OvA-LP: A Simple and Efficient Framework for Federated Learning on Non-IID Data

Anonymous authors
Paper under double-blind review

Abstract

Federated fine-tuning (FFT) adapts foundation models to decentralized data but remains fragile under heterogeneous client distributions due to local drift, i.e., client-level update divergences that induce systematic bias and amplified variance in the global model. Existing aggregation and personalization methods largely correct drift post hoc, which proves brittle under extreme non-IID conditions. We introduce OvA-LP, a minimalist framework that is, to our knowledge, the first explicitly designed to suppress drift at its source within the PEFT-based FFT paradigm. OvA-LP combines linear probing on a frozen encoder with a one-vs-all head and a simple two-stage procedure, preserving pretrained feature geometry and decoupling logits to prevent the mechanisms that amplify drift. On CIFAR-100 with 100 clients, averaged over shard-1, shard-2, and Bernoulli–Dirichlet partitions, OvA-LP retains 95.9% of its IID accuracy, whereas state-of-the-art FFT baselines retain only 10.1% (PFPT) and 34.5% (FFT-MoE) under the same conditions. OvA-LP further maintains resilience under both symmetric and asymmetric label noise. In addition, precomputing encoder features makes per-round cost nearly independent of encoder size. Together, these results demonstrate that OvA-LP provides a principled and efficient basis for robust FFT under heterogeneity.

1 Introduction

Foundation models (FMs) have reshaped machine learning by providing powerful pretrained representations that can be adapted to diverse downstream tasks. In federated learning (FL), this shift has given rise to federated fine-tuning (FFT), where clients adapt a shared encoder instead of training models from scratch (Zhuang et al., 2023). Parameter-efficient fine-tuning (PEFT) methods such as adapters, LoRA, and prompt tuning (Houlsby et al., 2019; Hu et al., 2022; Lester et al., 2021) further reduce computational and communication costs, making FFT a practical paradigm for large-scale decentralized adaptation. This paradigm is particularly critical in domains where data locality is a strict requirement, such as training on sensitive medical records across hospitals or financial data across institutions. However, despite its efficiency and practical importance, the robustness of FFT under heterogeneous client distributions remains a major challenge (Ren et al., 2025). This gap motivates a rethinking of how to make FFT robust to extreme heterogeneity.

The core difficulty lies in local drift: client updates diverge due to distributional differences, and when aggregated, these drifts bias and destabilize the global model. Specifically, client-level divergences manifest as both systematic bias and amplified variance in the aggregated update, degrading final accuracy and hindering convergence. In practice, under extreme non-IID conditions, state-of-the-art FFT methods often converge slowly and retain well below half of their IID accuracy within 50 rounds. This persistent relative gap highlights the need for approaches that directly mitigate drift at the client level, rather than merely compensating for it after aggregation.

Broadly, prior efforts fall into two families. Aggregation strategies modify the global update rule, from classical methods such as FedProx and Scaffold (Li et al., 2020; Karimireddy et al., 2020) to more recent LoRA-specific variants (Wang et al., 2024; Guo et al., 2024; Yan et al., 2025). Personalization frameworks attach client-specific modules to absorb drift locally, ranging from adapters and prompts to expert-based models (Zhang et al., 2024; Fan et al., 2024; Wang et al., 2025; Yang et al., 2025). In practice these families are not disjoint—many approaches combine both, such as FFT-MoE(Hu et al., 2025) with expert routing on top of FedAvg, or PFPT (Weng et al., 2024) with

redesigned prompts and aggregation. Yet despite their variety, they share a common philosophy: drift is treated as unavoidable and corrected only post-hoc, once it has already manifested at the client or global level. This reactive stance leaves them fragile under extreme heterogeneity, as no method so far has succeeded in preventing drift from arising in the first place.

In this paper, we present OvA-LP, a minimalist framework that suppresses client drift at its root. OvA-LP integrates three lightweight components—linear probing (LP) on a frozen encoder (Alain & Bengio, 2016), one-vs-all (OvA) binary heads, and a two-stage training schedule—each explored in isolation but never unified. By aligning them within a bias–variance decomposition of federated gradients, we systematically connect feature geometry, label decoupling, and variance control into a single source-level framework. This reframes drift mitigation from post-hoc correction to proactive prevention, suggesting shift in how robustness is pursued within PEFT-based FFT.

Our main findings are:

- OvA-LP consistently prevents local drift from arising at the client level, offering a principled foundation for robust FFT under heterogeneity.
- OvA-LP retains 95.9% of its IID accuracy on CIFAR-100 with 100 clients under shard-1, shard-2, and Bernoulli–Dirichlet ($p = 0.1, \alpha = 0.001$)(Xu et al., 2022), whereas FFT-MoE and PFPT retain only 34.5% and 10.1%, respectively.
- OvA-LP demonstrates innate robustness to label noise: it consistently reduces accuracy degradation under both symmetric and asymmetric corruption, maintaining resilience at higher noise levels and surpassing specialized noise-robust baselines.
- OvA-LP precomputes encoder features once, making per-round training nearly independent of encoder size and preserving modularity for integration with other FFT strategies.

2 Related Work

Aggregation strategies. A long line of work has sought to improve FL robustness by modifying the global update rule. Classical approaches such as FedProx and Scaffold reduce the variance of client updates and partially stabilize convergence. More recent extensions adapt these ideas to PEFT settings, for example FLoRA, FedSA-LoRA, and FRLoRA (Wang et al., 2024; Guo et al., 2024; Yan et al., 2025). While effective in mitigating some client drift, these methods still rely on aggregation at the server side, typically applied only after local divergence has already occurred.

Personalization frameworks. Another direction attaches client-specific modules to absorb drift locally. Examples include FedAdapter and FedPrompt (Cai et al., 2022; Zhao et al., 2023), as well as expert-based extensions such as FFT-MoE and PFPT. These approaches improve local adaptation, but global consistency remains limited because personalization cannot prevent drift from propagating into the shared model.

Classification heads for label imbalance. Another line of work modifies the classification head to mitigate skewed label distributions. FedRS (Li & Zhan, 2021) restricts softmax updates for missing classes, mitigating bias under label imbalance. OvA-based approaches such as FedOVA, FedABC, and ATHENA-FL (Zhu et al., 2021; Wang et al., 2023; de Souza et al., 2024) decompose multiclass tasks into binary classifiers to avoid softmax coupling and improve fairness. However, these methods are designed for scratch training and focus mainly on label imbalance, without addressing the broader challenge of feature drift.

Label noise robustness. A complementary line of research tackles noisy labels in FL. Methods such as FedCorr (Xu et al., 2022) and FedLTF (Zhan et al., 2025) design correction mechanisms or robust objectives to improve performance under corruption. While effective, these approaches explicitly address noise rather than the underlying drift mechanisms, and remain orthogonal to our focus.

Our positioning. Unlike prior OvA-based methods restricted to scratch training and label imbalance, OvA-LP is designed to prevent drift from arising by freezing the encoder and introducing a two-stage OvA head. Its minimalist design remains modular and in principle compatible with

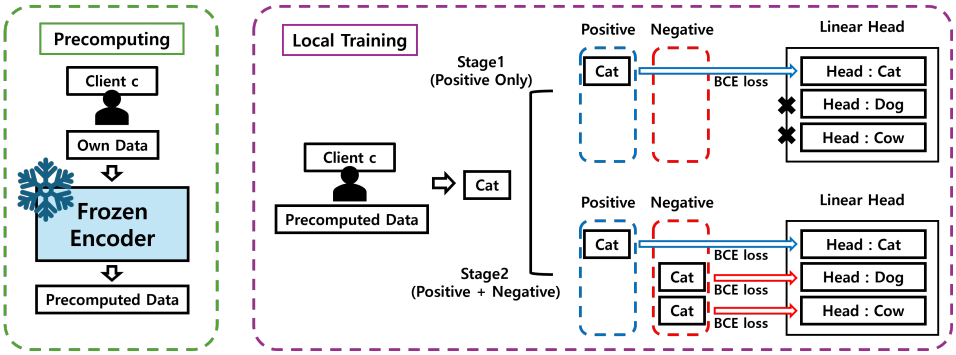

Figure 1: Overall structure of OvA-LP. Clients precompute encoder features once (left) and perform two-stage local training with one-vs-all heads (right).

aggregation and personalization families, suggesting potential for deployment across diverse FFT pipelines.

3 Methodology

Overview. OvA-LP is motivated by a source-level philosophy: preventing drift at its origin rather than correcting it post hoc. Fig. 1 summarizes the overall workflow. Clients first precompute encoder features with a frozen backbone, then train one-vs-all heads under a lightweight two-stage schedule. This design is guided by a bias–variance decomposition of federated gradients, which identifies local bias, global bias, and variance as the root causes of drift. OvA-LP targets each of these components with simple yet complementary mechanisms: feature geometry bounds the effect of feature skew, OvA heads eliminate label-skew bias and variance amplification, and the two-stage schedule stabilizes optimization under participation variance.

The remainder of this section develops these ideas step by step. Sec. 3.1 formalizes the bias–variance framework that motivates our design. Sec. 3.2 shows how pretrained geometry preserves alignment and separation, limiting bias from feature skew. Sec. 3.3 analyzes label skew, explaining how OvA decoupling removes the bias and variance amplification caused by softmax coupling. Finally, Sec. 3.4 addresses the remaining variance, demonstrating how the two-stage schedule achieves fast and stable convergence. Together, these analyses show how OvA-LP systematically aligns with the bias–variance view to bring Non-IID training close to the IID reference.

3.1 Bias–Variance Framework

We begin by formalizing drift through a bias–variance decomposition, which identifies local bias, global bias, and variance as the core sources of degradation.

Client drift under non-IID data can be understood through a bias–variance decomposition at both local and global levels. Let the stochastic gradient on client i be $g_i = \nabla \ell(w; x, y)$. Denote by $\mathcal{D}_i$ the local data distribution of client i and by $\mathcal{D}$ the global distribution. The local loss is $L_i(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\ell(w; x, y)]$ with expected gradient $\nabla L_i(w)$, and the global loss is $L(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(w; x, y)]$ with gradient $\nabla L(w)$.

Local bias. Each client's optimum deviates from the global one by $b_i = \nabla L_i(w) - \nabla L(w)$, arising from distributional differences across clients, in particular feature skew and label skew.

Global bias. Aggregating across clients yields $B = \mathbb{E}_i[\nabla L_i(w)] - \nabla L(w)$, which distorts the overall update direction and accumulates to reduce accuracy.

Local and global variance. Even within a single client, stochastic gradients fluctuate with variance $v_i = \mathrm{Var}[g_i]$. When aggregated with weights p_i (e.g., proportional to dataset sizes n_i), the update is $\hat{g} = \sum_i p_i g_i$ with variance $V = \mathrm{Var}[\hat{g}]$, which is further amplified by quantity skew.

Metric	Pretrained (T)	Pretrained (F)
Alignment ↓	1.366 ± 0.006	1.761 ± 0.008
Intra ↓	0.818 ± 0.002	0.917 ± 0.004
Inter ↑	0.840 ± 0.004	0.487 ± 0.008
Ratio ↓	0.974 ± 0.007	1.885 ± 0.038

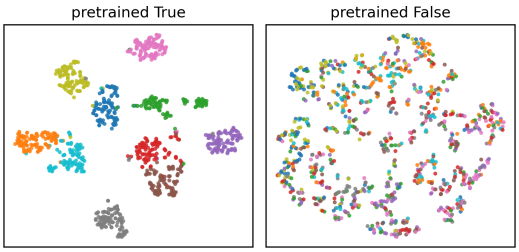

Figure 2: Feature geometry of pretrained vs randomly initialized encoders (CIFAR-10, ViT-L/16).

Takeaway. Local bias and variance are the primary contributors to drift, while global bias and variance are their aggregated manifestations. Variance is further exacerbated under label skew due to softmax coupling, which introduces cross-class covariance. This explains why aggregation-level fixes cannot fundamentally solve non-IID degradation: they address only the aggregate symptoms rather than the underlying local causes.

3.2 Linear Probing and Feature Geometry

Feature skew is bounded by pretrained geometry: alignment clusters same-class samples, and separation keeps classes apart.

We quantify feature geometry with four standard metrics. Following Wang & Isola (2020), alignment is defined as the expected squared distance between positive pairs:

$$\text{Alignment} = \mathbb{E}_{(x,y)\sim p_{\text{pos}}} \|f(x) - f(y)\|_2^2.$$

In addition, we report three well-known statistical measures of representation geometry:

$$\text{Intra} = \mathbb{E}_{(x,y)} \|f(x) - \mu_y\|_2^2, \qquad \text{Inter} = \mathbb{E}_{y\neq y'} \|\mu_y - \mu_{y'}\|_2^2, \qquad \text{Ratio} = \frac{\text{Intra}}{\text{Inter}}.$$

Here p_{pos} denotes the distribution over positive pairs, μ_y is the centroid of class y, and $f(\cdot)$ is the encoder representation. Alignment captures the closeness of positive pairs, Intra measures the compactness of each class cluster, Inter quantifies separation between class centroids, and Ratio summarizes the trade-off. Smaller Alignment, Intra, and Ratio and larger Inter indicate stronger feature geometry.

Fig. 2 compares pretrained and randomly initialized encoders. Across all four metrics, pretrained features show smaller Alignment and Intra, larger Inter, and a lower Ratio, confirming that they form compact, well-separated clusters.

From the bias–variance perspective, this structural geometry directly limits the bias induced by feature skew: alignment keeps same-class representations compact, while separation enforces clear boundaries across classes. As a result, client updates remain anchored to the global geometry, and the extent of local bias before aggregation is fundamentally bounded. In the ideal case of perfect alignment, feature-induced bias would vanish entirely.

3.3 OvA Head and Decoupling

The second major source of drift is label skew, which biases gradients and amplifies variance through softmax coupling. Let $h(x) = f_\theta(x) \in \mathbb{R}^d$ denote the encoder representation of input x, and let $w_c \in \mathbb{R}^d$ be the classifier weight vector for class c. We use $1[y = c]$ to denote the indicator for the ground-truth class.

Softmax coupling. For class c, the gradient of the cross-entropy loss with respect to w_c is

$$g_c(x, y) = (1[y = c] - p_c(x))h(x), \quad p_c(x) = \frac{\exp(w_c^\top h(x))}{\sum_{j=1}^K \exp(w_j^\top h(x))}.$$

Because all classes share a denominator, majority classes repeatedly dominate updates, while minority classes receive little signal. As analyzed in FedRS (Li & Zhan, 2021), this coupling introduces both bias and variance amplification under label skew. Replacing softmax with independent OvA heads removes this cross-class covariance, eliminating the mechanism behind label-skew drift. As analyzed in FedRS (Li & Zhan, 2021), majority classes dominate through repeated "pulls," while minority classes often receive only "pushes." This imbalance introduces bias, since updates are driven by probability-weighted terms $p_c(x)$ rather than purely class-specific targets. It also amplifies variance, because the shared denominator induces non-zero cross-class covariances $\mathrm{Cov}(g_c, g_j) \neq 0$. Together, these mechanisms destabilize training under heterogeneous distributions.

OvA decoupling. An OvA head replaces softmax with independent binary classifiers. The gradient of the logistic loss with respect to w_c is

$$g_c^{\mathrm{OvA}}(x, y) = (1[y = c] - q_c(x))h(x), \quad q_c(x) = \sigma(w_c^\top h(x)) = \tfrac{1}{1+\exp(-w_c^\top h(x))}.$$

$q_c(x)$ is the Bernoulli likelihood under a logistic regression head, and each head optimizes its binary logistic loss independently. As a result, the pull/push imbalance described in FedRS disappears: majority and minority classes are updated without mutual interference. This decoupling eliminates the mechanism of label-skew-induced bias and variance amplification, directly addressing the sources of drift at their origin.

3.4 Variance and Two-Stage Training

After bias terms are suppressed, variance remains the main source of drift. Variance cannot be eliminated entirely, but its destabilizing effect can be controlled through a two-stage curriculum aligned with the OvA structure.

Stage 1 (positive-only). When pretrained representations preserve alignment and separation, the global optimum of each OvA head lies near the class centroid at the point of maximum margin. Training only on positives thus pulls classifier weights toward these centroids, leading to rapid convergence without cross-class conflicts and helping to overcome the destabilizing effect of variance in the early rounds.

Stage 2 (positive+negative). After centroids are established, a large set of negatives is introduced to expand inter-class margins. At the same time, a small fraction of positives is retained as anchors, preventing the decision boundary from drifting away under the stronger influence of negatives. This combination enables efficient margin learning while preserving the stability achieved in Stage 1.

Takeaway. Together, the two stages implement an easy-first, hard-later curriculum. Stage 1 quickly aligns classifiers with class centroids under minimal variance, while Stage 2 leverages negatives for margin expansion without destabilizing the positive clusters. This design directly overcomes variance effects at their source, complementing OvA-LP's treatment of feature and label skew.

4 Experiments

4.1 Experimental Setup

Shared setting. Our primary experiments use CIFAR-100 with 100 clients for 50 rounds, a scale comparable to or larger than those adopted in recent FL benchmarks (see Appendix B.1 for survey). We fix five random seeds (0, 42, 777, 1337, 15254) across all runs for comparability. For the IID setting, data are split uniformly at random across clients. For the Non-IID setting, we use three representative configurations widely adopted in the literature: Shard-1 (one class per client), Shard-2 (two classes per client), and Dirichlet ($p = 0.1, \alpha = 0.001$) following the FedCorr construction (Xu et al., 2022). These serve as the standard Non-IID benchmarks throughout. Further analyses in Sec. 4.4 expand beyond this shared setting, including alternative partitions, encoder scaling, datasets like TinyImageNet, and robustness to label noise.

Our model. OvA-LP uses a frozen ViT-L/16 encoder and is trained with 100% client participation, three local epochs per round, batch size 50, learning rate 0.01, and AdamW optimizer with weight decay 1×10^{-4}. Client updates are aggregated by FedAvg (McMahan et al., 2017), with the first round conducted using Stage 1 (positive-only) training and all subsequent rounds using Stage 2 (positive+negative) training, as described in Sec. 3.4.

Baselines. Baseline methods are reproduced using their original model architectures and training protocols as specified in their papers. We preserve the encoders used in the original implementations (e.g., ViT-B/32, ViT-B/16), ensuring faithful reproduction; detailed configurations are reported in Appendix B.2.

Evaluation philosophy. We quantify non-IID robustness by comparing accuracy trajectories to the IID reference. For round t, we compute the relative ratio $R(t) = \text{Acc}_{\text{NonIID}}(t)/\text{Acc}_{\text{IID}}(t) \times 100$. We present results in two unified views: round-wise $R(t)$ curves showing how quickly and stably each method tracks the IID trajectory, and final $R(50)$ barplots summarizing the endpoint gap for each partition. This framing provides a consistent lens that captures convergence speed, stability, and final accuracy.

4.2 Ablation Study

We examine the contribution of each design component of OvA-LP by comparing three head configurations: (i) LP with softmax, (ii) OvA-LP without the two-stage design, and (iii) the full OvA-LP.

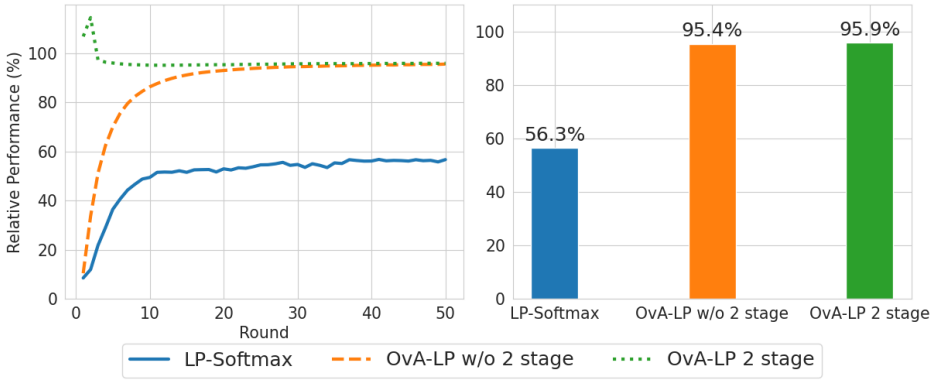

Figure 3: Ablation of OvA-LP components. Stepwise gains (56.3 → 95.4 → 95.9) illustrate the effects of OvA decoupling and two-stage training.

As shown in Fig. 3, the progression is stepwise. LP-softmax reaches only 56.3% under Non-IID, reflecting limited benefit from encoder freezing alone. Replacing the softmax with independent OvA classifiers stabilizes training and raises performance to 95.4%. Adding the two-stage design enables faster convergence to 95.9%, closely tracking the IID curve within only a few rounds.

A brief overshoot above the IID curve occurs in the first few rounds. This behavior arises from FedAvg's weighted averaging under imbalanced partitions and quickly settles.

In summary, OvA decoupling mitigates label-skew effects, and the two-stage procedure helps overcome variance, leading to faster and more stable convergence. These observations align with the bias–variance decomposition described in Sec. 3.4. Full accuracy curves and efficiency breakdowns are reported in Appendix A.1.

4.3 Comparison with Baselines

We next compare OvA-LP against two recent state-of-the-art baselines, FFT-MoE and PFPT.

Fig. 4 shows the outcome under the extreme Non-IID setting. Both baselines perform poorly compared to the IID reference: PFPT increases gradually but saturates at 34.5%, while FFT-MoE

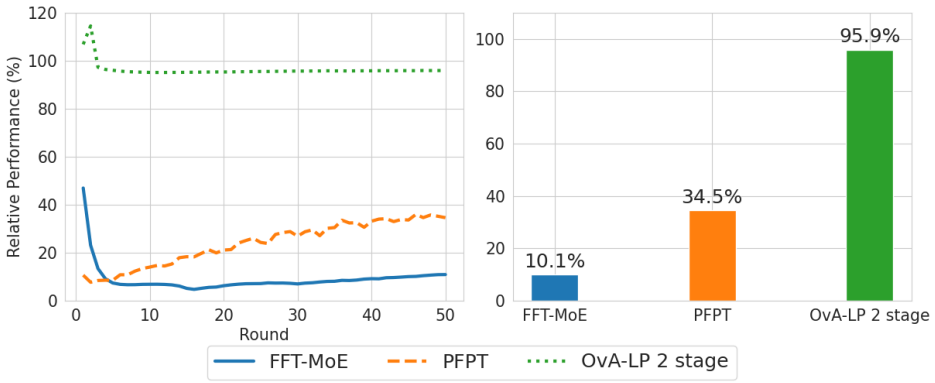

Figure 4: Comparison with state-of-the-art baselines. FFT-MoE plateaus near 10.1%, while PFPT rises slowly but saturates at 34.5%. OvA-LP remains stable and converges to 95.9%.

Method	Label Bias	Feature Bias	Var.	$R(50)(\%)$
FFT-MoE / PFPT	×	×	×	10.1 / 34.5
LP-softmax	×	△	×	56.3
OvA-LP (w/o 2-stage)	✓	△	×	95.4
OvA-LP (2-stage)	✓	△	✓	95.9

Table 1: Bias–variance view of methods. "✓" = removed/handled, "△" = partially removed, "×" = not addressed. As non-IID severity increases, leaving label bias intact results in low robustness, while OvA-LP progressively removes label bias and handles variance to reach near-IID performance.

plateaus early and remains near 10.1%. These results are consistent with their post-hoc philosophy, which seeks to mitigate drift only after aggregation.

OvA-LP, in contrast, maintains stability and converges to 95.9%, closely following the IID trajectory. As noted in Sec. 4.2, even LP-softmax, which partially reduces feature bias through encoder freezing, already surpasses post-hoc baselines in this setting. OvA-LP further improves upon this by also addressing label skew and variance effects, leading to accuracy near the IID level. Additional comparisons with FFT-MoE and PFPT are provided in Appendix A.2.

The ranking across methods follows a stepwise pattern. Post-hoc baselines leave both label and feature bias intact, resulting in poor robustness under strong heterogeneity. LP-softmax reduces feature bias but retains label bias, leading to moderate accuracy. OvA-LP without the two-stage procedure removes label bias and improves stability, and the full OvA-LP additionally addresses variance, reaching near-IID robustness. This stepwise progression aligns with the bias–variance decomposition and illustrates the benefit of addressing drift at its origin, as summarized in Table 1.

4.4 Additional Analyses

4.4.1 Partition-wise Robustness.

We evaluate OvA-LP under five representative heterogeneity patterns. Three of them—Shard-1, Shard-2, and Dirichlet ($\alpha = 0.001, p = 0.1$)—are the standard benchmarks already introduced in Sec. 4.1. We further include two additional settings. First, we adopt a Zipf distribution with exponent $s = 2.0$, a standard setup in FL for inducing quantity skew across clients (Piantadosi, 2014). Second, feature-based clustering, where each class is partitioned into K clusters (with K equal to the number of clients) using k-means, and each cluster is then assigned to a client.

Fig. 5 shows that across all five settings, the $R(t)$ curves remain aligned with the IID trajectory, and final $R(50)$ values range from 94.9% to 99.7%. This demonstrates that OvA-LP maintains robustness under diverse forms of skew, including label, feature, and quantity heterogeneity.

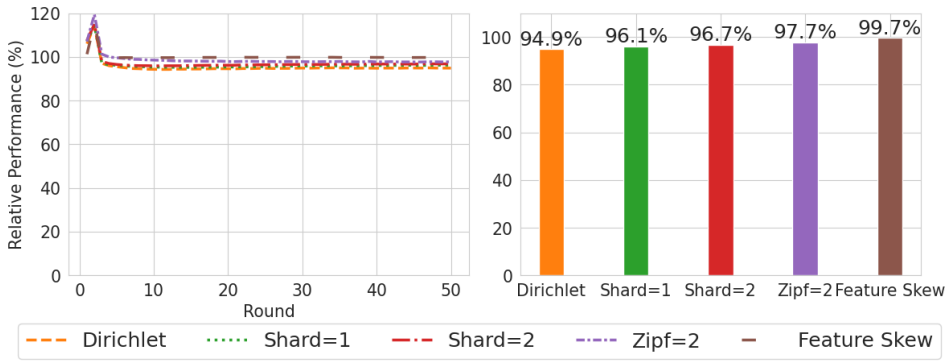

Figure 5: Partition-wise robustness of OvA-LP. Across five representative heterogeneity patterns, $R(t)$ curves (left) closely track the IID reference and final $R(50)$ values (right) remain above 94.9%. This confirms consistent robustness across diverse forms of skew, including label, feature, and quantity heterogeneity.

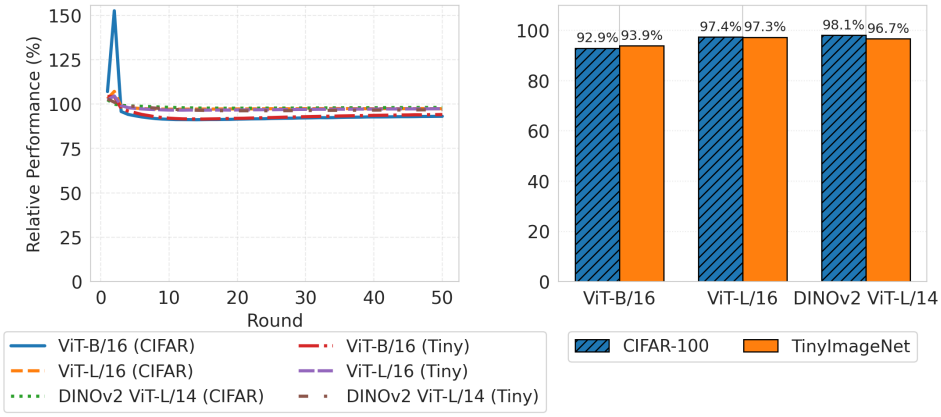

Figure 6: Encoder and task variations. OvA-LP is evaluated with different encoders (ViT-B/16, ViT-L/16, DINOv2-L/14) on CIFAR-100 and extended to TinyImageNet under Dirichlet partitioning.

4.4.2 Encoder and Task Variations.

We next test whether robustness depends on encoder scale or task domain. Fig. 6 compares ViT-B/16, ViT-L/16, and DINOv2-L/14 on CIFAR-100, and extends to TinyImageNet under Dirichlet partitioning. Absolute accuracy decreases for the smallest encoder (ViT-B/16), while ViT-L/16 and DINOv2-L/14 remain comparable, with minor fluctuations depending on the dataset. Crucially, in all cases $R(t)$ curves consistently track the IID trajectory, confirming that OvA-LP's robustness is agnostic to encoder scale, architecture, and task domain.

We note that the brief overshoot observed in the first round is a benign effect of size-weighted FedAvg, which appears more prominently with smaller encoders. It stabilizes quickly and does not affect final convergence.

4.4.3 Label Noise Robustness.

Following FedLTF (Zhan et al., 2025), we adopt the same label noise benchmarks and directly compare against the baselines it reports. FedLTF represented the prior state-of-the-art under label corruption. As Table 2 shows, OvA-LP achieves markedly smaller accuracy declines, surpassing FedLTF and all other reported methods.

Noise Type	Symmetric					Asymmetric		
Method	Baseline Acc (%)	Decline Rate(%) ↓				Baseline Acc (%)	Decline Rate(%) ↓	
Noise Ratio	0.30	0.40	0.50	0.60	0.70	0.20	0.30	0.40
FedAvg	16.75%	15.70%	23.70%	37.49%	51.46%	18.85%	13.37%	30.93%
Symmetric CE	16.99%	17.77%	25.66%	40.79%	49.79%	26.14%	17.71%	36.34%
Co-teaching	34.21%	8.68%	36.07%	51.04%	66.99%	34.19%	20.10%	33.23%
FedCorr	32.15%	13.50%	26.59%	41.65%	62.64%	41.12%	13.47%	30.81%
FedNoRo	38.58%	9.46%	19.57%	35.38%	43.36%	45.42%	9.82%	26.97%
FedLTF (Stage 2)	55.23%	3.73%	8.73%	12.18%	21.24%	52.63%	10.94%	26.51%
FedLTF (Stage 3)	58.43%	3.70%	9.24%	14.65%	20.91%	57.78%	8.71%	24.80%
OvA-LP (Ours)	88.78%	**0.76%**	**2.35%**	**4.52%**	**10.35%**	89.28%	**0.63%**	**1.53%**

Table 2: Robustness on CIFAR-100 with label noise, measured as accuracy decline rates (%) from baseline accuracy. Baseline results are taken from FedLTF (Zhan et al., 2025) (Table 2), which included standard and robust training schemes as well as its own variants. OvA-LP (2-stage) achieves the smallest decline, outperforming the prior state-of-the-art FedLTF.

Summary. Taken together, these analyses show that OvA-LP retains stability under diverse forms of heterogeneity, across encoder scales and task domains, and even in the presence of label corruption, demonstrating robustness across a broad range of conditions.

5 Limitations

We note two main limitations. First, OvA-LP relies heavily on the pretrained encoder: alignment and separation in the encoder's feature geometry are what reduce feature-skew bias and enable linear probing. If the encoder is weak, OvA-LP cannot compensate on its own. This is not unique to our method but reflects the broader trend in federated fine-tuning, where progress is fundamentally tied to advances in foundation models.

Second, all experiments assume full client participation. This choice highlights fast convergence under minimal variance but abstracts away from partial participation, which is common in practice. Indeed, as shown in Appendix A.3, reduced participation slows convergence, and our two-stage strategy alone cannot fully overcome this variance. However, Appendix A.2 demonstrates that OvA-LP remains highly efficient under full participation: it reaches Acc@95 within only 1–3 rounds, with both computation and communication costs substantially lower than prior methods, even when all 100 clients are active. Thus, while partial participation exposes a limitation, the lightweight design of OvA-LP makes full participation not only operationally feasible but also a practical advantage in real deployments.

In addition, our study is limited to vision benchmarks and does not yet combine with aggregation or personalization frameworks. We regard these as natural directions for future work.

6 Conclusion

We introduced OvA-LP, a minimalist framework for federated fine-tuning that addresses client drift at its origin. By combining linear probing with a one-vs-all head and a simple two-stage training strategy, OvA-LP shows that non-IID robustness can be achieved without architectural complexity. Despite its simplicity, it reaches near-IID accuracy across a wide range of non-IID settings, exhibits robustness to label noise consistent with its bias–variance suppression design, and maintains efficiency that scales favorably with encoder size.

Our perspective does not dismiss existing aggregation or personalization strategies; rather, it offers a complementary direction. Where prior approaches mitigate drift after it emerges, OvA-LP prevents its amplification at the source, making it especially effective under extreme heterogeneity. At the same time, post-hoc methods remain valuable for personalization and fine-grained corrections, suggesting a natural synergy with our framework.

OvA-LP stands as a strong new baseline and an initial step toward making source-level robustness a standard paradigm in federated fine-tuning.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. Fedadapter: Efficient federated learning for modern nlp. *arXiv preprint arXiv:2205.10162*, 2022. URL https://arxiv.org/abs/2205.10162.

Lucas Airam C de Souza, Gustavo F Camilo, Gabriel Antonio F Rebello, Matteo Sammarco, Miguel Elias M Campista, and Luís Henrique M. K. Costa. ATHENA-FL: Avoiding statistical heterogeneity with one-versus-all in federated learning. *Journal of Internet Services and Applications*, 15(1):273–288, 2024.

Dongyang Fan, Bettina Messmer, Nikita Doikov, and Martin Jaggi. On-device collaborative language modeling via a mixture of generalists and specialists. *arXiv preprint arXiv:2409.13931*, 2024.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. *arXiv preprint arXiv:2410.01463*, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gang Hu, Yinglei Teng, Pengfei Wu, and Nan Wang. Fft-MoE: Efficient federated fine-tuning for foundation models via large-scale sparse MoE under heterogeneous edge. *arXiv preprint arXiv:2508.18663*, 2025.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Xin-Chun Li and De-Chuan Zhan. FedRS: Federated learning with restricted softmax for label distribution non-IID data. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 995–1005, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282. PMLR, 2017.

Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, 2014. doi: 10.3758/s13423-014-0585-6.

Chao Ren, Han Yu, Hongyi Peng, Xiaoli Tang, Bo Zhao, Liping Yi, Alysa Ziying Tan, Yulan Gao, Anran Li, Xiaoxiao Li, et al. Advances and open challenges in federated foundation models. *IEEE Communications Surveys & Tutorials*, 2025.

Dui Wang, Li Shen, Yong Luo, Han Hu, Kehua Su, Yonggang Wen, and Dacheng Tao. FedABC: Targeting fair competition in personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10095–10103, 2023.

Jiaqi Wang, Jingtao Li, Weiming Zhuang, Chen Chen, Lingjuan Lyu, and Fenglong Ma. Enhancing foundation models with federated domain knowledge infusion. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28312c9491d60ed0c77f7fff4ad86dd1-Abstract-Conference.html.

Pei-Yau Weng, Minh Hoang, Lam Nguyen, My T. Thai, Lily Weng, and Nghia Hoang. Probabilistic federated prompt-tuning with non-IID and imbalanced data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/951877b24b376c5f4612e850251ee85b-Abstract-Conference.html.

Jingyi Xu, Zihan Chen, Tony Q. S. Quek, and Kai Fong Ernest Chong. FedCorr: Multi-stage federated learning for label noise correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10184–10193. IEEE, 2022.

Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Rick Siow Mong Goh, Yong Liu, and Lei Zhu. Federated residual low-rank adaptation of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/906c860f1b7515a8ffec02dcdac74048-Abstract-Conference.html.

Yihao Yang, Wenke Huang, Guancheng Wan, Bin Yang, and Mang Ye. Federated disentangled tuning with textual prior decoupling and visual dynamic adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025.

Shaojie Zhan, Lixing Yu, Hanqi Chen, and Tianxi Ji. FedLTF: Linear probing teaches fine-tuning to mitigate noisy labels in federated learning. In *Proceedings of the 16th Asian Conference on Machine Learning*, volume 260 of *Proceedings of Machine Learning Research*, pp. 1048–1063. PMLR, 2025. URL https://proceedings.mlr.press/v260/.

Yicheng Zhang, Zhen Qin, Zhaomin Wu, Jian Hou, and Shuiguang Deng. Personalized federated fine-tuning for LLMs via data-driven heterogeneous model architectures. *arXiv preprint arXiv:2411.19128*, 2024.

Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yuanshao Zhu, Christos Markos, Ruihui Zhao, Yefeng Zheng, and James J. Q. Yu. FedOVA: One-vs-all training method for federated learning with non-IID data. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2021.

Weiming Zhuang, Chen Chen, and Lingjuan Lyu. When foundation model meets federated learning: Motivations, challenges, and future directions. *arXiv preprint arXiv:2306.15546*, 2023.

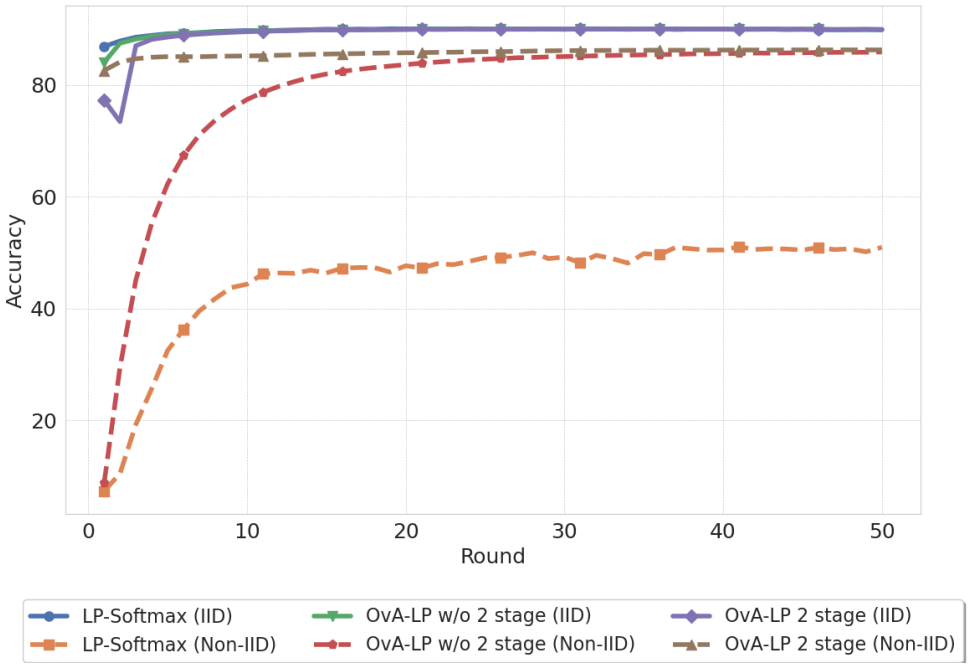

Figure 7: Ablation curves under IID and averaged Non-IID settings. Accuracy trajectories over 50 rounds.

Methodology	**Accuracy (%)**	**Acc@95 (Rounds)**	**Total Time (s)**	**Total Comm. (MB)**
LP-Softmax (IID)	90.13 ± 0.04	1 ± 0	0.03 ± 0.00	0.39 ± 0.00
LP-Softmax (Non-IID)	53.70 ± 6.79	27 ± 9	0.73 ± 0.23	10.71 ± 3.37
OvA-LP w/o 2 stage (IID)	90.03 ± 0.08	2 ± 0	0.05 ± 0.00	0.78 ± 0.00
OvA-LP w/o 2 stage (Non-IID)	85.89 ± 1.11	15 ± 3	0.37 ± 0.07	5.72 ± 1.02
OvA-LP 2 stage (IID)	90.04 ± 0.12	3 ± 0	0.05 ± 0.00	1.18 ± 0.00
OvA-LP 2 stage (Non-IID)	86.34 ± 0.73	1 ± 0	0.02 ± 0.00	0.42 ± 0.10

Table 3: Final performance metrics of ablation study, including accuracy, convergence rounds (Acc@95), total time, and total communication until convergence.

Methodology	**Time (Client, ms)**	**Time (Server, ms)**	**Comm. (Client)**	**Comm. (Server)**
LP-Softmax (IID)	23.18 ± 0.17	2.64 ± 0.04	401.20 KB	39.18 MB
LP-Softmax (Non-IID)	23.78 ± 0.60	2.97 ± 0.11	401.20 KB	39.18 MB
OvA-LP w/o 2 stage (IID)	21.67 ± 0.21	2.69 ± 0.06	401.20 KB	39.18 MB
OvA-LP w/o 2 stage (Non-IID)	22.27 ± 0.48	2.99 ± 0.03	401.20 KB	39.18 MB
OvA-LP 2 stage (IID)	14.15 ± 0.22	1.90 ± 0.04	401.20 KB	39.18 MB
OvA-LP 2 stage (Non-IID)	14.86 ± 0.51	1.91 ± 0.06	401.20 KB	39.18 MB

Table 4: Per-round computation and communication costs of ablation study.

A Additional Experimental Results

A.1 Ablation Results

Tables 3 and 4 summarize convergence and per-round costs. Under IID, all methods converge quickly with small differences; OvA-LP (2-stage) shows a slight increase in communication due to additional heads. The differences appear under Non-IID: LP-Softmax requires 27 rounds and about 0.73s to reach Acc@95, whereas OvA-LP (2-stage) reaches the same point in a single round, reducing total time by $\sim$36$\times$ and communication by $\sim$25$\times$. Per-round metrics show that OvA-

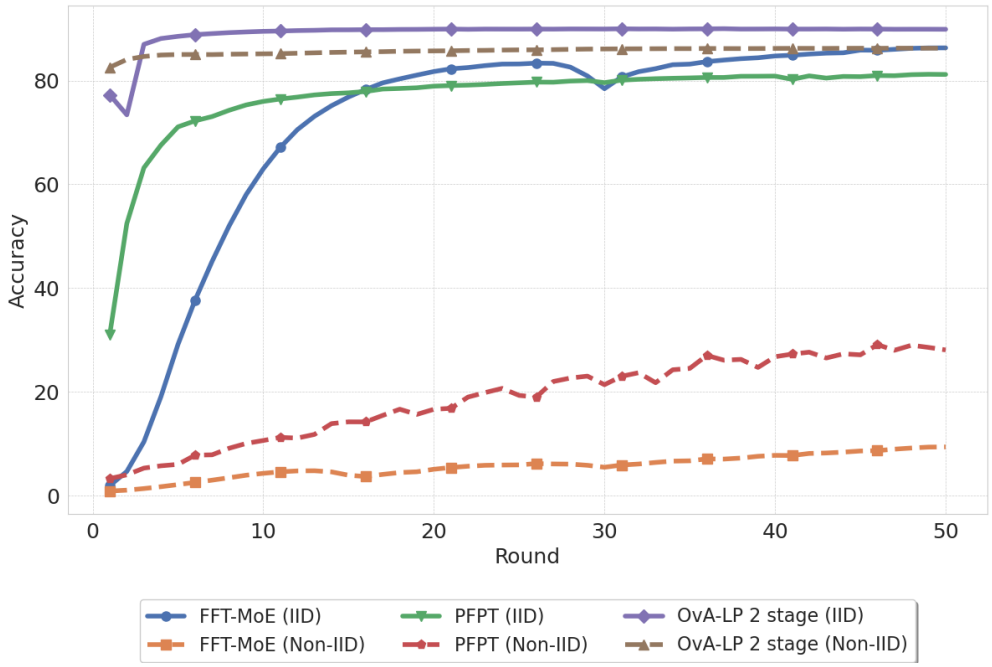

Figure 8: Baseline curves under IID and averaged Non-IID settings. Accuracy trajectories over 50 rounds.

LP (2-stage) is also more efficient, requiring 14.9ms on clients versus 23.8ms for LP-Softmax, and 1.9ms on the server versus 3.0ms.

In summary, the ablation indicates that while all methods behave similarly under IID, under Non-IID the full OvA-LP achieves near-IID efficiency with substantially fewer rounds and lower system cost.

A.2 Baseline Comparisons

Methodology	**Accuracy (%)**	**Acc@95 (Rounds)**	**Total Time (s)**	**Total Comm. (GB)**
FFT-MoE (IID)	96.39 ± 0.41	21 ± 0	34.58 ± 0.70	3.67 ± 0.08
FFT-MoE (Non-IID)	9.83 ± 7.96	37 ± 13	61.00 ± 21.59	6.48 ± 2.32
PFPT (IID)	80.27 ± 0.41	13 ± 1	432.75 ± 32.45	0.22 ± 0.02
PFPT (Non-IID)	33.17 ± 15.30	44 ± 6	1437.00 ± 182.03	0.74 ± 0.09
OvA-LP (IID)	90.04 ± 0.12	3 ± 0	0.05 ± 0.00	0.11 ± 0.00
OvA-LP (Non-IID)	86.34 ± 0.73	1 ± 0	0.02 ± 0.00	0.04 ± 0.01

Table 5: Final performance of baselines: accuracy, convergence rounds (Acc@95), and total costs until convergence.

Methodology	**Time (Client, ms)**	**Time (Server, ms)**	**Comm. (Client)**	**Comm. (Server)**
FFT-MoE (IID)	1560.27 ± 10.38	94.76 ± 0.42	1.7703 MB	177.03 MB
FFT-MoE (Non-IID)	1555.74 ± 18.46	95.49 ± 1.28	1.7703 MB	177.03 MB
PFPT (IID)	1860.98 ± 10.01	30941.43 ± 387.65	1.729 MB	17.29 MB
PFPT (Non-IID)	1855.49 ± 39.89	31002.22 ± 239.84	1.729 MB	17.29 MB
OvA-LP (IID)	14.16 ± 0.22	1.90 ± 0.04	0.3918 MB	39.18 MB
OvA-LP (Non-IID)	14.86 ± 0.51	1.91 ± 0.06	0.3918 MB	39.18 MB

Table 6: Per-round computation and communication costs of baselines.

Tables 5 and 6 compare FFT-MoE, PFPT, and OvA-LP under both IID and Non-IID. FFT-MoE achieves strong accuracy under IID (96%) but collapses under Non-IID, converging below 10% even after 37 rounds. PFPT is more stable across settings but converges slowly: its time-to-95% accuracy exceeds OvA-LP by over three orders of magnitude, despite using less communication. In contrast, OvA-LP converges within 3 rounds (IID) and 1 round (Non-IID), while its final accuracy remains above 86–90%. This corresponds to 10^2–10^4 reductions in time and communication compared to prior baselines. Per-round metrics further confirm the gap: OvA-LP requires only $\sim$14 ms on clients and 2 ms on the server, versus seconds or tens of seconds for PFPT and FFT-MoE.

Overall, OvA-LP attains comparable or better final accuracy while reaching convergence substantially faster and at far lower system cost.

A.3 Participation Rate Sweep

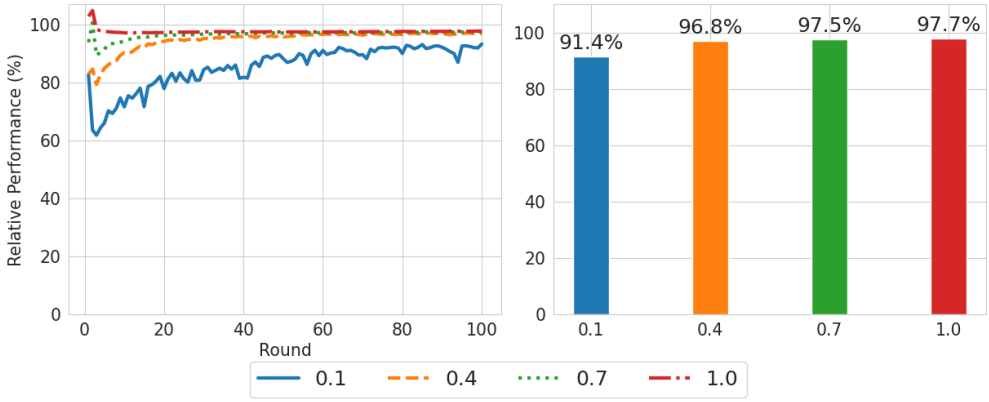

Figure 9: Results under Dirichlet($p = 0.1, \alpha = 0.001$) with participation ratios of 0.1, 0.4, 0.7, and 1.0. Left: accuracy trajectories $R(t)$ over 50 rounds. Right: final accuracy $R(50)$. Lower participation ratios lead to slower convergence, indicating that the two-stage method does not fully overcome participation-induced variance.

B Baseline Settings

B.1 Dataset Settings

Work	Dataset	# Clients
FedProx (Li et al., 2020)	MNIST, FEMNIST, Sent140, Shakespeare	10
SCAFFOLD (Karimireddy et al., 2020)	EMNIST	20
FedLTF (Zhan et al., 2025)	CIFAR-10/100, MNIST/FMNIST	20
PFPT (Weng et al., 2024)	CIFAR-10/100, TinyImageNet	10
FFT-MoE (Hu et al., 2025)	AgNews, CIFAR-10	4, 10
Our setup	CIFAR-100,TinyImageNet	**100**

Table 7: Survey of FL benchmarks in recent works. Our setup adopts 100 clients on CIFAR-100 and TinyImageNet which is comparable to or larger than prior scales.

	PFPT	FFT-MoE
Batch size	16	128
Encoder	ViT-B/32	ViT-B/16
Optimizer	Adam ($\beta = (0.9, 0.98), \epsilon = 1e^{-6}$)	Adam (weight decay=$1e^{-2}$)
Learning rate	$1e^{-4}$	$3e^{-4}$
Local epochs	5	1
Total rounds	50	50
Active client ratio	0.1 (10/100)	1.0 (full)

Table 8: Detailed training configurations of the baseline methods. PFPT: number of tokens = 10. FFT-MoE: num_experts = 8, rank_per_expert = 2, top-k = 1, aux loss $\lambda = 10^{-5}$.

B.2 Baseline Parameters

Use of Large Language Models

We used large language models (e.g., ChatGPT) in two limited ways: (i) to polish the writing and improve readability, and (ii) to aid in the discovery of related work. No parts of the conceptual design, experiments, or analysis were generated by LLMs.