FedLAGC: Towards High Performance System-Heterogeneous Federated Learning via Layer-Adaptive Submodel Extraction and Gradient Correction

Qing Hu[1], **Tianchi Liao**[1,2], **Shuyi Wu**[1], **Lei Yang**[1], **Chuan Chen**[1*]

[1]School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China
[2]School of Software Engineering, Sun Yat-Sen University, Zhuhai, China
huqing2023@163.com, {liaotch, wushy77}@mail2.sysu.edu.cn, {yanglei39, chenchuan}@mail.sysu.edu.cn

Abstract

Federated learning has emerged as a promising paradigm for collaborative model training while preserving data privacy. However, many existing FL methods implicitly assume that clients have sufficient computational and storage resources, making them less applicable in real-world scenarios with severe system heterogeneity. To address this, submodel extraction has recently gained attention as a promising strategy to tailor the global model to resource-constrained clients. Despite this progress, existing methods often suffer from noticeable performance gaps across clients and structural inconsistency in the extracted models, leading to degraded global performance and increased communication overhead. In this work, we propose FedLAGC, a novel federated framework that jointly tackles performance imbalance and communication inefficiency through Layer-Adaptive submodel extraction and Gradient Correction. Specifically, FedLAGC constructs client-specific submodels by selecting structurally important parameters according to layer-wise importance scores, ensuring both resource adaptiveness and architectural consistency. Additionally, we propose a lightweight correction mechanism that captures historical optimization drift, helping to align local updates with the global direction and reduce redundant communication. The rigorous convergence analysis of FedLAGC for system-heterogeneous federated learning under non-convex objectives is given. Extensive experiments on CIFAR-10 and CIFAR-100 with ResNet-18 and ResNet-34 under various system and data heterogeneity settings demonstrate the significant superiority of FedLAGC (up to 24% accuracy improvement and 3.66× communication efficiency) over state-of-the-art methods.

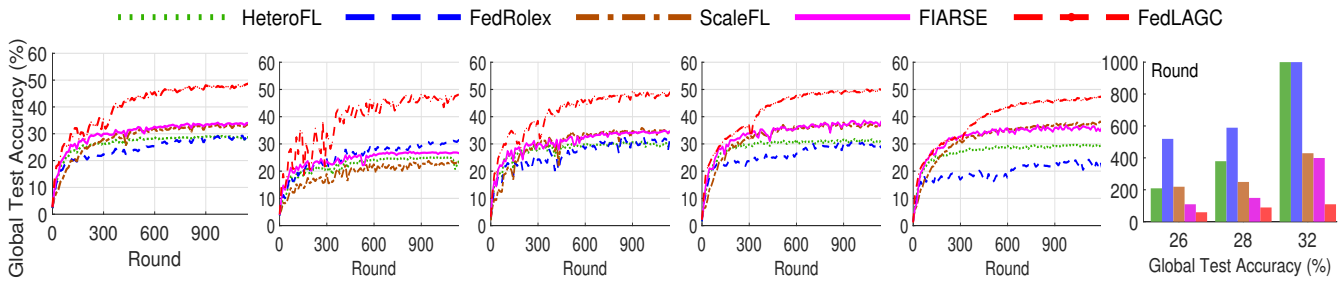
Code — https://github.com/huqing2023/FedLAGC26

Introduction

Federated Learning (FL) (McMahan et al. 2017; Imteaj and Amini 2022; Qi et al. 2022; Li et al. 2022, 2023, 2024; Huang et al. 2024; Qi et al. 2025) has gained increasing attention as a promising framework for privacy-preserving machine learning, allowing distributed clients to train models collaboratively without sharing raw data. This paradigm is particularly attractive for real-world applications such as smartphones, wearable devices, and edge IoT nodes, where data is inherently siloed and privacy-sensitive (Huang, Ye, and Du 2022; Qi et al. 2024; Li et al. 2025a; Liu et al. 2025; Xia et al. 2025; Yan et al. 2025; Liao et al. 2025a). Although FL has gained traction in research and practice, most existing methods are designed under the implicit assumption of homogeneous system conditions, i.e., all participating clients possess sufficient computational power and memory to store and train the full global model (Li et al. 2020; Qi et al. 2023; Huang et al. 2023; Fu et al. 2025; Li et al. 2025b). In practice, this assumption rarely holds. Client devices vary drastically in their hardware configurations, network connectivity, and energy availability, resulting in significant *system heterogeneity* (Liao et al. 2024; Xia et al. 2024). When low-resource clients are forced to train and store a large, uniform global model, it often leads to inefficient training, increased dropouts, and poor convergence. These challenges underscore the urgent need for more adaptive and resource-aware FL strategies that can effectively adapt to practical system heterogeneous scenarios.

To address this challenge, submodel extraction has gained traction as an effective strategy (Diao, Ding, and Tarokh 2021; Alam et al. 2022; Pfeiffer et al. 2023; Zhou et al. 2023; Liao et al. 2023; Wang et al. 2024a; Chen et al. 2025; Liao et al. 2025b). It enables each client to train a reduced model tailored to its resource budget. Initial efforts, such as federated dropout (Caldas et al. 2019), relied on random extracting to reduce model size. Subsequently, more advanced methods (Diao, Ding, and Tarokh 2021; Alam et al. 2022; Ilhan, Su, and Liu 2023) adopted structured submodel extraction to improve training stability and performance. However, these approaches often rely on heuristics or predefined slicing rules that treat all parameters equally, neglecting the fact that different parts of the model contribute unequally to performance. Recent efforts, such as FIARSE (Wu et al. 2024), improve the quality of submodels by ranking and extracting parameters based on global importance. However, they overlook the variation of parameter significance across layers, leading to structurally imbalanced submodels. This structural inconsistency causes inconsistent client updates and hinders convergence (as shown in Figure 1(a)-(e)). Moreover, our experiments show that many existing methods incur high communication overhead to reach the target accuracy, which limits their applicability in latency- or

*Corresponding author



(a) Average Accuracy. (b) Resource level 1. (c) Resource level $\frac{1}{4}$. (d) Resource level $\frac{1}{16}$.(e) Resource level $\frac{1}{64}$.(f) Communication cost.

Figure 1: Performance comparison of different methods on CIFAR-100 with ResNet-18 under the heterogeneous system {1-1/4-1/16-1/64}_{10-20-30-40}. FedLAGC consistently achieves higher global test accuracy across all client resource levels (see (a)-(e)) and achieves target accuracy with significantly fewer communication cost (see (f)).

* The heterogeneous system {1-1/4-1/16-1/64}_{10-20-30-40} has four distinct resource levels: 10 clients capable of running the full model (size 1), 20 clients operating with a reduced model of size 1/4, 30 clients using a smaller model of size 1/16, and 40 clients assigned the smallest model of size 1/64.

bandwidth-sensitive scenarios (see Figure 1(f))[1].

In this work, we propose FedLAGC, a novel high-performance FL framework that tackles the above challenges from two complementary perspectives. First, we introduce a *layer-adaptive submodel extraction* mechanism that quantifies parameter importance at the layer level and allocates model capacity accordingly. This enables each client to receive a submodel that is both resource-efficient and structurally consistent, thereby enhancing the overall convergence performance of the global model. Second, we propose a *lightweight gradient correction* mechanism that accumulates historical optimization drift and guides local updates in alignment with the global descent direction. This strategy improves communication efficiency by reducing the number of redundant or conflicting updates. Together, these components enable FedLAGC to support diverse client capabilities, accelerate convergence, and maintain strong performance even under severe system and data heterogeneity.

Our contributions are summarized as follows:

- We propose FedLAGC, a novel federated learning framework that jointly addresses performance imbalance and communication inefficiency in system-heterogeneous settings through layer-adaptive submodel extraction and gradient direction correction.
- We design a principled submodel extraction strategy that dynamically allocates capacity based on layer-wise importance, enabling structurally consistent submodels under tight client resource budgets.
- We introduce a lightweight gradient correction mechanism that tracks and compensates for accumulated update drift, accelerating convergence and reducing communication rounds.
- Extensive experiments on CIFAR-10 and CIFAR-100 using ResNet-18 and ResNet-34 across various heterogeneity settings show that FedLAGC significantly outperforms state-of-the-art methods in both accuracy and communication efficiency.

[1]For simplicity, we measure communication cost by the number of communication rounds required to reach a predefined accuracy.

Related Work

This section reviews related work on federated learning under homogeneous and heterogeneous models and highlights the limitations of existing submodel extraction strategies.

Federated Learning with Homogeneous Models

Traditional FL algorithms such as FedAvg (McMahan et al. 2017) and FedProx (Li et al. 2020) assume that all clients share an identical model architecture and are capable of training the full global model. These approaches focus primarily on mitigating statistical heterogeneity (e.g., non-IID data) through techniques like proximal terms (Li et al. 2020), adaptive aggregation (Karimireddy et al. 2020; Wang et al. 2020, 2024b), regularization (Li, He, and Song 2021), or alternating direction multipliers (Zhou and Li 2023; Wang et al. 2023; Song, Wang, and Zuazua 2025). While these methods have shown effectiveness in addressing data distribution challenges, they are difficult to directly and effectively apply to system-level heterogeneity due to the variations in computation power, memory, and energy among clients. As a result, directly deploying them in real-world scenarios with resource-limited edge devices may lead to severe performance degradation or even convergence failure.

Submodel Extraction for System-Heterogeneous Federated Learning

To address system heterogeneity, recent research has explored various submodel extraction strategies, allowing each client to train a resource-compatible subset of the global model. Early works like federated dropout (Caldas et al. 2019) randomly extracted neurons to form lightweight client models. However, such stochastic submodel selection often leads to unstable convergence and degraded performance, especially under high data and system heterogeneity. Structured submodel extraction approaches such as HeteroFL (Diao, Ding, and Tarokh 2021) and FjORD (Horváth et al. 2021) used fixed extraction patterns improve training stability but limit parameter sharing and global generalization. FedRolex (Alam et al. 2022) mitigated this issue via

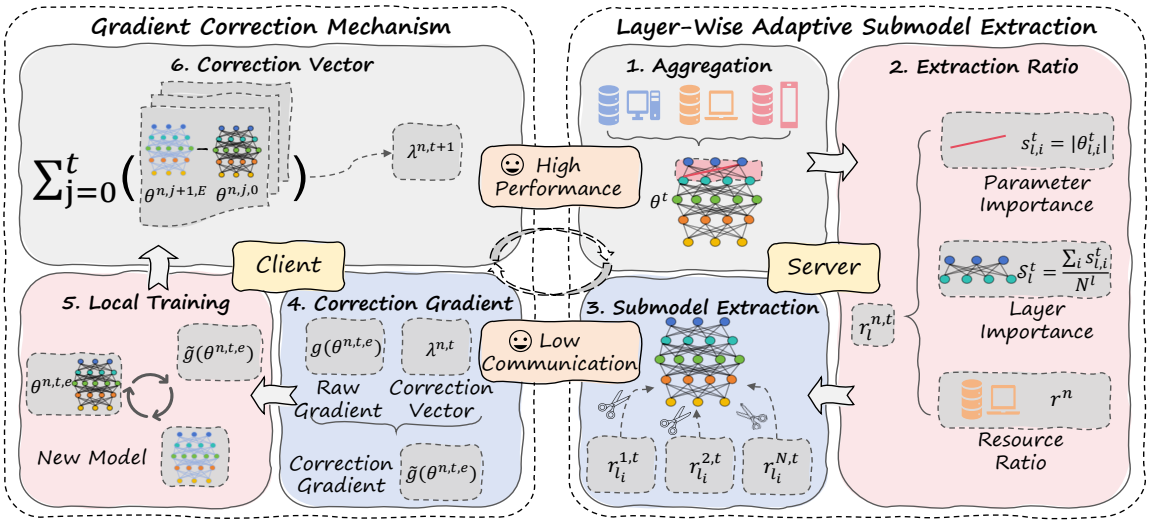


Figure 2: The framework diagram of FedLAGC.

rolling submodel regions, while ScaleFL (Ilhan, Su, and Liu 2023) and DepthFL (Kim et al. 2023) explored slicing the model by depth or width and apply self-distillation for better knowledge retention. Despite this progress, the aforementioned methods still heavily rely on predefined heuristics and treat all model parameters equally, overlooking the fact that parameters contribute unequally to the predictive power of a model. This oversight lead to inefficient submodel configurations that neglect critical information. Fortunately, FIARSE (Wu et al. 2024) recently introduced a global importance-aware extraction mechanism by ranking parameters based on magnitude. Although this strategy improves accuracy over uniform selection, it ignores layer-wise structural variance, leading to imbalanced submodels and inconsistent training signals. Moreover, existing methods often require excessive communication costs to reach satisfactory accuracy due to slow convergence, as demonstrated in our experiments.

Proposed FedLAGC

Overview

To address the challenges introduced by system heterogeneity, we propose a federated learning framework FedLAGC that integrates two core components: a layer-adaptive submodel extraction strategy and a gradient correction mechanism. These components work together to ensure that each client trains a submodel tailored to its resource budget while contributing stable and meaningful updates to the global model. Specifically, at the beginning of each communication round, the server estimates the importance of different model layers based on the current global parameters. Guided by this layer-wise importance and the resource constraint of each client, the server selects structurally essential components and informative parameters tailored to each client. During local training, each client optimizes its extracted submodel using a masked gradient strategy, while a lightweight correction vector is employed to refine the update direction by accumulating historical optimization drift. Once local training is completed, the updated submodels are sent back to the server, where an overlap-aware aggregation scheme (Li et al. 2021a,b; Diao, Ding, and Tarokh 2021) is applied to update only the parameters that are jointly trained across clients. The overall framework of our method is illustrated in Figure 2, and the detailed algorithmic steps are provided in Appendix due to page limitations.

In the sequel, we present a detailed explanation of the two core components of our proposed method.

Layer-adaptive Submodel Extraction

Although various submodel extraction methods have been proposed to accommodate client resource limitations in federated learning, most of them rely on uniform or heuristic extraction strategies. Moreover, directly extracting parameters without considering their structural importance significantly impacts the convergence and overall performance.

To address these, we design a layer-adaptive submodel construction strategy that allocates each client's limited capacity to the most critical regions of the model. Rather than relying on uniform extraction or global ranking, we evaluate both parameter-level and layer-level importance to guide the extraction process. Inspired by the magnitude-based pruning principle in (Han et al. 2015), we use the absolute value of each parameter as a proxy for its significance. Specifically, for the i-th parameter $\theta_{l,i}^t$ in layer l of the global model at round t, its importance is defined as:

$$s_{l,i}^t = |\theta_{l,i}^t|.$$

Then, the average parameter importance in each layer can be used as a measure of the overall significance for that layer:

$$\mathcal{S}_l^t = \frac{1}{N^l} \sum_i s_{l,i}^t,$$

where N^l denotes the total number of parameters in the l-th layer of the global model θ^t. To smooth inter-layer disparities and avoid dominance by layers with high variance, we

apply a logarithmic normalization:

$$\tilde{\mathcal{S}}_l^t = \frac{\log(1+\mathcal{S}_l^t)}{\sum_j \log(1+\mathcal{S}_j^t)}.$$

Based on these normalized importance scores, we then extract a submodel for each client by selecting a proportion of parameters from each layer, guided by both layer-level and parameter-level importance.

Practically speaking, given a client-specific resource budget r^n in $(0, 1]$, the target size of the submodel for client n is $d^n = r^n \cdot d$ with d being the global model size. To ensure functional integrity, we fully retain certain critical components, including the input and output layers, normalization layers, and all bias terms, whose total parameter count is denoted as $\tilde{d}$. The remaining budget, i.e., $d^n - \tilde{d}$, is allocated to the prunable layers $\{l_1, \ldots, l_L\}$ according to their layer importance. Here we assume $d^n - \tilde{d} > 0$. For each prunable layer l_i, the proportion of parameters retained by the client n in round t can be calculated by:

$$r_{l_i}^{n,t} = \tilde{\mathcal{S}}_{l_i}^t \cdot \frac{d^n - \tilde{d}}{\sum_{i=1}^{L} \tilde{\mathcal{S}}_{l_i}^t d_{l_i}},$$

where d_{l_i} is the number of parameters in layer l_i. For brevity, the derivation of this allocation formula is shown in Appendix. Based on the layer extraction ratio, we select the top $r_{l_i}^{n,t} \cdot d_{l_i}$ parameters in each layer according to their magnitude $s_{l,i}^t$. This yields a binary mask $M_{l_i}^{n,t}$ and a corresponding threshold $\tilde{\theta}_{l_i}^{n,t}$, which together define the sparse submodel $\theta^{n,t} = \theta^t \odot M^{n,t}$ sent to client n in round t.

By effectively utilizing client resources while preserving the model structure, our approach ensures balanced participation across heterogeneous clients and leads to more stable and efficient global training.

Gradient Correction

Existing submodel extraction approaches in federated learning frequently encounter inefficient optimization, stemming from biased and inconsistent local updates. This not only hinders convergence speed but also adversely affects the final model performance. To alleviate this, we further propose a carefully designed gradient correction mechanism that refines the update direction on each client by incorporating accumulated descent history, thereby accelerating convergence and enhancing global model quality.

More specifically, each client n maintains a correction vector $\lambda^{n,t}$ at communication round t, which accumulates the discrepancy between its local update trajectory and the global model path over time. During local training, the raw STE(Straight-Through Estimator)-based gradient $g(\theta^{n,t,e})$ [2] is modified by subtracting the correction vector:

$$\tilde{g}(\theta^{n,t,e}) = \Big(g(\theta^{n,t,e}) - h(t)\lambda^{n,t}\Big) \odot M^{n,t},$$

where $M^{n,t}$ is the binary mask indicating the parameters trained by client n in round t, and $h(t)$ is a scheduling function controlling the influence of the correction (e.g., in this paper, we heuristically set $h(t) = 1$ when $t < \lfloor \frac{T}{4} \rfloor$, and $h(t) = 0$ otherwise)[3]. This adjustment steers the optimization towards a more globally aligned direction. Inspired by the ADMM technique (Zhou and Li 2023), after local training, the correction vector is updated using the difference between the client's locally trained model $\theta^{n,t+1} := \theta^{n,t,E}$ (after E local epochs) and the initial model it received at the beginning of the round, denoted by $\theta^{n,t,0} := \theta^t \odot M^{n,t}$:

$$\lambda^{n,t+1} = \lambda^{n,t} + \beta \cdot \big(\theta^{n,t+1} - \theta^{n,t,0}\big) \odot M^{n,t},$$

where $\beta > 0$ is a tunable hyperparameter controlling the accumulation rate. The term $(\theta^{n,t+1} - \theta^{n,t,0})$ captures how far the client's local descent path diverges from the global model. By accumulating this deviation over time, $\lambda^{n,t}$ functions as a memory of useful descent directions, enabling the client to correct and stabilize its future updates.

This correction mechanism reduces the variance and inconsistency introduced by system heterogeneity, while also improving the effectiveness of each local update. As a result, clients can achieve faster convergence using fewer local steps, helping to alleviate communication overhead in resource-limited federated systems.

Convergence Analysis

This section focuses on the convergence analysis of FedLAGC under system-heterogeneous FL. Due to space constraints, the detailed problem formulation, assumptions and proof are provided in Appendix. Based on these preliminaries, we establish the following convergence theorem.

Theorem 1. *Suppose the assumptions shown in Appendix hold and the local learning rate satisfies* $\eta = \mathcal{O}(1/(E\sqrt{\tilde{T}}))$ *with* E *and* $T = \tilde{T} + \lceil \frac{T}{4} \rceil$ *being the number of local epoch and total round. Then FedLAGC converges to a small neighborhood of a stationary point of the standard FL under heterogeneous system* $\{level_1\text{-}level_2\text{-}\ldots\text{-}level_p\}\_\{N_1\text{-}N_2\text{-}\ldots\text{-}N_p\}$*:*

$$\frac{1}{\tilde{T}} \sum_{t=\lceil \frac{T}{4} \rceil}^{T-1} \sum_{i \in \mathcal{I}^t} \mathbb{E}\left[\Big(\nabla F(\theta^t)\Big)_i\right]^2$$

$$\leq \mathcal{O}\Big(\frac{1}{\sqrt{\tilde{T}}}\Big) + \mathcal{O}\Big(\frac{1}{\tilde{T}}\Big) \sum_{t=\lceil \frac{T}{4} \rceil}^{T-1} \sum_{i=1}^{p} N_i(1 - level_i)\delta_i^2 \|\theta^t\|^2$$

$$+ \mathcal{O}\Big(\frac{1}{\tilde{T}\sqrt{\tilde{T}}}\Big) \sum_{t=\lceil \frac{T}{4} \rceil}^{T-1} \sum_{i=1}^{p} N_i(1 - level_i)\delta_i^2 \|\theta^t\|^2,$$

where $\mathcal{I}^t$ *is the index set of elements updated in round* t *and* δ_i *is defined in Appendix.*

[2] Inspired by the recent work (Wu et al. 2024; Liu et al. 2022), the raw local gradient in the l-th layer for epoch e is defined as $\Big(g(\theta^{n,t,e})\Big)_l = \big(\nabla F^n(\theta^{n,t,e} \odot M^{n,t})\big)_l \odot M_l^{n,t} \odot \Big(1 + 2|\theta_l^{n,t,e}|\tilde{\theta}_l^{n,t}/(|\theta_l^{n,t,e}| + \tilde{\theta}_l^{n,t})^2\Big)$ with $F^n(\theta)$ being the local loss function.

[3] Our heuristic approach is simple and effective, and investigating more interpretable scheduling functions is a promising direction for future research.

DD	Scenario	1-1/4-1/16-1/64			1-16/25-9/25-4/25-1/25		
		5-10-25-60	10-20-30-40	25-25-25-25	5-5-10-20-60	5-10-15-20-50	20-20-20-20-20
		CIFAR-10/100	CIFAR-10/100	CIFAR-10/100	CIFAR-10/100	CIFAR-10/100	CIFAR-10/100
IID	HeteroFL	76.65/30.13	77.74/31.63	78.20/31.55	78.45/31.24	79.05/32.22	79.83/34.94
	FedRolex	76.48/29.29	78.24/33.51	79.00/33.69	80.35/40.72	81.71/36.07	82.94/38.18
	ScaleFL	78.20/36.42	79.87/38.24	80.34/40.59	79.14/36.75	79.37/38.19	82.85/41.58
	FIARSE	78.84/32.77	81.75/36.63	83.85/38.72	82.13/36.96	83.06/40.32	85.96/43.49
	FedLAGC	**86.03/47.94**	**87.06/50.40**	**87.09/52.24**	**88.67/53.39**	**88.76/56.06**	**89.32/59.00**
	△Improve	***7.19/11.52***	***5.31/12.16***	***3.24/11.65***	***6.54/12.67***	***5.70/15.74***	***3.36/15.51***
Dir(0.5)	HeteroFL	65.57/27.69	67.25/30.01	69.15/30.52	66.68/30.75	67.85/30.05	71.43/34.69
	FedRolex	59.95/26.64	65.60/31.52	65.99/32.10	68.76/32.41	71.29/35.85	72.88/39.05
	ScaleFL	59.78/33.07	63.31/35.62	64.62/36.31	61.98/34.00	63.73/35.26	68.09/40.14
	FIARSE	69.12/31.89	73.67/34.43	75.26/38.76	74.50/34.98	75.06/35.65	77.53/41.34
	FedLAGC	**73.48/45.79**	**75.66/47.95**	**77.42/51.32**	**78.44/51.61**	**81.54/54.16**	**82.93/56.53**
	△Improve	***4.36/12.72***	***1.99/12.33***	***2.16/12.56***	***3.94/16.63***	***6.48/18.31***	***5.40/15.19***
Dir(0.3)	HeteroFL	60.83/26.79	65.14/28.67	65.84/30.89	62.31/28.96	65.00/30.26	69.57/33.36
	FedRolex	53.43/26.16	58.63/29.11	61.83/30.41	62.45/32.78	65.21/34.18	70.38/37.22
	ScaleFL	52.80/30.29	57.10/33.39	59.56/35.21	53.62/30.86	55.39/32.87	61.97/36.51
	FIARSE	61.22/29.49	68.61/33.50	72.00/36.69	67.36/33.84	69.98/36.93	74.53/40.47
	FedLAGC	**68.08/43.87**	**69.65/48.14**	**73.13/50.74**	**73.27/50.43**	**76.96/52.34**	**81.20/53.68**
	△Improve	***6.86/13.58***	***1.04/14.64***	***1.13/14.05***	***5.91/16.59***	***6.98/15.41***	***6.67/13.21***

Table 1: Comparison of global test accuracy on CIFAR-10 and CIFAR-100 with ResNet-18 under various data distributions (DD) and heterogeneous systems.

* Resource levels: {1-1/4-1/16-1/64} and {1-16/25-9/25-4/25-1/25}.

* Client allocation schemes: {5-10-25-60}, {10-20-30-40}, {25-25-25-25}, {5-5-10-20-60}, {5-10-15-20-50}, {20-20-20-20-20}.

Experiments

This section presents extensive experiments to validate the effectiveness of our proposed FedLAGC framework in system-heterogeneous FL. We first outlines the datasets, architectures, heterogeneity settings, baselines, training configurations, and evaluation protocols used in our study.

Datasets and Architectures. Our experiments are carried out on two widely-used image classification benchmarks: CIFAR-10 and CIFAR-100 (Krizhevsky and Hinton 2009). We adopt ResNet-18 and ResNet-34 as backbone models and replace standard batch normalization (BN) layers with static BN (Wu et al. 2024; He et al. 2016).

Data Distribution Settings. To explore the influence of statistical heterogeneity, we consider IID and non-IID data partitions under Dirichlet distribution Dir(α) (Wu et al. 2024), where smaller values of α result in more skewed label distributions across clients.

System Heterogeneity Settings. To simulate realistic federated environments with varying client capabilities, we define two sets of client resource levels: {1-1/4-1/16-1/64} and {1-16/25-9/25-4/25-1/25}. Rather than assuming uniform distribution of clients over these levels as in previous work, we evaluate three different client allocation schemes to better reflect real-world resource asymmetries: i) Four-level scheme: {5-10-25-60}, {10-20-30-40}, and balanced {25-25-25-25}, ii) Five-level scheme: {5-5-10-20-60}, {5-10-15-20-50}, and balanced {20-20-20-20-20}.

Baselines. We compare our approach with recent state-of-the-art submodel extraction methods in heterogeneous FL, including HeteroFL (Diao, Ding, and Tarokh 2021), FedRolex (Alam et al. 2022), ScaleFL (Ilhan, Su, and Liu 2023), and FIARSE (Wu et al. 2024).

Training Settings. We apply the same settings across all methods to ensure fairness. The training process spans a total of 1000 communication rounds. In each round, 10 clients (10% of the total 100) are randomly selected to perform local training. Each selected client runs 5 local epochs per round with a batch size of 20, following the setup in (Wu et al. 2024). For optimization, we use stochastic gradient descent (SGD) with a learning rate of 0.01 and a momentum coefficient of 0.8. In our method, the hyperparameter β is set to 0.1. All experiments are conducted on a server equipped with two NVIDIA GeForce RTX 4090 GPUs.

Evaluation Metrics. The global test accuracy under a given client resource level refers to the accuracy of the corresponding submodel on the global test set. To mitigate variance across rounds, we report the average Top-1 accuracy over the last 20 rounds. Each experiment is repeated with three random seeds, and results are reported as the mean accuracy. Due to space constraints, local test accuracy on the local test set and additional results are shown in Appendix.

Superiority of FedLAGC in Convergence Performance and Communication Efficiency

We first evaluate the convergence performance and communication efficiency of FedLAGC under a challenging set-

Scenario	1-1/4-1/16-1/64			1-16/25-9/25-4/25-1/25		
	5-10-25-60	10-20-30-40	25-25-25-25	5-5-10-20-60	5-10-15-20-50	20-20-20-20-20
	ResNet-18/34	ResNet-18/34	ResNet-18/34	ResNet-18/34	ResNet-18/34	ResNet-18/34
HeteroFL	26.79/27.22	28.67/28.45	30.89/29.56	28.96/28.54	30.26/28.81	33.36/31.75
FedRolex	26.16/24.57	29.11/29.24	30.41/31.14	32.78/30.72	34.18/32.69	37.22/36.35
ScaleFL	30.29/28.77	33.39/31.71	35.21/35.34	30.86/31.71	32.87/33.51	36.51/39.34
FIARSE	29.49/29.50	33.50/34.01	36.69/37.59	33.84/31.72	36.93/35.81	40.47/41.30
FedLAGC w/o GC	**39.85/39.64**	**41.37/41.04**	**42.73/42.09**	**43.46/43.66**	**44.41/43.37**	**44.94/44.05**
FedLAGC	**43.87/44.73**	**48.14/48.60**	**50.74/48.96**	**50.43/51.00**	**52.34/52.16**	**53.68/53.75**

Table 2: Comparison of global test accuracy on CIFAR-100 with Dir(0.3) across different network architectures and heterogeneous systems.

ting: CIFAR-100 with ResNet-18, Dirichlet-distributed data ($\alpha = 0.3$), and a highly imbalanced system heterogeneity configuration $\{$1-1/4-1/16-1/64$\}_\{$10-20-30-40$\}$. As shown in Figure 1(a), FedLAGC yields a substantial improvement of 14%–20% in average global test accuracy over prior state-of-the-art methods. Figures 1(b)–(e) present the convergence curves for the global model under four different resource levels, showing that FedLAGC achieves stable and competitive accuracy across all clients. In contrast, existing methods exhibit large performance disparities under different resource levels, which compromises the effectiveness of global model training. In addition, FedLAGC requires significantly fewer communication rounds to reach the target accuracy (e.g., 26%, 28%, 32%) compared to all baselines, as shown in Figure 1(f). Specifically, it can achieve up to 3.66× communication efficiency than the best-performing baseline FIARSE. These results highlight the superiority of FedLAGC in terms of convergence performance and communication efficiency.

Superiority of FedLAGC under Various System Heterogeneous Scenarios

To validate the robustness of our method under different levels of system heterogeneity, we test FedLAGC and baseline methods across six client resource configurations on both CIFAR-10 and CIFAR-100 datasets with ResNet-18. As shown in Table 1, FedLAGC consistently outperforms the state-of-the-art methods in all settings, maintaining high global accuracy even when most clients have limited resources. This demonstrates that our layer-adaptive submodel extraction makes efficient use of client-side computational resources and ensures fair contribution from clients with diverse resource levels, enabling stable training under system heterogeneity.

Superiority of FedLAGC under Various Data Heterogeneous Scenarios

To further demonstrate the robustness of FedLAGC beyond system heterogeneity, beyond system heterogeneity, we evaluate it under three levels of data heterogeneity: IID, Dir(0.5) and Dir(0.3). As shown in Table 1, FedLAGC achieves consistent and superior performance across all settings on both CIFAR-10 and CIFAR-100. In particular, under the most challenging Dirichlet distribution scenario ($\alpha = 0.3$), where client distributions are highly skewed, our method maintains a substantial advantage over prior approaches, with performance gains reaching up to 21% in some cases. These results confirm the robustness of FedLAGC on both system and statistical heterogeneity.

Superiority of FedLAGC under Different Network Architectures

To evaluate the generality of FedLAGC, we compare its performance using ResNet-18 and ResNet-34 on CIFAR-100 under Dirichlet distribution with $\alpha = 0.3$ and six system heterogeneity settings. As shown in Table 2, FedLAGC consistently achieves the highest accuracy across all network architectures. For instance, on CIFAR-100 with ResNet-34 in the system heterogeneous scenario $\{$1-16/25-9/25-4/25-1/25$\}_\{$5-10-15-20-50$\}$, our method outperforms FIARSE, ScaleFL, FedRolex, and HeteroFL by 16.35%, 18.65%, 19.47%, and 23.35%, respectively. These results suggest that FedLAGC scales effectively to deeper and more complex models, making it a practical solution for real-world federated systems with heterogeneous client capabilities and demanding model complexity.

Effectiveness of Layer-Adaptive Submodel Extraction and Gradient Correction

To better understand and highlight the individual contributions of our two core components, we conduct an ablation study comparing the FedLAGC framework with a variant that excludes the gradient correction mechanism (FedLAGC w/o GC). As reported in Table 2, the variant still clearly outperforms all baseline methods, demonstrating the strong effectiveness of our layer-adaptive submodel extraction in handling resource heterogeneity. Furthermore, incorporating the gradient correction mechanism brings substantial additional gains over layer-adaptive submodel extraction alone. In some scenarios, this improvement reaches up to 9.7% compared to its variant without gradient correction, highlighting the effectiveness of aligning local updates with the global optimization direction. These results confirm that both components play essential and complementary roles:

the submodel extraction ensures efficient and balanced training across clients, while gradient correction stabilizes local updates and promotes faster convergence.

Ablation study for other components

To further investigate the contributions of different components in FedLAGC, we performed ablation experiments focusing on the STE-based gradient approximation and the logarithmic layer normalization used for computing layer importance. We also compared FedLAGC with two straightforward alternatives for predefined layer importance estimation $\mathcal{S}_l^t$, namely assigning equal importance to all layers (FedLAGC w eq) and using the maximum of parameter importance within each layer (FedLAGC w max). The experimental results are shown in Figure 3(a).

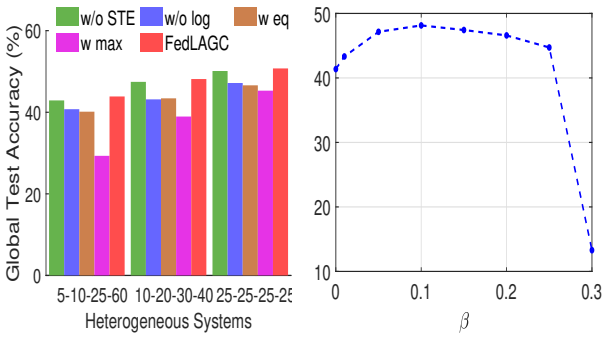
The comparison between *FedLAGC w/o log* and FedLAGC reveals that removing logarithmic normalization causes a clear performance decline, reducing its effectiveness to a level comparable to that of *FedLAGC w eq*. This suggests that logarithmic normalization is critical for stabilizing layer importance computation. In addition, examining the predefined layer importance strategies shows that FedLAGC consistently surpasses both the equal-weighting and maximum-based variants, indicating that using the mean of parameter importance provides a more balanced and reliable measure of layer importance. Finally, comparing *FedLAGC w/o STE* with FedLAGC demonstrates that incorporating STE technique leads to noticeable performance gains, underscoring its role in enhancing submodel training and overall accuracy. These findings collectively highlight that logarithmic normalization, mean-based layer importance estimation, and STE each play a crucial role in improving the effectiveness and robustness of FedLAGC under heterogeneous federated learning environments.

Impact of the Correction Factor β

In this subsection, we explore the sensitivity of FedLAGC to the correction factor β, which controls the strength of historical gradient adjustment during local updates. Figure 3(b) presents the performance of FedLAGC across different values of β on CIFAR-100 with ResNet-18, evaluated under the non-IID partition Dir(0.3) and a highly heterogeneous system configuration {1-1/4-1/16-1/64}_{10-20-30-40}. As shown in Figure 3(b), FedLAGC achieves strong and stable performance within a reasonable range of β values. However, overly large values may lead to performance degradation due to overcompensation in the gradient correction. Empirically, we find that setting $\beta = 0.1$ yields consistently good results across various scenarios, as shown in Tables 1-2. Designing more adaptive and theoretically grounded strategies for choosing β remains an important direction for future work.

Impact of Randomness

To further examine the impact of randomness on different methods, we conducted experiments using more random seeds in the heterogeneous system setting {1-1/4-1/16-1/64}_{10-20-30-40}. The global test accuracy for each individual seed is reported, along with the mean and standard deviation computed over all seeds, as shown in Table 3. From this table, one can see that the stability of our method is comparable to that of other baselines, which further confirms the reliability of the conclusions drawn above.



(a) Ablation. (b) Impact of β.

Figure 3: Ablation study for FedLAGC and the impact of the correction factor β for CIFAR-100 with Dir(0.3) under ResNet-18 and client resource levels {1-1/4-1/16-1/64}.

Seed	1	1999	2000	2008	2025	2026	mean	std
HeteroFL	28.43	28.51	28.55	27.83	27.69	28.70	28.29	0.38
FedRolex	28.98	29.76	29.61	28.96	28.74	29.69	29.29	0.41
ScaleFL	33.79	33.43	32.51	33.06	32.77	33.47	33.17	0.44
FIARSE	33.60	35.85	34.42	32.22	35.12	35.45	34.44	1.23
FedLAGC	**47.20**	**47.31**	**47.81**	**47.60**	**46.77**	**47.91**	***47.43***	***0.39***

Table 3: Impact of randomness for CIFAR-100 with Dir(0.3) under ResNet-18 and heterogeneous system {1-1/4-1/16-1/64}_{10-20-30-40}.

Conclusion

In this work, we propose a novel federated learning framework, FedLAGC, to address the challenges posed by system heterogeneity in FL. By integrating a novel layer-adaptive submodel extraction strategy with a carefully designed history-guided gradient correction mechanism, FedLAGC effectively leverages heterogeneous client resources while improving the stability of local updates. Extensive experiments conducted across multiple datasets, network architectures, and heterogeneous configurations demonstrate that our method achieves substantial gains in both convergence speed and communication efficiency over existing approaches. In future work, we aim to extend FedLAGC by incorporating advanced optimization acceleration techniques, which are often challenging to apply effectively in deep learning, to better accommodate more realistic and complex heterogeneous FL scenarios, thereby further reducing the communication burden through faster training.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2023YFB2703700), the National Natural Science Foundation of China (62176269) and GMCC-SYSU Joint Lab for Smart Applications.

References

Alam, S.; Liu, L.; Yan, M.; and Zhang, M. 2022. FedRolex: Model-Heterogeneous Federated Learning with Rolling Sub-Model Extraction. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 35, 29677–29690.

Caldas, S.; Konečny, J.; McMahan, H. B.; and Talwalkar, A. 2019. Expanding the Reach of Federated Learning by Reducing Client Resource Requirements. arXiv:1812.07210.

Chen, C.; Liao, T.; Deng, X.; Wu, Z.; Huang, S.; and Zheng, Z. 2025. Advances in Robust Federated Learning: A Survey with Heterogeneity Considerations. *IEEE Transactions on Big Data*, 11(3): 1548–1567.

Diao, E.; Ding, J.; and Tarokh, V. 2021. HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients. In *Proceedings of the International Conference on Learning Representations*.

Fu, L.; Huang, S.; Lai, Y.; Liao, T.; Zhang, C.; and Chen, C. 2025. Beyond Federated Prototype Learning: Learnable Semantic Anchors with Hyperspherical Contrast for Domain-Skewed Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16): 16648–16656.

Han, S.; Pool, J.; Tran, J.; and Dally, W. 2015. Learning both Weights and Connections for Efficient Neural Network. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 28.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 770–778.

Horváth, S.; Laskaridis, S.; Almeida, M.; Leontiadis, I.; Venieris, S.; and Lane, N. 2021. FjORD: Fair and Accurate Federated Learning under heterogeneous targets with Ordered Dropout. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 34, 12876–12889.

Huang, W.; Ye, M.; and Du, B. 2022. Learn from Others and Be Yourself in Heterogeneous Federated Learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10133–10143.

Huang, W.; Ye, M.; Shi, Z.; Li, H.; and Du, B. 2023. Rethinking Federated Learning with Domain Shift: A Prototype View. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16312–16322.

Huang, W.; Ye, M.; Shi, Z.; Wan, G.; Li, H.; Du, B.; and Yang, Q. 2024. Federated Learning for Generalization, Robustness, Fairness: A Survey and Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9387–9406.

Ilhan, F.; Su, G.; and Liu, L. 2023. ScaleFL: Resource-Adaptive Federated Learning with Heterogeneous Clients. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24532–24541.

Imteaj, A.; and Amini, M. H. 2022. Leveraging asynchronous federated learning to predict customers financial distress. *Intelligent Systems with Applications*, 14: 200064.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 5132–5143. PMLR.

Kim, M.; Yu, S.; Kim, S.; and Moon, S.-M. 2023. DepthFL : Depthwise Federated Learning for Heterogeneous Clients. In *Proceedings of the International Conference on Learning Representations*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4).

Li, A.; Sun, J.; Li, P.; Pu, Y.; Li, H.; and Chen, Y. 2021a. Hermes: an efficient federated learning framework for heterogeneous mobile clients. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*, 420–437.

Li, A.; Sun, J.; Wang, B.; Duan, L.; Li, S.; Chen, Y.; and Li, H. 2021b. LotteryFL: Empower Edge Intelligence with Personalized and Communication-Efficient Federated Learning. In *IEEE/ACM Symposium on Edge Computing*, 68–79.

Li, Q.; Diao, Y.; Chen, Q.; and He, B. 2022. Federated Learning on Non-IID Data Silos: An Experimental Study. In *IEEE International Conference on Data Engineering*, 965–978.

Li, Q.; He, B.; and Song, D. 2021. Model-Contrastive Federated Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10708–10717.

Li, Q.; Wen, Z.; Wu, Z.; Hu, S.; Wang, N.; Li, Y.; Liu, X.; and He, B. 2023. A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4): 3347–3366.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated Optimization in Heterogeneous Networks. In *Proceedings of Machine Learning and Systems*, volume 2, 429–450.

Li, Y.; Li, Q.; Wang, H.; Li, R.; Zhong, W.; and Zhang, G. 2024. Towards Efficient Replay in Federated Incremental Learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12820–12829.

Li, Y.; Shan, Y.; Liu, Y.; Wang, H.; Wang, W.; Wang, Y.; and Li, R. 2025a. Personalized Federated Recommendation for Cold-Start Users via Adaptive Knowledge Fusion. In *Proceedings of the ACM on Web Conference 2025*, WWW'25, 2700–2709. Association for Computing Machinery.

Li, Y.; Xu, W.; Wang, H.; Qi, Y.; Guo, J.; and Li, R. 2025b. Personalized Federated Domain-Incremental Learning Based on Adaptive Knowledge Matching. In *Proceedings of the European Conference on Computer Vision*, 127–144.

Liao, D.; Gao, X.; Zhao, Y.; and Xu, C. 2023. Adaptive Channel Sparsity for Federated Learning under System Heterogeneity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20432–20441.

Liao, T.; Fu, L.; Chen, J.; WANG, Z.; Zheng, Z.; and Chen, C. 2024. A Swiss Army Knife for Heterogeneous Federated Learning: Flexible Coupling via Trace Norm. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Liao, T.; Fu, L.; Zhang, L.; Yang, L.; Chen, C.; Ng, M. K.; Huang, H.; and Zheng, Z. 2025a. Privacy-Preserving Vertical Federated Learning With Tensor Decomposition for Data Missing Features. *IEEE Transactions on Information Forensics and Security*, 20: 3445–3460.

Liao, T.; Xu, Z.; Hu, Q.; Dai, H.-N.; Huang, H.; Zheng, Z.; and Chen, C. 2025b. FedBRB: A Solution to the Small-to-Large Scenario in Device-Heterogeneity Federated Learning. *IEEE Transactions on Mobile Computing*, 1–14.

Liu, R.; Hu, M.; Xia, Z.; Xie, X.; Xia, J.; Zhang, P.; Huang, Y.; and Chen, M. 2025. FedGraft: Memory-Aware Heterogeneous Federated Learning via Model Grafting. *IEEE Transactions on Mobile Computing*, 24(12): 13506–13519.

Liu, Z.; Cheng, K.-T.; Huang, D.; Xing, E.; and Shen, Z. 2022. Nonuniform-to-Uniform Quantization: Towards Accurate Quantization via Generalized Straight-Through Estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4932–4942.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. A. y. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 54, 1273–1282.

Pfeiffer, K.; Rapp, M.; Khalili, R.; and Henkel, J. 2023. Federated Learning for Computationally Constrained Heterogeneous Devices: A Survey. *ACM Computing Surveys*, 55(14s).

Qi, Z.; He, W.; Meng, X.; and Meng, L. 2024. Attentive Modeling and Distillation for Out-of-Distribution Generalization of Federated Learning. In *2024 IEEE International Conference on Multimedia and Expo*, 1–6.

Qi, Z.; Meng, L.; Chen, Z.; Hu, H.; Lin, H.; and Meng, X. 2023. Cross-Silo Prototypical Calibration for Federated Learning with Non-IID Data. MM'23, 3099–3107. Association for Computing Machinery.

Qi, Z.; Meng, L.; Li, Z.; Hu, H.; and Meng, X. 2025. Cross-Silo Feature Space Alignment for Federated Learning on Clients with Imbalanced Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19): 19986–19994.

Qi, Z.; Wang, Y.; Chen, Z.; Wang, R.; Meng, X.; and Meng, L. 2022. Clustering-based Curriculum Construction for Sample-Balanced Federated Learning. In *Artificial Intelligence*, 155–166. Cham: Springer Nature Switzerland. ISBN 978-3-031-20503-3.

Song, Y.; Wang, Z.; and Zuazua, E. 2025. FedADMM-InSa: An inexact and self-adaptive ADMM for federated learning. *Neural Networks*, 181: 106772.

Wang, H.; Jia, Y.; Zhang, M.; Hu, Q.; Ren, H.; Sun, P.; Wen, Y.; and Zhang, T. 2024a. FedDSE: Distribution-aware Submodel Extraction for Federated Learning over Resource-constrained Devices. In *Proceedings of the International Conference Companion on World Wide Web*, 2902–2913.

Wang, H.; Xu, H.; Li, Y.; Xu, Y.; Li, R.; and Zhang, T. 2024b. FedCDA: Federated Learning with Cross-rounds Divergence-aware Aggregation. In *Proceedings of the International Conference on Learning Representations*.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 7611–7623.

Wang, S.; Xu, Y.; Wang, Z.; Chang, T.-H.; Quek, T. Q. S.; and Sun, D. 2023. Beyond ADMM: A Unified Client-Variance-Reduced Adaptive Federated Learning Framework. 37: 10175–10183.

Wu, F.; Wang, X.; Wang, Y.; Liu, T.; Su, L.; and Gao, J. 2024. FIARSE: Model-Heterogeneous Federated Learning via Importance-Aware Submodel Extraction. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 37, 115615–115651.

Xia, Z.; Hu, M.; Yan, D.; Liu, R.; Li, A.; Xie, X.; and Chen, M. 2025. MultiSFL: Towards Accurate Split Federated Learning via Multi-Model Aggregation and Knowledge Replay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 914–922.

Xia, Z.; Hu, M.; Yan, D.; Xie, X.; Li, T.; Li, A.; Zhou, J.; and Chen, M. 2024. CaBaFL: Asynchronous Federated Learning via Hierarchical Cache and Feature Balance. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43(11): 4057–4068.

Yan, D.; Hu, M.; Xie, X.; Yang, Y.; and Chen, M. 2025. S2FL: Toward Efficient and Accurate Heterogeneous Split Federated Learning. *IEEE Transactions on Computers*, 1–14.

Zhou, H.; Lan, T.; Venkataramani, G. P.; and Ding, W. 2023. Every Parameter Matters: Ensuring the Convergence of Federated Learning with Dynamic Heterogeneous Models Reduction. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 36, 25991–26002.

Zhou, S.; and Li, G. Y. 2023. Federated Learning Via Inexact ADMM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9699–9708.