Fine-Tuning Impairs the Balancedness of Foundation Models in Long-tailed Personalized Federated Learning

Supplementary Material

This appendix provides further details, results, and analyses that could not be included in the main paper owing to space constraints. The content is organized as follows:

- Sec. A provides the notation table for FedPuReL.
- Sec. B presents the complete algorithm description.
- Sec. C provides additional analysis of FedPuReL, including complementary branch contributions and gradient alignment dynamics.
- Sec. D presents further experimental results, including robustness to data heterogeneity, hyperparameter studies, and convergence analysis.
- Sec. E discusses the key insights and implications of FedPuReL.

A. Notation Table

We provide the notation table in Tab. 5.

Table 5. Notation table for FedPuReL.

Notation	Description
Federated Learning Setup	
K	Number of clients
$\mathcal{D}_k$	Local dataset of client k
n_k	Number of samples in client k
$\mathcal{S}_t$	Set of selected clients at round t
C	Number of classes
IF	Imbalance factor (n_1/n_C)
α_{dir}	Dirichlet parameter for heterogeneity
Model Parameters	
$\boldsymbol{W}$	Frozen CLIP backbone parameters
$\boldsymbol{\phi}_g$	Global shared PEFT parameters
$\boldsymbol{\phi}_k$	Personalized PEFT parameters for client k
$f_{\text{img}}(\cdot)$, $f_{\text{text}}(\cdot)$	Image and text encoders
Predictions and Logits	
$\mathbf{x}$, y	Input image and label
$\mathbf{z}(\mathbf{x})$	Zero-shot logits
$\mathbf{f}(\mathbf{x})$	Fine-tuned logits
$\mathbf{l}_{\text{G}}(\mathbf{x})$	Global branch logits
$\mathbf{l}_{\text{P}}^k(\mathbf{x})$	Personalized branch logits for client k
Metrics and Temperature	
τ, τ_{zs}, τ_{ft}	Temperature parameters
$D_{\text{TKL}}(\cdot\|\|\cdot)$	Temperature-aligned KL divergence
$\beta(V)$	Balancedness metric
$H(\cdot)$	Entropy function
Gradients and Losses	
$\mathbf{g}_{\text{task}}$	Task gradient from cross-entropy loss
$\mathbf{g}_{\text{align}}$	Alignment gradient from TKL divergence
$\tilde{\mathbf{g}}_{\text{task}}$	Purified task gradient
$\mathcal{L}_{\text{align}}$	Alignment loss
$\mathcal{L}_{\text{fusion}}^k$	Fusion loss for additive combination
$\mathcal{L}_{\text{personal}}^k$	Personalization loss
λ	Weight balancing fusion and personal losses

B. Algorithm

We provide the algorithm description in Algorithm 1.

C. Additional Analysis of FedPuReL

Analysis 4: Global and personalized branches exhibit complementary contributions. Our design separates balanced global knowledge from personalized local adaptation, which should lead to different contribution patterns across class types. Specifically, head classes with abundant samples should rely more on global knowledge, while tail classes with scarce data should require stronger personalized corrections.

To evaluate this, we decompose each correct prediction to identify the dominant contributor between the global and personalized branches. For each correctly classified sample x with ground-truth label y, we compare the logits from both branches: $l_G^{(y)}(x)$ from the global branch and $l_P^{(y)}(x)$ from the personalized branch. We attribute the prediction to the branch with the higher logit value for the correct class, i.e., the prediction is attributed to the global branch if $l_G^{(y)}(x) > l_P^{(y)}(x)$, and to the personalized branch otherwise. We then compute the percentage of correct predictions attributed to each branch across all classes, grouped by their sample frequency.

As illustrated in Fig. 5b, as class frequency decreases from head to tail, the personalized branch contribution increases, while the global branch contribution decreases correspondingly. Notably, even for tail classes where personalization dominates, the global branch maintains a substantial 20-25% contribution. These results confirm our design principle: the global branch provides consistent balanced knowledge across all classes, while the personalized branch dynamically adjusts its contribution to compensate for data scarcity.

Analysis 5: Gradient Alignment Dynamics. Fig. 7 visualizes the angle between task gradient $\mathbf{g}_{\text{task}}$ and alignment gradient $\mathbf{g}_{\text{align}}$ during training. Baseline methods ex-

Algorithm 1 FedPuReL

Require: Number of clients K, global rounds T, personalization rounds T_p, balancedness weight λ, frozen CLIP weights $\boldsymbol{W}$.

Ensure: Global PEFT parameters $\boldsymbol{\phi}_g$, personalized parameters $\{\boldsymbol{\phi}_k\}_{k=1}^K$.

// Phase 1: Global Balanced Training
Initialize $\boldsymbol{\phi}_g^{(0)}$
for $t = 1, \ldots, T$ **do**
  Select active clients $\mathcal{S}_t \subseteq [K]$
  for all client $k \in \mathcal{S}_t$ **in parallel do**
    Receive $\boldsymbol{\phi}_g^{(t)}$ from server
    Compute $\mathbf{z}(\mathbf{x})$, $\mathbf{f}(\mathbf{x}; \boldsymbol{W}, \boldsymbol{\phi}_g)$
    $\mathbf{g}_{\text{align}} \leftarrow \nabla_{\boldsymbol{\phi}_g} D_{\text{TKL}}(\sigma_{\tau_{zs}}(\mathbf{z}) \| \sigma_{\tau_{ft}}(\mathbf{f}))$
    $\mathbf{g}_{\text{task}} \leftarrow \nabla_{\boldsymbol{\phi}_g} \text{CE}(y, \sigma(\mathbf{f}))$
    // Gradient Purification
    if $\langle \mathbf{g}_{\text{task}}, \mathbf{g}_{\text{align}} \rangle < 0$ **then**
      $\tilde{\mathbf{g}}_{\text{task}} \leftarrow \mathbf{g}_{\text{task}} - \frac{\langle \mathbf{g}_{\text{task}}, \mathbf{g}_{\text{align}} \rangle}{\|\mathbf{g}_{\text{align}}\|^2} \mathbf{g}_{\text{align}}$
    else
      $\tilde{\mathbf{g}}_{\text{task}} \leftarrow \mathbf{g}_{\text{task}}$
    end if
    Update $\boldsymbol{\phi}_g$ with $\tilde{\mathbf{g}}_{\text{task}}$
    Upload $\boldsymbol{\phi}_g^{k,(t)}$ to server
  end for
  $\boldsymbol{\phi}_g^{(t+1)} \leftarrow \sum_{k \in \mathcal{S}_t} \frac{n_k}{\sum_{j \in \mathcal{S}_t} n_j} \boldsymbol{\phi}_g^{k,(t)}$
end for
// Phase 2: Personalized Residual Learning
Freeze $\boldsymbol{\phi}_g^{(T)}$
for all client $k \in [K]$ **in parallel do**
  Initialize $\boldsymbol{\phi}_k$
  for $t_p = 1, \ldots, T_p$ **do**
    $\mathbf{l}_{\text{G}} \leftarrow f(\mathbf{x}; \boldsymbol{W}, \boldsymbol{\phi}_g^{(T)})$, $\mathbf{l}_{\text{P}}^k \leftarrow f(\mathbf{x}; \boldsymbol{W}, \boldsymbol{\phi}_k)$
    $\mathcal{L}^k \leftarrow (1-\lambda)\text{CE}(y, \sigma(\mathbf{l}_{\text{G}} + \mathbf{l}_{\text{P}}^k)) + \lambda \text{CE}(y, \sigma(\mathbf{l}_{\text{P}}^k))$
    Update $\boldsymbol{\phi}_k$ with $\nabla_{\boldsymbol{\phi}_k} \mathcal{L}^k$
  end for
end for
return $(\boldsymbol{W}, \boldsymbol{\phi}_g^{(T)})$, $\{(\boldsymbol{W}, \boldsymbol{\phi}_g^{(T)}, \boldsymbol{\phi}_k)\}_{k=1}^K$

hibit convergence to approximately 90° (orthogonal), reflecting the well-known property that high-dimensional random vectors tend to be mutually orthogonal [5]. This orthogonality explains balance degradation, as task optimization proceeds without consideration of zero-shot alignment. In contrast, FedPuReL consistently maintains obtuse angles (90°), revealing that the original task gradient actively conflicts with balance preservation and would otherwise compromise zero-shot knowledge. Our purification operation detects these conflicts and removes incompatible components, ensuring updates remain aligned with balanced representations.

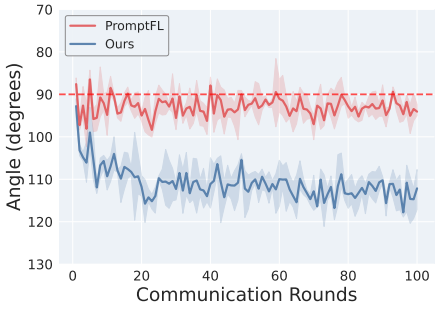


Figure 7. Angle between $\mathbf{g}_{\text{task}}$ and $\mathbf{g}_{\text{align}}$ during training on CIFAR-100-LT.

D. Additional Experiments

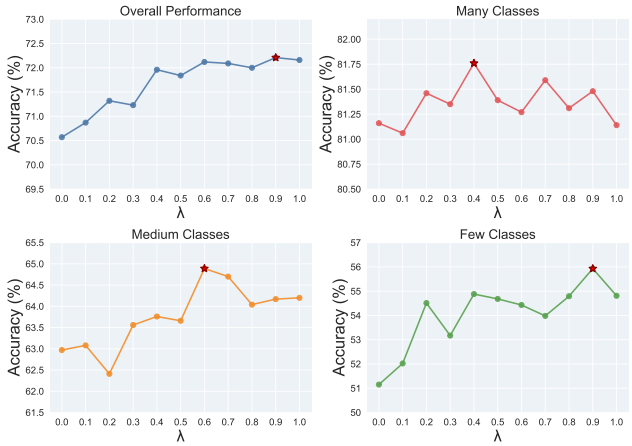
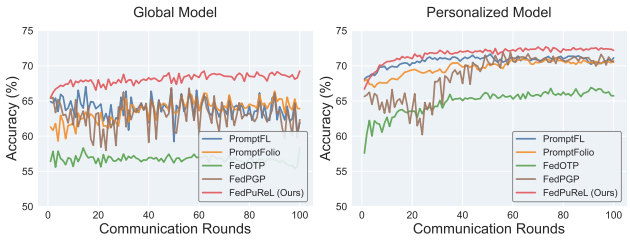


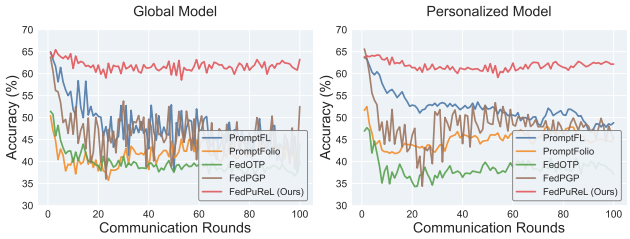
Figure 8. **Ablation study on λ in Eq. 12.** Performance across overall, Head, Mid, and Tail classes on CIFAR-100-LT with varying λ values. Red stars mark optimal points.

Ablation of TKL vs. Standard KL. Fig. 6 in the main paper qualitatively demonstrates that TKL distinguishes balanced adaptation from biased divergence, whereas standard KL exhibits large values even on balanced data due to confidence shifts. To further validate this quantitatively, we replace TKL with standard KL divergence in the gradient purification module and evaluate on CIFAR-100-LT. As shown in Table 6, TKL consistently outperforms standard KL in both accuracy and balancedness across global and personalized settings. This confirms that temperature alignment is essential for isolating structural distributional shift from benign confidence growth during fine-tuning, enabling more effective gradient purification.

Robustness to Data Heterogeneity. We evaluate FedPuReL's robustness across varying degrees of client heterogeneity by adjusting the Dirichlet parameter $\alpha \in \{0.1, 0.5, 1, 5\}$, where smaller α indicates higher non-IID heterogeneity. As shown in Table 7, FedPuReL consis-



(a) Overall Accuracy



(b) Few Class Accuracy

Figure 9. **Convergence of average accuracy** compared to Prompt-based SOTA method on CIAFR-100-LT

Table 6. **Ablation of TKL vs. standard KL** on CIFAR-100-LT.

Metric	Accuracy (%) GM	Accuracy (%) PM	Balancedness (%) GM	Balancedness (%) PM
KL	67.92	71.62	25.03	27.84
TKL (ours)	**69.77**	**73.37**	**27.87**	**30.02**

tently outperforms all baselines across different heterogeneity levels in both global and personalized settings. This robustness stems from our gradient purification mechanism, which anchors local updates to the zero-shot foundation model, effectively preventing divergence caused by heterogeneous local distributions. The consistent performance across all α values demonstrates that FedPuReL successfully preserves balanced knowledge while adapting to diverse client-specific distributions.

Hyperparameter Study. We analyze the impact of the balancing coefficient λ in Eq. 12, which controls the trade-off between fusion loss $\mathcal{L}^k_{\text{fusion}}$ and personalization loss $\mathcal{L}^k_{\text{personal}}$. As shown in Fig. 8, we observe that λ exhibits different optimal values across class groups: $\lambda = 0.9$ achieves the best overall and tail class performance, while moderate values favor head and mid classes. This demonstrates that higher λ values, which emphasize the personalized branch, are particularly beneficial for tail classes that require stronger client-specific corrections. Notably, head classes maintain relatively stable performance across different λ values, benefiting from the robust global branch. We set $\lambda = 0.9$ as the default in our experiments to balance overall accuracy with tail class performance.

Model Convergence Analysis. Fig. 9 visualizes the training dynamics of FedPuReL compared to state-of-the-art prompt-based methods on CIFAR-100-LT. For overall accuracy (Fig. 9a), FedPuReL maintains stable performance throughout training in both global and personalized settings. In contrast, baseline methods exhibit significant fluctuations and fail to achieve comparable performance. More critically, for few-shot class accuracy (Fig. 9b), FedPuReL maintains consistently superior performance on tail classes, while baselines show severe degradation or instability. This stability stems from our gradient purification mechanism, which prevents the model from drifting away from the balanced zero-shot anchor during adaptation. The consistent gap between FedPuReL and baselines throughout training validates that preserving balanced knowledge is essential for sustained performance on long-tailed federated data.

Table 7. **Comparison with state-of-the-art methods** on CIFAR-100-LT under different degrees of data heterogeneity (α) for global models (GM) and personalized models (PM).

CIFAR-100-LT / Method	$\alpha = 0.1$ GM	$\alpha = 0.1$ PM	$\alpha = 0.5$ GM	$\alpha = 0.5$ PM	$\alpha = 1$ GM	$\alpha = 1$ PM	$\alpha = 5$ GM	$\alpha = 5$ PM
Zero-shot	64.82	64.79	64.82	64.79	64.82	64.79	64.82	64.79
Prompt-based								
PromptFL [14]	63.72	72.18	65.25	71.97	61.98	70.52	64.95	71.38
PromptFolio [33]	64.32	72.83	64.79	72.19	64.10	70.54	65.68	71.10
FedOTP [22]	57.27	70.87	55.75	68.28	58.02	65.23	54.88	65.77
FedPGP [8]	61.66	73.10	62.12	72.56	62.96	67.83	63.41	70.41
PromptFolio+Fed-GraB [45]	64.97	75.13	65.43	73.26	65.14	71.79	66.23	72.18
FedPuReL (ours)	**68.88**	**77.72**	**69.05**	**74.97**	**69.77**	**73.37**	**70.54**	**73.71**
	↑ 3.91	↑ 2.59	↑ 3.62	↑ 1.71	↑ 4.63	↑ 1.58	↑ 4.31	↑ 1.53
LoRA-based								
CLIPLoRA [49]	74.83	78.53	75.50	78.75	75.02	77.80	76.12	79.12
FedSA-LoRA [12]	73.53	80.47	73.41	79.16	74.48	81.68	73.03	76.91
CLIPLoRA+Fed-GraB [45]	75.30	79.12	76.43	78.97	76.35	77.91	77.31	79.62
FedPuReL (ours)	**76.71**	**81.76**	**76.74**	**80.29**	**77.56**	**84.62**	**78.20**	**80.27**
	↑ 1.41	↑ 1.29	↑ 0.31	↑ 1.13	↑ 1.21	↑ 2.94	↑ 0.89	↑ 0.65
Adapter-based								
FedClip [28]	63.92	68.02	64.54	68.64	64.09	66.87	64.35	68.19
FedClip+Fed-GraB [45]	64.69	68.18	65.72	68.98	65.40	67.20	66.02	68.49
FedPuReL (ours)	**66.58**	**68.98**	**66.83**	**69.38**	**67.54**	**68.42**	**67.81**	**69.05**
	↑ 1.89	↑ 0.80	↑ 1.11	↑ 0.40	↑ 2.14	↑ 1.22	↑ 1.79	↑ 0.56

E. Discussion

FedPuReL preserves balanced knowledge in foundation models while enabling effective personalization. Gradient purification maintains balancedness close to zero-shot model throughout training (cf. Analysis 1 in Sec. 4.5). Residual learning then enables unbiased personalization through output-space offsets: Analysis 4 reveals complementary contributions where the global branch provides consistent knowledge across all classes while the personalized branch dynamically compensates for data scarcity in tail classes. Additionally, unlike methods requiring explicit class priors [17, 45], our approach leverages implicit balanced priors in foundation models, eliminating the need for class distribution that pose privacy risks while exceeding prior-based performance (Sec. 4.2). By reducing distributional divergence and client drift (Analysis 2–3), FedPuReL achieves superior convergence without explicit rebalancing.