasets. Federated training is conducted using AdamW as the base optimizer (noting that SAM-based approaches are also grounded in AdamW for text classification), with a learning rate of 2×10^{-5}, a weight decay of 0.01, and no learning rate decay. Each client processes 1 local epoch per communication round. We carry out 5 communication rounds per acquisition phase for both datasets, taking advantage of the pretrained DistilBERT model rather than training from scratch. Furthermore, since pretrained transformers generally require smaller perturbations compared to models trained from scratch, we tune the SAM radius over $\rho \in \{0.005, 0.01, 0.02, 0.05, 0.1\}$. We found that the optimal ρ for SAM-based algorithms is typically half of the optimal ρ specified for each method in Tab. 7.

F.3 Dataset Details

- **Fashion-MNIST (FMNIST)** (Xiao et al., 2017) is a grayscale image dataset featuring fashion items. It consists of 60,000 training images and 10,000 test images, all with a resolution of 28×28 pixels. The dataset includes 10 clothing categories and is commonly used as a lightweight alternative to MNIST in image classification benchmarks.
- **CIFAR-10** (Krizhevsky et al., 2009) comprises 60,000 natural images, each measuring 32×32 pixels and featuring three color channels. The dataset is divided into 50,000 training

[4] `https://huggingface.co/distilbert/distilbert-base-uncased`

samples and 10,000 test samples and encompasses 10 object classes. It serves as a standard benchmark for supervised and active learning.

- **CIFAR-100** (Krizhevsky et al., 2009) has same image resolution and training/testing split as CIFAR-10 but includes 100 fine-grained categories. Each category contains 500 training images and 100 test images. In comparison to CIFAR-10, this task poses a greater challenge for multi-class classification due to its larger label space and increased intra-class variability.
- **Tiny ImageNet** (Le et al., 2015) is derived from ImageNet, containing 200 object categories. Each category has 500 training images, 50 validation images, and 50 test images, with all images resized to 64×64 pixels. In total, the dataset consists of 100,000 training images, 10,000 validation images, and 10,000 test images. Due to its larger number of classes and greater visual diversity compared with CIFAR benchmarks, TinyImageNet provides a more challenging setting for evaluating FAL performance.
- **OCTMNIST** (Yang et al., 2023) is a medical imaging dataset derived from retinal optical coherence tomography (OCT) scans used for diagnosing retinal diseases. It has 109,309 grayscale images and is framed as a four-class classification task. All images are center-cropped and resized to $1 \times 28 \times 28$ pixels. Following MedMNIST protocol, the dataset is divided into training, validation, and test sets. Its label distribution reflects a naturally long-tailed imbalance, mirroring the real-world prevalence of various retinal conditions.
- **DermaMNIST** (Yang et al., 2023) is created from the HAM10000 collection of dermatoscopic skin lesion images. This dataset contains 10,015 color images across seven diagnostic categories, resized to $3 \times 28 \times 28$ pixels. The images are split into training, validation, and test sets using a 7:1:2 ratio. DermaMNIST exhibits a significant long-tailed distribution, making it an important benchmark for evaluating models.
- **AGNews** (Zhang et al., 2015) includes 120,000 training articles and 7,600 test articles. It consists of 4 categories: World, Sports, Business, and Science/Technology, each containing 30,000 training examples and 1,900 test examples. Each entry comprises a news title and a brief description, making this dataset a widely recognized benchmark for text classification.
- **DBpedia-14** (Zhang et al., 2015) is a large-scale ontology classification dataset derived from DBpedia. It contains 560,000 training examples and 70,000 test examples distributed evenly across 14 semantic classes, with 40,000 training and 5,000 test samples per class. The classes represent entity types such as Company, Educational Institution, Artist, Athlete, and Natural Place, and each example includes an entity title and a short textual description.

F.4 Experimental Categories

We evaluate `FalOTS` along four experimental dimensions:

1. *Heterogeneity level*: We evaluate the heterogeneity level of varying degrees of class imbalance using a Latent Dirichlet Allocation (LDA) (Li et al., 2021) strategy. For instance, a smaller α indicates a more heterogeneous data distribution.
2. *Global class imbalance ratio* of used datasets: We use $\gamma = 1$ and $\gamma = 20$ but mainly focus on severe global class imbalance ratio ($\gamma = 20$).
3. *Model architecture*: We use a 4-layer convolutional neural network as our base architecture. We also conduct experiments with ResNet-18 (He et al., 2016) and MobileNet (Howard et al., 2017) backbones.
4. *Experimental scope*: In contrast to existing FAL methods (Cao et al., 2023; Kim et al., 2023a; Zong et al., 2025; Zong & Huang, 2026), which focus solely on the image classification task, we broaden the experimental scope to include a different modality (language) with **text classification**. This allows us to better assess the robustness and generalization of our proposed `FalOTS`, demonstrating its ability to adapt and generalize to other modalities as well.

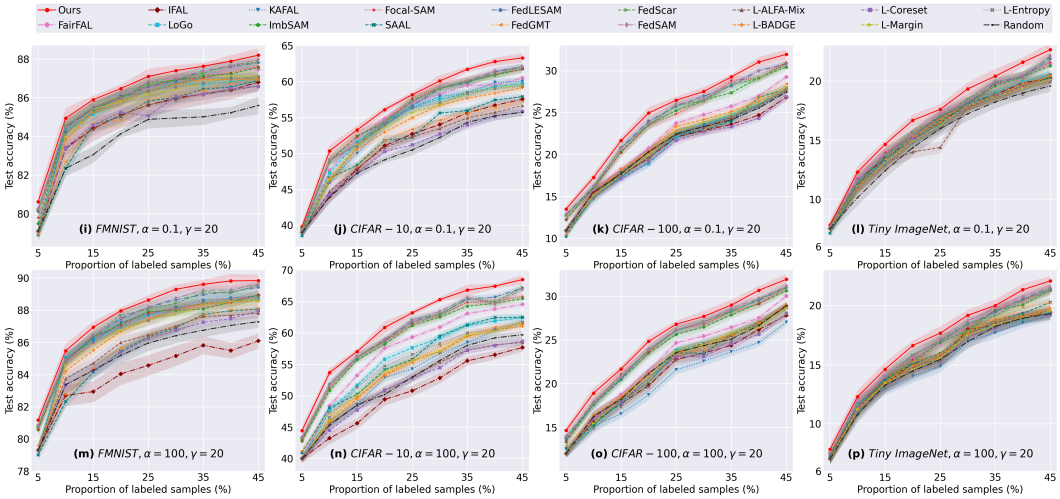


Figure 8: Test accuracy vs. proportion of labeled samples on FMNIST, CIFAR-10/100 and Tiny ImageNet under global imbalance ($\gamma = 20$) and two heterogeneity levels ($\alpha \in \{0.1, 100\}$). Results are averaged over five seeds. L denotes the local query-selector model. Numerical results appear in Tabs. 4– 5 in Appx. F.5.2.

F.5 Additional Results

F.5.1 Local Query-Selector Model Performance

Fig. 8 shows the performance comparison on FMNIST, CIFAR-10/100 and Tiny ImageNet under severe global imbalance $\gamma = 20$ and two heterogeneity levels ($\alpha = 0.1$ and $\alpha = 100$) using local query-selector model.

F.5.2 Numerical Results of Figures 2, 3, and 8

In relation to Figs. 2, 3 and 8, Tabs 4–6 present the complete numerical results for various labeled data ratios. `FalOTS` outperforms FAL baselines using SGD- and SAM-based optimizers across diverse benchmarks, and its advantages remain evident as dataset complexity increases on both image and text modalities.

F.6 FairFAL and SAM-based Baselines Implementation

FairFAL implementation. Since our experiment utilizes FairFAL (Zong & Huang, 2026) as the baseline, we detail the hyperparameters employed to reproduce the FairFAL baseline. We strictly follow the original paper and implementation of FairFAL. The model uses the SGD optimizer with a momentum of 0.9, a weight decay of 1×10^{-5}, and a batch size of 64. The learning rate is set to 0.01 and is decayed by a factor of 10 after round 75. All these hyperparameters are consistent with those reported in Appx. F.2.

SAM-based baselines implementation. Since there is no method using SAM for the federated active learning (FAL) problem, we evaluate the following seven advanced SAM-based methods in federated learning (FL), active learning (AL), and long-tailed problems. To make a fair comparison, we follow their hyperparameter search strategy as reported in Tab. 7. For the remaining baselines (non-SAM approach), we consistently use the original hyperparameters and implementations in our experiments.

- `FedSAM` (Qu et al., 2022; Caldarola et al., 2022)[5] directly applies the SAM objective in local learning.

[5] `https://github.com/debcaldarola/fedsam`

Table 4: Balanced test accuracy (%) on four standard benchmark datasets under different labeled budgets 15–45% with Dirichlet $\alpha = 0.1$ (strongly non-i.i.d.) and $\gamma = 20$. Results are averaged over five random seeds. Traditional AL baselines are evaluated using two query-selector models, G and L, denoting global and local selectors, respectively. **Bold**: best per column; underline: second best.

Method	Model	FMNIST				CIFAR-10				CIFAR-100				Tiny ImageNet			
		15%	25%	35%	45%	15%	25%	35%	45%	15%	25%	35%	45%	15%	25%	35%	45%
Random	–	83.06	84.88	85.01	85.60	47.24	50.46	54.29	55.70	17.69	22.44	24.13	27.44	12.41	15.97	18.22	19.58
Traditional Active Learning-based Baselines																	
Entropy	G	85.20	85.71	86.48	86.72	50.15	54.79	56.76	58.54	16.89	22.01	23.74	27.15	12.83	15.88	18.14	19.60
	L	85.49	86.36	86.76	87.06	50.93	56.26	58.34	59.35	17.44	22.52	24.56	27.70	12.90	16.08	18.72	19.89
Margin	G	85.21	85.71	86.23	86.53	49.66	53.90	56.89	58.24	17.52	22.46	24.47	27.70	12.96	16.00	18.41	19.92
	L	85.43	86.16	86.76	87.04	51.16	55.64	58.28	59.49	17.51	22.71	24.80	27.92	13.06	16.22	18.67	20.11
Coreset	G	84.77	85.52	86.09	86.80	47.66	52.69	54.89	56.69	17.36	22.03	23.65	26.75	12.74	15.83	18.14	19.85
	L	84.58	85.06	86.18	86.57	47.23	51.17	53.92	55.82	17.08	21.66	23.30	26.87	12.90	16.10	18.75	20.07
BADGE	G	85.49	85.77	86.13	86.70	49.43	53.98	57.10	58.68	17.92	23.16	24.62	27.37	12.96	16.24	18.73	20.03
	L	85.25	86.30	86.95	86.94	50.53	54.93	57.65	59.13	18.20	23.36	25.12	27.95	12.93	16.49	18.88	20.45
ALFA-Mix	G	84.88	85.86	85.88	86.37	49.20	54.31	57.27	58.75	17.86	22.95	24.90	28.43	12.86	16.17	18.56	20.12
	L	84.38	85.63	86.22	86.61	47.87	52.27	54.90	56.62	17.73	22.39	24.35	27.58	13.07	14.38	18.71	20.43
Federated Active Learning Baselines																	
KAFAL	G, L	85.27	86.27	87.04	87.05	49.99	56.34	58.41	60.01	17.36	22.52	24.46	27.84	13.19	16.38	18.81	20.46
LoGo	G, L	85.15	86.31	86.70	86.98	51.56	56.35	58.30	59.68	17.33	22.08	24.74	27.95	13.25	16.34	18.97	20.53
IFAL	G, L	84.41	85.62	86.18	86.80	47.76	52.67	55.62	57.51	17.56	22.10	23.58	26.82	13.38	16.42	18.86	20.24
FairFAL	G, L	85.39	86.52	87.09	87.47	52.20	57.02	59.46	60.46	18.23	23.74	25.74	29.24	13.99	17.25	19.65	21.47
SAM-based Class-aware Baselines																	
ImbSAM	G, L	85.41	86.77	87.13	87.56	52.37	57.37	60.03	61.74	20.28	25.23	27.37	30.42	13.73	16.97	19.58	21.30
Focal-SAM	G, L	85.47	86.63	87.02	87.62	52.40	57.62	60.15	61.68	20.22	24.84	28.85	30.98	13.92	16.88	19.71	21.58
SAM-based Active Learning Baseline																	
SAAL	G, L	84.59	85.83	86.44	86.90	48.35	52.19	55.94	57.87	17.84	22.36	23.97	27.85	13.41	16.72	18.90	20.29
SAM-based Federated Learning Baselines																	
FedLESAM	G, L	85.46	86.57	87.40	87.81	51.17	56.63	59.80	61.99	20.81	25.73	28.16	30.82	13.70	17.02	19.81	21.87
FedGMT	G, L	84.66	85.90	86.38	87.11	48.12	53.34	55.58	57.23	17.65	22.81	24.42	28.41	13.37	16.79	18.63	20.04
FedScar	G, L	85.49	86.65	87.28	87.86	51.22	57.19	59.96	62.10	20.81	25.87	28.31	30.77	13.80	17.05	19.89	22.04
FedSAM	G, L	85.54	**86.86**	87.39	87.97	51.28	57.15	60.11	62.04	20.70	25.98	28.50	30.73	13.78	17.20	19.98	22.14
FalOTS (ours)	G, L	**85.69**	**86.86**	**87.63**	**88.00**	**52.64**	**58.15**	**61.69**	**63.28**	**21.45**	**26.47**	**28.87**	**31.95**	**14.25**	**17.41**	**20.21**	**22.64**

- FedLESAM (Fan et al., 2024)[6] estimates the global perturbation as the difference between the locally stored historical model from the activation round and the global model received in the current round.
- FedGMT (Li et al., 2025b)[7] incorporates a local SAM optimizer with a dynamic regularizer.
- FedScar (Lu et al., 2026)[8] corrects heterogeneity-induced geometric bias in local optimization to align client-side flatness with the global loss landscape
- SAAL (Kim et al., 2023b)[9] incorporates loss-landscape sharpness into active learning by selecting unlabeled samples with large potential perturbed loss, using pseudo-labels to estimate this sharpness before annotation.
- ImbSAM (Zhou et al., 2023)[10] leverages class priors to make SAM class-aware, focusing sharpness minimization on tail classes to alleviate their severe overfitting under class-imbalanced training
- Focal-SAM (Li et al., 2025a)[11] assigns class-dependent penalties to sharpness, enabling fine-grained loss-landscape control for long-tailed recognition without requiring additional backpropagation.

[6] https://github.com/MediaBrain-SJTU/FedLESAM
[7] https://github.com/harrylee999/FL-SAM
[8] https://github.com/Ultraman6/FedScar
[9] https://github.com/aailabkaist/SAAL
[10] https://github.com/cool-xuan/Imbalanced_SAM
[11] https://github.com/scongl/Focal-SAM

Table 5: Balanced test accuracy (%) on four standard benchmark datasets under different labeled budgets 15–45% with Dirichlet $\alpha = 100$ (strongly non-i.i.d.) and $\gamma = 20$. Results are averaged over five random seeds. Traditional AL baselines are evaluated using two query-selector models, G and L, denoting global and local selectors, respectively. **Bold**: best per column; underline: second best.

Method	Model	FMNIST				CIFAR-10				CIFAR-100				Tiny ImageNet			
		15%	25%	35%	45%	15%	25%	35%	45%	15%	25%	35%	45%	15%	25%	35%	45%
Random	–	84.26	85.96	86.76	87.29	48.51	52.96	58.00	59.70	18.23	23.57	25.09	28.93	13.27	15.44	18.23	19.30
Traditional Active Learning-based Baselines																	
Entropy	G	85.88	87.28	88.14	88.46	52.43	57.47	61.53	63.83	17.56	22.98	24.80	28.77	13.04	15.29	18.20	19.22
	L	85.98	87.34	88.50	88.80	51.37	56.59	60.01	61.74	17.33	22.66	24.88	27.97	13.14	15.21	18.17	19.32
Margin	G	85.84	86.92	88.18	88.38	51.26	57.93	62.13	63.72	18.68	23.92	25.76	28.91	13.47	15.62	18.44	19.58
	L	85.89	87.39	88.29	88.61	50.09	55.59	59.58	61.68	18.27	23.71	25.78	28.70	13.41	15.52	18.36	19.54
Coreset	G	85.18	86.67	87.98	88.23	48.81	54.26	57.97	59.57	17.93	22.95	24.70	27.82	13.09	15.30	18.16	19.25
	L	84.24	86.24	87.25	87.96	47.71	52.81	57.24	58.66	17.87	23.08	24.56	28.04	13.17	15.24	18.07	19.21
BADGE	G	85.89	87.32	88.15	88.50	51.54	58.02	61.65	63.83	18.76	24.42	25.78	28.91	13.53	15.70	18.64	19.78
	L	85.51	87.43	88.40	88.58	49.52	55.59	59.42	61.06	18.47	23.53	25.65	29.05	13.46	15.63	18.83	19.72
ALFA-Mix	G	85.28	86.86	87.78	88.15	51.39	57.37	61.61	63.46	18.65	24.22	26.32	29.42	13.47	15.86	18.77	19.61
	L	84.71	86.47	87.62	87.81	48.49	52.90	57.23	58.47	18.35	23.56	25.55	28.84	13.38	15.74	18.58	19.56
Federated Active Learning Baselines																	
KAFAL	G, L	86.10	87.73	88.60	88.90	48.56	54.25	58.51	61.61	16.55	21.59	23.64	27.02	12.98	14.87	17.69	19.16
LoGo	G, L	86.15	87.67	88.45	88.65	51.64	57.62	61.19	62.44	17.79	23.82	25.71	28.91	13.24	15.41	18.29	19.47
IFAL	G, L	82.94	84.57	85.82	86.09	45.60	50.78	55.55	57.67	17.72	22.77	24.37	27.81	13.36	15.34	18.27	19.40
FairFAL	G, L	85.96	87.64	88.14	88.91	53.24	59.31	63.07	64.53	18.76	24.68	26.42	30.03	14.06	16.77	19.45	21.25
SAM-based Class-aware Baselines																	
ImbSAM	G, L	86.34	87.91	88.07	88.95	55.72	61.18	64.23	65.47	20.46	25.62	27.87	30.63	13.82	16.89	19.61	21.30
Focal-SAM	G, L	86.30	87.83	88.41	88.97	56.12	61.47	64.85	65.78	21.11	26.14	28.22	30.94	13.88	16.97	19.62	21.43
SAM-based Active Learning Baseline																	
SAAL	G, L	84.15	86.34	87.73	88.06	50.22	55.79	61.27	62.48	17.98	23.52	25.60	28.79	13.68	15.94	18.74	20.26
SAM-based Federated Learning Baseline																	
FedLESAM	G, L	86.32	88.17	88.97	89.43	56.09	61.77	65.36	67.14	20.58	26.11	28.36	31.03	13.91	16.98	19.44	21.40
FedGMT	G, L	84.38	86.41	87.69	88.10	49.78	55.22	59.58	61.38	17.81	23.37	25.48	28.66	13.53	15.90	18.65	20.23
FedScar	G, L	86.41	88.08	89.13	89.56	56.15	62.12	65.49	67.20	20.80	26.07	28.27	31.01	14.00	17.02	19.14	21.28
FedSAM	G, L	86.46	88.20	89.24	89.63	56.20	62.04	65.60	66.91	20.79	26.19	28.62	31.18	13.96	17.13	19.76	21.55
FalOTS (ours)	G, L	**86.94**	**88.63**	**89.41**	**89.84**	**57.04**	**63.19**	**66.50**	**68.28**	**21.49**	**26.60**	**28.81**	**31.63**	**14.21**	**17.36**	**19.88**	**22.08**

For the algorithms mentioned above, we search for hyperparameters and select the best candidates. All hyperparameters for the methods are referenced from their official implementations or research papers. In Tab. 7, β represents the coefficient for the penalty term. The selection of this hyperparameter has been extensively studied in previous research, which confirms its effectiveness. ρ is the perturbation radius for the nearby neighborhood in parameter space. The choice of ρ is often associated with the local learning rate η (Sun et al., 2023b). Additionally, γ and τ represent the strength and temperature parameters in the KL function within FedGMT. In FedScar, γ is the factor that controls the history or accumulation behavior used in the geometric correction. Meanwhile, in Focal-SAM, γ also serves as the factor that dictates the extent to which sharpness regularization shifts toward tail classes and λ is the sharpness coefficient. In FedGMT, α represents the decay coefficient of the EMA, while in FedScar, it serves as the geometric-correction control parameter. In ImbSAM, $\eta_{imbalance}$ is the class-frequency threshold separating head and tail classes.

We note that some notation in Tab. 7 is similar to that in Tab. 3, but we retain them to maintain consistency with existing SAM-based optimizers.

G FalOTS Algorithm

Algorithm 1 outlines a client-side update process for FalOTS. Given a local mini-batch, we begin by calculating the standard empirical gradient to determine the ascent direction of SAM. Simultaneously, we construct a target-aware gradient by reweighting examples based on the global train-target class-prior ratio. We then subtract from the target-aware gradient the component that is already aligned with the SAM direction, resulting in a normalized orthogonal correction. This correction

Table 6: Test accuracy (%) on two text classification benchmark datasets under different labeled budgets from 15% to 45% with Dirichlet $\alpha = 0.1$ (strongly non-i.i.d.) and $\gamma = 20$. Results are averaged over five random seeds. **Bold**: best per column; underline: second best.

Method	Model	AG News				DBpedia-14			
		15%	25%	35%	45%	15%	25%	35%	45%
Federated Active Learning Baseline									
FairFAL	G, L	85.58	86.38	87.26	87.88	97.84	97.84	98.43	98.57
SAM-based Class-aware Baselines									
ImbSAM	G, L	85.90	86.76	87.00	87.87	98.07	98.18	98.12	98.33
Focal-SAM	G, L	85.97	86.92	87.02	87.76	98.12	98.34	98.26	98.41
SAM-based Active Learning Baseline									
SAAL	G, L	85.24	86.24	86.97	87.33	97.93	98.10	98.15	98.22
SAM-based Federated Learning Baselines									
FedLESAM	G, L	86.35	87.59	87.44	88.20	98.54	98.67	98.63	98.60
FedGMT	G, L	85.48	86.67	87.12	87.96	97.99	98.34	98.40	98.43
FedScar	G, L	86.70	87.79	87.62	88.24	98.61	98.72	98.72	98.77
FedSAM	G, L	86.64	87.89	87.88	88.30	98.78	98.81	98.70	98.80
FalOTS (ours)	G, L	**87.53**	**88.55**	**89.18**	**89.37**	**98.85**	**98.87**	**98.91**	**99.02**

Table 7: Algorithm-specific hyperparameters.

Method	**Searched Candidates**	**Best Selection**
FedSAM (Qu et al., 2022; Caldarola et al., 2022)	$\rho \in \{0.5\eta, \eta, 2\eta, 5\eta, 10\eta, 20\eta\}$	$\rho = 0.1$
FedLESAM (Fan et al., 2024)	$\rho \in \{0.5\eta, \eta, 2\eta, 5\eta, 10\eta, 20\eta\}$ $\beta \in \{1, 0.1, 0.01, 0.001\}$	$\rho = 0.02$ $\beta = 0.1$
FedGMT (Li et al., 2025b)	$\gamma \in \{0.5, 1.0, 2.0\}$ $\beta \in \{1, 0.1, 0.01, 0.001\}$ $\alpha \in \{0.5, 0.95, 0.995, 0.998\}$ $\tau \in \{1.0, 2.0, 3.0, 5.0\}$	$\gamma = 1.0$ $\beta = 0.1$ $\alpha = 0.95$ $\tau = 3.0$
FedScar (Lu et al., 2026)	$\rho \in \{0.5\eta, \eta, 2\eta, 5\eta, 10\eta, 20\eta\}$ $\alpha \in \{0.5\eta, \eta, 2\eta, 5\eta, 10\eta, 20\eta\}$ $\beta \in \{1, 0.1, 0.01, 0.001\}$ $\gamma \in \{0.5, 0.9, 0.95, 0.995\}$	$\rho = 0.1$ $\alpha = 0.1$ $\beta = 0.06$ $\gamma = 0.9$
ImbSAM (Zhou et al., 2023)	$\rho \in \{0.5\eta, \eta, 2\eta, 5\eta, 10\eta, 20\eta\}$ $\eta_{imbalance} \in \{10, 20, 50, 60, 100\}$	$\rho = 0.05$ $\eta_{imbalance} = 20$
Focal-SAM (Li et al., 2025a)	$\rho \in \{0.5\eta, \eta, 2\eta, 5\eta, 10\eta, 20\eta\}$ $\gamma \in \{0, 0.2, 0.5, 1.0, 2, 5, 10\}$ $\lambda \in \{0, 0.2, 0.5, 0.8, 1.0, 1.5, 2.0\}$	$\rho = 0.2$ $\gamma = 1.0$ $\lambda = 0.5$
SAAL (Kim et al., 2023b)	$\rho \in \{0.5\eta, \eta, 2\eta, 5\eta, 10\eta, 20\eta\}$	$\rho = 0.05$
FalOTS (ours)	$\rho \in \{0.5\eta, \eta, 2\eta, 5\eta, 10\eta, 20\eta\}$	$\rho = 0.1$

captures complementary target information without modifying the original SAM component. The final perturbation merges the empirical SAM direction and the orthogonal target-aware direction, maintaining an equal component budget, denoted as ρ. Notably, this correction specifically affects the perturbation; the descent step continues to be computed using the original unweighted empirical objective assessed at the perturbed parameters. This approach effectively preserves the standard SAM update structure while incorporating target-aware information only where the neighborhood is explored.

The FAL procedure utilizing FalOTS is summarized in Algorithm 2. At the beginning of each training phase, the server aggregates the counts of labeled classes and broadcasts target-aware class weights, which remain constant throughout that phase. Clients then proceed to train on their current labeled datasets using Algorithm 1, alternating between local updates and server aggregations over

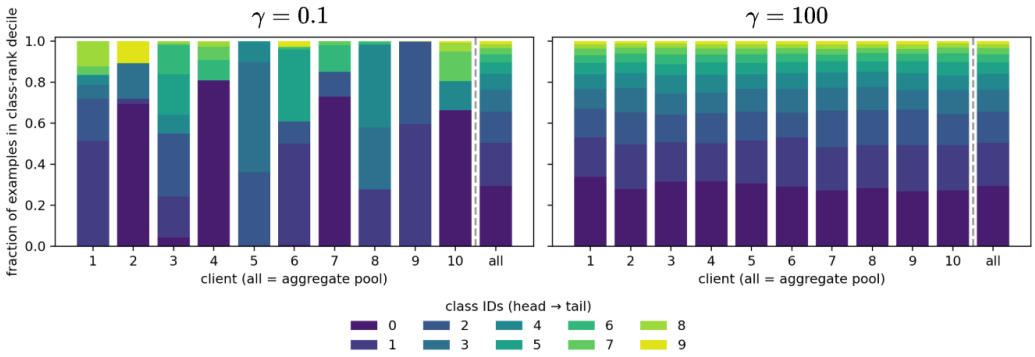


Figure 9: Global class imbalance persists under both client heterogeneity settings $\alpha = 0.1$ and $\alpha = 100$

T communication rounds. Between phases, the acquisition rule identifies additional examples for annotation, thereby expanding the labeled datasets before refreshing class weights and reinitializing the model for the next phase.

H Verification of the Target-Aware Sharpness Motivation

We perform two complementary checks: whether client heterogeneity changes the global long-tail prior and whether active acquisition solve the global class imbalance issue.

Reducing client heterogeneity does not remove the global long tail. We partition the same CIFAR-10 training pool among ten clients using Dirichlet parameters $\alpha = 0.1$ and $\alpha = 100$. Let π_k^{pool} denote the class prior of client k's full data pool $\mathcal{P}_k$, and π_{pool} the aggregate pool prior. Figure 9 shows that $\alpha = 0.1$ produces substantially different client priors, whereas $\alpha = 100$ yields nearly identical priors across clients. Both partitions contain the same 17,023 examples and preserve exactly the same aggregate class counts. The mean client divergence, $K^{-1}\sum_{k=1}^{K} D_{\text{KL}}(\pi_k^{\text{pool}} \| \pi_{\text{pool}})$, decreases from 0.88444 nats at $\alpha = 0.1$ to 0.00388 nats at $\alpha = 100$. Thus, increasing α can reduce client heterogeneity while leaving the global long-tail distribution unchanged, demonstrating that nearly i.i.d. client partitions can still exhibit global class imbalance.

Persistent Global Labeled-Target Prior Mismatch after Acquisition. Fig. 10 examines whether `FairFAL`'s acquisition strategy under the SGD-based FAL baseline is sufficient to remove the global labeled–target class-prior mismatch. The full LT pool denotes the entire long-tailed CIFAR-10 training set before any samples are selected for labeling; it contains both the initially labeled samples and the unlabeled candidates available across all clients. Its class prior, π_{pool}, follows the predefined long-tailed imbalance ratio and therefore provides a reference for the underlying global dataset skew. At each labeled budget, we aggregate the labeled samples from all clients and compute the aggregate labeled prior as $\pi_{\text{emp}}(c) = N_c / \sum_{j=1}^{C} N_j$, where N_c is the total number of globally labeled examples belonging to class c. The target prior is uniform, i.e., $\pi_{\text{tgt}}(c) = 1/C$. We quantify the remaining prior mismatch using $D_{\text{KL}}(\pi_{\text{emp}} \| \pi_{\text{tgt}}) = \sum_{c=1}^{C} \pi_{\text{emp}}(c) \log(\pi_{\text{emp}}(c)/\pi_{\text{tgt}}(c))$, measured in nats, where zero indicates that the aggregate labeled prior matches the target prior. The dashed line reports the correspond-

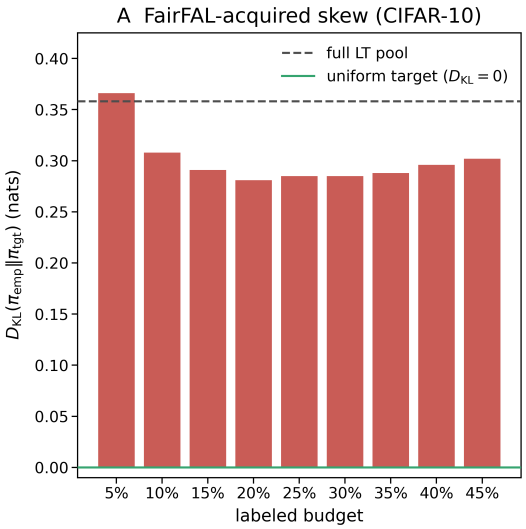


Figure 10: Persistent aggregate labeled-prior skew after `FairFAL` acquisition on CIFAR-10 under $\gamma = 20$ and $\alpha = 0.1$

.

Algorithm 1 Client-side update of `FalOTS`.

Require: parameters θ, nonempty mini-batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{n}$, radius $\rho \geq 0$, learning rate $\eta > 0$, nonnegative class weights w from Eq. 5
Ensure: updated parameters θ^{+}

$L(\theta) \leftarrow \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i, y_i)$ ▷ Empirical objective
$g_{\mathrm{emp}} \leftarrow \nabla_\theta L(\theta)$
if $\|g_{\mathrm{emp}}\| > 0$ **then**
  $\widehat{g} \leftarrow g_{\mathrm{emp}} / \|g_{\mathrm{emp}}\|$ ▷ Standard SAM direction
else
  $\widehat{g} \leftarrow 0$ ▷ Zero empirical gradient
end if
$S_w \leftarrow \sum_{i=1}^{n} w(y_i)$
if $S_w > 0$ **then**
  $L_w(\theta) \leftarrow \frac{1}{S_w}\sum_{i=1}^{n} w(y_i)\ell(\theta; x_i, y_i)$ ▷ Target-prior reweighting
  $g_{\mathrm{tgt}} \leftarrow \nabla_\theta L_w(\theta)$ ▷ Target-aware reweighted gradient
else
  $g_{\mathrm{tgt}} \leftarrow 0$ ▷ Omit target correction for this batch
end if
$v \leftarrow g_{\mathrm{tgt}} - \langle g_{\mathrm{tgt}}, \widehat{g}\rangle\, \widehat{g}$ ▷ Remove the SAM-aligned component
if $\|v\| > 0$ **then**
  $u \leftarrow v / \|v\|$ ▷ Unit orthogonal correction
else
  $u \leftarrow 0$ ▷ No complementary component
end if
$\epsilon_{\texttt{FalOTS}} \leftarrow \rho\widehat{g} + \rho u$ ▷ Corrected SAM perturbation
$\widetilde{\theta} \leftarrow \theta + \epsilon_{\texttt{FalOTS}}$ ▷ Hold the perturbation fixed during descent
$g_{\mathrm{pert}} \leftarrow \nabla_\vartheta L(\vartheta)|_{\vartheta=\widetilde{\theta}}$ ▷ Unweighted empirical descent
$\theta^{+} \leftarrow \theta - \eta g_{\mathrm{pert}}$ ▷ Update the base parameters
return θ^{+}

ing divergence for the full LT pool, $D_{\mathrm{KL}}(\pi_{\mathrm{pool}} \| \pi_{\mathrm{tgt}})$, which is approximately 0.36 nats. The initial randomly selected 5% labeled set exhibits a similar divergence of approximately 0.37 nats. `FairFAL` reduces this value to 0.31, 0.29, and 0.28 at the 10%, 15%, and 20% budgets, respectively, showing that acquisition partially reduces the global class-prior mismatch relative to the original pool. Nevertheless, the divergence remains around 0.28–0.30 nats from the 20% to 45% budgets. Thus, acquisition alone does not eliminate the mismatch between π_{emp} and π_{tgt}. This residual mismatch directly motivates the optimization-level correction in `FalOTS`, which constructs target-aware class weights from $\pi_{\mathrm{tgt}}/\pi_{\mathrm{emp}}$ when forming the SAM perturbation.

I Additional Ablation Studies

In this appendix, we present additional ablation studies due to the page limitations of the main paper. All ablations are primarily conducted using the CIFAR-100 dataset, under conditions of significant global class imbalance (with a parameter of $\gamma = 20$) and local heterogeneity (with $\alpha = 0.1$) on a single seed to reduce the training time. We will specify the settings for each ablation when they differ from this standard.

I.1 Generalization across FL frameworks

We investigate `FalOTS` within two additional widely used federated learning (FL) frameworks: `FedProx` (Li et al., 2020a) and `SCAFFOLD` (Karimireddy et al., 2020) on CIFAR-10, to assess its robustness against different optimization dynamics. As shown in Fig. 11, `FalOTS` consistently outperforms all baseline methods across both frameworks and all heterogeneity settings, including algorithms based on SAM. These findings suggest that `FalOTS` functions independently of the underlying FL optimizer and can be effortlessly integrated into a variety of federated learning pipelines.

Algorithm 2 Federated active learning with `FalOTS`.

Require: client pools $\{\mathcal{P}_k\}_{k=1}^K$; initial nonempty labeled sets $\{\mathcal{L}_k^{(0)}\}_{k=1}^K$; target prior π_{tgt}; S training phases, T communication rounds per phase, E local epochs per round; radius ρ, learning rate η; acquisition rule $\mathcal{A}$, server aggregation rule $\mathcal{G}$; feasible query budgets $\{m_k^{(s)}\}_{s=0}^{S-2}$

Ensure: final global model $\theta^{(S-1,T)}$

for $s = 0, \dots, S-1$ **do**
  Refresh class weights at the phase boundary
  for each client k in parallel **do**
    $n_k^{(s)} \leftarrow |\mathcal{L}_k^{(s)}|$
    $c_{k,y}^{(s)} \leftarrow \sum_{(x,\bar{y})\in\mathcal{L}_k^{(s)}} \mathbf{1}\{\bar{y} = y\}$ for each class y
    Send $\{c_{k,y}^{(s)}\}_{y=1}^C$ to the server
  end for
  $\pi_{\mathrm{emp}}^{(s)}(y) \leftarrow \frac{\sum_k c_{k,y}^{(s)}}{\sum_k n_k^{(s)}}$ for each class y
  Compute $w^{(s)}$ using Eq. 5 with $\pi_{\mathrm{emp}}^{(s)}$ and π_{tgt}
  Broadcast $w^{(s)}$ to clients ▷ Fixed throughout this phase
  Federated training on fixed labeled sets
  Initialize $\theta^{(s,0)}$ and reset training states ▷ Phase-wise reinitialization
  for $t = 0, \dots, T-1$ **do**
    Broadcast $\theta^{(s,t)}$ to clients
    for each client k in parallel **do**
      $\theta_k \leftarrow \theta^{(s,t)}$
      for $e = 1, \dots, E$ **do**
        for each mini-batch $\mathcal{B}$ from $\mathcal{L}_k^{(s)}$ **do**
          Update θ_k using Algorithm 1 with $(\mathcal{B}, \rho, \eta, w^{(s)})$
        end for
      end for
      Send the local model θ_k to the server
    end for
    $\theta^{(s,t+1)} \leftarrow \mathcal{G}\Big(\theta^{(s,t)}, \{(\theta_k, n_k^{(s)})\}_{k=1}^K\Big)$ ▷ Server aggregation
  end for
  Acquire labels for the next training phase
  if $s < S-1$ **then**
    Distribute $\theta^{(s,T)}$ for acquisition
    for each client k in parallel **do**
      $\mathcal{U}_k^{(s)} \leftarrow \mathcal{P}_k \setminus \mathcal{L}_k^{(s)}$
      $\mathcal{Q}_k^{(s)} \leftarrow \mathcal{A}\Big(\theta^{(s,T)}, \mathcal{L}_k^{(s)}, \mathcal{U}_k^{(s)}, m_k^{(s)}\Big)$ ▷ Select $m_k^{(s)}$ unlabeled examples
      Obtain annotations for $\mathcal{Q}_k^{(s)}$
      $\mathcal{L}_k^{(s+1)} \leftarrow \mathcal{L}_k^{(s)} \cup \mathcal{Q}_k^{(s)}$
    end for
  end if
end for
return $\theta^{(S-1,T)}$

This compatibility underscores its practical significance for real-world applications, where different federated learning frameworks may be chosen based on system constraints or stability requirements.

I.2 Generalization across acquisition strategies

We assess `FalOTS` using four acquisition strategies, including `KAFAL`, `LoGo`, `IFAL`, and `FairFAL`, on the CIFAR-100 dataset, focusing on a scenario of global class imbalance ($\gamma = 20$). As illustrated in Tab. 8, all four combinations demonstrate enhanced accuracy as the labeled budget

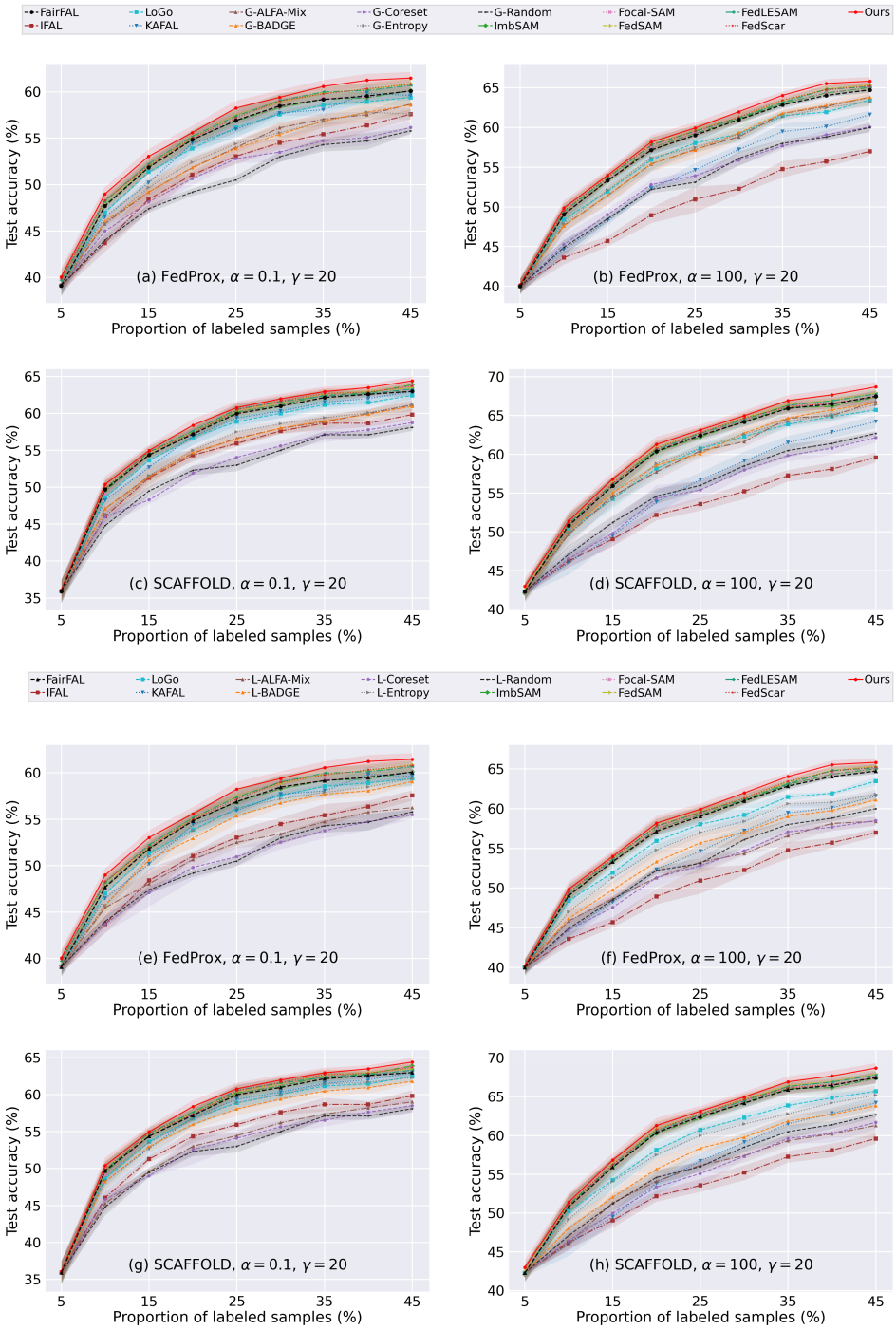


Figure 11: Test accuracy vs. proportion of labeled samples on CIFAR-10 under global imbalance ($\gamma = 20$) and two heterogeneity levels ($\alpha = 0.1$ and $\alpha = 100$), evaluated with `FedProx` (Li et al., 2020a) and `SCAFFOLD` (Karimireddy et al., 2020). Each subplot shows the corresponding FL framework and heterogeneity setting, comparing `FalOTS` with FAL and AL baselines, as well as SAM-based algorithms in long-tailed, AL, and FL problems using either the global (G-) or local (L-) query model.

increases, with results observed under both heterogeneous ($\alpha = 0.1$) and nearly i.i.d. ($\alpha = 100$) client partitions. At the 45% budget mark, accuracy varies from 30.91% to 31.76% for the heterogeneous partition and from 31.03% to 32.06% for the nearly i.i.d. partition. Notably, the acquisition

Table 8: Test accuracy versus labeled-data ratio of FAL baseline with our FalOTS algorithm under different acquisition strategies on CIFAR-100 with $\alpha = 0.1$ and $\alpha = 100$ (global class imbalance $\gamma = 20$).

Acquisition Strategy	**CIFAR-100**							
	$\alpha = 0.1,\ \rho = 20$				$\alpha = 100,\ \rho = 20$			
	15%	25%	35%	45%	15%	25%	35%	45%
KAFAL	21.34	26.23	28.09	30.91	21.27	26.30	28.02	31.03
LoGo	21.43	<u>26.49</u>	<u>28.41</u>	30.94	<u>21.58</u>	26.77	<u>28.12</u>	31.35
IFAL	<u>21.56</u>	26.42	28.37	<u>31.01</u>	21.50	<u>26.83</u>	28.04	<u>31.48</u>
FairFAL	**21.62**	**26.75**	**28.65**	**31.76**	**21.80**	**27.14**	**28.56**	**32.06**

Table 9: Final-round test accuracy comparison under different backbone architectures on CIFAR-10 with heterogeneity levels $\alpha = 0.1$ and $\alpha = 100$ (global imbalance $\rho = 20$). The default 4-layer CNN backbone is replaced with MobileNet and ResNet-18, respectively.

Method	**MobileNet**		**ResNet-18**	
	$\alpha = 0.1$	$\alpha = 100$	$\alpha = 0.1$	$\alpha = 100$
KAFAL	41.16	43.14	37.96	42.57
LoGo	39.08	45.87	37.10	44.95
IFAL	36.64	41.80	36.89	42.10
FairFAL	41.01	46.73	38.45	45.50
ImbSAM	41.63	47.02	38.51	45.89
Focal-SAM	41.78	47.34	38.54	45.98
SAAL	40.85	46.54	38.22	45.37
FedLESAM	42.74	47.90	39.51	46.72
FedGMT	40.96	46.82	38.48	45.54
FedScar	42.81	48.04	39.69	46.78
FedSAM	42.86	47.91	39.88	46.85
FalOTS	**43.27**	**48.51**	**40.28**	**47.59**

strategy from FairFAL achieves the highest accuracy across these settings, while the other strategies also remain competitive. These findings underscore the versatility of FalOTS across different acquisition rules, highlighting the complementary roles of sample selection and target-aware local optimization within the FAL framework.

I.3 Impact of backbone architecture

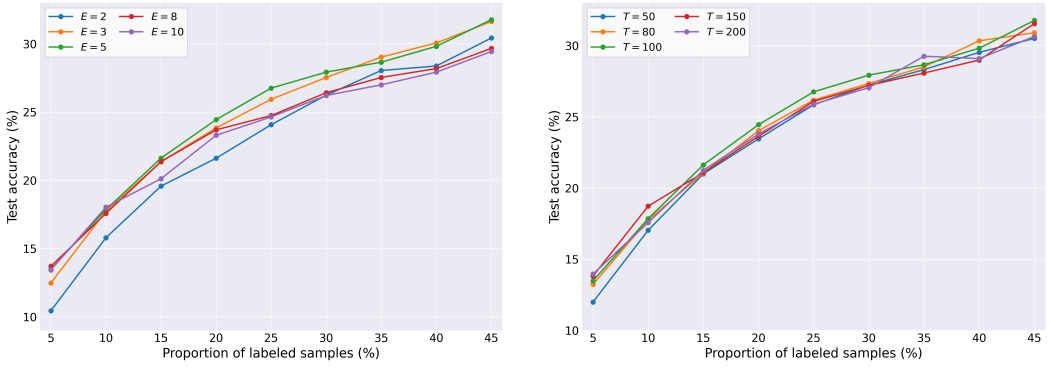
To evaluate the architectural robustness of FalOTS, we replace the default 4-layer CNN with two stronger backbones, MobileNet (Howard et al., 2017) and ResNet-18 He et al. (2016), and compare its performance against baseline methods. As presented in Tab. 9, FalOTS consistently matches or outperforms all other baselines across both architectures and both heterogeneity levels. FalOTS's stable performance demonstrates that its core mechanisms do not depend on any particular network design and generalize effectively to modern, deeper architectures commonly used in practical federated systems. Note that we replace the 4-layer CNN with MobileNet and ResNet-18 without changing any other hyperparameters (which we expect to fine-tune for better performance) to match the FairFAL baseline.

I.4 Ablation for ρ sweep

Tab. 10 examines the sensitivity of FalOTS and FedSAM to the perturbation radius ρ, under a fixed local heterogeneity level of $\alpha = 0.1$ and a global imbalance ratio of $\gamma = 20$. Across all evalu-

Table 10: Test accuracy (%) under different ρ values under local imbalance ratios $\alpha = 0.1$ and global imbalance $\gamma = 20$ for our FalOTS and FedSAM on CIFAR-100.

ρ	Method	Proportion of labeled samples (%)			
		15	25	35	45
0.02	FalOTS	**19.88**	**24.73**	**26.59**	**29.68**
	FedSAM	18.7	23.13	26.12	29.37
0.05	FalOTS	**20.18**	**25.51**	**27.39**	**29.96**
	FedSAM	19.43	24.67	26.81	29.25
0.1	FalOTS	**21.62**	**26.75**	**28.65**	**31.76**
	FedSAM	20.17	25.32	27.62	30.79
0.15	FalOTS	**22.24**	**26.26**	**28.31**	**31.44**
	FedSAM	21.21	25.07	27.50	30.65
0.2	FalOTS	**21.66**	**26.45**	**28.27**	**31.43**
	FedSAM	20.09	25.29	27.49	30.48



(a) Impact of local epochs E

(b) Impact of communication rounds T

Figure 12: Ablation studies for local training epochs and communication rounds on CIFAR-100 dataset

ated values of ρ and labeling budgets, FalOTS consistently surpasses FedSAM. For both methods, increasing ρ from 0.02–0.05 to moderate values generally enhances performance, while further increasing it beyond 0.1 does not yield any consistent additional benefits. Notably, FalOTS reaches its highest accuracy with $\rho = 0.1$ at labeling budgets of 25%, 35%, and 45%, whereas $\rho = 0.15$ shows the best performance at 15%. These results suggest that FalOTS remains robust across a reasonably wide range of perturbation radii, with $\rho = 0.1$ serving as a stable operating point; hence, we adopt it as the default setting in our experiments.

I.5 Scalability across client counts

We further evaluate the scalability of FalOTS by varying the number of participating clients to 5 and 20 on CIFAR-10 dataset. As illustrated in Fig. 13, FalOTS consistently delivers strong performance in both scenarios, surpassing all baseline methods, irrespective of the level of data fragmentation. These findings suggest that FalOTS effectively corrects the original SAM perturbation using an orthogonal target-aware correction mechanism.

Table 11: Test accuracy (%) under different local imbalance ratios α and proportions of labeled samples with global imbalance $\gamma = 20$. Bold values indicate the better performance between FalOTS and FedSAM under the same setting.

Heterogeneity Level α	Method	Proportion of labeled samples (%)			
		15	25	35	45
0.1	FalOTS	**21.62**	**26.75**	**28.65**	**31.76**
	FedSAM	20.17	25.32	27.62	30.79
1	FalOTS	**21.13**	**27.08**	**28.51**	**31.29**
	FedSAM	21.06	26.23	28.18	30.90
10	FalOTS	**21.98**	**26.20**	**28.00**	**31.12**
	FedSAM	21.59	25.76	**28.00**	31.01
20	FalOTS	**21.64**	**26.58**	**29.13**	**32.48**
	FedSAM	20.61	25.96	28.07	31.65
50	FalOTS	**21.33**	**26.31**	**29.15**	**31.47**
	FedSAM	20.22	25.58	28.17	30.95
80	FalOTS	**21.54**	**25.98**	**28.50**	**32.11**
	FedSAM	20.55	25.48	28.01	30.88
100	FalOTS	**21.87**	**26.91**	**28.49**	**30.98**
	FedSAM	20.78	25.79	27.83	30.26

I.6 Impact of local heterogeneity

Tab. 11 compares FalOTS with FedSAM baseline under varying heterogeneity levels α, while fixing the global imbalance at $\gamma = 20$. Across all labeled-data budgets and heterogeneity settings, FalOTS consistently achieves higher test accuracy than FedSAM, with the only tie occurring at $\alpha = 10$ and 35% labeled samples. Notably, the advantage of FalOTS remains evident across both low and high heterogeneity levels, indicating that its target-aware perturbation is robust to varying degrees of client heterogeneity and provides consistent benefits over conventional federated SAM.

I.7 Impact of local epochs

We further examine the sensitivity of FalOTS to the number of local training epochs E by varying E across the set $\{2, 3, 5, 8, 10\}$, with $E = 5$ serving as the default configuration in all primary experiments, consistent with recent works (Zong et al., 2025; Kim et al., 2023a; Zong & Huang, 2026). As illustrated in Fig. 12a, FalOTS demonstrates relatively stable performance across different values of E, indicating that its effectiveness is not significantly affected by this hyperparameter. Notably, the default setting of $E = 5$ achieves the strongest or highly competitive results across most labeling budgets, securing the best outcomes at 15%, 20%, 25%, 30%, and 45% of labeled data. While marginal improvements can occasionally be observed with $E = 8$ or $E = 10$ at very low labeling budgets, and with $E = 3$ at the 35%–40% range, these enhancements are limited and do not present consistent advantages across different budgets. Conversely, utilizing too few local epochs ($E = 2$) typically results in subpar performance, whereas increasing the number of local epochs beyond 5 tends to gradually diminish accuracy at larger labeling budgets. These findings reinforce $E = 5$ as a robust and well-balanced default choice for FalOTS and other FAL methods as well.

Table 12: Impact of correction strength on test accuracy (%) under different labeled-data ratios.

Correction Factor	Proportion of labeled samples (%)			
	15	25	35	45
0.5	20.89	25.91	28.02	30.87
0.8	<u>21.16</u>	<u>26.37</u>	<u>28.44</u>	<u>31.30</u>
1.0	**21.62**	**26.75**	**28.65**	**31.76**
1.5	20.23	25.33	27.89	30.82
2.0	20.19	25.38	27.68	30.85

Table 13: Impact of the target distribution used to construct the SAM ascent direction. Values report balanced test accuracy (%) at different labeled-data budgets.

Target distribution	15%	25%	35%	45%
None (Standard SAM)	20.17	25.32	27.62	30.79
Pool	21.53	<u>26.06</u>	28.03	30.83
Partial uniform	**21.69**	26.04	<u>28.43</u>	<u>31.04</u>
Uniform	<u>21.62</u>	**26.75**	**28.65**	**31.76**

I.8 Impact of communication rounds

We further assess the sensitivity of FalOTS to the number of communication rounds T by varying T within the set $\{50, 80, 100, 150, 200\}$, with $T = 100$ designated as the default setting for our main experiments, consistent with recent work (Zong et al., 2025; Kim et al., 2023a; Zong & Huang, 2026). As illustrated in Fig. 12b, FalOTS demonstrates generally stable performance across a broad range of communication budgets. Notably, $T = 100$ strikes an optimal balance between predictive performance and communication cost. Specifically, $T = 100$ achieves the highest accuracy at several intermediate labeling budgets (15%–30%) and also reaches the maximum final accuracy with 45% labeled data. Using fewer communication rounds, particularly $T = 50$, tends to result in lower performance, suggesting that limited communication can hinder global model convergence. On the other hand, increasing the communication budget beyond 100 rounds does not consistently yield improvements; although $T = 150$ or $T = 200$ may perform better at certain labeling budgets, these gains are not maintained throughout the acquisition process and come with significantly higher communication overhead. Overall, these findings indicate that FalOTS is not highly sensitive to the communication budget, supporting the choice of $T = 100$ as an effective and computationally balanced default setting.

I.9 Impact of selected target distribution in FalOTS

Tab. 13 examines how the target class prior used to construct the orthogonal correction affects balanced test accuracy across labeled-data budgets. Standard SAM applies no target correction, corresponding to $\pi_{\mathrm{tgt}} = \pi_{\mathrm{emp}}$ and unit weights on the labeled support. The *Pool* variant uses $\pi_{\mathrm{tgt}} = \widehat{\pi}_{\mathrm{pool}}$, a label-free estimate of π_{pool} obtained by averaging the global model's softmax predictions over examples in the full client pools $\{\mathcal{P}_k\}$ and applying an exponential moving average. For this ablation, this estimate is refreshed at the start of each communication round. The *Partial uniform* variant uses $\pi_{\mathrm{tgt}}(y) = (1-\lambda)\pi_{\mathrm{emp}}(y) + \lambda/C$ with $\lambda = 0.5$, whereas the *Uniform* variant sets $\pi_{\mathrm{tgt}}(y) = 1/C$ for every class. Each corrected variant computes $w(y)$ according to Eq. 5, with $w(y) = 0$ when $\pi_{\mathrm{emp}}(y) = 0$, and uses the resulting weighted gradient to construct the orthogonal correction while retaining the unweighted descent loss. All three variants outperform standard SAM at every reported budget. Partial uniform achieves the highest accuracy at 15%, exceeding SAM by 1.52pp, although its advantage over Uniform is only 0.07pp. Uniform performs best at the remaining budgets, improving over SAM by 1.43, 1.03, and 0.97pp at 25%, 35%, and 45%, respectively. These results support using a target prior aligned with the class-balanced evaluation objective, with the uniform prior providing the strongest performance at the three larger budgets.

I.10 Sensitivity to prior reweighting strength

We generalize Eq. 5 to

$$w_\psi(y) = \left(\frac{\pi_{\mathrm{tgt}}(y)}{\pi_{\mathrm{emp}}(y)} \right)^{\psi}$$

when $\pi_{\mathrm{emp}}(y) > 0$, and zero otherwise, where ψ controls the strength of class-prior reweighting and $\psi = 1$ recovers our default formulation. We evaluate $\psi \in \{0.5, 0.8, 1.0, 1.5, 2.0\}$ across labeled budgets of 15–45%. As shown in Tab. 12, the default $\psi = 1$ achieves the highest accuracy at every reported budget, with $\psi = 0.8$ consistently ranking second. Both weaker and stronger reweighting

yield lower accuracy; at the 45% budget, accuracy decreases from 31.76% at $\psi = 1$ to 30.87% and 30.85% at $\psi = 0.5$ and $\psi = 2$, respectively. These results support the unmodified prior ratio as the default choice within the evaluated settings.

J Limitations and Discussion

Target specification and gradient estimation. `FalOTS` employs a predefined target class prior to guide local optimization. This prior should represent the intended evaluation objective; for instance, a uniform prior assigns equal significance to each class. Its behavior under target-prior misspecification remains to be studied. The population target-gradient identity, as outlined in Proposition 1, holds true under the label-shift and support-overlap conditions specified in Assumption 1, along with the stated regularity conditions. Active acquisition may affect within-class distributions, which means that prior weighting can provide a target-informed training signal without necessarily recovering the precise target-risk gradient. Additionally, the population identity does not guarantee the finite-batch unbiasedness of the self-normalized gradient used in local training, and weighting alone cannot retrieve gradient information for classes that are absent from the global labeled set. These factors highlight the need for further exploration into the interplay between target-aware optimization and acquisition strategies that enhance class coverage.

Computational considerations. The target-aware correction necessitates an extra gradient evaluation compared to standard SAM or `FedSAM` as discussed in Sec. 5.1, along with class-count communication at acquisition boundaries. The practical trade-offs vary based on model size, client resources, and the respective costs of annotation, training, and communication. Our proposed method assists in minimizing communication rounds when communication is costly during practical deployment.

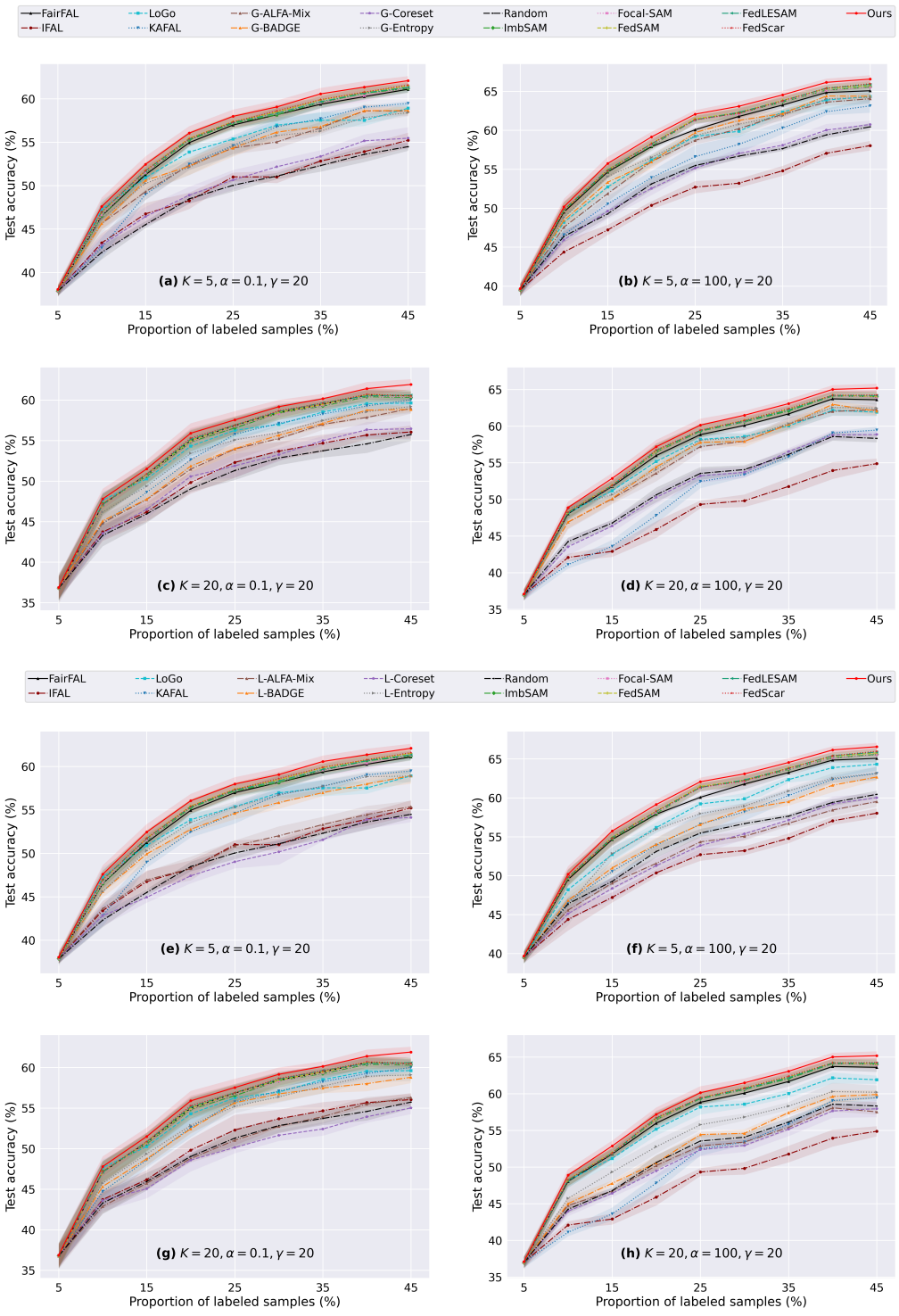


Figure 13: Test accuracy vs. proportion of labeled samples on CIFAR-10 under global imbalance ($\gamma = 20$) and two heterogeneity levels ($\alpha = 0.1$ and $\alpha = 100$), evaluated with either 5 or 20 clients. Each subplot shows the corresponding specific combination of client count and heterogeneity level and compares `FalOTS` with FAL and AL baselines, as well as SAM-based algorithms in long-tailed, AL, and FL problems using either the global (G-) or local (L-) query model.