

A PROOF OF THEOREM

In the following, we proof Proposition 1. For convenience, we restate the assumptions and proposition.

Assumptions 1. *The following conditions hold for each client $i \in [m]$ at round t :*

1. *Each loss function $\mathcal{L}_i^{\text{local}}$ and $\mathcal{L}_i^{\text{global},t}$ is $L(1+e)^{-1}$ -smooth.*

2. *The gradient estimator g_i^t is unbiased and has bounded variance:*

$$\mathbb{E}[g_i^t] = \nabla \mathcal{L}_i^t(\theta_t), \quad \mathbb{E}[\|g_i^t - \nabla \mathcal{L}_i^t(\theta_t)\|^2] \leq \sigma^2.$$

3. *The global loss has bounded gradients: $\|\nabla \mathcal{L}_i^{\text{global},t}(\theta)\| \leq G$ for all θ and t .*

4. *The objective drift is bounded:*

$$|\mathcal{L}_i^{t+1}(\theta) - \mathcal{L}_i^t(\theta)| \leq \delta, \quad \forall \theta.$$

5. *The per-sample gradient variance is bounded:*

$$\begin{aligned} \mathbb{E}_{x \in D_i} \left[\left\| \nabla_{\theta} \ell(\theta, x, \hat{y}^t) - \nabla \mathcal{L}_i^{\text{local},t} \right\|^2 \right] &\leq \bar{\sigma}^2 \\ \mathbb{E}_{x \in U} \left[\left\| \nabla_{\theta} \ell(\theta, x, \hat{y}^t) - \nabla \mathcal{L}_i^{\text{global},t} \right\|^2 \right] &\leq \tilde{\sigma}^2 \end{aligned}$$

With these assumptions, FEDMOSAIC converges to a stationary point.

Proposition 1 (Convergence of FEDMOSAIC). *Let each client's objective at round t be*

$$\mathcal{L}_i^t(\theta) = \mathcal{L}_i^{\text{local}}(\theta) + \lambda_i^t \mathcal{L}_i^{\text{global},t}(\theta), \text{ where } \lambda_i^t = \exp \left(-\frac{\mathcal{L}_i^{\text{global}}(\theta_t) - \mathcal{L}_i^{\text{local},t}(\theta_t)}{\mathcal{L}_i^{\text{local},t}(\theta_t)} \right),$$

and $\mathcal{L}_i^{\text{global},t}$ may change at each round due to pseudo-label updates. Under Assumptions 1-5, for a fixed step size $0 < \eta \leq (2L)^{-1}$ and $\min_i |D_i| = d$, after T rounds of FEDMOSAIC, it holds that

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2] \leq \frac{4L(\mathcal{L}_i^0 - \mathcal{L}_i^*)}{T} + \frac{\bar{\sigma}^2}{2Ld} + \frac{e^2 \tilde{\sigma}^2}{2L|U|} + 2\delta.$$

Proof. Since $\mathcal{L}_i^{\text{local}}$ and $\mathcal{L}_i^{\text{global},t}$ are $L(1+e)^{-1}$ -smooth, and since during optimization steps $\lambda_i^t < e$ is fixed, the Lipschitz constant of \mathcal{L}_i^t is

$$L(1+e)^{-1} + \lambda_i^t L(1+e)^{-1} \leq L(1+e)^{-1} + eL(1+e)^{-1} = L.$$

Thus, the standard descent lemma (Bottou et al., 2018) gives:

$$\mathbb{E}[\mathcal{L}_i^t(\theta_{t+1})] \leq \mathbb{E}[\mathcal{L}_i^t(\theta_t)] - \eta \mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2] + \frac{L\eta^2}{2} \mathbb{E}[\|g_i^t\|^2].$$

To bound $\mathbb{E}[\|g_i^t\|^2]$, expand

$$\mathbb{E}[\|g_i^t\|^2] = \mathbb{E}[\|g_i^t - \nabla \mathcal{L}_i^t(\theta_t) + \nabla \mathcal{L}_i^t(\theta_t)\|^2] \leq 2\sigma^2 + 2\mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2],$$

and substitute into the descent inequality to obtain

$$\mathbb{E}[\mathcal{L}_i^t(\theta_{t+1})] \leq \mathbb{E}[\mathcal{L}_i^t(\theta_t)] - \eta \mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2] + L\eta^2 (\sigma^2 + \mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2]).$$

Rearranging terms yields

$$\mathbb{E}[\mathcal{L}_i^t(\theta_{t+1})] \leq \mathbb{E}[\mathcal{L}_i^t(\theta_t)] - \eta(1 - L\eta) \mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2] + L\eta^2 \sigma^2.$$

This step requires $\eta \leq (2L)^{-1} < L^{-1}$ to ensure that the coefficient $(1 - L\eta)$ is positive. We now account for the fact that the function changes between rounds, i.e.,

$$\mathbb{E}[\mathcal{L}_i^{t+1}(\theta_{t+1})] \leq \mathbb{E}[\mathcal{L}_i^t(\theta_{t+1})] + \delta,$$

which gives

$$\mathbb{E}[\mathcal{L}_i^{t+1}(\theta_{t+1})] \leq \mathbb{E}[\mathcal{L}_i^t(\theta_t)] - \eta(1 - L\eta)\mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2] + L\eta^2\sigma^2 + \delta.$$

Summing from $t = 0$ to $T - 1$ and rearranging yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2] \leq \frac{\mathcal{L}_i^0 - \mathcal{L}_i^T}{(1 - L\eta)\eta T} + \frac{L\eta^2\sigma^2}{1 - L\eta} + \frac{\delta}{1 - L\eta}.$$

Denoting the minimum loss as \mathcal{L}_i^* , i.e., $\forall t, \mathcal{L}_i^t \geq \mathcal{L}_i^*$ yields the formal result

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2] \leq \frac{\mathcal{L}_i^0 - \mathcal{L}_i^*}{(1 - L\eta)\eta T} + \frac{L\eta^2\sigma^2}{1 - L\eta} + \frac{\delta}{1 - L\eta}.$$

Since $((1 - L\eta)\eta)^{-1}$, $L\eta^2/(1 - L\eta)$, and $(1 - L\eta)^{-1}$ have a maximum at $(2L)^{-1}$ for $\eta \leq (2L)^{-1}$, we can upper bound this by

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2] \leq \frac{4L(\mathcal{L}_i^0 - \mathcal{L}_i^*)}{T} + \frac{\sigma^2}{4L} + 2\delta.$$

Since $g_i^t = g_i^{local,t} + \lambda_i^t g_i^{global,t}$, we decompose σ^2 in round t at client i as $2\bar{\sigma}^2 + 2(\lambda_i^t)^2 \tilde{\sigma}^2$, and further bound

$$\begin{aligned} \sigma_{global}^2 &\leq \frac{\mathbb{E}_{x \in D_i} [\|\nabla_{\theta} \ell(\theta, x, \hat{y}^t) - \nabla \mathcal{L}_i^{local,t}\|^2]}{\min_i |D_i|} \\ &\quad + \sup_{i,t} (\lambda_i^t)^2 \frac{\mathbb{E}_{x \in U} [\|\nabla_{\theta} \ell(\theta, x, \hat{y}^t) - \nabla \mathcal{L}_i^{global,t}\|^2]}{|U|} \\ &\leq \frac{2\bar{\sigma}^2}{d} + \frac{2e^2 \tilde{\sigma}^2}{|U|}, \end{aligned}$$

since $\sup_{i,t} (\lambda_i^t)^2 = e^2$ and using Assumption 5. With this, we obtain

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2] \leq \frac{4L(\mathcal{L}_i^0 - \mathcal{L}_i^*)}{T} + \frac{\bar{\sigma}^2}{2Ld} + \frac{e^2 \tilde{\sigma}^2}{2L|U|} + 2\delta.$$

□

B ADDITIONAL EMPIRICAL EVALUATION

Robustness under Misleading Global Knowledge To further evaluate FEDMOSAIC’s adaptivity, we conducted an experiment designed to test its behavior when the global consensus signal is actively misleading for a particular client. We constructed a scenario using CIFAR-10 dataset with 5 clients, where client 0 was assigned flipped labels so effectively training on corrupted data. This setup results in the global pseudo labels being systematically misaligned with this client’s local distribution. As expected the client’s local model suffers a significantly higher loss when trained using the global pseudo labels compared to its own data, leading to a near zero value of λ . This confirms the intended behavior of FEDMOSAIC: when the global signal is detrimental, the client autonomously reduces its reliance on it, effectively opting out of harmful collaboration. Fig.3 illustrates this behavior by showing the divergence between global and local loss for the corrupted client (client 0) in comparison to a non-corrupted one (client 1). Fig. 4 shows the evolution of the adaptive weight λ across communication rounds for all 5 clients.

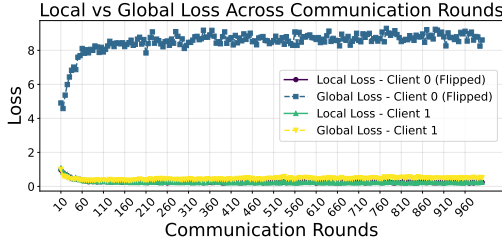


Figure 3: Local Vs Global loss across communication rounds on CIFAR-10.

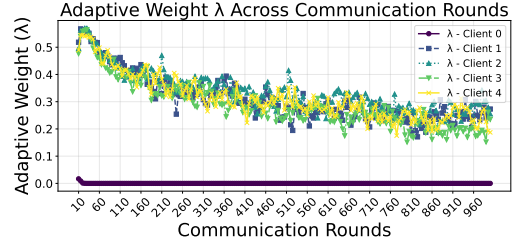


Figure 4: Adaptive weight λ across communication rounds on CIFAR-10.

Personalization vs. Local training In Low-collaboration Regimes While FEDMOSAIC consistently archives the highest accuracy across both pathological and practical label skew settings (Table 2), the margin between its performance and that of local training is notably small. This observation raises a critical insight. In such scenarios, where each client’s local distribution is highly disjoint and local alignment provides limited benefit, personalization through collaboration may be unnecessary or even detrimental. Indeed, FEDMOSAIC’s adaptive mechanism reflects this reality. The per-client weighting strategy reduces reliance on the global information when it does not align with local data. This is evident in Fig. 5 and Fig. 6, which show that the global loss remains consistently higher than the local loss for many clients, leading to near zero value of the adaptive weight λ as seen in Fig. 7. In such cases, FEDMOSAIC defaults to local training behavior, effectively opting out of collaboration when it offers no advantage. This reinforces the methods’ robustness as it personalizes only when beneficial, and falls back to local training when collaboration yields little or a negative return. To ensure numerical stability in the computation of the adaptive coefficient

$$\lambda_i^t = \exp \left(- \frac{\mathcal{L}_i^{\text{global}}(\theta_t) - \mathcal{L}_i^{\text{local},t}(\theta_t)}{\mathcal{L}_i^{\text{local},t}(\theta_t)} \right),$$

, we add a small constant ϵ to the denominator to prevent division by zero.

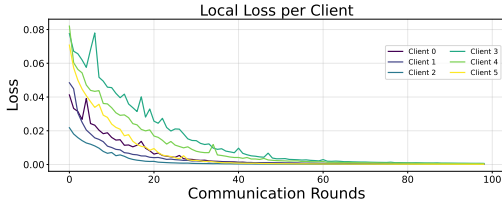


Figure 5: Local loss across communication rounds on Fashion-MNIST for the first 6 clients.

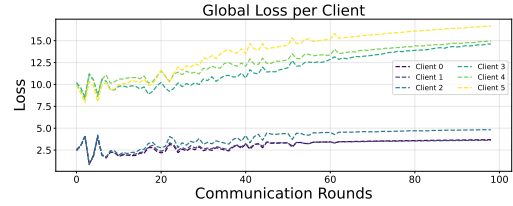


Figure 6: Global loss across communication rounds on Fashion-MINST for the first 6 clients.

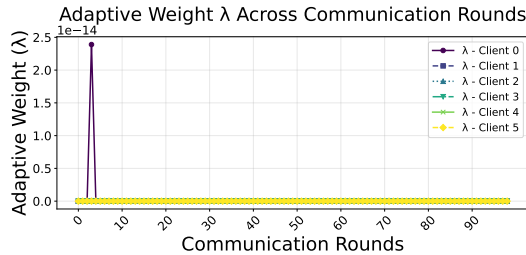


Figure 7: Adaptive weight λ across communication rounds on Fashion-MINST for the first 6 clients.

The Effect of the Unlabeled Dataset : As mentioned in sec. 4, FEDMOSAIC relies heavily on a shared, unlabeled public dataset $|U|$. To understand how sensitive FEDMOSAIC is to this dataset’s

characteristics, we conducted a study on CIFAR-10 dataset focusing on two critical questions: First, how does the amount of available data affect performance? Second, does it matter whether the class distribution is balanced (IID) or heavily skewed?

IMPACT OF PUBLIC DATASET SIZE : We evaluated the performance of FEDMOSAIC using different sizes of the public unlabeled dataset, with $|U|$ set to 3000, 2000, 1000, 500 and 250. For this experiment, the public dataset was always sampled in an IID fashion to ensure all classes were present. The results, summarized in Fig.8, show that the performance of FEDMOSAIC is remarkably stable. Even as the size of the public dataset is reduced by over 90% (from 3000 to 250 samples), the drop in final test accuracy is minimal. This finding suggests that the collaboration mechanism does not require a large volume of public unlabeled data. As long as a small class-representative set of examples is available, clients can effectively share knowledge and build high-quality personalized models.

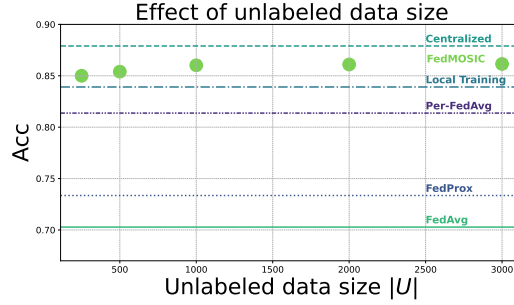


Figure 8: Test accuracy (ACC) of FEDMOSAIC under different unlabeled dataset size $|U|$

IMPACT OF PUBLIC DATASET DISTRIBUTION : Next, we studied the effect of the public unlabeled dataset distribution. We simulated varying degrees of distribution skew by sampling $|U|$ (with a fixed size of 3,000) using a Dirichlet distribution. We tested different values of the concentration parameter $\alpha = 1, 0.7, 0.5, 0.3, 0.1$, where $\alpha = 1$ corresponds to a perfectly IID distribution and lower values induce increasingly severe skew.

As shown in Fig.9 and Fig.10, we observe a degradation in performance as the public dataset become more skewed. The most significant drop occurs at very low α values (e.g., 0.3, 0.1), where some classes are absent from U . In such cases, the global consensus offers no useful information for clients whose private data contains the missing classes.

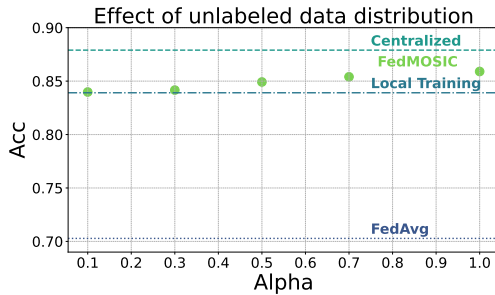


Figure 9: Test accuracy (ACC) of FEDMOSAIC under different distribution of U .

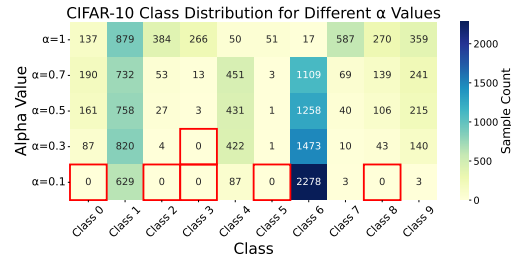


Figure 10: Class distribution of U under different values of alpha.

However, the most crucial finding is that the performance of FEDMOSAIC never drops below the local training baseline. This demonstrates the robustness of the adaptive aggregation scheme. When the global signal becomes irrelevant or misleading, the dynamic loss weight λ automatically steers clients to disregard it, effectively defaulting to local training. This acts as a critical fail-safe, ensuring that collaboration is never actively detrimental, even when the public data is of poor quality.

A Note on the Byzantine Resilience of FEDMOSAIC Following the argument by (Jiang et al., 2020), who show that federated semi-supervised learning with soft labels sharing (e.g., FedDistill) is more Byzantine resilient than FEDAVG due to the bounded nature of the threat vector on the probability simplex, we argue that FEDMOSAIC exhibits similar (if not stronger) resilience properties. Like FedCT (Abourayya et al., 2025), FEDMOSAIC relies on hard label sharing, further constraining the threat vector to a binary classification decision per example. Moreover, FEDMOSAIC incorporates confidence-based aggregation, which naturally downweights unreliable predictions. This mechanism provides an additional layer of robustness by reducing the influence of low confidence (and potentially malicious) clients. While a formal analysis remains open, these properties suggest that FEDMOSAIC may be at least as Byzantine resilient as FedDistill and FedCT. Exploring this direction further is promising for future work.

C DETAILS ON EXPERIMENTS

All experiments are conducted for a sufficient number of communication rounds until convergence, using three different random seeds. While the standard deviation across the three runs with different seeds is consistently small, this observation aligns with prior work Zhang et al. (2023d), Zhang et al. (2023c), Zhang et al. (2023b).

Label Skew Fashion-Minst and CIFAR-10 datasets have been used for label skew experiments. In Fashion-Minst, we converted the raw grayscale 28×28 images into Pytorch tensors and normalized pixel values to the range $[-1, 1]$ using a mean of 0.5 and standard deviation of 0.5. In CIFAR-10, we converted RGB 32×32 images into Pytorch tensors of shape $[3, 32, 32]$ and normalizes each color channel independently to the range of $[-1, 1]$, using a mean of 0.5 and standard deviation of 0.5. The data is partitioned across 15 clients. In a pathological non-IID setting, each client receives data from only 2 out of 10 classes. In a practical non-IID setting, data is distributed across 15 clients using a Dirichlet distribution. This creates naturally overlapping, imbalanced label distributions among clients. Training data distribution of each scenario of CIFAR-10 are showing in Fig.11 and Fig.12.

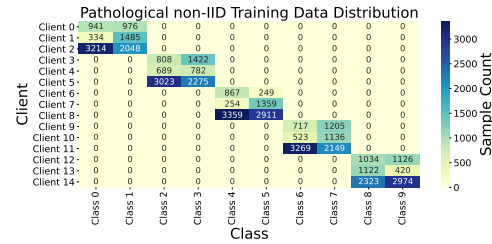


Figure 11: CIFAR-10 clients data distribution in Pathological non-IID setting

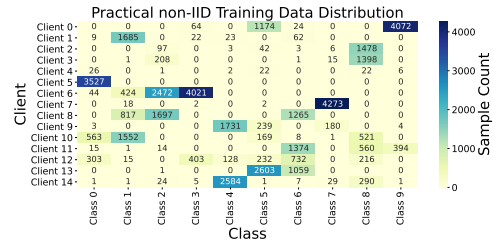


Figure 12: CIFAR-10 clients data distribution in Practical non-IID setting

Feature Shift we used the Office-10 and DomainNet datasets. For both, we adopt AlexNet as a neural network architecture. Input images are resized to $256 \times 256 \times 3$. Training is performed till convergence using the corss-entropy loss and Adam optimizer with learning rate of 10^{-2} . We use a batch size of 32 for Office-10 dataset and 64 for DomainNet. For DomainNet, which originally contains 345 categories, we restrict the label space to the top 10 most frequent classes to reduce complexity, The selected categories are: bird, feather, headphones, icecream, teapot, tiger, whale, windmill, wineglass, zebra. For Office-10, each client get one of the 4 domains and For DomainNet dataset, each client get one of the 6 domains. The distribution of each client training data are showing in Fig.13 and Fig.14.

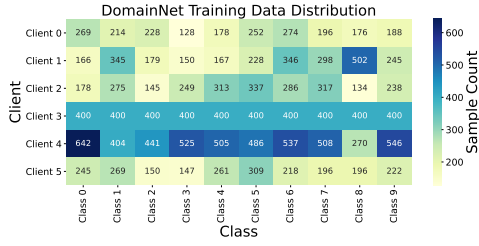


Figure 13: DomainNet clients data distribution.

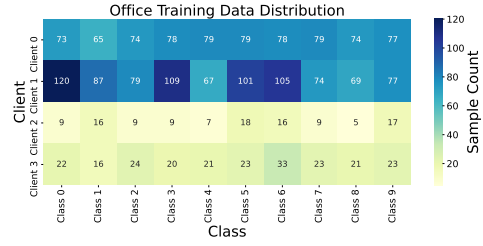


Figure 14: Office-10 clients data distribution.

Hybrid Distribution We simulate the hybrid data distribution by combining both label distribution skew and feature distribution shift. We use the same two datasets as in feature shift experiments: Office-10 and DomainNet. To introduce label skew, for each domain, we randomly sample 5 clients and assign to each client only 2 out of 10 total classes. This results in 20 clients for the Office-Caltech10 dataset (4 domains \times 5 clients) and 30 clients for DomainNet (6 domains \times 5 clients). This creates a hybrid non-IID setting where clients differ significantly in both input distribution and output distribution. We use the same preprocessing and training configurations as the feature shift experiments. All input images are resized to $256 \times 256 \times 3$ before being fed into *AlexNet*. Models are trained using cross-entropy loss and Adam optimizer with learning rate of 10^{-2} . The batch size is set to 32 for Office-10 and 64 for DomainNet. For DomainNet, we selected the 10 most frequent as feature shift experiments. To effectively visualize the distribution of local training data across 30 clients, we used a dot matrix plot, which offers a compact and intuitive representation of client-level variation. The visualization of the Clients distribution of DomainNet and Office-10 datasets are shown in Fig.15 and Fig.16

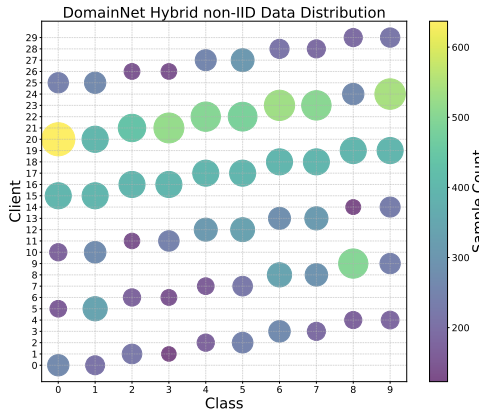


Figure 15: DomainNet clients Hybrid data distribution.

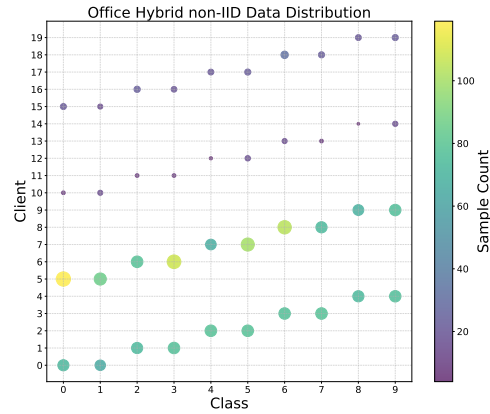


Figure 16: Office-10 clients Hybrid data distribution.

D PRACTICAL IMPACT OF FEDMOSAIC

FEDMOSAIC addresses data heterogeneity in personalized federated learning (PFL) via a fine-grained collaboration mechanism that lets each client selectively rely on collective expertise, aiming to improve accuracy and robustness. This is particularly relevant in domains with substantial variability (e.g., healthcare, finance, recommendation), where traditional federated methods can struggle. Empirically, FEDMOSAIC often outperforms strong PFL baselines and, in our evaluated settings, local and centralized training across label skew, feature shift, and hybrid heterogeneity; where margins are small, it performs comparably. Its design limits disclosure by sharing only hard predictions on a shared unlabeled dataset, reducing potential privacy leakage relative to parameter sharing. This follows “share as little as possible” (Mian et al., 2023; Tan et al., 2022) and aligns with privacy-by-design (Cavoukian et al., 2009). In addition, our differentially private variant (DP-

FEDMOSAIC) illustrates how to obtain formal (ϵ, δ) -DP guarantees for the released signals (labels and expertise), with the privacy accounting provided and empirical calibration left to future work. Finally, federated co-training is communication-efficient for large models: when parameter counts vastly exceed $|U|$, sending hard labels (and one expertise scalar per example) can reduce uplink by orders of magnitude. Combining this with communication-efficient protocols (Kamp et al., 2016; Kamp, 2019) has the potential to reduce communication by several orders of magnitude, in particular for large transformer-based models, such as LLMs.

E NOTATION

Federated Learning Setup

m	Number of participating clients
$i \in [m]$	Index of a client
D_i	Private dataset of client i
U	Shared public unlabeled dataset used for co-training
T	Total number of communication rounds
b	Communication period (local steps between rounds)
A_i	Local learning algorithm used by client i

Models and Predictions

h_i^t	Local model of client i at round t
$L(h, D)$	Loss of model h on dataset D
$\ell_{\text{priv}} = L(h_i^{t-1}, D_i)$	Private loss on client i 's local data
$\ell_{\text{pseudo}} = L(h_i^{t-1}, P^t)$	Loss on pseudo-labeled public data P^t
$L_i^t \in \{0, 1\}^{ U \times C}$	One-hot prediction matrix from client i on public data
$E_i^t \in (0, \infty)^{ U }$	Confidence (expertise) vector from client i on public data
$S^t = \sum_{i=1}^m \text{diag}(E_i^t) \cdot L_i^t$	Weighted score matrix used for consensus aggregation
$L^t[j] = \arg \max_{c \in [C]} S^t[j, c]$	Consensus pseudo-label for public example $x_j \in U$

Adaptive Weighting Mechanism

λ_i^t	Adaptive weight controlling trust in global signal for client i at round t
$\ell = \ell_{\text{priv}} + \lambda_i^t \cdot \ell_{\text{pseudo}}$	Total loss used for local model update at round t

Optimization and Convergence

θ	Model parameters
$\nabla L(\theta)$	Gradient of loss with respect to model parameters
σ^2	Bounded variance of local gradient estimator
$\tilde{\sigma}^2$	Bounded variance of global gradient estimator (pseudo-label noise)
δ	Bounded drift in local objectives across rounds
L	Smoothness constant (Lipschitz constant of the gradient)

Sets and Indexing

1026	$[m] = \{1, \dots, m\}$	Index set of all clients
1027		
1028	$[C] = \{1, \dots, C\}$	Index set of all classes
1029	$x_j \in U$	j -th public unlabeled sample
1030		
1031	y_j	True (unknown) label of public sample x_j
1032	$ U $	Number of samples in the public dataset U
1033	$ D_i $	Number of samples in the local dataset of client i
1034		
1035	$L_i^t[j, c]$	(j, c) -th entry of prediction matrix L_i^t
1036	$E_i^t[j]$	Confidence of client i on public example x_j
1037		
1038		
1039		
1040		
1041		
1042		
1043		
1044		
1045		
1046		
1047		
1048		
1049		
1050		
1051		
1052		
1053		
1054		
1055		
1056		
1057		
1058		
1059		
1060		
1061		
1062		
1063		
1064		
1065		
1066		
1067		
1068		
1069		
1070		
1071		
1072		
1073		
1074		
1075		
1076		
1077		
1078		
1079		