

FedMerge: Federated Personalization via Model Merging

Shutong Chen¹ Tianyi Zhou² Guodong Long¹ Jing Jiang¹ Chengqi Zhang¹

Abstract

One global model in federated learning (FL) might not be sufficient to serve many clients with non-IID tasks and distributions. While there has been advances in FL to train multiple global models for better personalization, they only provide limited choices to clients so local finetuning is still indispensable. In this paper, we propose a novel “FedMerge” approach that can create a personalized model per client by simply merging multiple global models with automatically optimized and customized weights. In FedMerge, a few global models can serve many non-IID clients, even without further local finetuning. We formulate this problem as a joint optimization of global models and the merging weights for each client. Unlike existing FL approaches where the server broadcasts one or multiple global models to all clients, the server only needs to send a customized, merged model to each client. Moreover, instead of periodically interrupting the local training and re-initializing it to a global model, the merged model aligns better with each client’s task and data distribution, smoothening the local-global gap between consecutive rounds caused by client drift. We evaluate FedMerge on three different non-IID settings applied to different domains with diverse tasks and data types, in which FedMerge consistently outperforms existing FL approaches, including clustering-based and mixture-of-experts (MoE) based methods.

1. Introduction

Federated learning (FL) enables decentralized and collaborative learning of models across various clients without sharing their local data. By exchanging only local models instead of raw data, FL ensures data privacy and security while

¹Australian Artificial Intelligence Institute, FEIT, University of Technology Sydney ²University of Maryland. Correspondence to: Shutong Chen <shutong.chen@student.uts.edu.au>, Tianyi Zhou <tianyi@umd.edu>.

leveraging distributed datasets for improved performance. In particular, local datasets across clients are often non-identically and independently distributed (non-IID), making the design of algorithms to address such non-IIDness a fundamental challenge.

Model Merging (Yadav et al., 2024) is a knowledge ensemble technique that integrates multiple models into a single unified model. It has recently gained popularity in applications involving foundation models (Yang et al., 2024a; Huang et al., 2023; Chronopoulou et al., 2023). Given a large number of foundation models with diverse knowledge domains, model merging facilitates the combination of their knowledge into a new, enriched knowledge space. This approach effectively reuses pre-trained models while avoiding the heavy computational cost of training models from scratch. These reused models can encompass a wide range of techniques, including Parameter-Efficient Fine-Tuning (PEFT) methods such as Low-Rank Adaptation (LoRA). Due to its low computational overhead, LoRA is particularly well-suited for FL, and several studies have explored its applications in this domain (Yi et al., 2023; Yang et al., 2024b; Sun et al., 2024; Cho et al., 2024).

Model merging is especially relevant to FL because it supports both the similarities and differences among the merged models. In FL, clients share commonalities that enable collaborative training while maintaining unique data distributions that reflect their personalized knowledge. This characteristic makes model merging an ideal approach for FL, as it facilitates the integration of shared knowledge across clients by sharing groups of models while preserving individual differences through personalized merging weights.

A number of FL methods with multiple models have been proposed (Yi et al., 2024; Wu et al., 2023; Marfoq et al., 2021a). However, these methods share similar optimization strategies and fail to address the contradiction between the large number of models and the limited resources available to clients. Consider a “model soup” on the server containing a diverse set of models. Each client utilizes models from this model soup to enhance its personalized training. As shown in Fig. 1a, previous methods require transmitting the entire model soup to each client, which is often impractical due to the large number of global models. Selecting a subset of models from the model soup (Chronopoulou et al., 2023)

can alleviate this issue, but it typically requires task-specific knowledge, making it less generalizable. In contrast, model merging enables each client to leverage the model soup by applying personalized merging weights, thereby forming its own customized model.

In this paper, we propose a novel and general framework for Federated Learning with Model Merging. We introduce a simple merging strategy that decouples the size of the model soup from the constraints of limited client resources. Specifically, instead of transmitting the entire model soup and merging weights to clients, we transmit only a single merged model for each client. As shown in Fig. 1b, the model merging process is performed on the server rather than on each client, and only the merged models are communicated, eliminating the need to share the original models with clients. In this way, all the computational and storage costs introduced by model merging are shifted to the server, leaving each client to communicate and optimize only a single model. This approach allows us to cheaply increase the size of the model soup on the server to boost performance, while each client only handles a single model.

FedMerge is an end-to-end framework that jointly optimizes the merging weights and global models from scratch. However, FedMerge introduces a unique challenge to FL: model collapse. When the parameters of multiple models are collapsed into a single merged model through weight averaging, clients are optimized only on the merged models, making it impossible to directly update the merging weights or global models on the clients. To address this challenge, we leverage local gradients as intermediate variables and derive a backpropagation-like update strategy for both the merging weights and the global models. This novel update formula forms the basis of the FedMerge algorithm. Experimental results demonstrate that the FedMerge algorithm surpasses traditional FL methods as well as Parameter-Efficient Fine-Tuning (PEFT)-based FL methods on synthetic data distributions and more realistic scenarios.

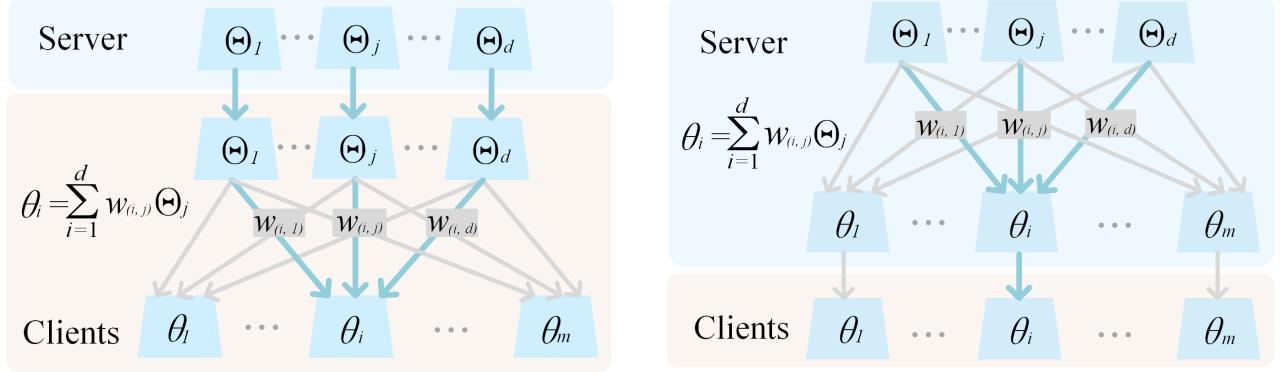
We summarize the main contributions of this paper:

- We propose a novel framework for Federated Learning with multiple models, addressing the conflict between large parameter scales and limited client resources.
- We introduce the Federated Merging (FedMerge) framework, which provides personalized models for each client by merging global models with personalized combining weights.
- Experimental results demonstrate that FedMerge enables a cost-effective way to improve performance by increasing the number of global models without significantly affecting system costs.

2. Related Work

Model Merging Model merging refers to methods aimed at reusing the knowledge contained in multiple models. The fundamental technique of model merging is to integrate existing models to form new ones. There have been a number of works focusing on merging methodologies, such as simple averaging (Wortsman et al., 2022), merging with Fisher Information (Matena & Raffel, 2022), and merging with task vectors (Ilharco et al., 2022). Recently, model merging has become closely related to the development of foundation models (Sukhbaatar et al., 2024; Lu et al., 2023; 2024; Ostapenko et al., 2024a), as such developments have brought numerous reusable expert foundation models online (Face, 2023). For example, LoraHub (Huang et al., 2023) trains one LoRA expert per task on a collection of 200 tasks and combines these experts to evaluate downstream tasks. Due to the large number of expert models, model selection plays a practical role in choosing the most relevant experts for each downstream task (Zhao et al., 2024; Chronopoulou et al., 2023). Model merging is also related to Mixture of Experts (MoE)-based methods, where the design of routing weights is a key factor. The router can be neural network-based (Lu et al., 2023), feature similarity-based (Wang et al., 2024), or even SVD-decomposed (Ostapenko et al., 2024b). In this paper, rather than directly adopting the various model merging techniques mentioned above in FL, we leverage the fundamental concept of model merging to address the primary challenge in Federated Learning with Multiple Models, enabling a large model soup to be efficiently reused on the server while communicating only a single model to each client.

Federated Learning with Multiple Models The use of multiple models in FL arises from the fact that sharing a single model cannot effectively address the Non-IID problem. Personalized FL methods (Tan et al., 2021; Li et al., 2021; T Dinh et al., 2020; Collins et al., 2021) apply the concept of multiple models by designing a unique model for each client’s personalized data distribution. Our method is more related to a type of multi-model FL, where multiple models are shared on the server (Marfoq et al., 2021a; Wu et al., 2023). In FedEM (Marfoq et al., 2021a), a group of global models is transmitted to each client, and each client optimizes its personalized combination weights. pFedMoE (Yi et al., 2024) employs one local expert and one global expert for each client, while deriving routing weights using personalized router neural networks. Recently, a few methods have incorporated merging concepts into FL (Chen et al., 2024; Salami et al., 2024). However, these methods mainly use model merging as part of their solutions and do not provide a unified framework for model merging in the context of federated learning.



(a) The architecture of **Federated Learning with Mixture of Experts**. Each client receives d global models from the server and optimizes them separately. The client's resource consumption is proportional to the number of global models, making it inefficient.

(b) The architecture of the proposed **Federated Learning with Model Merging (FedMerge)**. The merging operation (weight averaging) is performed on the server side instead of on each client. Only the merged models, θ_i , are communicated and optimized, making the cost irrelevant to the number of global models.

Figure 1: Comparison between MoE-like FL methods and the proposed FedMerge. Both method perform weight averaging over global models. MoE-like FL perform weight averaging on each client, while FedMerge perform weight averaging on the server.

3. Methodology

3.1. Problem Formulation

In Federated Learning, clients collaboratively learn from each other via a central server. Let (Y_i, X_i) represent client- i 's personalized data. Each client trains local models using (Y_i, X_i) . Models carrying diverse local knowledge are transmitted from each client to the server, and the server aggregates these local models to centralize the knowledge. The objective of a fundamental FL method, FedAvg, is as follows:

$$\min_{\Theta} \sum_{i=1}^m \frac{n_i}{n} \ell(Y_i, f(X_i; \Theta)) \quad (1)$$

The m clients' objective functions are combined based on the proportion of each client's data n_i relative to the total data n , collaboratively optimizing a single global model Θ .

In practice, local data across clients are often non-IID (non-identically independently distributed). If we assume that each client owns a personalized model, the personalized FL objective becomes:

$$\min_{\theta} \sum_{i=1}^m \frac{n_i}{n} \ell(Y_i, f(X_i; \theta_i)) \quad (2)$$

where each client optimizes its personalized model θ_i on its own data. The personalized models are often related to some global knowledge.

3.2. Objective Function

In this section, we present the objective and learning algorithm of FedMerge.

The core idea of model merging is to reuse a set of models by combining them to create new ones. Let $\{\Theta_1, \Theta_2, \dots, \Theta_d\}$ be a set of d global models, where each model is learned with respect to its own knowledge space. The merging process for creating model θ_i is defined as:

$$\theta_i = \sum_{j=1}^d w_{(i,j)} \cdot \Theta_j.$$

Here, $w_{i,1:d}$ represents the task-specific merging weights used to create θ_i , determining the contribution of each global model to the merged model. By merging global models into new models, their knowledge spaces are reused and integrated, contributing to various downstream tasks.

In federated learning, model merging can be used to address non-IIDness. A model soup containing a set of global models is maintained on the server and reused across all clients. Each client's personalized model is created by selecting and weight-combining models from the model soup. The objective of FedMerge is given as:

$$\min_{\theta} \sum_{i=1}^m \frac{n_i}{n} \ell(Y_i, f(X_i; \theta_i)) \quad (3)$$

$$\text{s.t. } \theta_i = \sum_{j=1}^d w_{(i,j)} \cdot \Theta_j \quad (4)$$

where each client's personalized model is created by combining a set of global models using its personalized merging weights.

A commonly employed MoE-like (Mixture of Experts) strategy (Yi et al., 2024; Wu et al., 2023; Marfoq et al., 2021a) in previous methods is to substitute Eq (4) into Eq (3), resulting in directly optimizing w and Θ on each client. As

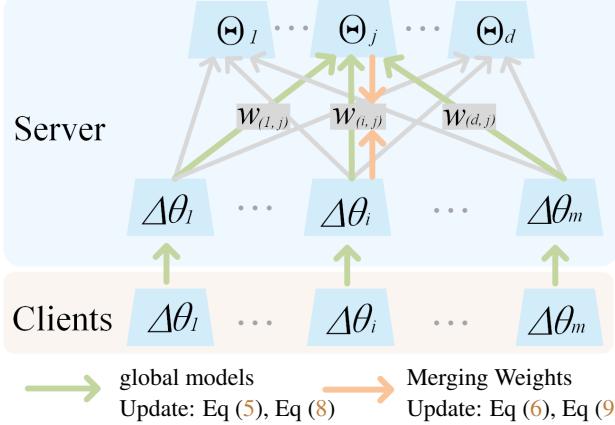


Figure 2: The update information flow of global models and merging weights for FedMerge. The update information for each global model is derived from the updates of all clients. The update information for each merging weight comes from the global model and merged model it connects.

shown in Fig. 1a, at each communication round, the MoE-like strategy transmits all the global models to each client, and each client optimizes these global models alongside its personalized merging weights. On the server side, each global model performs FedAvg-like aggregation separately. However, this operation requires broadcasting all the global models to each client, which is impractical when the model soup is large. A potential solution is to use a model selection strategy (Bhuyan & Moharir, 2022), selecting a subset of models from the model soup. While model selection can help alleviate communication costs, it still requires broadcasting multiple models and does not fundamentally solve the problem. The key issue with MoE-like methods is that they omit the merged model θ .

In our method, we assign the merged model θ the significance it deserves. Specifically, as shown in Fig. 1b, the merging operation $\theta_i = \sum_{j=1}^d w_{i,j} \cdot \Theta_j$ is performed on the server rather than on each client, with both the global models and personalized merging weights stored on the server. Only the merged model θ is transmitted between the client and the server.

Model merging offers several benefits to federated learning. First, by adjusting the merging weights, it enables the creation of different models tailored to each client’s needs, making it well-suited for heterogeneous federated learning where personalized models are required. Second, by sharing the same group of models, FedMerge allows for a balance between personalization and knowledge sharing, achieving both a certain degree of variation and a certain degree of commonality across clients. Finally, only the merged models are transmitted between the client and the server, which means FedMerge is client-resource friendly. The communicated parameters per round and the optimized parameters per client are equivalent to those in FedAvg. Personalized models are dynamically created on the server, and no local

models are remained on each client.

3.3. Optimizing via Backpropagation-like Strategy

Solving Eq (3) with Eq (4) requires optimizing w and Θ . For MoE-like methods, the gradients with respect to w and Θ are direct and clear: all the global models are sent to each client, and their gradients can be directly computed on each client. However, in the proposed FedMerge, each client only receives the merged model, and only the gradient of the merged model $\frac{\partial L}{\partial \theta}$ is directly available. The weights w and global models Θ remain on the server, and their gradients cannot be computed directly.

The update of merging weights w and global models Θ is conceptually similar to backpropagation in neural networks. If we consider FedMerge as a single-layer fully connected network, where local models θ are high-level nodes and global models Θ are low-level nodes, then w acts as the propagation weights. The gradient of w and Θ depends on the gradient of θ . Following the chain rule, we compute $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial \Theta}$ using $\frac{\partial L}{\partial \theta}$ as an intermediate variable:

$$\frac{\partial L}{\partial \Theta_j} = \sum_{i=1}^m \frac{n_i}{n} w_{i,j} \frac{\partial \ell}{\partial \theta_i}. \quad (5)$$

$$\frac{\partial L}{\partial w_{i,j}} = \frac{n_i}{n} \langle \Theta_j, \frac{\partial \ell}{\partial \theta_i} \rangle. \quad (6)$$

Eq (5) and Eq (6) are derived directly from the objective of FedMerge in Eq (3) with Eq (4). The key factor is the use of local updates $\frac{\partial L}{\partial \theta}$ as an intermediate variable. For the detailed derivation process, please refer to Appendix C. To better illustrate Eq (5) and Eq (6), Fig. 2 highlights the information flows from local models to global models and merging weights. It shows that the updates of each global model are influenced by all client updates, while the updates for each merging weight depend on both the global model and the merged model it connects.

We summarize the training procedure as follows:

- 1. Forward Propagation:** The server computes the merged models for each client using Eq. (4) and distributes them accordingly. Each client then updates its merged model based on its local objective function:

$$\theta_i^{(t)} - \theta_i^{(t-1)} \leftarrow -\eta_\theta \nabla_{\theta_i^{(t-1)}} \ell(Y_i, f(X_i; \theta_i^{(t-1)})), \quad (7)$$

where η_θ is the learning rate for θ . Notably, the local update procedure in FedMerge follows the same approach as in FedAvg.

- 2. Backward Propagation:** After the local update, each client transmits its gradient $\frac{\partial \ell}{\partial \theta_i}$ or the updated parameters $\Delta \theta_i = \theta_i^{(t)} - \theta_i^{(0)}$ back to the server. The server

then updates both the global models Θ and the merging weights w . Using the chain rule, the gradients for the merging weights and global models are computed based on local updates, as formulated in Eq. (5) and Eq. (6).

3. **Parameter Update:** Once the gradients for the global models and merging weights are obtained, the server iteratively updates Θ and w via gradient descent:

$$\Theta_j \leftarrow \Theta_j - \eta_\Theta \frac{\partial L}{\partial \Theta_j}, \quad (8)$$

$$w_{(i,j)} \leftarrow w_{(i,j)} - \eta_w \frac{\partial L}{\partial w_{(i,j)}}, \quad (9)$$

where η_Θ and η_w are the learning rates for Θ and w , respectively. The overall FedMerge process is summarized in Algorithm 1.

3.4. Technical Implementations

Constraint on Merging Weights Values The original objective of FedMerge in Eq (4) has no constraints on merging weights, allowing the merging weight values to be negative. However, neural network models typically require regularized parameters. To address this, we apply a softmax function to the merging weights to obtain normalized weights. This introduces slight modifications to the gradients in Eq (5) and Eq (6). Detailed derivations can be found in Appendix D.

Clarification of Merging Weights Gradient In Eq (6), the update of merging weights involves the inner product of model parameters and gradients. For large models, the effective update directions can be overwhelmed by the inner products of a large number of parameters. To mitigate this issue, model compression techniques, such as pruning parameters before computing the inner product, can be employed to highlight the most significant gradient directions.

4. Experiments

4.1. Classical FL Settings

Baselines FedMerge is related to Cluster FL, Multi-model FL, and MoE-like FL. To analyze the performance of FedMerge, we consider popular methods as well as state-of-the-art methods as follows: 1) **Single-model methods**: These FL methods share one model on the server, including Local, FedAvg (McMahan et al., 2017), FedAvg+, and pFedMoE (Yi et al., 2024). Local denotes training local models separately on each client, while FedAvg+ denotes personalized fine-tuning of the global model from FedAvg. Although pFedMoE includes "MoE" in its name, it uses only one

global model. 2) **Multi-model methods**: These FL methods have flexible number of global models on the server, such as IFCA (Ghosh et al., 2020) and FedEM (Marfoq et al., 2021b). IFCA is a Cluster FL method, and FedEM is a MoE-like FL method.

Algorithm 1: Federated Merging (FedMerge)

```

Initialize : global models  $\{\Theta_1, \Theta_2, \dots, \Theta_d\}$ ,
merging weight  $w \in \mathbb{R}^{m \times d}$ 
for each communication round do
    Randomly select a subset  $A$  of clients;
    Merge models by Eq (4) for each client  $i \in A$ ;
    Send the merged model  $\theta_i$  to each client- $i$ ;
    for each selected client in parallel do
        Update  $\theta_i$  using local gradient descent via
        Eq (7) for  $t$  steps;
        Send local update  $\theta_i^{(t)} - \theta_i^{(0)}$  to the server;
    for each global iteration do
        Compute gradients for  $\Theta$  via Eq (5);
        Update  $\Theta$  via Eq (8);
        Compute gradients for  $w$  via Eq (6);
        Update  $w$  via Eq (9);

```

Parameter Scale Multi-model FL requires multiple global models. For a fair comparison, we vary the parameter scale across all baselines. For Single-model methods (i.e., Local, FedAvg, FedAvg+, and pFedMoE), we use ResNet-9 (He et al., 2016) as the basic model and vary the number of global models chosen from the set {5, 10, 15, 20, 25, 30}. For Multi-model methods (i.e., IFCA, FedEM, and FedMerge), we vary the model architecture by selecting from ResNet-9, ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. These models have parameter scales relative to ResNet-9 as follows: {1, 2.27, 4.32, 4.82, 8.73, 11.95}. We summarize our parameter scale settings in Table 1.

Non-IID Settings FedMerge adapts to different Non-IID settings by adjusting the merging weights adaptively. To validate this adaptability, we consider three data distributions using the CIFAR-100 (Krizhevsky et al., 2009) and PACS (Li et al., 2017) datasets: **Cluster Non-IID**: We split the original CIFAR-100 dataset into 5 clusters based on label categories. Each cluster is evenly assigned to a subgroup of clients out of 50 clients. The ground-truth clustering is $\{\{1, \dots, 6\}, \{7, \dots, 11\}, \{12, \dots, 19\}, \{20, \dots, 32\}, \{33, \dots, 50\}\}$. **Personalize Non-IID**: A commonly used Non-IID setting proposed by (Hsu et al., 2019), where the CIFAR-100 dataset is partitioned using a Dirichlet distribution into 50 clients. We set the concentration parameter $\alpha = 0.1$ for high Non-IIDness. **Domain Non-IID**: We use the PACS (Li et al., 2017) dataset, assigning each client to one domain, resulting in 4 clients.

Table 1: Parameter scale settings. Model Num. denotes number of models and Model Arch. is model architecture. The scale of parameters is measured as a multiple of the parameter count in ResNet-9.

Multi-Model	Model Num. Parameter Scale	ResNet-9 × 5 5	ResNet-9 × 10 10	ResNet-9 × 15 15	ResNet-9 × 20 20	ResNet-9 × 25 25	ResNet-9 × 30 30
Single-Model	Model Arch. Parameter Scale	ResNet-9 1	ResNet-18 2.27	ResNet-34 4.32	ResNet-50 4.82	ResNet-101 8.73	ResNet-152 11.95

Table 2: Comparison with Single-model Methods on Client-resource Constrained Scenarios. We use ResNet-9 for all methods.

	Non-IID	Local	FedAvg	FedAvg+	pFedMoE	FedMerge					
						5	10	15	20	25	30
Personal	38.76	30.29	53.54	54.28	37.88	37.88	46.62	51.18	53.52	53.99	54.38
Cluster	23.13	29.20	46.53	49.03	51.69	51.69	51.38	53.27	51.16	51.69	51.96
Domain	54.97	60.19	60.39	63.59	62.12	62.12	64.64	65.80	65.32	65.26	64.96

Table 3: Numerical Results for Multi-Model Methods by Varying global model Numbers. We use ResNet-9 for all methods.

Non-IID	Method	Global Model Num					
		5	10	15	20	25	30
Personal	IFCA	25.68	26.45	28.86	27.35	29.68	30.80
	FedEM	32.50	33.67	34.59	37.08	38.12	39.48
	FedMerge	37.88	46.62	51.18	53.52	53.99	54.38
Cluster	IFCA	42.64	41.06	44.97	40.34	37.53	38.04
	FedEM	48.47	48.34	46.35	46.16	46.51	46.65
	FedMerge	51.69	51.38	53.27	51.16	51.69	51.96
Domain	FedEM	61.77	61.33	64.63	61.23	62.93	63.14
	FedMerge	62.12	64.64	65.80	65.32	65.26	64.96

FedMerge Settings The learning rate of the global models Θ is set to 1 for all clients, meaning that at each round, the global model is replaced with the local update model. For updating the merging weights, we use a fixed-magnitude update strategy to ensure smooth parameter updates. Specifically, at each round, we perform a grid search on the gradient direction with small steps and stop when the update of the merging weight reaches 0.01.

System Settings For all methods, we split the dataset into training, validation, and test sets, selecting the checkpoint with the minimum validation error. The number of communication rounds is set to 500 to ensure convergence. Each method is executed three times, and the reported results are the average of these runs.

4.2. Numerical Comparison for Multi-Model Methods

As shown in Table 1, for IFCA, FedEM, and FedMerge, we vary the number of global models chosen from the set $\{5, 10, 15, 20, 25, 30\}$, using ResNet-9 (He et al., 2016) as the basic model. Note that we do not run IFCA on Domain Non-IID since having only 4 clients is too few for Cluster FL methods.

Table 3 suggests the following: 1) Multi-Model methods can

boost their performance by using a large number of models. For example, on Personal Non-IID, the performance continues to improve for all methods as the number of models increases. On Cluster Non-IID and Domain Non-IID, FedMerge achieves optimal performance with 15 global models. 2) FedMerge consistently surpasses FedEM and IFCA across various global model numbers, demonstrating the effectiveness of FedMerge compared to other Multi-Model FL methods.

4.3. Numerical Comparison for Single-Model Methods in Client-Resource-Constrained Scenarios

For Single-model methods (Local, FedAvg, FedAvg+, pFedMoE), performance can be improved by using larger models. However, increasing model scale significantly raises client resource consumption, which is a critical constraint in FL. To ensure a fair comparison, we first analyze the performance of Single-model methods under client-resource-constrained scenarios. Then, we analyze all methods under unlimited resource conditions.

FedMerge can cheaply boost performance by sharing a large number of models on the server without requiring much client resources. In Table 2, we compare FedMerge with Single-model methods under constrained client resources. For all baselines, each client communicates one ResNet-9 model with the server. All methods, except pFedMoE, optimize one model on each client, while pFedMoE optimizes two models. We assume server resources are unlimited, allowing FedMerge to boost performance by using a large number of global models. As shown in Table 2, on Cluster and Domain Non-IID, FedMerge consistently outperforms all baselines across all numbers of global models. On Personal Non-IID, although FedMerge performs poorly with a small number of global models (5), its performance consistently increases as the number of global models grows, eventually surpassing pFedMoE with 30 global models.

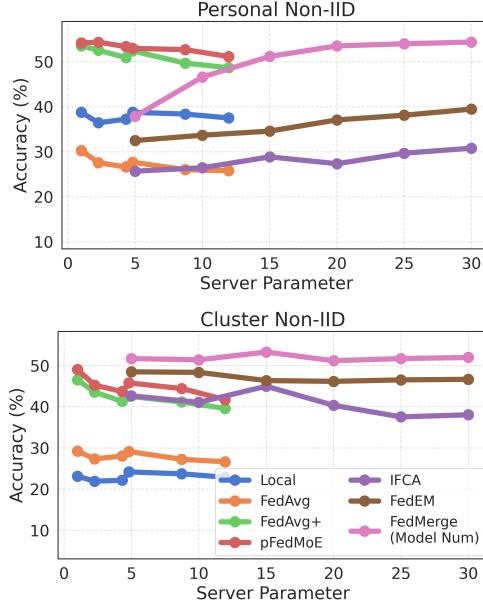


Figure 3: Server Parameter Usage vs. Accuracy. The x-axis represents multiples of ResNet-9’s parameter count. In FedMerge, we fix the architecture to ResNet-9 and vary the number of global models from [5, 10, 15, 20, 25, 30].

4.4. Server Parameter Usage vs. Performance

We analyze the parameter efficiency of both single-model and multi-model methods under the assumption of unlimited server resources. All methods follow the parameter scaling configurations outlined in Table 1.

Fig. 3 presents server parameter usage versus accuracy for Personal Non-IID and Cluster Non-IID settings. A comprehensive comparison across all Non-IID scenarios is provided in the Appendix. The results indicate the following:

FedMerge trades off higher server-side resource consumption for improved client-side efficiency. In Fig. 3, parameter-efficient methods, which achieve high performance with minimal server parameter usage, appear in the top-left corner. Multi-Model methods are less server resource-efficient compared to Single-model methods, as they utilize more parameters on the server while achieving comparable performance. However, FedMerge and IFCA require transmitting and optimizing only a single ResNet-9 per client. Despite FedMerge utilizing substantial server resources in Fig. 3, it remains client-efficient and achieves the highest performance among all baselines. Notably, pFedMoE assigns two models and an additional routing network to each client, whereas FedMerge allocates only a single model. As a result, FedMerge outperforms pFedMoE in Personal Non-IID with 30 global models and surpasses pFedMoE across all settings in Cluster Non-IID.

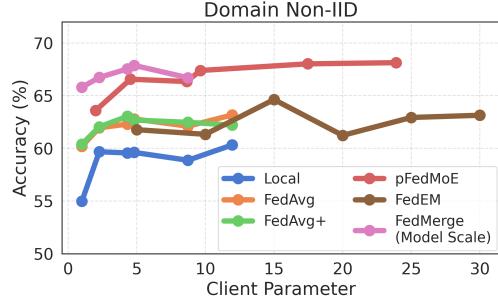


Figure 4: Client Parameter Usage vs. Accuracy. The x-axis represents multiples of ResNet-9’s parameter count. In FedMerge, we fix the number of global models at 15, all using the same architecture, selected from [ResNet-9, ResNet-18, ResNet-34, ResNet-50, and ResNet-101].

4.5. Client Parameter Usage vs. Performance

From Fig. 3, the performance of single-model methods degrades as model scale increases. This is because both Personal and Cluster Non-IID settings involve 50 clients, where the large number of clients with sparse client data limits performance gains from scaling up the model. In contrast, Domain Non-IID is more conducive to model scaling, as it consists of only four clients with abundant local data.

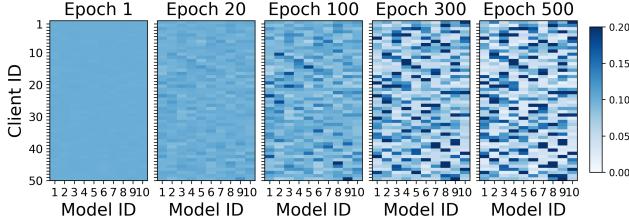
In Fig. 4, we present client parameter usage versus accuracy for Domain Non-IID. As shown, the performance of single-model methods improves significantly as model scale increases.

FedMerge further enhances performance through model scaling while remaining client resource-efficient.

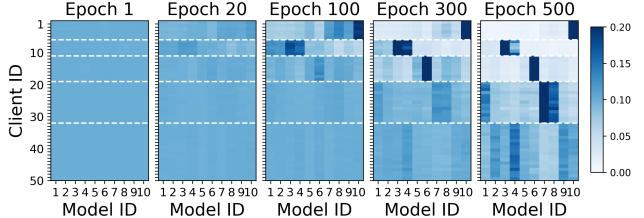
In Fig. 4, we evaluate FedMerge with a fixed set of 15 global models while varying the model architecture across [ResNet-9, ResNet-18, ResNet-34, ResNet-50, and ResNet-101]. Since FedMerge requires only a single model to be transmitted and optimized per client, it remains highly resource-efficient even as model scale increases. As demonstrated in Fig. 4, FedMerge outperforms pFedMoE by leveraging model scaling while incurring lower client-side resource consumption.

4.6. Ablation Study

FedMerge adapts to Non-IID settings with dynamic merging weights: To understand the behavior of FedMerge under different Non-IID settings, we visualized the heatmaps of merging weight matrices under different communication rounds for Personal Non-IID and Cluster Non-IID in Fig. 5. Each row represents the normalized merging weights for a single client. Initially, the merging weight matrix is uniformly initialized, and as the epoch progresses, it begins to exhibit specific distribution patterns. We separate the ground-truth clusters of Cluster Non-IID with white dashed lines. In Cluster Non-IID, rows of merging weights



(a) Heat map of merge weights for Personal Non-IID.



(b) Heat map of merge weights for Cluster Non-IID.

Figure 5: Visualization of merge weights with 10 global models and 50 clients. Each row denotes one normalized weight vector for each client. In Cluster Non-IID, we separate ground-truth clusters with white dashed lines.

are similar within each ground-truth cluster and distinct between different clusters. This indicates that FedMerge automatically allows clients with similar data distributions to prefer similar subgroups of global models, achieving a soft clustering. In contrast, for Personal Non-IID, rows of merging weights are relatively distinct from one another, and the merging weight matrix does not exhibit a clear structure. This irregularity aligns with the nature of Personal

Table 4: Ablation study on merging weights. Fixed indicates that each client randomly select some global models. The corresponding merging weight is uniformly initialized and frozen during training. ResNet-9 is used for all methods.

Non-IID	FedAvg	FedMerge (15 global models)			
		Fixed (1/15)	Fixed (5/15)	Fixed (10/15)	Dynamic
Personal	29.20	40.94	44.7	45.48	51.18
Cluster	30.29	32.55	32.52	34.06	53.27

Table 5: Numerical results on foundation models with multi-language datasets.

Dataset	Local	FedIT	FedLoRA	FedMerge
XNLI	0.698	0.770	0.769	0.783
XQUAD	0.561	0.559	0.556	0.566

Non-IID, where each client has a unique data distribution.

Effect of different weight merging strategies: We further analyze different weight merging strategies in Table 4. Specifically, each client randomly select a subset of models out of 15 global models. The corresponding merging weights are randomly initialized and frozen. As observed, The performance of Fixed strategy decreases compared with dynamic merging, and Cluster FL decreases more than Personal Non-IID because random selection conflicts with the structural information of the cluster.

4.7. Foundation-model FL Settings and Results

In federated learning (FL) with foundation models, we validate FedMerge with a group of tasks, where each client is assigned a distinct task. While these clients share commonalities, they also have distinct features that reflect their specific

styles. Specifically, we use datasets from multi-language tasks, XNLI (Conneau et al., 2018) and XQUAD (Artetxe et al., 2019). Each client is assigned a single language, and the content of the client data is identical across languages, differing only in the language used. For XNLI, we randomly select 300 samples for training and 200 samples for evaluation for each client. For XQUAD, we skip samples that exceed the maximum token length, randomly select 600 samples for training, and use the remaining samples for evaluation. The batch size is set to 8, and the maximum token length is 512.

We compare FedMerge with three baselines: Local, FedIT (Zhang et al., 2024), and FedLoRA (Yi et al., 2023). The “Local” baseline refers to models trained exclusively on local data without any communication between clients or the server. We adapt Alpaca-LoRA¹ as the base model, initialized with LLaMA3-8B (Dubey et al., 2024). The rank of LoRA is set to 8. For FedMerge, we use 5 global adapters. ROGUE-1 is used as the evaluation metric. The number of communication rounds is set to 20, and the checkpoint with the minimum validation error is selected.

As shown in Table 5, on XNLI, sharing information among clients significantly enhances performance (e.g., FedAvg vs. Local), indicating that knowledge from common language clients (e.g., English) helps the model understand rare language clients. On XQUAD, sharing information among clients using a single model does not enhance performance (FedAvg vs. Local). FedMerge consistently outperforms the compared methods on both multi-language datasets, demonstrating the effectiveness of FedMerge in FL with foundation models for multi-language tasks.

4.8. Conclusion

In this paper, we propose an end-to-end Model Merging framework for Federated Learning, where merging weights and global models are optimized jointly from scratch. The server manages all global models and merging weights, customizing personalized models for each client by weight-averaging the global models. FedMerge can boost perfor-

¹<https://github.com/tloen/alpaca-lora>

mance with a large model soup on the server, while still being client-resource friendly by only communicate and optimize a single model for each client.

FedMerge offers a promising direction for Federated Foundation Models: A large server knowledge pool can match foundation model scales, while each client selects and combines a small subset of parameters without transmitting the full model.

References

- Artetxe, M., Ruder, S., and Yogatama, D. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2019.
- Bhuyan, N. and Moharir, S. Multi-model federated learning. In *2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pp. 779–783. IEEE, 2022.
- Chen, M., Jiang, M., Zhang, X., Dou, Q., Wang, Z., and Li, X. Local superior soups: A catalyst for model merging in cross-silo federated learning. *arXiv preprint arXiv:2410.23660*, 2024.
- Cho, Y. J., Liu, L., Xu, Z., Fahrezi, A., and Joshi, G. Heterogeneous lora for federated fine-tuning of on-device foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12903–12913, 2024.
- Chronopoulou, A., Peters, M. E., Fraser, A., and Dodge, J. Adaptersoup: Weight averaging to improve generalization of pretrained language models. *arXiv preprint arXiv:2302.07027*, 2023.
- Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.
- Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Face, H. Hugging face - the ai community building the future. <https://huggingface.co>, 2023. Accessed: October 15, 2023.
- Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33: 19586–19597, 2020.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

- Huang, C., Liu, Q., Lin, B. Y., Pang, T., Du, C., and Lin, M. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023.
- Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.
- Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.
- Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.
- Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.
- Lu, K., Yuan, H., Lin, R., Lin, J., Yuan, Z., Zhou, C., and Zhou, J. Routing to the expert: Efficient reward-guided ensemble of large language models. *arXiv preprint arXiv:2311.08692*, 2023.
- Lu, Z., Fan, C., Wei, W., Qu, X., Chen, D., and Cheng, Y. Twin-merging: Dynamic integration of modular expertise in model merging. *arXiv preprint arXiv:2406.15479*, 2024.
- Marfoq, O., Neglia, G., Bellet, A., Kameni, L., and Vidal, R. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34:15434–15447, 2021a.
- Marfoq, O., Neglia, G., Bellet, A., Kameni, L., and Vidal, R. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34:15434–15447, 2021b.
- Matena, M. S. and Raffel, C. A. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.
- Ostapenko, O., Su, Z., Ponti, E. M., Charlin, L., Roux, N. L., Pereira, M., Caccia, L., and Sordoni, A. Towards modular llms by building and reusing a library of loras. *arXiv preprint arXiv:2405.11157*, 2024a.
- Ostapenko, O., Su, Z., Ponti, E. M., Charlin, L., Roux, N. L., Pereira, M., Caccia, L., and Sordoni, A. Towards modular llms by building and reusing a library of loras. *arXiv preprint arXiv:2405.11157*, 2024b.
- Salami, R., Buzzega, P., Mosconi, M., Bonato, J., Sabetta, L., and Calderara, S. Closed-form merging of parameter-efficient modules for federated continual learning. *arXiv preprint arXiv:2410.17961*, 2024.
- Sukhbaatar, S., Golovneva, O., Sharma, V., Xu, H., Lin, X. V., Rozière, B., Kahn, J., Li, D., Yih, W.-t., Weston, J., et al. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. *arXiv preprint arXiv:2403.07816*, 2024.
- Sun, Y., Li, Z., Li, Y., and Ding, B. Improving lora in privacy-preserving federated learning. *arXiv preprint arXiv:2403.12313*, 2024.
- T Dinh, C., Tran, N., and Nguyen, J. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.
- Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., and Zhang, C. Fedproto: Federated prototype learning over heterogeneous devices. *arXiv preprint arXiv:2105.00243*, 2021.
- Wang, H., Ping, B., Wang, S., Han, X., Chen, Y., Liu, Z., and Sun, M. Lora-flow: Dynamic lora fusion for large language models in generative tasks. *arXiv preprint arXiv:2402.11455*, 2024.
- Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022.
- Wu, Y., Zhang, S., Yu, W., Liu, Y., Gu, Q., Zhou, D., Chen, H., and Cheng, W. Personalized federated learning under mixture of distributions. In *International Conference on Machine Learning*, pp. 37860–37879. PMLR, 2023.
- Yadav, P., Raffel, C., Muqeeth, M., Caccia, L., Liu, H., Chen, T., Bansal, M., Choshen, L., and Sordoni, A. A survey on model merging: Recycling and routing among specialized experts for collaborative learning. *arXiv preprint arXiv:2408.07057*, 2024.
- Yang, E., Shen, L., Guo, G., Wang, X., Cao, X., Zhang, J., and Tao, D. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*, 2024a.

Yang, Y., Long, G., Shen, T., Jiang, J., and Blumenstein, M. Dual-personalizing adapter for federated foundation models. *arXiv preprint arXiv:2403.19211*, 2024b.

Yi, L., Yu, H., Wang, G., and Liu, X. Fedlora: Model-heterogeneous personalized federated learning with lora tuning. *arXiv preprint arXiv:2310.13283*, 2023.

Yi, L., Yu, H., Ren, C., Zhang, H., Wang, G., Liu, X., and Li, X. Fedmoe: Data-level personalization with mixture of experts for model-heterogeneous personalized federated learning. *arXiv preprint arXiv:2402.01350*, 2024.

Zhang, J., Vahidian, S., Kuo, M., Li, C., Zhang, R., Yu, T., Wang, G., and Chen, Y. Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919. IEEE, 2024.

Zhao, Z., Gan, L., Wang, G., Hu, Y., Shen, T., Yang, H., Kuang, K., and Wu, F. Retrieval-augmented mixture of lora experts for uploadable machine learning. *arXiv preprint arXiv:2406.16989*, 2024.

A. More Discussion on Parameter Usage vs. Accuracy

We analyze three different parameter usage standards: global parameter scale on the server (Server), communicated parameter scale for each client (Communicate), and stored parameter scale for each client (Client). We adopt the parameter scale settings in Table 1.

As a federated learning (FL) method with multiple models, FedMerge requires storing and optimizing a large number of parameters on the server, which limits its parameter efficiency in terms of global models compared to single-model FL methods. As shown in the first row of Fig. 6, FedMerge is less parameter-efficient on the server side than single-model methods. However, in FL, client resources are typically more valuable than server resources. FedMerge can utilize a large number of global models without significantly increasing client resource usage.

To illustrate this, we plot communicated parameter usage vs. accuracy and client-optimized parameter usage vs. accuracy in Fig. 6. From the second and third rows in Fig. 6, FedMerge and IFCA exhibit straight lines at the top-left corner. This is because these methods use only a single model for each client, enabling performance improvement without increasing client resource requirements.

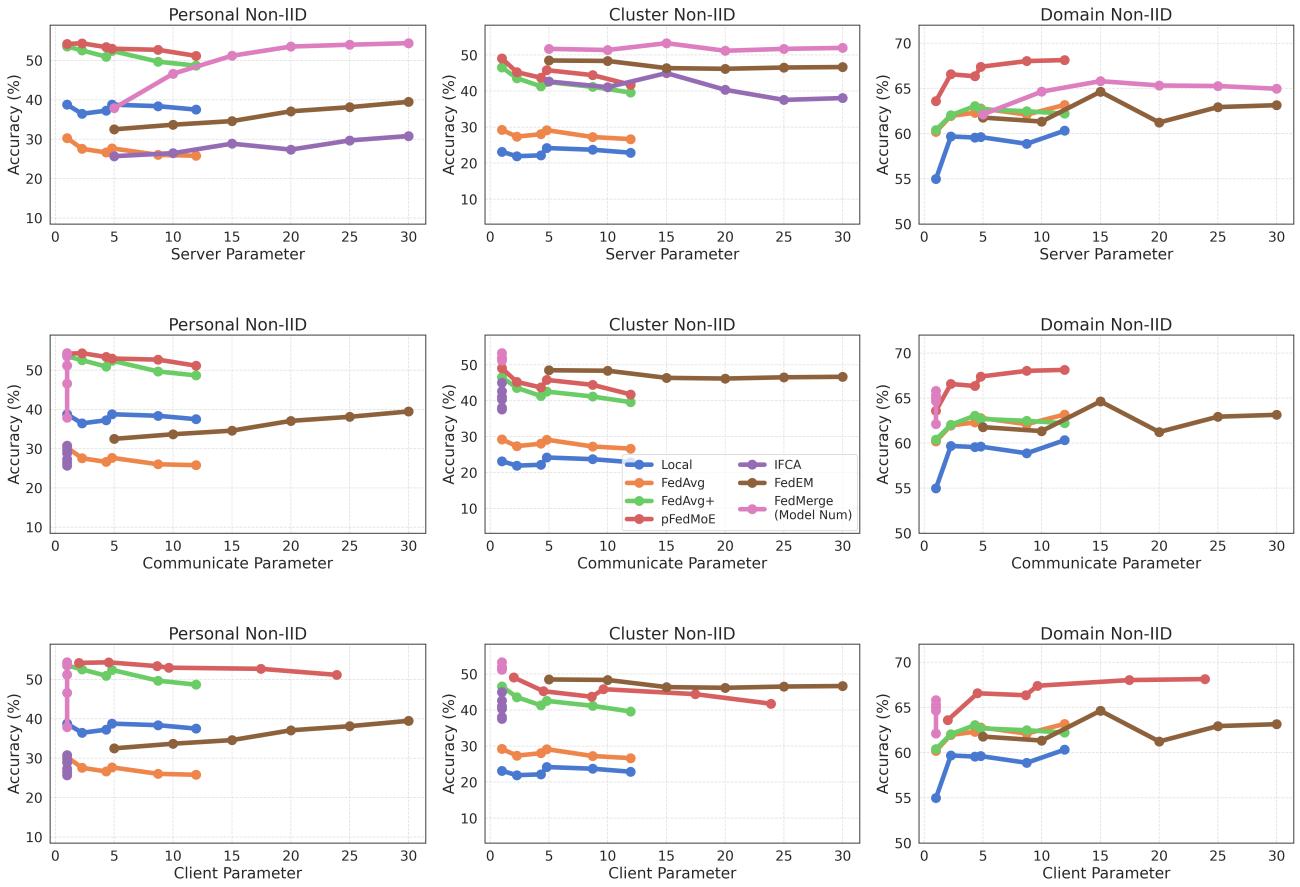


Figure 6: Parameter usage vs. accuracy across all Non-IID settings and parameter standards. "Server" refers to parameter usage on the server, "Communicate" denotes the number of parameters communicated by each client in a single round, and "Client" represents parameter usage for each client. Points closer to the top-left corner indicate higher performance with a smaller parameter scale, which is more efficient.

B. More Discussion on Merging Weight Matrix

We provide additional visualizations of the merging weight matrix in Fig. 7. Specifically, in Fig. 7b, the evolution speed of different clusters varies. Smaller clusters with fewer clients evolve faster and demonstrate distinct patterns within 200

epochs, as observed in the first three clusters with small client IDs. However, for the largest cluster (clients with IDs [32, ..., 50]), the merging weights evolve more slowly, even after 500 epochs. This is because the largest cluster has the highest number of clients (19 clients) and requires more global models, resulting in a more uniform merging weight distribution.

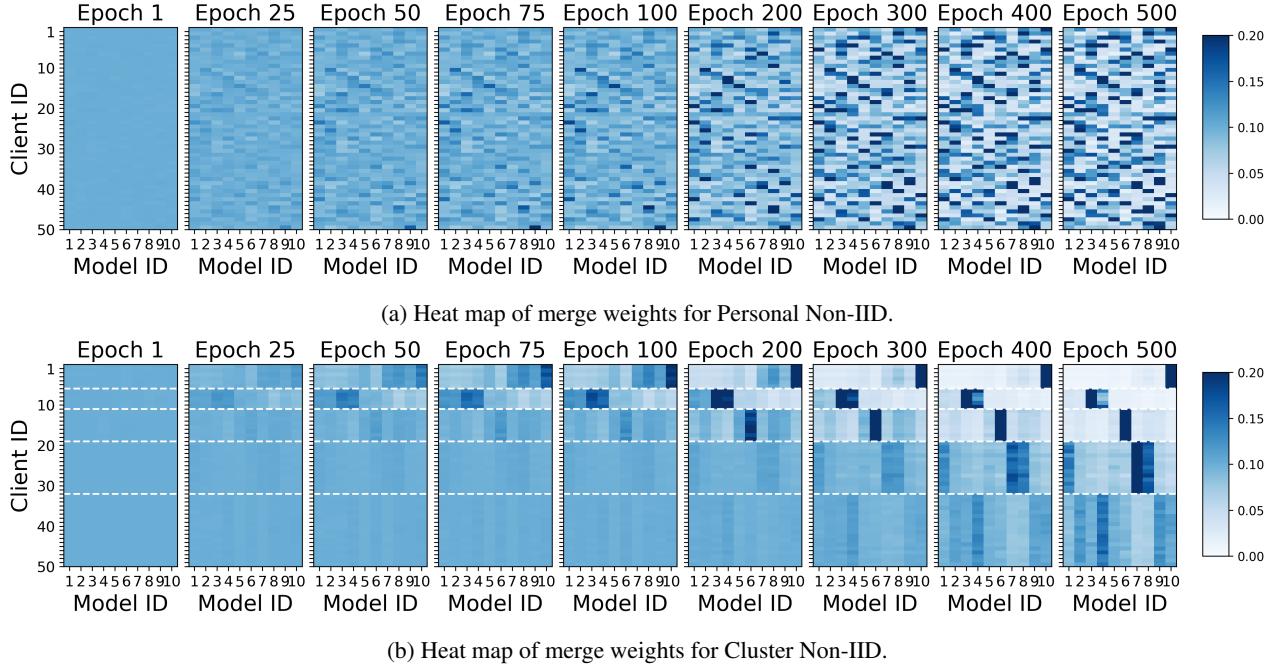


Figure 7: Additional visualizations of merge weights with 10 global models and 50 clients. For Cluster Non-IID, the convergence speed varies across clusters.

C. Gradient Derivations for Θ (Eq (5)) and w (Eq (6))

Given the objective function:

$$L(\{\Theta_j\}, \{w_{(i,j)}\}) = \sum_{i=1}^m \ell(Y_i, f(X_i, \theta_i)),$$

where

$$\theta_i = \sum_{j=1}^d w_{(i,j)} \Theta_j.$$

The goal is to derive gradient for Θ and w with θ as intermediate variable.

C.1. Gradient with respect to Θ_j

1. Compute $\frac{\partial \theta_i}{\partial \Theta_j}$:

$$\frac{\partial \theta_i}{\partial \Theta_j} = w_{(i,j)}.$$

2. Use the chain rule to compute $\frac{\partial L}{\partial \Theta_j}$:

$$\frac{\partial L}{\partial \Theta_j} = \sum_{i=1}^m \frac{\partial \ell(Y_i, f(X_i, \theta_i))}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \Theta_j}.$$

3. Substitute $\frac{\partial \theta_i}{\partial \Theta_j}$:

$$\frac{\partial L}{\partial \Theta_j} = \sum_{i=1}^m w_{(i,j)} \cdot \frac{\partial \ell(Y_i, f(X_i, \theta_i))}{\partial \theta_i}.$$

C.2. Gradient with respect to $w_{(i,j)}$

1. Compute $\frac{\partial \theta_i}{\partial w_{(i,j)}}$:

$$\frac{\partial \theta_i}{\partial w_{(i,j)}} = \Theta_j.$$

2. Use the chain rule to compute $\frac{\partial L}{\partial w_{(i,j)}}$:

$$\frac{\partial L}{\partial w_{(i,j)}} = \frac{\partial \ell(Y_i, f(X_i, \theta_i))}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial w_{(i,j)}}.$$

3. Substitute $\frac{\partial \theta_i}{\partial w_{(i,j)}}$:

$$\frac{\partial L}{\partial w_{(i,j)}} = \langle \Theta_j, \frac{\partial \ell(Y_i, f(X_i, \theta_i))}{\partial \theta_i} \rangle.$$

C.3. Final Expressions

Gradient with respect to Θ_j :

$$\frac{\partial L}{\partial \Theta_j} = \sum_{i=1}^m w_{(i,j)} \cdot \frac{\partial \ell}{\partial \theta_i}.$$

Gradient with respect to $w_{(i,j)}$:

$$\frac{\partial L}{\partial w_{(i,j)}} = \langle \Theta_j, \frac{\partial \ell}{\partial \theta_i} \rangle.$$

These expressions use θ as an intermediate variable, facilitating efficient computation during implementation.

D. Gradient Derivations for Θ and w with Softmax Constraint

For clearer notation, we write the full merging weights as a weight matrix:

$$\mathbf{w} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,d} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,d} \end{bmatrix}$$

where each row represents the personalized merging weights for each client. In the original form of the objective in Eq (3) and Eq (4), w is unconstrained and can take arbitrary values. In practice, this can result in an unstable training process as well as meaningless merging operations. For example, when the global models are neural networks, it is necessary to regularize the parameters before and after merging to prevent the parameters from becoming excessively large.

We introduce a row-wise softmax function for w to ensure each element is non-negative and the sum of each row in w equals 1. Formally, $w_{(i,j)}$ is obtained via a softmax function over parameters $a_{(i,j)}$:

$$w_{(i,j)} = \frac{\exp(a_{(i,j)})}{\sum_{j=1}^d \exp(a_{(i,j)})}.$$

This ensures that each $w_{(i,j)}$ satisfies $0 \leq w_{(i,j)} \leq 1$ and $\sum_{j=1}^d w_{(i,j)} = 1$ for each i . Next we give out the gradient with respect to a .

D.1. Gradient with respect to Θ_j

The gradient of L with respect to Θ_j remains similar to the original derivation because Θ_j appears directly in θ_i :

$$\frac{\partial L}{\partial \Theta_j} = \sum_{i=1}^m w_{(i,j)} \cdot \frac{\partial \ell}{\partial \theta_i}.$$

D.2. Gradient with respect to $a_{(i,j)}$

Since $w_{(i,j)}$ now depends on $a_{(i,j)}$ via the softmax function, we need to compute the gradient of L with respect to $a_{(i,j)}$.

1. Compute $\frac{\partial w_{(i,k)}}{\partial a_{(i,j)}}$:

The derivative of the softmax function is:

$$\frac{\partial w_{(i,k)}}{\partial a_{(i,j)}} = w_{(i,k)} (\delta_{k,j} - w_{(i,j)}) ,$$

where $\delta_{k,j}$ is the Kronecker delta, which is 1 if $k = j$ and 0 otherwise.

2. Compute $\frac{\partial \theta_i}{\partial a_{(i,j)}}$:

Since θ_i depends on $w_{(i,j)}$:

$$\frac{\partial \theta_i}{\partial a_{(i,j)}} = \sum_{k=1}^d \frac{\partial \theta_i}{\partial w_{(i,k)}} \cdot \frac{\partial w_{(i,k)}}{\partial a_{(i,j)}} = \sum_{k=1}^d \Theta_k \cdot \frac{\partial w_{(i,k)}}{\partial a_{(i,j)}}.$$

3. Use the chain rule to compute $\frac{\partial L}{\partial a_{(i,j)}}$:

$$\frac{\partial L}{\partial a_{(i,j)}} = \left\langle \frac{\partial \ell}{\partial \theta_i}, \frac{\partial \theta_i}{\partial a_{(i,j)}} \right\rangle = \left\langle \frac{\partial \ell}{\partial \theta_i}, \sum_{k=1}^d \Theta_k \cdot \frac{\partial w_{(i,k)}}{\partial a_{(i,j)}} \right\rangle.$$

4. Substitute $\frac{\partial w_{(i,k)}}{\partial a_{(i,j)}}$:

$$\frac{\partial L}{\partial a_{(i,j)}} = \left\langle \frac{\partial \ell}{\partial \theta_i}, \sum_{k=1}^d \Theta_k \cdot w_{(i,k)} (\delta_{k,j} - w_{(i,j)}) \right\rangle.$$

Simplify the expression:

$$\begin{aligned} \frac{\partial L}{\partial a_{(i,j)}} &= \left\langle \frac{\partial \ell}{\partial \theta_i}, \left[\Theta_j w_{(i,j)} (1 - w_{(i,j)}) - \sum_{k \neq j} \Theta_k w_{(i,k)} w_{(i,j)} \right] \right\rangle \\ &= \left\langle \frac{\partial \ell}{\partial \theta_i}, w_{(i,j)} \left(\Theta_j - \sum_{k=1}^d \Theta_k w_{(i,k)} \right) \right\rangle. \end{aligned}$$

Note that $\sum_{k=1}^d \Theta_k w_{(i,k)} = \theta_i$, so:

$$\frac{\partial L}{\partial a_{(i,j)}} = \left\langle \frac{\partial \ell}{\partial \theta_i}, w_{(i,j)} (\Theta_j - \theta_i) \right\rangle.$$

D.3. Final Expressions

Gradient with respect to Θ_j :

$$\frac{\partial L}{\partial \Theta_j} = \sum_{i=1}^m w_{(i,j)} \cdot \frac{\partial \ell}{\partial \theta_i}.$$

Gradient with respect to $a_{(i,j)}$:

$$\frac{\partial L}{\partial a_{(i,j)}} = w_{(i,j)} \cdot \left\langle \frac{\partial \ell}{\partial \theta_i}, (\Theta_j - \theta_i) \right\rangle.$$

These expressions efficiently compute the gradients while respecting the constraint on $w_{(i,j)}$. The term $\Theta_j - \theta_i$ represents the difference between the parameter Θ_j and the weighted average θ_i , scaled by $w_{(i,j)}$ and the derivative of the loss with respect to θ_i . We replace Eq (5) and Eq (6) with the above to equations to perform a row softmax constraint on merging weights.