FEDCASA: FEDERATED CONSENSUS ALIGNMENT VIA SELECTIVE AGGREGATION IN LOW-RANK SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid evolution of Large Language Models (LLMs) has driven widespread adoption of LoRA-based Federated Learning (FL), accompanied by increasingly diverse application demands. While some scenarios require a generalized model equipped with broad domain knowledge, others demand highly specialized models tailored to specific local tasks. However, existing federated LoRA frameworks struggle to simultaneously achieve both robust global generalization and effective local personalization. To address this gap, we propose **FedCASA** (**Fed**erated **C**onsensus **A**lignment via **S**elective **A**ggregation). By exploiting the inherent rank independence and directional discrepancy of LoRA, our approach decouples the client-side LoRA into global consensus component and local personalized component. We evaluate both the derived global generalized model and the client-specific personalized models on multi-task instruction-tuning datasets across varying degrees of data heterogeneity. Extensive experiments demonstrate that FedCASA consistently outperforms existing baselines in both global and local performance, while simultaneously maintaining low communication costs. Our code is available at `https://anonymous.4open.science/r/Phrolova-72`.

1 INTRODUCTION

The rapid evolution of Large Language Models (LLMs) has revolutionized various Natural Language Processing (NLP) tasks, demonstrating remarkable capabilities in comprehension, reasoning, and generation (Brown et al., 2020; Touvron et al., 2023; Team et al., 2023; Liu et al., 2024). However, adapting these powerful pre-trained models to specialized downstream domains typically requires extensive fine-tuning (Wei et al., 2021; Ouyang et al., 2022). In real-world scenarios, centralizing diverse, domain-specific data to fine-tune LLMs is often impractical, and in many cases prohibited, due to increasingly strict data privacy regulations and substantial computational overhead (Yang et al., 2019).

To address these challenges, Federated Learning (FL) has emerged as a promising paradigm, enabling multiple clients to collaboratively train a model without sharing their raw, sensitive data (McMahan et al., 2017; Kairouz & McMahan, 2021). Concurrently, Parameter-Efficient Fine-Tuning (PEFT) techniques (Houlsby et al., 2019), particularly Low-Rank Adaptation (LoRA) (Hu et al., 2021), have been proposed to significantly reduce the computational and memory costs of LLM adaptation. The seamless combination of these two advances, commonly referred to as Federated LoRA, has rapidly become the standard framework for privacy-preserving and efficient LLM fine-tuning across distributed networks (Kuang et al., 2024). As Federated LoRA frameworks are increasingly deployed across highly heterogeneous clients, ranging from edge devices to enterprise data silos, they must satisfy diverse application requirements. Some scenarios require a generalized foundation model with robust broad-domain knowledge (Babakniya et al., 2023), whereas others demand highly specialized models tailored to specific local tasks (Tan et al., 2022).

However, under severe data heterogeneity, existing Fed-LoRA frameworks struggle to reconcile global and local objectives. Conventional approaches attempt to aggregate all local updates into a single global model, which inevitably suffers from severe client drift. Such forced aggregation dilutes domain-specific knowledge, degrading the global model's performance across shifted domains. In contrast, Personalized Federated Learning (pFL) paradigms retain client-specific parameters to improve local accuracy, but they often sacrifice global consensus. By fragmenting the learned

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

knowledge across, pFL fails to produce a fully functional and deployable global model at the server. Consequently, a fundamental dilemma remains: existing architectures cannot simultaneously obtain a robust generalized global model and highly effective personalized local models.

To break this bottleneck, we take a closer look at the internal structure of LoRA matrices. Existing methods typically treat these local updates as solid black boxes, and directly average them during aggregation. However, we find that the internal components of LoRA actually exhibit **rank independence**, meaning that different ranks can capture distinct information. This observation leads to two key insights. First, within the same client, different ranks may serve distinct roles: some lean toward shared global features, while others focus on specific local features. Second, across different clients, the directional information represented by the same rank can also vary substantially. Some ranks align with the global trend, while others deviate from it. These insights reveal that simply averaging all parameters forces conflicting knowledge to be mixed. It is therefore highly intuitive and necessary to decouple these local updates based on their directions: separating the shared global consensus from the distinct local variations.

Motivated by these findings, we propose **FedCASA** (**Fed**erated **C**onsensus **A**lignment via **S**elective **A**ggregation), a framework designed to resolve the conflict between global generalization and local personalization. The core idea of FedCASA is to explicitly separate shared knowledge from private knowledge based on rank-level directional alignment. Specifically, our framework evaluates the individual ranks of the local LoRA matrix and calculates their similarities to the corresponding global rank. Based on this similarity, the local LoRA module is decoupled into a global component and a local component. During each communication round, only the global components that align with the shared consensus, are uploaded and aggregated at the server. Conversely, the local components, which encode client-specific variations, are strictly excluded from global aggregation and retained locally. This selective aggregation mechanism prevents biased local updates from degrading the global foundation model while preserving personalized expertise on each client. Our main contributions are summarized as follow:

- To the best of our knowledge, we are the first to simultaneously train a full functional global foundation model and highly accurate personalized local models in Federated Lora, thereby breaking the conventional trade-off between global generalization and local personalization.
- We propose FedCASA, which dynamically decomposes the local LoRA module into global and local components based on the rank-wise similarity in the B matrix. By aggregating only the global components, FedCASA prevents knowledge interference and reduces communication overhead.
- Extensive multi-task instruction-tuning experiments on LLMs demonstrate that FedCASA consistently outperforms existing baselines in both global generalization and local personalization across diverse Non-IID settings.

2 Related work

2.1 LoRA in Federated Learning

Low-Rank Adaptation (LoRA) (Hu et al., 2021) is a parameter-efficient fine-tuning method that substantially reduces training costs by introducing low-rank trainable matrices while keeping the pre-trained weight matrices frozen. Yi et al. (2023) proposed FedLoRA, which integrates LoRA into Federated Learning (FL) and significantly reduces both communication overhead and resource consumption. To further improve efficiency, subsequent studies have sought to reduce the communication and computational costs of Federated LoRA. For example, Sun et al. (2024) introduced FFA-LoRA, which reduces costs by freezing matrix A. Similarly, LoRA-A^2 (Koo et al., 2025) and RoLoRA (Chen et al., 2026a) achieve cost reductions by alternately freezing different LoRA components.

Beyond resource efficiency, another line of work focuses on mitigating server-side aggregation bias. FlexLoRA (Bai et al., 2024) transforms heterogeneous LoRA modules into homogeneous ones via SVD prior to aggregation, thereby alleviating the impact of client drift. FLoRA (Wang et al., 2024) employs stacked aggregation to achieve noise-free training, enabling unbiased aggregation

of heterogeneous adapters. To address the high communication overhead of FLoRA, FedExLoRA (Singhal et al., 2024) introduces full-matrix residual terms to correct bias while utilizing SVD to compress communication costs. LoRA-FAIR (Bian et al., 2025) goes a step further by simplifying the residual terms to lower the communication overhead. FedLEASE (Wang et al., 2026a) optimizes federated LoRA fine-tuning by dynamically clustering clients to allocate and select tailored domain-specific experts, thereby effectively addressing data heterogeneity while preserving communication efficiency. FedOTAB (Chen et al., 2026b) mitigates aggregation noise by alternately fixing and optimizing LoRA matrix pairs, enabling effective large language model adaptation across heterogeneous, resource-constrained devices.

2.2 Global-local Decoupling in Federated LoRA

Although the aforementioned work has minimized client-side drift, client drift induced by the inherent local characteristics of client data remains inevitable. Simultaneously, the demand for client-side personalization has increasingly come to the forefront. Consequently, disentangling global and local features has emerged as an alternative paradigm to enhance the performance of federated models, giving rise to Personalized Federated Learning (PFL). To this end, FedDPA (Yang et al., 2024) employs dual adapters to learn global and local features separately. FedALT (Bian et al., 2026) introduces Individual and Rest-of-the-World adapters, leveraging the Mixture-of-Experts (MoE) concept to integrate local and global representations. Furthermore, FedSA-LoRA (Guo et al., 2025) mitigates communication overhead by uploading matrix A for server-side aggregation while retaining matrix B locally for personalization. FedDynMask (Wang et al., 2026b) further reduces communication costs by exclusively uploading parameters with large absolute magnitudes for aggregation, thereby simultaneously preserving local features.

However, existing methods either overlook personalization requirements or fail to provide a complete global model. To address these limitations, we propose FedCASA, which decouples LoRA into multiple rank-1 matrices. By selectively aggregating the components that capture global features while retaining the local-specific components, FedCASA simultaneously trains a comprehensive global model and personalized models tailored to the individual demands of each client.

3 Motivate Example

Preliminary Low-Rank Adaptation (LoRA) projects model updates onto a low-dimensional basis vector space, with its rank independence originating from the additive nature of low-rank matrix decomposition. Within the LoRA framework, the pre-trained weights $W_0 \in R^{d\times k}$ remain frozen, and the fine-tuning process updates only the incremental weight matrix. By definition, this update can be expressed as the product of two low-rank matrices: $\Delta W = B \cdot A$. By further decomposing this matrix multiplication, the update can be rewritten as the sum of r rank-1 matrices: $\Delta W = \sum_{i=1}^{r} b_i a_i^\top$, where b_i is the i-th column vector of matrix B and a_j is the i-th row vector of matrix A. Each outer product term $u_i = b_i a_i^\top$ can be conceptualized as an independent training unit, encapsulating a specific update dimension or rank direction within the model's parameter space. In other words, a LoRA module of rank r is mathematically equivalent to the superposition of r independent rank-1 LoRA modules. Given an input x, the output h following the LoRA update is computed as follows:

$$h = W_0 x + \Delta W x = W_0 x + \sum_{i=1}^{r} b_i (a_i^\top x). \tag{1}$$

Evidently, during the forward pass, the computations across different ranks are strictly additive. Thus, they operate independently without mutual interference. Furthermore, examining the back-propagation phase:

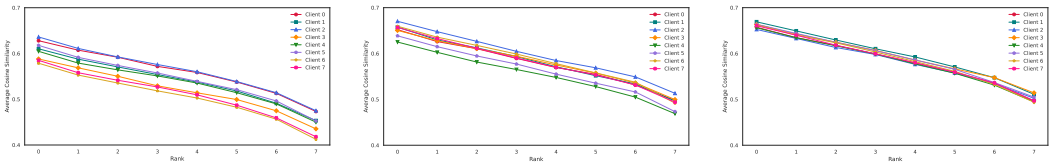
$$\frac{\partial L}{\partial b_i} = \frac{\partial L}{\partial h} \cdot \frac{\partial h}{\partial b_i} = \frac{\partial L}{\partial h} \cdot (a_i^\top x), \qquad \frac{\partial L}{\partial a_i} = \frac{\partial L}{\partial h} \cdot \frac{\partial h}{\partial a_i} = \frac{\partial L}{\partial h} \cdot (b_i^\top x), \tag{2}$$

it is apparent that the gradient computation for each rank remains entirely isolated as well. Consequently, we can conclude that the individual ranks within LoRA are strictly independent—interacting with neither one another during the forward pass nor the backward pass. This intrinsic independence establishes the theoretical foundation for our proposed algorithm.

Numerous studies have demonstrated that the slow convergence of traditional federated learning primarily stems from data heterogeneity across clients, which induces divergent and conflicting gradient update directions (Huang et al., 2024; Li et al., 2020; 2021; Wang et al., 2020). Consequently, naive element-wise aggregation significantly retards the model's learning process. To address this, we investigate the relationship between local and global optimization directions within each rank of the LoRA modules during the instruction tuning of text generation models. Specifically, we use Dolly dataset (databricks-dolly-15k)[1], an instruction fine-tuning dataset containing eight task types, and partition them among 8 clients. We also set three configurations of varying degrees of data heterogeneity.(See Section 5 for details) Under these three configurations, we conduct one communication round of local training using LLaMA-2 7B Touvron et al. (2023). Following the methodology in FedIT (Zhang et al., 2024), we directly aggregate matrices A and B independently via FedAvg (McMahan et al., 2017) to obtain the global LoRA weights for the initial round.

Guo et al. (2025) observed that in federated learning for text classification models using LoRA, matrix A predominantly captures global features, with its update directions aligning across all clients. Conversely, matrix B is more reflective of the characteristics of client-specific local data, exhibiting pronounced inter-client variance. We compute the average cosine similarity across the aforementioned clients for each rank within the LoRA modules (See the details in A.2). We found that similar conclusions hold for instruction tuning tasks in text generation models, even when analyzed at the granularity of individual ranks. To reduce computational overhead, we restrict our subsequent calculations of rank-wise directional similarity exclusively to matrix B.

Furthermore, we compare each rank of every client's LoRA module with the corresponding rank in the global LoRA, computing the cosine similarities, sorting them in descending order, and averaging the results across clients, as illustrated in Figure 1. We observe that across all scenarios, a clear discriminative margin exists between the ranks most aligned with the global direction and those least aligned, with a consistent difference of approximately 0.1 to 0.2. In addition, overall similarity decreases as the degree of data heterogeneity increases. This demonstrates that our approach of decomposing the rank into global and local components is feasible across heterogeneous scenarios of varying degrees.



(a) high data heterogeneity (b) moderate data heterogeneity (c) iid data

Figure 1: similarity between local and global under diverse degrees of data heterogeneity

4 Our Method

4.1 FedCASA: Consensus-Aware Selective Aggregation

Motivated by our observations, we propose Federated Consensus Alignment via Selective Aggregation (FedCASA), which partitions the ranks of each client-side LoRA module into global and local components. The global components are updated for server-side aggregation, whereas the local components are retained by their respective clients, as illustrated in Figure 2. We denote the local LoRA adapter matrices of client i at communication round t as $A_i^{(t)}$ and $B_i^{(t)}$. To promote greater orthogonality and independence among the individual ranks, we apply Singular Value Decomposition (SVD) to the pre-trained weight matrix before local training. The resulting decomposed matrices are then utilized to initialize the LoRA modules distributed to each client:

$$U_r, \Sigma_r, V_r = \text{SVD}_r(W_{base}), A_g^{(0)} = \Sigma_r^{\frac{1}{2}} V_r^\top, B_g^{(0)} = U_r \Sigma_r^{\frac{1}{2}},$$

where $\text{SVD}_r(\cdot)$ denotes the rank-r truncated SVD process applied to the pre-trained matrix. Also, the LoRA weights need to be subtracted from the base model weights: $W_{base} = W_{base} - B_g^{(0)} A_g^{(0)}$.

[1] https://huggingface.co/datasets/databricks/databricks-dolly-15k

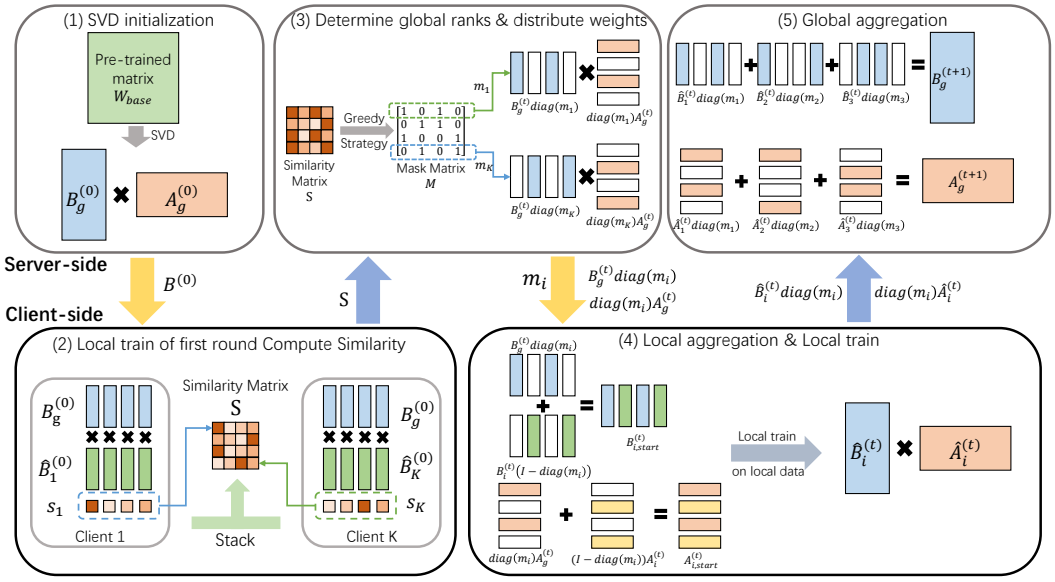


Figure 2: Overview of FedCASA. The workflow consists of five steps, with server-side operations shown in the upper part and client-side operations shown in the lower part. After Step (5) is completed in the first communication round, the process returns to Step (3), where the updated weights are distributed to clients. This cycle is repeated until training converges.

Upon the first round of local training, we represent the updated matrix $B_i^{(1)}$ of client i as a collection of r column vectors. Let the initial matrix $B^{(0)}$ and the matrix $B_i^{(0)}$ after local training of the first round on client i be denoted by their respective column vectors:

$$B_g^{(0)} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r], \quad B_i^{(0)} = [\mathbf{b}_{i,1}, \mathbf{b}_{i,2}, \ldots, \mathbf{b}_{i,r}].$$

We then compute the cosine similarity between each of these vectors and the corresponding initial vectors from the global initialization $B^{(0)}$. The cosine similarity for the k-th rank is computed as: $s_{i,k} = \cos(\mathbf{b}_{i,k}, \mathbf{u}_k)$, for $\forall k \in \{1, 2, \ldots, r\}$. Accordingly, the similarity vector representing the divergence of the ranks for client i is given by $\mathbf{s}_i = [s_{i,1}, s_{i,2}, \ldots, s_{i,r}]^\top$.

Following the computation of the local similarity, the clients upload their respective similarity vectors to the server, which are then stacked to construct the global similarity matrix $S \in \mathbb{R}^{K \times r}$, where K denotes the total number of clients:

$$S = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K]^\top.$$

Let $\beta \in (0, 1)$ be the predefined aggregation ratio. To determine which ranks participate in the global aggregation and which are retained locally, we introduce a binary mask matrix $M \in \{0, 1\}^{K \times r}$, initialized with all zeros. Specifically, $M_{i,j} = 1$ indicates that the j-th rank of client i is selected for global aggregation. To ensure balanced aggregation across clients and ranks, the selection must satisfy two capacity constraints:

1. Client-wise Balance: Each client contributes exactly $m = \lfloor \beta r \rfloor$ ranks to the global model.
2. Rank-wise Balance: Each rank is aggregated across exactly $c = \lfloor \beta K \rfloor$ clients.

Subject to these constraints, we aim to maximize the overall consensus of the selected ranks. To this end, we design a greedy selection algorithm based on the global similarity matrix S. Specifically, we iteratively pick the maximum unassigned similarity score in S. Suppose the current maximum value is $S_{i,j}$. The corresponding rank j of client i is selected, and we update $M_{i,j} = 1$, provided that neither of the following capacity constraints is violated:

$$\forall i, \sum_{k=1}^{r} M_{i,k} < m, \qquad \forall j, \sum_{k=1}^{K} M_{k,j} < c.$$

If either constraint is already reached, which means client i has already selected $\lfloor \beta r \rfloor$ ranks, or rank j has already been selected by $\lfloor \beta K \rfloor$ clients, we skip $S_{i,j}$ and proceed to the next largest similarity score in S. This iterative process continues until all balance conditions are met. Upon determining the mask matrix M, the server broadcasts the respective binary mask row $\mathbf{m}_i^\top$ (where $\mathbf{m}_i \in \{0,1\}^r$ corresponds to the i-th row of M) to each client i. To ensure the stability of FL (see Section for experimental verification), the fixed mask matrix is maintained across all subsequent training rounds, meaning that clients only upload and update their designated global components.

After the global ranks for each client are determined, we start the federated training cycle. First, the server sends the current globally aggregated matrices $diag(\mathbf{m}_i))A_g^{(t)}$ and $B_g^{(t)}diag(\mathbf{m}_i))$, which are masked by $\mathbf{m}_i$. Client i merges these global components with its locally retained weights using its mask vector $\mathbf{m}_i$:

$$\begin{aligned} A_{i,\text{start}}^{(t)} &= \text{diag}(\mathbf{m}_i)A_g^{(t)} + (I_r - \text{diag}(\mathbf{m}_i))A_i^{(t)}, \\ B_{i,\text{start}}^{(t)} &= B_g^{(t)}\text{diag}(\mathbf{m}_i) + B_i^{(t)}(I_r - \text{diag}(\mathbf{m}_i)). \end{aligned} \tag{3}$$

where I_r is the $r \times r$ identity matrix, and $A_i^{(t)}, B_i^{(t)}$ are the local matrices from the previous round. After optimizing these merged adapters on its private dataset to yield the updated matrices $\hat{A}_i^{(t)}$ and $\hat{B}_i^{(t)}$, client i uploads only its designated global components to the server. This extraction is algebraically represented as $\text{diag}(\mathbf{m}_i)\hat{A}_i^{(t)}$ and $\hat{B}_i^{(t)}\text{diag}(\mathbf{m}_i)$. Finally, the server aggregates these updates to construct the global model for round $t+1$. For each rank $j \in \{1, \ldots, r\}$, the averaged global ranks are computed as:

$$\mathbf{a}_{g,j}^{(t+1)} = \frac{1}{c}\sum_{k=1}^{K} M_{k,j}\hat{\mathbf{a}}_{k,j}^{(t)}, \quad \mathbf{b}_{g,j}^{(t+1)} = \frac{1}{c}\sum_{k=1}^{K} M_{k,j}\hat{\mathbf{b}}_{k,j}^{(t)}, \tag{4}$$

where $\hat{\mathbf{a}}_{k,j}^{(t)}$ and $\hat{\mathbf{b}}_{k,j}^{(t)}$ denote the j-th row of $\hat{A}_k^{(t)}$ and the j-th column of $\hat{B}_k^{(t)}$, respectively. Then we get the new global LoRA matrices, $A_g^{(t+1)}$ and $B_g^{(t+1)}$. This completes one full training iteration, maintaining a strict boundary between global consensus and local personalization.

4.2 Theoretical Analysis

We next analyze the convergence behavior of FedCASA. Details of the analysis are in Appendix A.5. This convergence rate is comparable to that of traditional FedAvg in the non-convex scenario (Yu et al., 2019).

Theorem 1 (Convergence of FedCASA)**.** *Under smooth and non-convex conditions, FedCASA achieves an* $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ *convergence rate to a stationary solution for both the global model and the local personalized models.*

5 Experiments

To evaluate the effectiveness of our proposed method, we conduct experiments on the Dolly dataset using LLaMA-2 7B. We evaluate the performance of FedCASA in terms of both global model generalization and local model personalization. We set 8 clients, each allocated 480 training samples and 120 test samples. We use the same template as FedIT (see Appendix A.1). For parameter-efficient fine-tuning, we integrate LoRA into the query and value projection matrices (W_q and W_v) with a rank of $r = 8$. Specifically, we set the scaling factor α to 16 and the LoRA dropout rate to 0.05. To balance model convergence and communication overhead (as discussed in Section 5.4), we set the client participation ratio to 0.5. All experiments are conducted on a single NVIDIA A800 GPU using BF16 mixed precision, with a batch size of 32 and a learning rate of 5×10^{-4}. All the experiments metrics are ROUGE-1 (Lin, 2004). Our implementation is based on the code released for FedIT (Zhang et al., 2024).

5.1 Global model generalization

We establish three data heterogeneity scenarios: two non-IID settings with varying degrees of heterogeneity and one IID setting. In the high heterogeneity scenario, each client is assigned an in-

Table 1: Global model performance comparison with FedCASA

Method	High Heterogeneity		Moderate Heterogeneity		iid	
	ROUGE-1	Impr.	ROUGE-1	Impr.	ROUGE-1	Impr.
FedIT	38.97	+1.41%	38.83	+3.37%	38.39	+4.32%
FFA-LoRA	38.73	+2.04%	38.74	+3.61%	38.39	+4.32%
FLoRA	38.85	+1.72%	37.87	+5.99%	37.48	+6.86%
FedEx-LoRA	38.45	+2.78%	38.43	+4.45%	39.10	+2.43%
FedOTAB	38.16	+3.56%	39.76	+0.96%	39.20	+2.17%
FedCASA	**39.52**	-	**40.14**	-	**40.05**	-

dependent task type. In moderate heterogeneity scenario each client holds data for four different task types, with equal quantities for each. In the IID scenario, each client gets data of all eight task types. We train the model for 10 rounds with 1 epoch per round and test the model on a test dataset combined the test datasets of all the clients.

Baseline To evaluate the global generalization performance of our proposed method, we compare it with several state-of-the-art methods in federated fine-tuning global model with LoRA. **FedIT** (Zhang et al., 2024) aggregates the LoRA by averaging the matrices A and B separately and represents the vanilla FedAvg + LoRA approach. **FFA-LoRA** (Sun et al., 2024) fixes matrix A and fine-tuning only matrix B which addresses server-side aggregation bias and reduces computation and conmmunication cost. **FLoRA** (Wang et al., 2024) addresses server-side aggregation bias by stacking local LoRA modules as the global matrices. **FedEx-LoRA** (Singhal et al., 2024) add the residual error $\Delta W_g^* - \Delta W_g$ to the base model to achieve exact aggregation and uses SVD to compress the residual error matrix for communication efficiency. **FedOTAB** (Chen et al., 2026b) accurately aggregates local LoRA modules and then minimizes induced noise by alternately fixing matrix A or B to optimize their optimal approximations.

As shown in Table 1, FedCASA consistently outperforms all baseline methods across various data heterogeneity settings, achieving state-of-the-art performance. Under the IID setting, FedCASA attains 40.05, surpassing the second-best baseline FedOTAB by 2.17%. More importantly, FedCASA demonstrates remarkable robustness against data heterogeneity. While baseline methods like FedOTAB suffer from severe performance drops under high heterogeneity (falling from 39.20 to 38.16), FedCASA maintains superior stability, achieving 39.52 and outperforming the strongest baseline FedIT by 1.41%. These results validate the exceptional efficacy of FedCASA in handling complex non-IID distributions.

5.2 Local model personalization

We test the local personalization performance of our proposed method under the high heterogeneity scenario set in 5.1. We train the model for 5 rounds with 2 epochs per round and test the local models on the local test dataset.

Baseline To evaluate the local personalization performance of our proposed method, we compare it with several state-of-the-art methods in fine-tuning personalized federated model with LoRA. **Local** means that each client trains its own model locally, without server-side aggregation. **FedDPA** uses two adapters to represent global and local information separately and trains them alternatively. **FedSA** trains A matrix and aggregates on server while keeping B matrix local. **FedDynMask** only uploads the most significant parameters while keeping the others within clients.

As shown in Table 2, our FedCASA consistently outperforms local training and all state-of-the-art federated baseline methods across a wide spectrum of tasks, achieving the highest overall average score of 41.82. Specifically, FedCASA demonstrates exceptional superiority in six tasks. While baseline methods exhibit performance drops or trade-offs in complex generative and open-ended scenarios, FedCASA maintains a robust and well-balanced performance profile across all dimensions. These comprehensive improvements strongly validate the effectiveness of our design in mitigating statistical heterogeneity and boosting multi-task generalization under federated constraints.

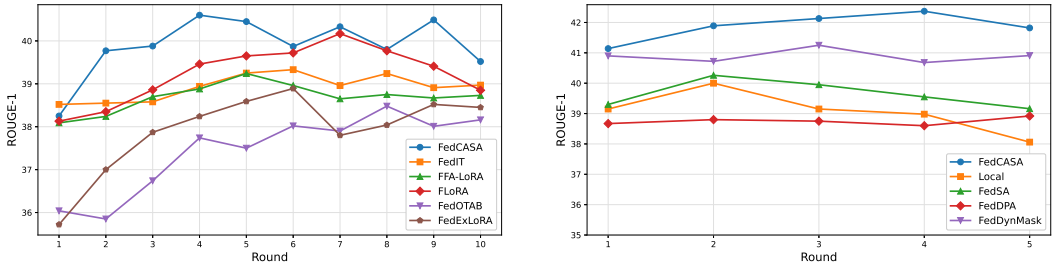
Table 2: Local personalization performance comparison with FedCASA

Method	Close QA	Classification	Open QA	Info Extraction	Brainstorming	General QA	Summarization	Creative Writing	Average
Local	48.10	71.51	30.75	50.06	24.02	24.16	42.54	24.88	39.50
FedDPA	**48.69**	71.09	33.84	46.53	23.36	25.49	41.51	20.87	38.92
FedSA	44.33	69.64	28.84	47.73	27.00	28.66	41.82	25.24	39.16
FedDynMask	48.29	72.20	**36.39**	48.08	26.91	28.89	40.39	26.13	40.91
FedCASA	47.72	**74.45**	33.54	**50.19**	**27.48**	**29.63**	**44.38**	**27.20**	**41.82**

5.3 In-depth analysis

5.3.1 Effect of training rounds

To investigate the training efficiency and convergence performance of FedCASA compared to other baseline methods, we compared the performance of various algorithms across different numbers of training rounds. The results are shown in Figure 3.



(a) global model

(b) local personalized model

Figure 3: Performance of methods under different training rounds

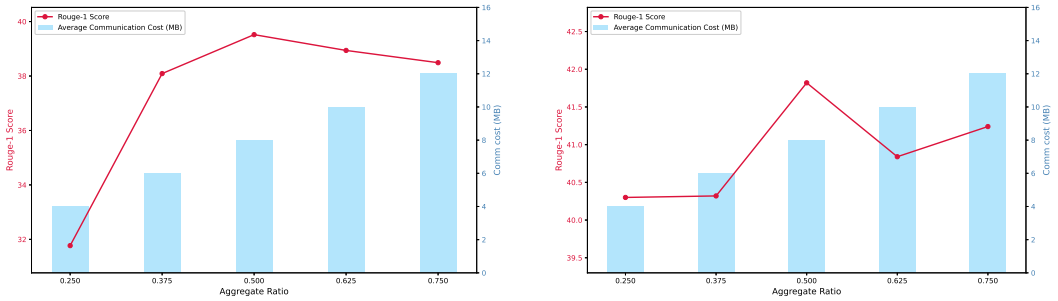
Figure 3a illustrates the performance of the global model. As observed, FedCASA exhibits significantly faster convergence compared to all baselines, rapidly reaching a stable plateau. Specifically, FedCASA achieves its peak performance (40.60) as early as the 4th communication round, after which it remains highly stable with only marginal fluctuations. In contrast, among the baselines, FedEx-LoRA is the fastest to reach its optimum (at round 6), yet it suffers from severe performance degradation in subsequent rounds. Notably, FedCASA consistently outperforms all competing methods across all communication rounds, firmly validating its superior stability and efficiency in training the global model.

Figure 3b illustrates the performance of the local personalized models. In contrast to the global model dynamics, all baseline methods peak prematurely often within the first or second round and subsequently converge at a suboptimal level. FedCASA, however, steadily attains its peak performance (42.37) at round 4, substantially outperforming all other baselines. This further demonstrates the exceptional stability and efficacy of FedCASA in the context of personalized federated learning.

5.3.2 Effect of aggregation ratio

As shown in Figure 4, we also investigated the performance of FedCASA at aggregation ratio $\beta = \{0.25, 0.375, 0.5, 0.625, 0.75\}$, as well as the corresponding communication costs.

As observed, $\beta = 0.5$ consistently emerges as the optimal choice for both global and local personalized models. When the aggregation ratio is excessively high, performance degrades marginally, likely because a large number of conflicting rank directions are forcibly aggregated. Conversely, a lower aggregation ratio leads to a pronounced performance drop, which is particularly evident in the global model. This can be attributed to insufficient client feedback transmitted to the global model, resulting in under-training. Furthermore, $\beta = 0.5$ represents a moderate choice regarding communication overhead. Therefore, an aggregation ratio of 0.5 strikes an optimal trade-off between model performance and communication cost for FedCASA.



(a) global model

(b) local personalized model

Figure 4: Performance and communication cost of methods under different aggregation ratios

5.4 Communication Efficiency

Table 3 in Appendix A.3 shows the comparison of the communication cost excluding initial model allocation between our method and baselines. It can be seen that our communication overhead is lower than or equal to that of most methods, demonstrating the advantage of our approach in terms of communication cost. Communication overheads other than weight transmission in our method, such as similarity uploads, amount to only about 1 KB and are therefore not taken into account.

5.5 Ablation Study

The default FedCASA utilizes SVD initialization, B-matrix similarity for consensus alignment, similarity-based selective aggregation, and a static updating mask. To evaluate the core mechanisms of FedCASA, we conduct an ablation study (Table 4 in Appendix A.4) demonstrating that each component is vital. Removing SVD initialization in favor of standard LoRA initialization causes severe performance drop, as zero-initializing matrix B strips the model of the directional information required for the critical first-round similarity assessment. Furthermore, incorporating matrix A into the similarity calculation with a 50% weight degrades performance, confirming that matrix B alone serves as a more accurate and efficient proxy for local feature projections without introducing alignment noise. Replacing our similarity-based selection with random rank selection also reduces the score by blindly mixing parameters, which overwrites personalized knowledge and pollutes the global model. Finally, dynamically updating the aggregation mask at each communication round yields a suboptimal result compared to our static approach, which shows retaining a static mask from the initial round effectively prevents parameter thrashing and maintains a stable, structural boundary between the shared global sub-space and the personalized local sub-space throughout training.

6 Conclusion

In this paper, we proposed FedCASA, a novel federated learning framework designed to resolve the persistent conflict between global generalization and local personalization in Federated LoRA. By investigating the internal structure of low-rank spaces, we revealed that local LoRA updates exhibit rank independence and directional discrepancy. Building upon this insight, FedCASA dynamically decouples the local LoRA matrix into global and local components based on rank-wise similarity, selectively aggregating only the consensus-aligned knowledge. Extensive experiments on multi-task LLM instruction tuning demonstrate that our approach successfully breaks the conventional trade-off. FedCASA simultaneously achieves state-of-the-art performance for both the global foundation model and client-specific personalized models, all while maintaining high communication efficiency. Ultimately, this work provides a highly practical and scalable paradigm for deploying tailored LLMs across privacy-sensitive and heterogeneous environments.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and grammar correction. We have not used generative AI tools for core scientific ideation or mathematical proof derivation, and experimental design are not applicable to this work. Additionally, we used generative AI tools for assistance with code writing and mathematical formulas writing. We have reviewed all AI-assisted work. All LLM-generated or polished text was carefully checked by the authors for factual accuracy and academic tone. Furthermore, any AI-suggested code snippets were manually tested, debugged, and verified for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our complete experimental code is at `https://anonymous.4open.science/r/Phrolova-72`. The complete convergence analysis of FedCASA is provided in the Appendix A.5. The datasets used in our experiments are all publicly available. Detailed experimental hyperparameters are listed in the Section 5.

REFERENCES

Sara Babakniya, Ahmed Roushdy Elkordy, Yahya H Ezzeldin, Qingfeng Liu, Kee-Bong Song, Mostafa El-Khamy, and Salman Avestimehr. Slora: Federated parameter efficient fine-tuning of language models. *arXiv preprint arXiv:2308.06522*, 2023.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. *Advances in Neural Information Processing Systems*, 37:14457–14483, 2024.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3737–3746. IEEE, 2025.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Fedalt: Federated fine-tuning through adaptive local training with rest-of-world lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19728–19736, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Shuangyi Chen, Yuanxin Guo, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish Khisti. Robust federated finetuning of llms via alternating optimization of lora. *Advances in Neural Information Processing Systems*, 38:120038–120090, 2026a.

Yiming Chen, Ruanjun Li, Jiawei Shao, Lifeng Sun, Chi Zhang, Jun Zhang, and Xuelong Li. Federated lora fine-tuning of llms with only transmitting matrix a or b. *IEEE Transactions on Mobile Computing*, 2026b.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *International Conference on Learning Representations*, volume 2025, pp. 99003–99027, 2025.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Xinmeng Huang, Ping Li, and Xiaoyun Li. Stochastic controlled averaging for federated learning with communication compression. In *International Conference on Learning Representations*, volume 2024, pp. 27956–27994, 2024.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Jabin Koo, Minwoo Jang, and Jungseul Ok. Towards robust and efficient federated low-rank adaptation with heterogeneous clients. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 416–429, 2025.

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5260–5271, 2024.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10708–10717. IEEE, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. Fedex-lora: Exact aggregation for federated and efficient fine-tuning of foundation models. *arXiv preprint arXiv:2410.09432*, 2024.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning. In *International conference on learning representations*, volume 2024, pp. 17978–17994, 2024.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Lei Wang, Jieming Bian, Letian Zhang, and Jie Xu. Adaptive lora experts allocation and selection for federated fine-tuning. *Advances in Neural Information Processing Systems*, 38:76018–76045, 2026a.

Yan Wang, Ziyi Gao, Yida Zhang, and Rui Wang. Feddynmask: Efficient federated fine-tuning for edge llms via dynamic sparse masking. In *IEEE INFOCOM 2026-IEEE Conference on Computer Communications*, pp. 1–10. IEEE, 2026b.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *Advances in Neural Information Processing Systems*, 37:22513–22533, 2024.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. Dual-personalizing adapter for federated foundation models. *Advances in Neural Information Processing Systems*, 37:39409–39433, 2024.

Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. pfedlora: model-heterogeneous personalized federated learning with lora tuning. *arXiv preprint arXiv:2310.13283*, 2023.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 5693–5700, 2019.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. 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.

A Appendix

A.1 Prompt template

The prompt template we use is Figure 5.

A.2 Similarity across clients Analysis

We computed the cross-client cosine similarity for the same rank in matrices A and B, respectively, with the average values plotted in Figure 6. As shown, the similarities for matrix A approach 1, indicating highly aligned directions. Conversely, matrix B exhibits significantly lower similarities, demonstrating substantial directional discrepancy.

A.3 Communication Cost Comparison of FedCASA

See Table 3.

A.4 Ablation Study Results

See Table 4.

Prompt Template (with input)

```
Below is an instruction that describes a task, paired with an input
that provides further context.  Write a response that appropriately
completes the request.

### Instruction:
{instruction}

### Input:
{input}

### Response:
```

Prompt Template (without input)

```
Below is an instruction that describes a task.  Write a response
that appropriately completes the request.

### Instruction:
{instruction}

### Response:
```

Figure 5: The Alpaca-LoRA prompt templates used in our experiments, showing configurations for tasks with and without additional input context.

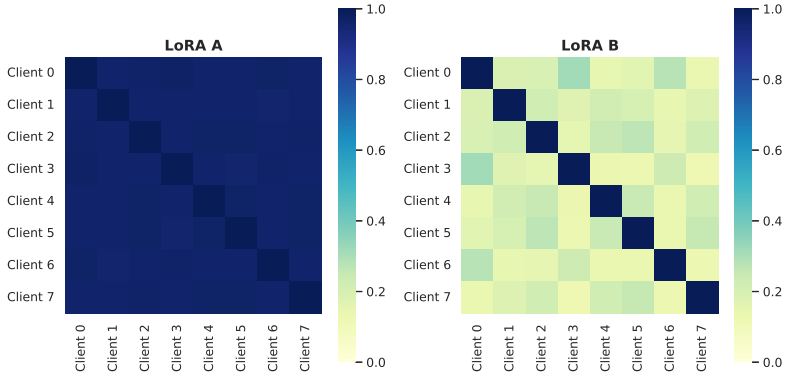


Figure 6: Heatmap of the similarity across clients

A.5 Convergence Analysis

FedCASA separates each LoRA adapter into global and local components through the rank selection mask. In this section we analyze the convergence behavior of the global consensus model and the personalized client models. The assumptions we used is the same as those used in Guo et al. (2025).

Assumption 1 (Smoothness)**.** *For every client $i \in \{1, \dots, K\}$, the local objective $L_i(\cdot)$ is L-smooth. For any matrices X and Y,*

$$L_i(Y) \leq L_i(X) + \langle \nabla_W L_i(X), Y - X \rangle_F + \frac{L}{2}\|Y - X\|_F^2. \tag{5}$$

Assumption 2 (Stochastic Gradient)**.** *For client i at step t, let*

$$G_i^{(t)} = \nabla_W L_i\left(W_i^{(t)}\right), \qquad g_i^{(t)} = \nabla_W L_i\left(W_i^{(t)}; \xi_i^{(t)}\right). \tag{6}$$

The stochastic gradient is unbiased, bounded, and satisfies a relative variance condition:

$$\mathbb{E}_t\left[g_i^{(t)}\right] = G_i^{(t)}, \qquad \|g_i^{(t)}\|_F \leq G, \qquad \mathbb{E}_t\left[\|g_i^{(t)} - G_i^{(t)}\|_F^2\right] \leq \nu^2 \|G_i^{(t)}\|_F^2. \tag{7}$$

Table 3: Communication Cost Comparison ($\beta = 0.5$)

Method	Round Average Communication Cost per client (MB)
FedIT	16.00
FFA-LoRA	8.00
FLoRA	72.00
FedOTAB	8.00
FedEx-LoRA	72.00
FedDPA	16.00
FedSA	8.00
FedDynMask	0.80
FedCASA (Ours)	8.00

Table 4: Ablation study on the key components of FedCASA

Method Variant	Score
FedCASA	**39.52**
w/o SVD Initialization	36.11
w/ Sim(A, B)	38.91
w/ Random Selection	38.59
w/ Dynamic Masking	38.86

Define

$$C_\nu := 1 + \nu^2. \tag{8}$$

Then

$$\mathbb{E}_t\left[\|g_i^{(t)}\|_F^2\right] \le C_\nu \|G_i^{(t)}\|_F^2. \tag{9}$$

Assumption 3 (Bounded LoRA Factors and Gradient Alignment)**.** *The global LoRA factors satisfy*

$$\|A_g^{(t)}\|_F \le C_A, \qquad \|B_g^{(t)}\|_F \le C_B. \tag{10}$$

Moreover, there exist constants $c_A > 0$ and $c_B > 0$ such that

$$\left\langle A_g^{(t)\top} D_i A_g^{(t)}, G_i^{(t)\top} G_i^{(t)} \right\rangle_F \ge c_A \|G_i^{(t)}\|_F^2, \tag{11}$$

and

$$\left\langle B_g^{(t)} D_i B_g^{(t)\top}, G_i^{(t)} G_i^{(t)\top} \right\rangle_F \ge c_B \|G_i^{(t)}\|_F^2. \tag{12}$$

Let the mask matrix of client i be

$$D_i = \mathrm{diag}(\mathbf{m}_i), \tag{13}$$

where $D_{i,jj} = 1$ indicates that rank j is selected as a global rank by client i. Each client selects

$$m = \lfloor \beta r \rfloor \tag{14}$$

global ranks. For rank j, define its participating client set as

$$\mathcal{C}_j = \{i : M_{ij} = 1\}, \qquad |\mathcal{C}_j| = c, \tag{15}$$

where $c = \lfloor \beta K \rfloor$. The client model before local optimization is

$$W_i^{(t)} = W_0 + B_g^{(t)} D_i A_g^{(t)} + B_i^{l,(t)} (I - D_i) A_i^{l,(t)}. \tag{16}$$

The global-rank factors after one local update are

$$\widehat{A}_{i,g}^{(t+1)} = D_i A_g^{(t)} - \eta D_i B_g^{(t)\top} g_i^{(t)}, \qquad \widehat{B}_{i,g}^{(t+1)} = B_g^{(t)} D_i - \eta g_i^{(t)} A_g^{(t)\top} D_i. \tag{17}$$

The corresponding model after the local update is

$$U_i^{(t)} = W_0 + \widehat{B}_{i,g}^{(t+1)} \widehat{A}_{i,g}^{(t+1)} + B_i^{l,(t)} (I - D_i) A_i^{l,(t)}. \tag{18}$$

After selective aggregation, define

$$A_{i,g}^{(t+1)} = D_i A_g^{(t+1)}, \qquad B_{i,g}^{(t+1)} = B_g^{(t+1)} D_i, \tag{19}$$

and

$$V_i^{(t)} = W_0 + B_{i,g}^{(t+1)} A_{i,g}^{(t+1)} + B_i^{l,(t)} (I - D_i) A_i^{l,(t)}. \tag{20}$$

The next model is

$$W_i^{(t+1)} = V_i^{(t)}. \tag{21}$$

Define the average gradient measure

$$S_t = \frac{1}{K} \sum_{i=1}^{K} \|G_i^{(t)}\|_F^2. \tag{22}$$

Since $\|G_i^{(t)}\|_F = \|\mathbb{E}_t[g_i^{(t)}]\|_F \le G$, we have

$$S_t \le G^2. \tag{23}$$

Also define

$$c_g = c_A + c_B. \tag{24}$$

Lemma 1 (Descent After the Local Update)**.** *Under Assumptions 1–3, if $\eta G \le 1$, the local update satisfies*

$$\frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_t \left[L_i(U_i^{(t)}) \right] \le \frac{1}{K} \sum_{i=1}^{K} L_i(W_i^{(t)}) - c_g \eta S_t + M_{\text{loc}} \eta^2, \tag{25}$$

where

$$M_{\text{loc}} = C_A C_B G^3 + \frac{3L}{2} \left(C_A^4 + C_B^4 + C_A^2 C_B^2 \right) G^2. \tag{26}$$

Proof. By equation 5,

$$L_i(U_i^{(t)}) \le L_i(W_i^{(t)}) + \left\langle U_i^{(t)} - W_i^{(t)}, G_i^{(t)} \right\rangle_F + \frac{L}{2} \|U_i^{(t)} - W_i^{(t)}\|_F^2. \tag{27}$$

Substituting equation 17 into equation 18,

$$\begin{aligned} U_i^{(t)} - W_i^{(t)} = & -\eta g_i^{(t)} A_g^{(t)\top} D_i A_g^{(t)} - \eta B_g^{(t)} D_i B_g^{(t)\top} g_i^{(t)} \\ & + \eta^2 g_i^{(t)} A_g^{(t)\top} D_i B_g^{(t)\top} g_i^{(t)}. \end{aligned} \tag{28}$$

Taking the Frobenius inner product with $G_i^{(t)}$ and then taking conditional expectation,

$$\begin{aligned} \mathbb{E}_t \left[\left\langle U_i^{(t)} - W_i^{(t)}, G_i^{(t)} \right\rangle_F \right] = & -\eta \left\langle A_g^{(t)\top} D_i A_g^{(t)}, G_i^{(t)\top} G_i^{(t)} \right\rangle_F \\ & - \eta \left\langle B_g^{(t)} D_i B_g^{(t)\top}, G_i^{(t)} G_i^{(t)\top} \right\rangle_F \\ & + \eta^2 \mathbb{E}_t \left[\left\langle g_i^{(t)} A_g^{(t)\top} D_i B_g^{(t)\top} g_i^{(t)}, G_i^{(t)} \right\rangle_F \right]. \end{aligned} \tag{29}$$

By equation 11 and equation 12,

$$\mathbb{E}_t \left[\left\langle U_i^{(t)} - W_i^{(t)}, G_i^{(t)} \right\rangle_F \right] \le -c_g \eta \|G_i^{(t)}\|_F^2 + C_A C_B G^3 \eta^2. \tag{30}$$

Averaging equation 30 over all clients gives

$$\frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_t \left[\left\langle U_i^{(t)} - W_i^{(t)}, G_i^{(t)} \right\rangle_F \right] \le -c_g \eta S_t + C_A C_B G^3 \eta^2. \tag{31}$$

For the squared update, define

$$X_1 = -\eta g_i^{(t)} A_g^{(t)\top} D_i A_g^{(t)}, \quad X_2 = -\eta B_g^{(t)} D_i B_g^{(t)\top} g_i^{(t)}, \quad X_3 = \eta^2 g_i^{(t)} A_g^{(t)\top} D_i B_g^{(t)\top} g_i^{(t)}. \tag{32}$$

Then

$$U_i^{(t)} - W_i^{(t)} = X_1 + X_2 + X_3. \tag{33}$$

Using $\|X_1 + X_2 + X_3\|_F^2 \le 3(\|X_1\|_F^2 + \|X_2\|_F^2 + \|X_3\|_F^2)$,

$$\|X_1\|_F^2 \le \eta^2 C_A^4 \|g_i^{(t)}\|_F^2, \qquad \|X_2\|_F^2 \le \eta^2 C_B^4 \|g_i^{(t)}\|_F^2, \tag{34}$$

and

$$\|X_3\|_F^2 \le \eta^4 C_A^2 C_B^2 \|g_i^{(t)}\|_F^4. \tag{35}$$

Since $\eta G \le 1$,

$$\eta^4 \|g_i^{(t)}\|_F^4 \le \eta^2 \|g_i^{(t)}\|_F^2. \tag{36}$$

Therefore,

$$\mathbb{E}_t\left[\|U_i^{(t)} - W_i^{(t)}\|_F^2\right] \le 3\eta^2 \left(C_A^4 + C_B^4 + C_A^2 C_B^2\right) \mathbb{E}_t\left[\|g_i^{(t)}\|_F^2\right]. \tag{37}$$

Using equation 9 and equation 22,

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_t\left[\|U_i^{(t)} - W_i^{(t)}\|_F^2\right] \le 3C_\nu \left(C_A^4 + C_B^4 + C_A^2 C_B^2\right) \eta^2 S_t \le C_{\text{loc}}^{(d)} \eta^2 G^2, \tag{38}$$

where

$$C_{\text{loc}}^{(d)} = 3C_\nu \left(C_A^4 + C_B^4 + C_A^2 C_B^2\right). \tag{39}$$

Substituting equation 31 and equation 38 into equation 27 and averaging over all clients proves equation 25. □

Lemma 2 (Descent Bound After Selective Aggregation)**.** *Under Assumptions 2 and 3, if $\eta G \le 1$, the selective aggregation satisfies*

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_t\left[L_i(V_i^{(t)})\right] \le \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_t\left[L_i(U_i^{(t)})\right] + \eta\sqrt{C_{\text{agg}}}\, S_t + M_{\text{agg}}\eta^2, \tag{40}$$

where

$$C_{\text{agg}} = 8mC_\nu (C_A + C_B)^2 (C_A^2 + C_B^2), \tag{41}$$

and

$$M_{\text{agg}} = L\sqrt{C_{\text{loc}}^{(d)} C_{\text{agg}}}\, G^2 + \frac{L}{2} C_{\text{agg}} G^2. \tag{42}$$

Proof. For rank j, define the local A-factor gradient

$$q_{A,i,j}^{(t)} = b_{g,j}^{(t)\top} g_i^{(t)}, \tag{43}$$

and its average over the clients participating in rank j:

$$\bar{q}_{A,j}^{(t)} = \frac{1}{c}\sum_{k\in\mathcal{C}_j} q_{A,k,j}^{(t)}. \tag{44}$$

The local and aggregated rank updates are

$$\widehat{a}_{i,j}^{(t+1)} = a_{g,j}^{(t)} - \eta q_{A,i,j}^{(t)}, \qquad a_{g,j}^{(t+1)} = a_{g,j}^{(t)} - \eta \bar{q}_{A,j}^{(t)}. \tag{45}$$

Hence,

$$a_{g,j}^{(t+1)} - \widehat{a}_{i,j}^{(t+1)} = \eta\left(q_{A,i,j}^{(t)} - \bar{q}_{A,j}^{(t)}\right). \tag{46}$$

Using $\|X - Y\|_2^2 \le 2\|X\|_2^2 + 2\|Y\|_2^2$ and Jensen's inequality,

$$\mathbb{E}_t\left[\|q_{A,i,j}^{(t)} - \bar{q}_{A,j}^{(t)}\|_2^2\right] \le 2\mathbb{E}_t\left[\|q_{A,i,j}^{(t)}\|_2^2\right] + \frac{2}{c}\sum_{k\in\mathcal{C}_j} \mathbb{E}_t\left[\|q_{A,k,j}^{(t)}\|_2^2\right]. \tag{47}$$

From equation 43, equation 9, and equation 10,

$$\mathbb{E}_t\left[\|q_{A,i,j}^{(t)}\|_2^2\right] \le C_\nu C_B^2 \|G_i^{(t)}\|_F^2. \tag{48}$$

Therefore,

$$\mathbb{E}_t\left[\|a_{g,j}^{(t+1)} - \widehat{a}_{i,j}^{(t+1)}\|_2^2\right] \le 4C_\nu C_B^2 \eta^2 \|G_i^{(t)}\|_F^2. \tag{49}$$

Since client i contains exactly m global ranks,

$$\|A_{i,g}^{(t+1)} - \widehat{A}_{i,g}^{(t+1)}\|_F^2 = \sum_{j:M_{ij}=1} \|a_{g,j}^{(t+1)} - \widehat{a}_{i,j}^{(t+1)}\|_2^2. \tag{50}$$

Thus,

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_t\left[\|A_{i,g}^{(t+1)} - \widehat{A}_{i,g}^{(t+1)}\|_F^2\right] \le 4mC_\nu C_B^2 \eta^2 S_t. \tag{51}$$

Similarly, defining

$$q_{B,i,j}^{(t)} = g_i^{(t)} a_{g,j}^{(t)\top}, \qquad \bar{q}_{B,j}^{(t)} = \frac{1}{c}\sum_{k\in\mathcal{C}_j} q_{B,k,j}^{(t)}, \tag{52}$$

we obtain

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_t\left[\|B_{i,g}^{(t+1)} - \widehat{B}_{i,g}^{(t+1)}\|_F^2\right] \le 4mC_\nu C_A^2 \eta^2 S_t. \tag{53}$$

Define

$$\Delta A_i^{(t)} = A_{i,g}^{(t+1)} - \widehat{A}_{i,g}^{(t+1)}, \qquad \Delta B_i^{(t)} = B_{i,g}^{(t+1)} - \widehat{B}_{i,g}^{(t+1)}. \tag{54}$$

Because the local-rank components are unchanged by communication,

$$V_i^{(t)} - U_i^{(t)} = B_{i,g}^{(t+1)} A_{i,g}^{(t+1)} - \widehat{B}_{i,g}^{(t+1)} \widehat{A}_{i,g}^{(t+1)}. \tag{55}$$

Using $XY - X'Y' = (X - X')Y + X'(Y - Y')$,

$$V_i^{(t)} - U_i^{(t)} = \Delta B_i^{(t)} \widehat{A}_{i,g}^{(t+1)} + B_{i,g}^{(t+1)} \Delta A_i^{(t)}. \tag{56}$$

From equation 17, equation 10, and $\eta G \le 1$,

$$\|\widehat{A}_{i,g}^{(t+1)}\|_F \le C_A + \eta C_B G \le C_A + C_B, \qquad \|B_{i,g}^{(t+1)}\|_F \le C_B. \tag{57}$$

Therefore, using $\|X+Y\|_F^2 \le 2\|X\|_F^2 + 2\|Y\|_F^2$,

$$\begin{aligned}
\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_t\left[\|V_i^{(t)} - U_i^{(t)}\|_F^2\right] &\le 2(C_A + C_B)^2 \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_t\left[\|\Delta B_i^{(t)}\|_F^2\right] \\
&\quad + 2C_B^2 \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_t\left[\|\Delta A_i^{(t)}\|_F^2\right] \\
&\le 8mC_\nu\left[C_A^2(C_A + C_B)^2 + C_B^4\right]\eta^2 S_t \\
&\le C_{\text{agg}} \eta^2 S_t,
\end{aligned} \tag{58}$$

where C_{agg} is defined in equation 41. Let

$$H_i^{(t)} = \nabla_W L_i(U_i^{(t)}). \tag{59}$$

By equation 5,

$$L_i(V_i^{(t)}) \le L_i(U_i^{(t)}) + \left\langle H_i^{(t)}, V_i^{(t)} - U_i^{(t)} \right\rangle_F + \frac{L}{2}\|V_i^{(t)} - U_i^{(t)}\|_F^2. \tag{60}$$

Write

$$H_i^{(t)} = G_i^{(t)} + \left(H_i^{(t)} - G_i^{(t)}\right). \tag{61}$$

By L-smoothness,

$$\|H_i^{(t)} - G_i^{(t)}\|_F \le L\|U_i^{(t)} - W_i^{(t)}\|_F. \tag{62}$$

Therefore, by Cauchy–Schwarz and equation 55,

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_t\left[\left\langle G_i^{(t)}, V_i^{(t)}-U_i^{(t)}\right\rangle_F\right] \le \left(\frac{1}{K}\sum_{i=1}^{K}\|G_i^{(t)}\|_F^2\right)^{1/2}\left(\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_t\|V_i^{(t)}-U_i^{(t)}\|_F^2\right)^{1/2} \le \eta\sqrt{C_{\mathrm{agg}}}\,S_t. \tag{63}$$

For the remaining gradient difference term, equation 62, equation 38, and equation 55 give

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_t\left[\left\langle H_i^{(t)}-G_i^{(t)}, V_i^{(t)}-U_i^{(t)}\right\rangle_F\right] \le L\sqrt{\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_t\|U_i^{(t)}-W_i^{(t)}\|_F^2}\sqrt{\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_t\|V_i^{(t)}-U_i^{(t)}\|_F^2} \le L\sqrt{C_{\mathrm{loc}}^{(d)}C_{\mathrm{agg}}}\,\eta^2G^2. \tag{64}$$

Finally, from equation 55,

$$\frac{L}{2K}\sum_{i=1}^{K}\mathbb{E}_t\left[\|V_i^{(t)}-U_i^{(t)}\|_F^2\right] \le \frac{L}{2}C_{\mathrm{agg}}\eta^2G^2. \tag{65}$$

Combining equation 60, equation 63, equation 64, and equation 65,

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_t\left[L_i(V_i^{(t)})\right] \le \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_t\left[L_i(U_i^{(t)})\right] + \eta\sqrt{C_{\mathrm{agg}}}\,S_t + M_{\mathrm{agg}}\eta^2. \tag{66}$$

This proves the lemma. □

Theorem 2 (Convergence of FedCASA)**.** *Under Assumptions 1–3, suppose $\eta G \le 1$ and define*

$$c_{\mathrm{eff}} = c_g - \sqrt{C_{\mathrm{agg}}}. \tag{67}$$

Assume

$$c_{\mathrm{eff}} > 0. \tag{68}$$

Define

$$M = M_{\mathrm{loc}} + M_{\mathrm{agg}}, \tag{69}$$

and

$$D = \frac{1}{K}\sum_{i=1}^{K}L_i(W_i^{(0)}) - \mathcal{L}_*, \qquad \mathcal{L}_* = \inf_t \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}\left[L_i(W_i^{(t)})\right]. \tag{70}$$

Then

$$\frac{1}{KT}\sum_{i=1}^{K}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_W L_i(W_i^{(t)})\|_F^2\right] \le \frac{D}{c_{\mathrm{eff}}\eta T} + \frac{M\eta}{c_{\mathrm{eff}}}. \tag{71}$$

Choosing

$$\eta = \sqrt{\frac{D}{MT}}, \tag{72}$$

whenever $\eta G \le 1$, gives

$$\frac{1}{KT}\sum_{i=1}^{K}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_W L_i(W_i^{(t)})\|_F^2\right] \le \frac{2}{c_{\mathrm{eff}}}\sqrt{\frac{DM}{T}} = O\left(T^{-1/2}\right). \tag{73}$$

Proof. Define the averaged local objective

$$\mathcal{L}_t = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}\left[L_i(W_i^{(t)})\right]. \tag{74}$$

By Lemma 1,

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_t\left[L_i(U_i^{(t)})\right] \le \mathcal{L}_t - c_g\eta S_t + M_{\mathrm{loc}}\eta^2. \tag{75}$$

By Lemma 2,

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_t\left[L_i(V_i^{(t)})\right] \le \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_t\left[L_i(U_i^{(t)})\right] + \eta\sqrt{C_{\text{agg}}}\,S_t + M_{\text{agg}}\eta^2. \tag{76}$$

Since $W_i^{(t+1)} = V_i^{(t)}$, combining equation 75 and equation 76 gives

$$\mathcal{L}_{t+1} \le \mathcal{L}_t - \left(c_g - \sqrt{C_{\text{agg}}}\right)\eta S_t + (M_{\text{loc}} + M_{\text{agg}})\,\eta^2 = \mathcal{L}_t - c_{\text{eff}}\eta S_t + M\eta^2. \tag{77}$$

Rearranging,

$$c_{\text{eff}}\eta S_t \le \mathcal{L}_t - \mathcal{L}_{t+1} + M\eta^2. \tag{78}$$

Summing equation 78 over $t = 0, \dots, T-1$,

$$c_{\text{eff}}\eta \sum_{t=0}^{T-1} S_t \le \mathcal{L}_0 - \mathcal{L}_T + M\eta^2 T. \tag{79}$$

Since $\mathcal{L}_T \ge \mathcal{L}_*$,

$$\mathcal{L}_0 - \mathcal{L}_T \le \mathcal{L}_0 - \mathcal{L}_* = D. \tag{80}$$

Therefore,

$$c_{\text{eff}}\eta \sum_{t=0}^{T-1} S_t \le D + M\eta^2 T. \tag{81}$$

Dividing by $c_{\text{eff}}\eta T$,

$$\frac{1}{T}\sum_{t=0}^{T-1} S_t \le \frac{D}{c_{\text{eff}}\eta T} + \frac{M\eta}{c_{\text{eff}}}. \tag{82}$$

Using the definition of S_t in equation 22,

$$\frac{1}{KT}\sum_{i=1}^{K}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_W L_i(W_i^{(t)})\|_F^2\right] \le \frac{D}{c_{\text{eff}}\eta T} + \frac{M\eta}{c_{\text{eff}}}. \tag{83}$$

Substituting equation 72,

$$\frac{D}{c_{\text{eff}}\eta T} + \frac{M\eta}{c_{\text{eff}}} = \frac{2}{c_{\text{eff}}}\sqrt{\frac{DM}{T}}. \tag{84}$$

Hence,

$$\frac{1}{KT}\sum_{i=1}^{K}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_W L_i(W_i^{(t)})\|_F^2\right] = O\left(T^{-1/2}\right). \tag{85}$$

This completes the proof. □

Corollary 1 (Convergence of the Global Consensus Components)**.** *Define the global consensus descent measure*

$$\Gamma_{i,g}^{(t)} = \left\langle A_g^{(t)\top} D_i A_g^{(t)}, G_i^{(t)\top} G_i^{(t)} \right\rangle_F + \left\langle B_g^{(t)} D_i B_g^{(t)\top}, G_i^{(t)} G_i^{(t)\top} \right\rangle_F \tag{86}$$

and its average

$$\Gamma_{g,t} = \frac{1}{K}\sum_{i=1}^{K}\Gamma_{i,g}^{(t)}. \tag{87}$$

By Assumption 3,

$$c_g S_t \le \Gamma_{g,t} \le (C_A^2 + C_B^2) S_t. \tag{88}$$

Consequently,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\Gamma_{g,t}\right] \le (C_A^2 + C_B^2)\left[\frac{D}{c_{\text{eff}}\eta T} + \frac{M\eta}{c_{\text{eff}}}\right]. \tag{89}$$

With $\eta = \sqrt{D/(MT)}$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\Gamma_{g,t}\right] \le \frac{2(C_A^2 + C_B^2)}{c_{\text{eff}}}\sqrt{\frac{DM}{T}} = O\left(T^{-1/2}\right). \tag{90}$$

Proof. The lower bound in equation 88 follows directly from equation 11, equation 12, and equation 24. For the upper bound, by the Cauchy–Schwarz inequality and equation 10,

$$\begin{aligned}\left\langle A_g^{(t)\top} D_i A_g^{(t)}, G_i^{(t)\top} G_i^{(t)} \right\rangle_F &\le \left\| A_g^{(t)\top} D_i A_g^{(t)} \right\|_F \left\| G_i^{(t)\top} G_i^{(t)} \right\|_F \\ &\le C_A^2 \left\| G_i^{(t)} \right\|_F^2, \\ \left\langle B_g^{(t)} D_i B_g^{(t)\top}, G_i^{(t)} G_i^{(t)\top} \right\rangle_F &\le C_B^2 \left\| G_i^{(t)} \right\|_F^2. \end{aligned} \tag{91}$$

Therefore,

$$\Gamma_{i,g}^{(t)} \le (C_A^2 + C_B^2) \left\| G_i^{(t)} \right\|_F^2. \tag{92}$$

Averaging over all clients gives

$$\Gamma_{g,t} \le (C_A^2 + C_B^2) S_t. \tag{93}$$

Taking expectation and averaging over t, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\Gamma_{g,t}\right] \le (C_A^2 + C_B^2) \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[S_t\right]. \tag{94}$$

Using equation 82,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\Gamma_{g,t}\right] \le (C_A^2 + C_B^2) \left(\frac{D}{c_{\text{eff}} \eta T} + \frac{M\eta}{c_{\text{eff}}} \right). \tag{95}$$

Substituting $\eta = \sqrt{D/(MT)}$ gives

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\Gamma_{g,t}\right] \le \frac{2(C_A^2 + C_B^2)}{c_{\text{eff}}} \sqrt{\frac{DM}{T}} = O\Big(T^{-1/2}\Big), \tag{96}$$

which completes the proof. □