CFLoRA: Federated Fine-tuning of LLMs with Complementary Factors for Error-free Aggregation

Anonymous authors
Paper under double-blind review

Abstract

Federated low-rank adaptation (LoRA) enables collaborative fine-tuning of large language models without centralizing private client data. Its factorized update, however, creates a structural mismatch in federated averaging: averaging the two LoRA factors separately does not equal averaging their products. Existing exact methods resolve this issue mainly by freezing an entire factor or alternating factors across rounds, but none can update factors simultaneously without aggregation errors or expanding communication ranks. To address this fundamental problem, we present `CFLoRA`, a federated LoRA scheme that partitions latent LoRA channels into two complementary sets in every communication round. By ensuring that columns and rows are complementary across factors, we eliminate bilinear terms in matrix multiplications, making federated aggregation exact. Crucially, our framework also supports clients with heterogeneous rank budgets. Convergence analysis validates `CFLoRA` achieves $\mathcal{O}(1/\sqrt{T})$ convergence rate of the *original* LoRA objective in homogeneous-rank cases. Extensive experiments with RoBERTa on the GLUE benchmark and with LLaMA-3.2-3B-Instruct on commonsense reasoning tasks demonstrate that `CFLoRA` achieves superior performance and training efficiency compared to state-of-the-art federated LoRA baselines.

1 Introduction

Federated learning (FL) enables clients to learn from private, distributed data without transferring raw data to a central server (McMahan et al., 2017; Kairouz et al., 2021), which is important for privacy-sensitive applications. However, full-parameter fine-tuning imposes substantial memory, computation, and communication costs on participating clients. Low-rank adaptation (LoRA) reduces these costs by freezing pretrained weights and representing each trainable weight update as the product of two matrices (Hu et al., 2022). Federated LoRA, which integrates FL with LoRA, therefore offers a practical approach to distributed LLM adaptation subject to clients' resource constraints (Zhang et al., 2024).

Despite these benefits, it is well-known that LoRA's factorized parameterization introduces a fundamental mismatch in federated aggregation. For client factors $(\mathbf{B}_i, \mathbf{A}_i)$ and normalized aggregation weights w_i, in general,

$$\left(\sum_i w_i \mathbf{B}_i\right)\left(\sum_i w_i \mathbf{A}_i\right) \neq \sum_i w_i \mathbf{B}_i \mathbf{A}_i. \tag{1}$$

The resulting aggregation error can impair federated optimization, particularly when heterogeneous data drives client factors apart (Sun et al., 2024). A detailed formulation and analysis is provided in Section 3.2.

Existing approaches navigate this mismatch through various trade-offs. Freezing one factor throughout training or alternating entire factors across rounds ensures exact aggregation, but prevents simultaneous adaptation within a single round (Sun et al., 2024; Chen et al., 2025; Koo et al., 2025). Restricting updates in this way limits the model's expressive capacity and slows convergence, as both

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.

matrices are supposed to co-adapt to capture knowledge. Alternatively, stacking client adapters preserves exact aggregation but undesirably expands the rank to the sum of all client ranks (Wang et al., 2024). Other methods adopt approximate decompositions (Bai et al., 2024; Bian et al., 2025; Wang et al., 2026), incorporate residual corrections into the pretrained weights (Singhal et al., 2025), or utilize alternative parameterizations with fixed outer bases (Raje et al., 2025). However, these either only approximate or diverge from the original LoRA learning objective. These limitations raise a still unresolved question: *Can two-factor federated LoRA achieve exact factor-wise aggregation and distribution while simultaneously updating both factors?*

We answer this question with *complementary-channel federated LoRA* (CFLoRA). Our central insight is that exact aggregation can be enforced by a simple modification to LoRA: in every round, if a column of $\mathbf{B}$ is trainable, its corresponding row in $\mathbf{A}$ can be frozen, and vice versa. This mutually exclusive update ensures $\Delta\mathbf{B}_i\Delta\mathbf{A}_j = \mathbf{0}$ for every client pair (i, j), thus leading to exact factor-wise aggregation. CFLoRA can be treated as a generalized version of alternative factor updates, but enables simultaneous factor updating with flexible choices of trainable channels. Compared with standard federated LoRA, our scheme *halves trainable parameters* and *lowers communication costs*. More importantly, benefiting from exact aggregation and the co-adaptation of both factors, we show that CFLoRA achieves an $\mathcal{O}(1/\sqrt{T})$ convergence rate for the *original* LoRA objective under homogeneous client ranks and a surrogate objective under heterogeneous ranks.

Our main contributions are as follows:

- We introduce CFLoRA, a federated LoRA fine-tuning scheme with complementary factors that ensure exact factor-wise aggregation and seamlessly support heterogeneous ranks.
- We establish an $\mathcal{O}(1/\sqrt{T})$ convergence rate for CFLoRA under both homogeneous and heterogeneous rank settings, and theoretically reveals the importance of factor co-adaptation.
- We demonstrate the performance superiority and resource efficiency of CFLoRA over other federated LoRA baselines by fine-tuning RoBERTa and LLaMA-3.2-3B-Instruct models on GLUE and commonsense reasoning tasks, respectively.

2 Related Work

Federated LoRA Fine-tuning. Research on Federated LoRA investigated how to preserve the parameter efficiency of low-rank adaptation while aggregating updates learned from decentralized data. FedIT directly combines LoRA with federated averaging and communicates both factors (Zhang et al., 2024); this simple baseline produces an error between the product of factor averages and the desired average of products. FedEx-LoRA computes the aggregation error and absorbs this residual into the nominally frozen backbone (Singhal et al., 2025). However, this residual has a rank proportional to the aggregated rank and changes the base model across rounds. LoRA-FAIR optimizes a correction to an averaged factor against the product-space target, whereas iFLoRA reconstructs local products and applies SVD in a pipelined aggregation procedure (Bian et al., 2025; Wang et al., 2026). However, these schemes introduce server-side reconstruction complexity (e.g., SVD of high-dimensional models) without entirely avoiding approximation errors.

Beyond error correction, another line of work modifies the factor-updating rule. FFA-LoRA fixes the randomly initialized $\mathbf{A}$ and updates only the zero-initialized $\mathbf{B}$ (Sun et al., 2024). While the aggregation is exact, it introduces a permanently frozen factor that restricts the learned subspace. FedSA-LoRA instead shares only the factor associated with general knowledge and retains the other factor locally for personalization (Guo et al., 2025). RoLoRA alternates the updated factor across communication rounds (Chen et al., 2025). AS-LoRA (Kim et al., 2026) extends the alternating-update idea to the layer level, allowing each layer to select one factor to update at a time. These alternating schemes eliminate aggregation errors, yet still fail to update $\mathbf{A}$ and $\mathbf{B}$ concurrently. Consequently, none of the aforementioned works can *simultaneously update both factors within the same round with exact factor-wise aggregation and distribution.*

Federated LoRA with Heterogeneous Ranks. Rank-heterogeneous federated LoRA assigns client-specific adapter capacity to accommodate differences in memory, computation, communication, and task complexity (Fang et al., 2025). Several approaches address this through padding or sampling mechanisms (Cho et al., 2024; Byun & Lee, 2025). Alternatively, FSLoRA samples

submatrices from a server-side LoRA model based on client-specific sketching ratios and provides a convergence analysis (Fang et al., 2026). Another line of work relies on product-space decomposition. FlexLoRA constructs a weighted product-space aggregate and employs SVD to distribute rank-specific adapters to clients (Bai et al., 2024). FedPA-LoRA applies SVD solely on a small core matrix obtained via reduced QR factorization (Jeon et al., 2026). Nevertheless, repeated model decompositions incur significant overhead, and crucially, these schemes fail to preserve exact factor-wise aggregation. To preserve exact aggregation, FLoRA stacks heterogeneous client factors so that their product equals the sum of local products (Wang et al., 2024). GLoRA aggregates factors in consensus coordinates (Chen et al., 2026). However, to eliminate aggregation error, FLoRA requires a broadcast rank of $\sum_i r_i$, and GLoRA also requires the server rank to cover the union span across clients. Other works achieve exact aggregation, yet with alternative updates or different parameterizations (e.g., updating only middle matrices) (Raje et al., 2025; Koo et al., 2025; Meng et al., 2026). In a nutshell, none of the existing FedLoRA methods, which update $\mathbf{A}$ and $\mathbf{B}$ concurrently, can support heterogeneous client ranks with exact factor-wise aggregation and distributions.

3 Preliminaries and Problem Formulation

3.1 Federated LoRA Fine-Tuning

We consider a federated LoRA fine-tuning system comprising N clients, which can be formulated as follows:

$$F(\mathbf{B}, \mathbf{A}) = \sum_i w_i f_i(\mathbf{B}, \mathbf{A}), \qquad f_i(\mathbf{B}, \mathbf{A}) = \mathbb{E}_{\xi \sim \mathcal{D}_i}[\ell(\mathbf{W}_0 + \mathbf{B}\mathbf{A}; \xi)], \tag{2}$$

where ℓ, f_i, and F denote the sample loss, the local loss for client i, and the global loss, respectively. ξ represents a data sample drawn from the local dataset $\mathcal{D}_i$ on client i, and $w_i > 0$ is the weight assigned to client i with $\sum_i w_i = 1$. Furthermore, $\mathbf{W}_0 \in \mathbb{R}^{d_o \times d_i}$ denotes the frozen base model, while $\mathbf{B} \in \mathbb{R}^{d_o \times R}$ and $\mathbf{A} \in \mathbb{R}^{R \times d_i}$ are the trainable rank-R LoRA modules.

3.2 Failures in Factor-Wise Aggregation

A fundamental challenge in LoRA-based federated fine-tuning is that averaging and distributing $\mathbf{A}$ and $\mathbf{B}$ *separately* following FedAvg (McMahan et al., 2017; Sun et al., 2024; Chen et al., 2025) is not equal to averaging and distributing their products. Formally, suppose $\mathbf{B}$ and $\mathbf{A}$ denote the global factors at the beginning of the round. Each client i produces local endpoints $\mathbf{B}_i = \mathbf{B} + \Delta\mathbf{B}_i$, $\mathbf{A}_i = \mathbf{A} + \Delta\mathbf{A}_i$. Factor-wise averaging across N clients yields the update:

$$\begin{aligned} &\left(\mathbf{B} + \sum_i w_i \Delta\mathbf{B}_i\right)\left(\mathbf{A} + \sum_i w_i \Delta\mathbf{A}_i\right) \\ =&\mathbf{B}\mathbf{A} + \sum_i w_i \Delta\mathbf{B}_i\mathbf{A} + \mathbf{B}\sum_i w_i \Delta\mathbf{A}_i + \underbrace{\left(\sum_i w_i \Delta\mathbf{B}_i\right)\left(\sum_i w_i \Delta\mathbf{A}_i\right)}_{\text{cross-client term}}, \end{aligned} \tag{3}$$

whereas the ideal aggregation target is:

$$\sum_i w_i \mathbf{B}_i\mathbf{A}_i = \mathbf{B}\mathbf{A} + \sum_i w_i \Delta\mathbf{B}_i\mathbf{A} + \mathbf{B}\sum_i w_i \Delta\mathbf{A}_i + \underbrace{\sum_i w_i \Delta\mathbf{B}_i\Delta\mathbf{A}_i}_{\text{client-wise interaction term}}. \tag{4}$$

The discrepancy between Equations 3 and 4 lies in the final terms, where the difference defines the *aggregation error* of factor-wise averaging.

This aggregation error compounds over successive rounds and degrades convergence. Its magnitude is governed by two key factors. First, it may grow with the *number of local update steps*: more local steps can cause $\Delta\mathbf{B}_i$ and $\Delta\mathbf{A}_i$ to diverge further from their global mean. Second, it grows with the degree of *data heterogeneity* across clients: when local datasets follow substantially different distributions. Taken together, these two factors motivate the design of new federated LoRA schemes.

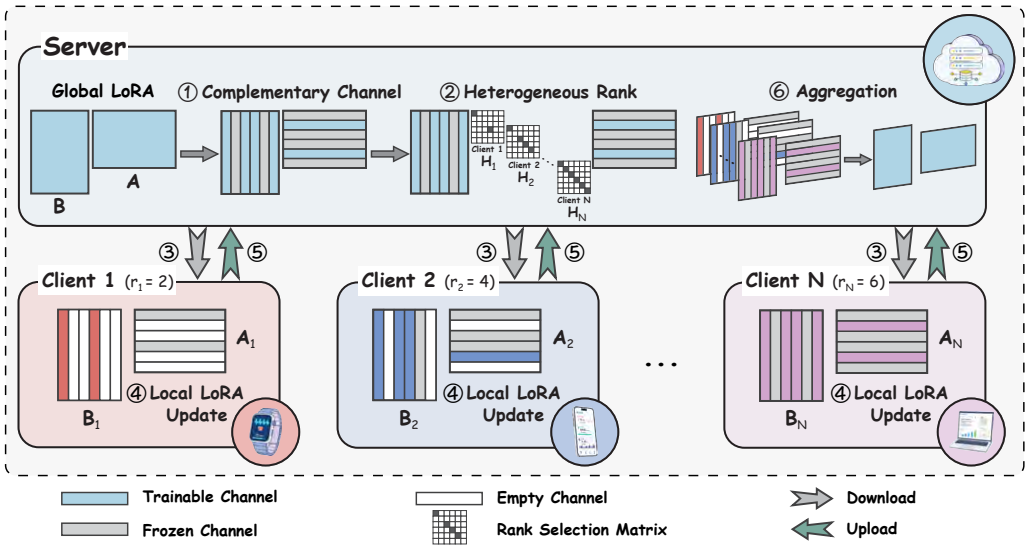


Figure 1: The illustration of the proposed CFLoRA framework. CFLoRA separates two decisions: which factor may change on each latent channel and which channels a client can afford to train. A shared complementary selection mask ensures exact factor aggregation, while client-specific supports accommodate heterogeneous ranks.

4 The CFLoRA Framework

4.1 Complementary-Channel Selection

The aggregation inconsistency discussed in Section 3.2 arises from the bilinear interaction term $\Delta\mathbf{B}_i\Delta\mathbf{A}_i$. To eliminate this, we propose a *complementary channel partitioning* strategy. In every round, if a column of $\mathbf{B}$ is trainable, its corresponding row in $\mathbf{A}$ can be frozen, and vice versa[1]. Formally, at round t, the server constructs a diagonal channel-selection mask, defined as

$$\mathbf{Z}^t = \mathrm{diag}(z_1^t, \ldots, z_R^t), \qquad z_j^t \in \{0, 1\}, \tag{5}$$

where z_j^t is sampled independently with $\Pr(z_j^t = 1) = p$. For each channel j, the update rule is defined as follows.

- If $z_j = 1$: update column $\mathbf{B}(:, j)$ and freeze row $\mathbf{A}(j, :)$;
- If $z_j = 0$: freeze column $\mathbf{B}(:, j)$ and update row $\mathbf{A}(j, :)$.

The server broadcasts the same mask to all participating clients in each round, thereby ensuring a common update structure within each round while allowing the channel partition to vary across rounds.

During local training, client i initializes as $\mathbf{B}_i^{t,0} = \mathbf{B}^t$ and $\mathbf{A}_i^{t,0} = \mathbf{A}^t$ and updates at local step k according to $\mathbf{B}_i^{t,k+1} = \mathbf{B}_i^{t,k} + \Delta\mathbf{B}_i^{t,k}\mathbf{Z}^t$ and $\mathbf{A}_i^{t,k+1} = \mathbf{A}_i^{t,k} + (\mathbf{I}_R - \mathbf{Z}^t)\Delta\mathbf{A}_i^{t,k}$, where $\Delta\mathbf{B}_i^{t,k}$ and $\Delta\mathbf{A}_i^{t,k}$ denote client i's unmasked local updates at step k. Note that this full-matrix formulation is introduced primarily for clarity; in practice, clients can update only the active parameters determined by $\mathbf{Z}^t$. After K local steps, the endpoint of client i is $\mathbf{B}_i^{t+1} = \mathbf{B}^t + \Delta\mathbf{B}_i^t\mathbf{Z}^t$ and $\mathbf{A}_i^{t+1} = \mathbf{A}^t + (\mathbf{I}_R - \mathbf{Z}^t)\Delta\mathbf{A}_i^t$, where the accumulated unmasked local updates over the K steps are given by

$$\Delta\mathbf{B}_i^t = \sum_{k=0}^{K-1} \Delta\mathbf{B}_i^{t,k}, \qquad \Delta\mathbf{A}_i^t = \sum_{k=0}^{K-1} \Delta\mathbf{A}_i^{t,k}. \tag{6}$$

[1]Here, we describe a single LoRA layer for notational simplicity. For multiple layers, CFLoRA applies the same rule to each layer, with its complementary mask shared by participating clients.

In each round, suppose that the server randomly samples a subset $\mathcal{S}_t$ of m clients from the total N clients[2]. Then, the server performs factor-wise aggregation:

$$\mathbf{B}^{t+1} = \mathbf{B}^t + \sum_i \frac{w_i}{q} \Delta \mathbf{B}_i^t \mathbf{Z}^t, \qquad \mathbf{A}^{t+1} = \mathbf{A}^t + \sum_i \frac{w_i}{q} (\mathbf{I}_R - \mathbf{Z}^t) \Delta \mathbf{A}_i^t. \tag{7}$$

where $q = m/N$ denotes the client participation ratio. The scaling $1/q$ compensates for the partial participation, ensuring an unbiased estimator of the full-client update. Since the complementary masks satisfy $\mathbf{Z}^t(\mathbf{I}_R - \mathbf{Z}^t) = \mathbf{0}$, we can establish Proposition 1 for the aggregation exactness.

Proposition 1 (Error-Free Aggregation of CFLoRA)**.** *Factor-wise aggregation in equation 7 satisfies*

$$\begin{aligned} &\left(\mathbf{B}^t + \sum_i \frac{w_i}{q} \Delta \mathbf{B}_i^t \mathbf{Z}^t \right) \left(\mathbf{A}^t + \sum_i \frac{w_i}{q} (\mathbf{I}_R - \mathbf{Z}^t) \Delta \mathbf{A}_i^t \right) \\ &\qquad = \mathbf{B}^t \mathbf{A}^t + \sum_i \frac{w_i}{q} \left(\mathbf{B}_i^{t+1} \mathbf{A}_i^{t+1} - \mathbf{B}^t \mathbf{A}^t \right). \end{aligned} \tag{8}$$

Proposition 1 shows that CFLoRA achieves exact federated LoRA aggregation. The detailed proof is provided in Appendix A.1.

4.2 Generalization to Heterogeneous Client Ranks

As shown in Fig. 1, the proposed CFLoRA framework naturally accommodates clients with heterogeneous computational and memory budgets. Specifically, the server maintains LoRA factors of rank R, while client i activates only $r_i \le R$ channels subject to its resource budget. Specifically, in round t, the server constructs a selection matrix $\mathbf{E}_i^t \in \{0,1\}^{R \times r_i}$ for each client i, defined as

$$\mathbf{E}_i^t = [\mathbf{e}_{s_{i,1}^t}, \ldots, \mathbf{e}_{s_{i,r_i}^t}] \in \{0,1\}^{R \times r_i}, \qquad (\mathbf{E}_i^t)^\top \mathbf{E}_i^t = \mathbf{I}_{r_i}, \tag{9}$$

where $\mathbf{e}_j$ is the j-th standard basis vector of $\mathbb{R}^R$. The corresponding global activation mask for client i is then defined as

$$\mathbf{H}_i^t = \mathbf{E}_i^t (\mathbf{E}_i^t)^\top \in \{0,1\}^{R \times R}, \qquad \operatorname{rank}(\mathbf{H}_i^t) = r_i. \tag{10}$$

Client i then initializes its LoRA factors by extracting the selected channels from the global factors $\widetilde{\mathbf{B}}_i^{t,0} = \mathbf{B}^t \mathbf{E}_i^t \in \mathbb{R}^{d_{\text{out}} \times r_i}$ and $\widetilde{\mathbf{A}}_i^{t,0} = (\mathbf{E}_i^t)^\top \mathbf{A}^t \in \mathbb{R}^{r_i \times d_{\text{in}}}$. Their product is therefore $\widetilde{\mathbf{B}}_i^{t,0} \widetilde{\mathbf{A}}_i^{t,0} = \mathbf{B}^t \mathbf{H}_i^t \mathbf{A}^t$, which is the rank-$r_i$ submodel to initialize local training. At local step k, client i updates according to

$$\widetilde{\mathbf{B}}_i^{t,k+1} = \widetilde{\mathbf{B}}_i^{t,k} + \Delta \widetilde{\mathbf{B}}_i^{t,k} (\mathbf{E}_i^t)^\top \mathbf{Z}^t \mathbf{E}_i^t, \qquad \widetilde{\mathbf{A}}_i^{t,k+1} = \widetilde{\mathbf{A}}_i^{t,k} + (\mathbf{I}_{r_i} - (\mathbf{E}_i^t)^\top \mathbf{Z}^t \mathbf{E}_i^t) \Delta \widetilde{\mathbf{A}}_i^{t,k}, \tag{11}$$

where $\Delta \widetilde{\mathbf{B}}_i^{t,k}$ and $\Delta \widetilde{\mathbf{A}}_i^{t,k}$ denote client i's unmasked local updates at step k. After K local steps, the client uploads the updates, which are given by

$$\Delta \widetilde{\mathbf{B}}_i^t = \widetilde{\mathbf{B}}_i^{t,K} - \widetilde{\mathbf{B}}_i^{t,0} \qquad \Delta \widetilde{\mathbf{A}}_i^t = \widetilde{\mathbf{A}}_i^{t,K} - \widetilde{\mathbf{A}}_i^{t,0}. \tag{12}$$

Then, the server embeds these updates into the R-rank space and aggregates them by:

$$\mathbf{B}^{t+1} = \mathbf{B}^t + \sum_i \frac{w_i}{q} \Delta \widetilde{\mathbf{B}}_i^t (\mathbf{E}_i^t)^\top, \qquad \mathbf{A}^{t+1} = \mathbf{A}^t + \sum_i \frac{w_i}{q} \mathbf{E}_i^t \Delta \widetilde{\mathbf{A}}_i^t. \tag{13}$$

We establish the following proposition to show the exactness of aggregation under heterogeneous client ranks.

Proposition 2 (Error-Free Aggregation of CFLoRA with Heterogeneous Client Ranks)**.** *Suppose all participating clients use the common complementary mask $\mathbf{Z}^t$, while their ranks r_i and selection matrices $\mathbf{E}_i^t$ may differ. The factor-wise aggregation rule in equation 13 satisfies*

$$\left(\mathbf{B}^t + \sum_i \frac{w_i}{q} \Delta \widetilde{\mathbf{B}}_i^t (\mathbf{E}_i^t)^\top \right) \left(\mathbf{A}^t + \sum_i \frac{w_i}{q} \mathbf{E}_i^t \Delta \widetilde{\mathbf{A}}_i^t \right) = \mathbf{B}^t \mathbf{A}^t + \sum_i \frac{w_i}{q} \left(\widetilde{\mathbf{B}}_i^{t,K} \widetilde{\mathbf{A}}_i^{t,K} - \mathbf{B}^t \mathbf{H}_i^t \mathbf{A}^t \right). \tag{14}$$

Proposition 2 shows that CFLoRA still achieves exact federated LoRA aggregation under heterogeneous client ranks. The proof can be found in Appendix A.2.

[2]We consider uniform sampling rate q for notational simplicity. Our analysis can be easily extended to consider client-specific sampling probability.

4.3 COMPLEMENTARY-CHANNEL SCALING

To prevent biased updates arising from channel selection, the protocol introduces a unified scaling compensation factor. The numbers of trainable channels assigned to $\mathbf{B}$ and $\mathbf{A}$ are

$$k_{B,i}^t = \mathrm{tr}(\widetilde{\mathbf{Z}}_i^t), \qquad k_{A,i}^t = r_i - k_{B,i}^t, \tag{15}$$

where $\widetilde{\mathbf{Z}}_i^t = (\mathbf{E}_i^t)^\top \mathbf{Z}^t \mathbf{E}_i^t$, and we have $\mathbb{E}[k_{B,i}^t] = r_i p$ and $\mathbb{E}[k_{A,i}^t] = r_i(1-p)$. Then, to compensate for the unequal probabilities of selecting the two factors, we scale the gradients by $1/p$ and $1/(1-p)$, respectively. Specifically, for $0 < p < 1$, the factors are updated as

$$\begin{aligned}
\widetilde{\mathbf{B}}_i^{t,k+1} &= \widetilde{\mathbf{B}}_i^{t,k} - \frac{\eta}{p} \nabla_{\widetilde{\mathbf{B}}} \ell\Big(\mathbf{W}_0 + \widetilde{\mathbf{B}}_i^{t,k} \widetilde{\mathbf{A}}_i^{t,k}; \xi\Big)\, \widetilde{\mathbf{Z}}_i^t, \\
\widetilde{\mathbf{A}}_i^{t,k+1} &= \widetilde{\mathbf{A}}_i^{t,k} - \frac{\eta}{1-p} (\mathbf{I}_{r_i} - \widetilde{\mathbf{Z}}_i^t) \nabla_{\widetilde{\mathbf{A}}} \ell\Big(\mathbf{W}_0 + \widetilde{\mathbf{B}}_i^{t,k} \widetilde{\mathbf{A}}_i^{t,k}; \xi\Big),
\end{aligned} \tag{16}$$

where η is the learning rate. Note that setting $r_i = R$, which implies $\mathbf{E}_i^t = \mathbf{I}_R$ and $\widetilde{\mathbf{Z}}_i^t = \mathbf{Z}^t$, recovers the homogeneous update rule. The activation matrix $\mathbf{E}_i^t$ and complementary mask $\widetilde{\mathbf{Z}}_i^t$ remain fixed throughout the round. Since the updates defined in Eq. 6 and Eq. 12 can naturally absorb this scaling, Propositions 1 and 2 continue to guarantee exact aggregation.

Considering practical partial participation (selecting m out of N clients per round), the procedure of CFLoRA is summarized in Algorithm B.1 for homogeneous cases and Algorithm B.2 for heterogeneous cases.

4.4 CONVERGENCE ANALYSIS

Next, we theoretically analyze the convergence of CFLoRA under partial client participation. Our analysis naturally encompasses full participation as a special case.

Notations. For each client i, let $\mathcal{H}_i$ denote the uniform distribution over diagonal binary masks with r_i active channels, satisfying $\mathbb{E}_{\mathbf{H}\sim\mathcal{H}_i}[\mathbf{H}] = \frac{r_i}{R}\mathbf{I}_R$. The activation mask $\mathbf{H}_i^t = \mathbf{E}_i^t(\mathbf{E}_i^t)^\top \sim \mathcal{H}_i$ is sampled independently across clients and independently of $\mathbf{Z}^t$. Let $\mathbf{X} = (\mathbf{B}, \mathbf{A})$, for a given activation mask $\mathbf{H}$, the mask-conditioned and expected local objectives for client i are respectively defined as

$$f_i^{\mathbf{H}}(\mathbf{X}) = \mathbb{E}_{\xi\sim\mathcal{D}_i}\big[\ell(\mathbf{W}_0 + \mathbf{BHA}; \xi)\big], \qquad \phi_i(\mathbf{X}) = \mathbb{E}_{\mathbf{H}\sim\mathcal{H}_i}\big[f_i^{\mathbf{H}}(\mathbf{X})\big], \tag{17}$$

and therefore the global objective is formulated as

$$\Phi(\mathbf{X}) = \sum_{i=1}^N w_i \phi_i(\mathbf{X}). \tag{18}$$

Notably, when $r_i = R$ for all clients, $\mathbf{H}_i = \mathbf{I}_R$ and $\Phi(\mathbf{X})$ reduces to the standard objective $F(\mathbf{X})$.

We assume that each client's local objective is L-smooth and the global objective is bounded (Assumption A.1), local stochastic gradients are unbiased with bounded variance (Assumption A.2), which are detailed in Appendix A.3. With notations above and Lemma A.1 on unbiased estimators of the scaled aggregated gradients (detailed in Appendix A.4), we establish the following convergence analysis.

Theorem 3 (Convergence of CFLoRA with Heterogeneous Client Ranks)**.** *Suppose that Assumptions A.1–A.2 hold and $\Delta_0 > 0$. With the learning rate*

$$\eta_T = \frac{1}{K} \min\left\{ \frac{1}{4L\kappa}, \sqrt{\frac{\Delta_0}{L\mathcal{V}_K T}}, \left(\frac{2\Delta_0}{\mathcal{D}_K T}\right)^{1/3} \right\}, \tag{19}$$

CFLoRA satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla\Phi(\mathbf{X}^t)\|^2\right] \le 4\sqrt{\frac{L\Delta_0\mathcal{V}_K}{T}} + 2\left(\frac{4\Delta_0^2\mathcal{D}_K}{T^2}\right)^{1/3} + \frac{8L\kappa\Delta_0}{T}, \tag{20}$$

where $\mathcal{V}_K = \kappa\left(\frac{N\tau_w^2}{m} + \frac{N-m}{m(N-1)}\zeta_w^2\right) + \frac{\kappa^2\sigma_w^2}{qK}$, $\mathcal{D}_K = \frac{4+\chi_m}{8}L^2\kappa^3 M\left(1-\frac{1}{K}\right)^2$, $\chi_m = 1 + \frac{N-m}{m(N-1)}\left(N\sum_i w_i^2 - 1\right)$, $\sigma_w^2 = \sum_{i=1}^N w_i^2\sigma_i^2$, $\tau_w^2 = \sum_{i=1}^N w_i^2\tau_i^2$, $\kappa = p_{\min}^{-1}$, $p_{\min} = \min\{p, 1-p\}$, $\Delta_0 = \mathbb{E}[\Phi(\mathbf{X}^0)] - \Phi^\star$, $\zeta_w^2 = \sup_{\mathbf{X}\in\mathcal{C}}\left\{N\sum_i w_i^2\|\nabla\phi_i(\mathbf{X})\|^2 - \|\nabla\Phi(\mathbf{X})\|^2\right\}$, *and* $\mathbf{X}^0$ *denotes the initial iterate. For* $N = 1$, *set* $\frac{N-m}{m(N-1)} = 0$.

The proof is provided in Appendix A.5. Due to the exact aggregation, the convergence bound in Theorem 3 achieves an $\mathcal{O}(1/\sqrt{T})$ convergence rate without non-vanishing bilinear aggregation-error terms. Heterogeneous client ranks affect both the objective Φ and the activation-mask variance τ_w^2, which vanishes when $r_i = R$ for every client. Partial participation introduces the client-sampling term $\frac{N-m}{m(N-1)}\zeta_w^2$. More discussions are provided in Appendix A.7.

Corollary 4 (Convergence of CFLoRA with Homogeneous Client Ranks)**.** *Under the conditions of Theorem 3, suppose* $r_i = R$ *for every client. Then* $\Phi = F$ *and* $\tau_w^2 = 0$. *Choosing* $\eta_T = (4L\kappa K\sqrt{T})^{-1}$, *the server iterates satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(\mathbf{X}^t)\|^2\right] \le \frac{8L\kappa\Delta_0 + \frac{N-m}{m(N-1)}\zeta_w^2/2 + \kappa\sigma_w^2/(2qK)}{\sqrt{T}} + \frac{(4+\chi_m)\kappa M}{128T}\left(1-\frac{1}{K}\right)^2. \tag{21}$$

The proof is provided in Appendix A.6.

Significance of Corollary 4. Under homogeneous ranks, there is no objective discrepancy induced by lower client ranks. As a result, the $\mathcal{O}(1/\sqrt{T})$ convergence rate applies directly to the original LoRA objective F in equation 2. This results from the unique advantage of CFLoRA: it co-adapts both factors while entirely avoiding the aggregation errors in standard federated LoRA.

Determination of channel sampling probability p**.** The parameter $\kappa = p_{\min}^{-1}$ reveals that low selection probabilities for either factor amplify the upper bound, underscoring the need to update both factors simultaneously. Since the upper bounds in Theorem 3 and Corollary 4 decrease as $p_{\min} = \min\{p, 1-p\}$ increases, they are minimized at $p = 0.5$. However, this stems from a worst-case theoretical bound, which may not be optimal empirically.

We conduct extensive empirical studies on p (Tables 3-4 and Tables D.5-D.6), yielding two intriguing insights. On the one hand, increasing p from 0.3 to 0.9 generally improves performance. This aligns with prior findings (Hayou et al., 2024) that the two low-rank matrices contribute asymmetrically to the optimization process, and prioritizing the $\mathbf{B}$ matrix update is beneficial. On the other hand, setting $p = 1.0$ degrades performance, which is consistent with our theoretical analysis that the co-adaptation of both factors remains crucial [3]. Consequently, we adopt $p = 0.9$ as the default setting across all experiments.

4.5 Resource Efficiency of CFLoRA

Compared to standard federated LoRA at the same active rank, complementary updates provide efficiency gains across several key aspects: i) **Trainable Parameters**: By freezing the rows and/or columns, CFLoRA can reduce the number of trainable parameters by 50% per round [4]. ii) **Communication Cost**: By transmitting only the trainable parameters, CFLoRA reduces the adapter uploading cost by 50%, though downlink of full factors is still needed for clients to perform forward passes. A comprehensive analysis is provided in Appendix C.

5 Experiments

Implementation Details and Baselines. Our evaluation focuses on RoBERTa (125M) (Liu et al., 2019) on the GLUE benchmark (Wang et al., 2018) and LLaMA-3.2-3B-Instruct (Grattafiori et al.,

[3] Note that for $p = 1.0$, the scaling step described in Section 4.3 is omitted.

[4] The 50% reduction refers to square attention projection matrices, so each trainable row of $\mathbf{A}$ and column of $\mathbf{B}$ contains the same number of parameters. For rectangular matrices, the reduction depends on their dimensions and the selected channels.

Table 1: Test accuracy for fine-tuning the RoBERTa model on the GLUE benchmark under *homogeneous* client ranks.

Method	QNLI	MRPC	CoLA	MNLI-M	MNLI-MM	QQP	Avg.
FedIT	86.61 ± 0.26	72.27 ± 2.32	70.49 ± 0.90	79.10 ± 2.21	**79.89** ± 2.28	71.75 ± 1.35	76.69
FFA-LoRA	88.49 ± 0.23	70.58 ± 1.56	73.81 ± 1.74	72.42 ± 3.27	73.39 ± 3.08	73.26 ± 1.38	75.33
RoLoRA	87.60 ± 2.21	73.16 ± 2.11	74.40 ± 0.89	72.01 ± 1.55	72.50 ± 3.56	73.47 ± 0.41	75.52
AS-LoRA	88.26 ± 0.67	72.30 ± 1.06	73.57 ± 2.15	71.41 ± 1.88	72.60 ± 1.51	75.23 ± 2.57	75.56
FLoRA	89.27 ± 0.20	71.46 ± 1.24	76.81 ± 2.74	70.56 ± 0.82	71.52 ± 0.21	**76.43** ± 2.74	76.01
CFLoRA	**90.61** ± 0.25	**73.53** ± 1.59	**78.95** ± 1.48	**79.16** ± 1.22	79.85 ± 1.01	76.30 ± 0.36	**79.73**

Table 2: Testing accuracy for fine-tuning the RoBERTa model on the GLUE benchmark under *heterogeneous* client ranks.

Method	QNLI	MRPC	CoLA	MNLI-M	MNLI-MM	QQP	Avg.
FedIT	70.34 ± 2.01	69.60 ± 2.43	73.56 ± 3.67	58.34 ± 1.03	58.22 ± 0.93	73.12 ± 1.41	67.20
FFA-LoRA	86.86 ± 2.15	70.67 ± 2.31	73.09 ± 2.97	63.74 ± 5.02	64.04 ± 4.88	75.91 ± 1.01	72.39
RoLoRA	84.87 ± 3.10	68.46 ± 3.12	73.14 ± 3.40	51.19 ± 3.95	51.60 ± 2.84	74.73 ± 1.29	67.33
AS-LoRA	80.71 ± 2.52	71.12 ± 1.33	73.09 ± 2.97	63.60 ± 4.86	63.95 ± 5.94	70.73 ± 1.94	70.53
FLoRA	87.35 ± 1.84	69.80 ± 1.71	75.75 ± 1.34	72.22 ± 1.26	70.97 ± 1.14	76.61 ± 0.53	75.45
FlexLoRA	86.66 ± 1.55	71.92 ± 1.19	77.60 ± 0.58	71.35 ± 1.49	72.25 ± 1.48	76.40 ± 1.32	76.03
FSLoRA	86.68 ± 2.27	**75.95** ± 1.34	74.95 ± 0.52	72.55 ± 1.57	73.49 ± 1.62	**79.58** ± 1.29	77.20
CFLoRA	**91.29** ± 0.45	74.86 ± 0.69	**77.80** ± 1.26	**76.82** ± 0.80	**77.32** ± 1.14	79.34 ± 0.87	**79.57**

2024) on commonsense reasoning tasks (Hu et al., 2023). Following prior work (Zhang et al., 2024; Fang et al., 2026), we adopt Dirichlet-based partitioning for non-IID dataset splits. In our default evaluation, we consider a pool of 25 clients and randomly select 10 per round for training. We also present an extended analysis by varying the number of clients in Appendix D.4. All results are averaged over 4 independent seeds. Further experimental details are provided in Appendix D.1.

We compare CFLoRA against several state-of-the-art federated LoRA baselines, including FedIT (Zhang et al., 2024), the standard factor-wise averaging baseline; FFA-LoRA (Sun et al., 2024), RoLoRA (Chen et al., 2025), FLoRA (Wang et al., 2024), and AS-LoRA (Kim et al., 2026), which are designed to provide error-free aggregation; and FlexLoRA (Bai et al., 2024) and FSLoRA (Fang et al., 2026), which are explicitly designed to accommodate heterogeneous client ranks.

5.1 Main Results

Comparison under Homogeneous Client Ranks. Table 1 summarizes the results for fine-tuning RoBERTa on the GLUE benchmark under homogeneous client ranks. CFLoRA achieves the highest average accuracy of 79.73%, outperforming the strongest baseline, FedIT, by 3.04 percentage points. It obtains the highest scores on QNLI, MRPC, CoLA, and MNLI-M, with improvements of 1.34 and 2.14 points over the best baseline scores on QNLI and CoLA, respectively. It also remains competitive on MNLI-MM and QQP. These results demonstrate that complementary updates enable effective adaptation despite training fewer adapter parameters per round.

Comparison under Heterogeneous Client Ranks. Table 2 reports the corresponding results under heterogeneous client ranks. CFLoRA achieves the highest average accuracy of 79.57%, exceeding FSLoRA, the strongest baseline, by 2.37 percentage points. It achieves the highest scores on QNLI, CoLA, MNLI-M, and MNLI-MM, with gains of 3.94, 4.27, and 3.83 points over the best baseline scores on QNLI, MNLI-M, and MNLI-MM, respectively. These results support the effectiveness of complementary updates when clients train compact adapters with different rank budgets. Additional results for LLaMA-3.2-3B-Instruct under homogeneous and heterogeneous ranks are reported in Tables D.3 and D.4, respectively, and discussed in Appendix D.2.

Parameter and Communication Efficiency. Importantly, these accuracy improvements of CFLoRA are accompanied by reduced resource costs. Figure C.1 (in Appendix C) illustrates the trade-off between accuracy and communication cost over 200 rounds. On MNLI, CFLoRA reaches

Table 3: Ablation study on selection probability p for RoBERTa on GLUE tasks under homogeneous client ranks.

Schedule	QNLI	MRPC	CoLA	MNLI-M	MNLI-MM	QQP	Avg.
$p = 0.3$	87.80 ± 0.02	70.83 ± 1.73	73.33 ± 2.37	71.15 ± 1.07	72.73 ± 1.03	72.45 ± 1.39	74.72
$p = 0.5$	90.64 ± 0.25	72.43 ± 2.50	78.72 ± 1.73	77.82 ± 0.49	78.00 ± 0.12	74.57 ± 0.43	78.70
$p = 0.7$	**90.69** ± 0.24	72.55 ± 2.06	**79.81** ± 1.52	77.47 ± 0.61	78.02 ± 0.65	74.52 ± 1.43	78.84
$p = 0.9$	90.61 ± 0.25	**73.53** ± 1.59	78.95 ± 1.48	**79.16** ± 1.22	**79.85** ± 1.01	**76.30** ± 0.36	**79.73**
$p = 1.0$	88.61 ± 0.27	70.47 ± 1.54	73.89 ± 1.77	72.36 ± 3.23	73.54 ± 3.13	73.17 ± 1.39	75.34

Table 4: Ablation study on selection probability p for RoBERTa on GLUE tasks under heterogeneous client ranks.

Schedule	QNLI	MRPC	CoLA	MNLI-M	MNLI-MM	QQP	Avg.
$p = 0.3$	88.31 ± 0.99	68.79 ± 0.35	71.46 ± 1.27	68.78 ± 0.49	69.66 ± 1.29	73.21 ± 1.15	73.37
$p = 0.5$	91.16 ± 0.10	73.98 ± 2.26	76.90 ± 2.38	76.37 ± 1.02	76.85 ± 1.30	77.16 ± 1.28	78.74
$p = 0.7$	**91.41** ± 0.87	74.13 ± 1.04	77.32 ± 0.85	76.74 ± 1.37	**77.51** ± 0.73	**79.61** ± 1.07	79.45
$p = 0.9$	91.29 ± 0.45	**74.86** ± 0.69	**77.80** ± 1.26	**76.82** ± 0.80	77.32 ± 1.14	79.34 ± 0.87	**79.57**
$p = 1.0$	89.90 ± 1.39	73.00 ± 2.02	74.51 ± 1.42	75.25 ± 1.57	75.81 ± 0.91	73.63 ± 1.95	77.02

approximately 70% matched accuracy with only about 8 GB of cumulative communication, whereas FSLoRA requires roughly 14 GB to attain comparable performance. On QNLI, CFLoRA exceeds 90% accuracy at approximately 8 GB, also making it significantly more communication-efficient than the baselines. In comparison, FLoRA is not communication-efficient because its effective rank is the sum of all client ranks. Overall, these results demonstrate that CFLoRA achieves strong accuracy within smaller communication budget. A detailed cost analysis is provided in Appendix C.

5.2 Ablation Study: Impact of channel sampling probability p

In Table 3-Table 4 and Table D.5 and Table D.6 (in Appendix D.3), we conduct experiments with different p on various models in both homogeneous and heterogeneous settings, which reveal a similar phenomenon. In addition to the theoretical finding discussed in Section 4.4 that low selection probabilities for either factor amplify the convergence bound, we empirically observe that CFLoRA remains good for different values of $p > 0.5$ (excluding the boundary case of $p = 1.0$). While $p = 0.9$ is generally the best, the performance robustness across a wide range of p demonstrates CFLoRA's general insensitivity to this hyperparameter, making it easy to configure in practice. More discussions are provided in Section D.3.

6 Conclusion

We presented CFLoRA, a federated LoRA framework that can update both factors in each round while preserving exact factor-wise aggregation. Our key idea is to utilize a shared complementary mask to assign each latent channel to either the $\mathbf{A}$ or $\mathbf{B}$ factor, eliminating the bilinear terms that cause the aggregation mismatch. Furthermore, CFLoRA seamlessly extends to settings with heterogeneous client ranks by applying additional activation masks. Due to the exact factor-wise aggregation, our analysis establishes an $\mathcal{O}(1/\sqrt{T})$ stationarity rate for the original LoRA objective with homogeneous ranks and for a surrogate objective with heterogeneous ranks. Experiments with RoBERTa and LLaMA-3.2-3B-Instruct show strong performance while reducing trainable parameters and uplink communications.

Limitations. While our complementary-channel design preserves an $\mathcal{O}(1/\sqrt{T})$ convergence rate for the original LoRA objective under homogeneous ranks, the result only applies to a surrogate objective in heterogeneous scenarios. This theoretical gap remains open. Additionally, there is a discrepancy between the worst-case convergence bound, which is minimized at channel selection probability $p = 0.5$, and our empirical findings, which favor an asymmetric allocation (e.g., $p = 0.9$). Developing a tighter theoretical framework that fully captures the asymmetric roles of factors remains an area for future exploration.

AI USE STATEMENT

In this work, we used generative AI tools specifically to polish the manuscript text, assist with code development, and check mathematical proofs. We did not use generative AI tools for any other tasks that require disclosure. All AI-assisted work was carefully reviewed by the authors. Specifically, the polished text was manually verified to ensure it accurately reflects our original scientific intent without altering any technical meaning, while all AI-assisted code and math were checked for accuracy. We take full responsibility for the final content of this work, including any text, code, or mathematical checking produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 4 details the complementary masks, client channel selection, local updates, and server aggregation mechanisms central to `CFLoRA`. Appendix A provides complete proofs for the aggregation and convergence guarantees, while Appendix C presents a comprehensive resource analysis. Finally, Section 5 and Appendix D thoroughly document the datasets, client partitions, baselines, hyperparameters, and evaluation protocols required to reproduce the experimental results.

REFERENCES

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *Advances in Neural Information Processing Systems*, volume 37, pp. 14457–14483, 2024. doi: 10.52202/079017-0461. URL https://openreview.net/forum?id=gkOzoHBXUw.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. LoRA-FAIR: Federated LoRA fine-tuning with aggregation and initialization refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3737–3746, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Bian_LoRA-FAIR_Federated_LoRA_Fine-Tuning_with_Aggregation_and_Initialization_Refinement_ICCV_2025_paper.html.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. PIQA: reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Yuji Byun and Jaeho Lee. Towards federated low-rank adaptation of language models with rank heterogeneity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 356–362, 2025. doi: 10.18653/v1/2025.naacl-short.30. URL https://aclanthology.org/2025.naacl-short.30/.

Jinqian Chen, Chang Liu, and Jihua Zhu. Beyond factor aggregation: Gauge-aware low-rank server representations for federated LoRA. *arXiv preprint arXiv:2605.06733*, 2026. doi: 10.48550/arXiv.2605.06733. URL https://arxiv.org/abs/2605.06733.

Shuangyi Chen, Yuanxin Guo, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish Khisti. Robust federated finetuning of LLMs via alternating optimization of LoRA. In *Advances in Neural Information Processing Systems*, volume 38, pp. 120038–120090, 2025. doi: 10.52202/085713-4009. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/adf17e7346b6be4c2f2bb40de572e5bc-Abstract-Conference.html.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous LoRA for federated fine-tuning of on-device foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12903–12913, 2024. doi: 10.18653/v1/2024.emnlp-main.717. URL https://aclanthology.org/2024.emnlp-main.717/.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Wenzhi Fang, Dong-Jun Han, Liangqi Yuan, Seyyedali Hosseinalipour, and Christopher G. Brinton. Federated sketching LoRA: A flexible framework for heterogeneous collaborative fine-tuning of LLMs. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=qFKFSVvff5.

Zihan Fang, Zheng Lin, Zhe Chen, Xianhao Chen, Yue Gao, and Yuguang Fang. Automated federated pipeline for parameter-efficient fine-tuning of large language models. *IEEE Transactions on Mobile Computing*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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*, 2025. URL https://openreview.net/forum?id=iX3uESGdsO.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 5254–5276, 2023.

Juseok Jeon, Ramy E. Ali, Doyun Kwon, Myungbeom Her, Jinhwi Kim, and Jinhyun So. FedPA-LoRA: Product-aligned framework for mitigating aggregation and initialization errors in heterogeneous federated LoRA. *arXiv preprint arXiv:2608.15381*, 2026. URL https://arxiv.org/abs/2608.15381.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021. doi: 10.1561/2200000083.

Myoungjun Kim, Sangwoo Park, Yoseob Han, and Jin-Hyun Ahn. Adaptive selection of LoRA components in privacy-preserving federated learning. *arXiv preprint arXiv:2605.05769*, 2026.

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*, pp. 416–429, 2025. doi: 10.18653/v1/2025.acl-long.19. URL https://aclanthology.org/2025.acl-long.19/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Chuiyang Meng, Ming Tang, and Vincent Wong. Florg: Federated fine-tuning with low-rank gram matrices and procrustes alignment. In *International Conference on Learning Representations*, volume 2026, pp. 82804–82824, 2026.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Arian Raje, Baris Askin, Divyansh Jhunjhunwala, and Gauri Joshi. Ravan: Multi-head low-rank adaptation for federated fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 48564–48592, 2025. doi: 10.52202/085713-1622. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/456a6eed524d1d647c3366f1ef96d0ea-Abstract-Conference.html.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. SocialIQA: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1316–1336, 2025. doi: 10.18653/v1/2025.acl-long.67. URL https://aclanthology.org/2025.acl-long.67/.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *International Conference on Learning Representations*, pp. 17978–17994, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4e243e95c913b367775d71d7182b99d9-Abstract-Conference.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, pp. 353–355, 2018. doi: 10.18653/v1/W18-5446.

Haoran Wang, Xiong Wang, Yuqing Li, Jing Chen, Junyi Zhang, Nan Yan, Kun He, and Wei Wang. Federated LoRA fine-tuning with pipelined error-mitigated aggregation and matrix-wise freezing. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5749–5762, 2026. doi: 10.18653/v1/2026.findings-acl.284. URL https://aclanthology.org/2026.findings-acl.284/.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 22513–22533, 2024. doi: 10.52202/079017-0708. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28312c9491d60ed0c77f7fff4ad86dd1-Abstract-Conference.html.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6915–6919, 2024. doi: 10.1109/ICASSP48485.2024.10446656.

APPENDIX

A Proof of the Theoretical Results **14**

A.1 Proof of Proposition 1 14

A.2 Proof of Proposition 2 14

A.3 Assumptions 14

A.4 Lemma A.1 and Its Proof 15

A.5 Proof of Theorem 3 15

A.6 Proof of Corollary 4 18

A.7 Insights from Theorem 3. 18

B Algorithm of `CFLoRA` **19**

C Implementation and Complexity Analysis **19**

D Experiments **21**

D.1 Implementation Details 21

D.2 Experiments of LLaMA-3.2-3B-Instruct 22

D.3 Further Experiments on Ablation Study on p 23

D.4 Experiments with Varying Client Participation 24

A Proof of the Theoretical Results

A.1 Proof of Proposition 1

Expanding the product of the aggregated factors yields

$$\begin{aligned}
&\left(\mathbf{B}^t + \sum_i \frac{w_i}{q}\Delta\mathbf{B}_i^t\mathbf{Z}^t\right)\left(\mathbf{A}^t + \sum_i \frac{w_i}{q}(\mathbf{I}_R - \mathbf{Z}^t)\Delta\mathbf{A}_i^t\right)\\
=&\mathbf{B}^t\mathbf{A}^t + \sum_i \frac{w_i}{q}\Delta\mathbf{B}_i^t\mathbf{Z}^t\mathbf{A}^t + \sum_i \frac{w_i}{q}\mathbf{B}^t(\mathbf{I}_R - \mathbf{Z}^t)\Delta\mathbf{A}_i^t + \sum_{i,k}\frac{w_i w_k}{q^2}\Delta\mathbf{B}_i^t\mathbf{Z}^t(\mathbf{I}_R - \mathbf{Z}^t)\Delta\mathbf{A}_k^t\\
=&\mathbf{B}^t\mathbf{A}^t + \sum_i \frac{w_i}{q}\Delta\mathbf{B}_i^t\mathbf{Z}^t\mathbf{A}^t + \sum_i \frac{w_i}{q}\mathbf{B}^t(\mathbf{I}_R - \mathbf{Z}^t)\Delta\mathbf{A}_i^t.
\end{aligned} \tag{22}$$

Each client endpoint satisfies

$$\mathbf{B}_i^{t+1}\mathbf{A}_i^{t+1} = (\mathbf{B}^t + \Delta\mathbf{B}_i^t\mathbf{Z}^t)(\mathbf{A}^t + (\mathbf{I} - \mathbf{Z}^t)\Delta\mathbf{A}_i^t) = \mathbf{B}^t\mathbf{A}^t + \Delta\mathbf{B}_i^t\mathbf{Z}^t\mathbf{A}^t + \mathbf{B}^t(\mathbf{I}_R - \mathbf{Z}^t)\Delta\mathbf{A}_i^t, \tag{23}$$

Subtracting $\mathbf{B}^t\mathbf{A}^t$ from each client endpoint product and summing over $i \in \mathcal{S}_t$ with weights w_i/q yields exactly the two linear terms in the expansion of the aggregated factors. This completes the proof.

A.2 Proof of Proposition 2

Expanding the product of the aggregated factors therefore yields

$$\begin{aligned}
&\left(\mathbf{B}^t + \sum_i \frac{w_i}{q}\Delta\widetilde{\mathbf{B}}_i^t(\mathbf{E}_i^t)^\top\right)\left(\mathbf{A}^t + \sum_i \frac{w_i}{q}\mathbf{E}_i^t\Delta\widetilde{\mathbf{A}}_i^t\right)\\
=&\mathbf{B}^t\mathbf{A}^t + \sum_i \frac{w_i}{q}\Delta\widetilde{\mathbf{B}}_i^t(\mathbf{E}_i^t)^\top\mathbf{A}^t + \sum_i \frac{w_i}{q}\mathbf{B}^t\mathbf{E}_i^t\Delta\widetilde{\mathbf{A}}_i^t + \sum_{i,k}\frac{w_i w_k}{q^2}\Delta\widetilde{\mathbf{B}}_i^t(\mathbf{E}_i^t)^\top\mathbf{Z}^t(\mathbf{I}_R - \mathbf{Z}^t)\mathbf{E}_k^t\Delta\widetilde{\mathbf{A}}_k^t\\
=&\mathbf{B}^t\mathbf{A}^t + \sum_i \frac{w_i}{q}\Delta\widetilde{\mathbf{B}}_i^t(\mathbf{E}_i^t)^\top\mathbf{A}^t + \sum_i \frac{w_i}{q}\mathbf{B}^t\mathbf{E}_i^t\Delta\widetilde{\mathbf{A}}_i^t.
\end{aligned} \tag{24}$$

For the local models, each client's compact endpoint satisfies

$$\begin{aligned}
\widetilde{\mathbf{B}}_i^{t,K}\widetilde{\mathbf{A}}_i^{t,K} &= (\mathbf{B}^t\mathbf{E}_i^t + \Delta\widetilde{\mathbf{B}}_i^t)((\mathbf{E}_i^t)^\top\mathbf{A}^t + \Delta\widetilde{\mathbf{A}}_i^t)\\
&= \mathbf{B}^t\mathbf{H}_i^t\mathbf{A}^t + \Delta\widetilde{\mathbf{B}}_i^t(\mathbf{E}_i^t)^\top\mathbf{A}^t + \mathbf{B}^t\mathbf{E}_i^t\Delta\widetilde{\mathbf{A}}_i^t + \Delta\widetilde{\mathbf{B}}_i^t(\mathbf{E}_i^t)^\top\mathbf{Z}^t(\mathbf{I}_R - \mathbf{Z}^t)\mathbf{E}_i^t\Delta\widetilde{\mathbf{A}}_i^t\\
&= \mathbf{B}^t\mathbf{H}_i^t\mathbf{A}^t + \Delta\widetilde{\mathbf{B}}_i^t(\mathbf{E}_i^t)^\top\mathbf{A}^t + \mathbf{B}^t\mathbf{E}_i^t\Delta\widetilde{\mathbf{A}}_i^t.
\end{aligned} \tag{25}$$

Subtracting $\mathbf{B}^t\mathbf{H}_i^t\mathbf{A}^t$ and taking the weighted summation over all clients yields exactly the two linear terms in equation 24. Substituting this weighted sum back into equation 24 reconstructs the right-hand side of equation 14, which concludes the proof.

A.3 Assumptions

We adopt standard assumptions widely used in the analysis of non-convex federated optimization.

Assumption A.1. *For each client i and activation mask $\mathbf{H} \in \mathrm{supp}(\mathcal{H}_i)$, the local objective $f_i^{\mathbf{H}}(\mathbf{X})$ is differentiable and L-smooth over a region $\mathcal{C}$ containing the iterates; that is, there exists a constant $L > 0$ such that for all $\mathbf{X}, \mathbf{X}' \in \mathcal{C}$,*

$$\|\nabla f_i^{\mathbf{H}}(\mathbf{X}) - \nabla f_i^{\mathbf{H}}(\mathbf{X}')\| \le L\|\mathbf{X} - \mathbf{X}'\|, \quad \forall i. \tag{26}$$

Consequently, the global objective $\Phi(\mathbf{X})$ is also L-smooth on $\mathcal{C}$. Moreover, there exists a finite lower bound $\Phi^\star$ such that $\Phi(\mathbf{X}) \ge \Phi^\star$ for all $\mathbf{X} \in \mathcal{C}$.

Assumption A.2. *For each client i and any activation mask $\mathbf{H} \in \mathrm{supp}(\mathcal{H}_i)$, the stochastic gradient is unbiased, i.e.,*

$$\mathbb{E}_{\xi\sim\mathcal{D}_i}\left[\nabla_{\mathbf{X}}\ell(\mathbf{W}_0 + \mathbf{BHA}; \xi)\right] = \nabla f_i^{\mathbf{H}}(\mathbf{X}). \tag{27}$$

Furthermore, there exist constants $\sigma_i^2, \tau_i^2 \geq 0$ bounding the variances of minibatch sampling and mask sampling, respectively, and a uniformly bounded second moment $M > 0$:

$$\sup_{\mathbf{X}\in\mathcal{C},\, \mathbf{H}\in\mathrm{supp}(\mathcal{H}_i)} \mathbb{E}_{\xi\sim\mathcal{D}_i}\left[\left\|\nabla_{\mathbf{X}}\ell(\mathbf{W}_0+\mathbf{BHA};\xi)-\nabla f_i^{\mathbf{H}}(\mathbf{X})\right\|^2\right] \leq \sigma_i^2, \tag{28}$$

$$\sup_{\mathbf{X}\in\mathcal{C}} \mathbb{E}_{\mathbf{H}\sim\mathcal{H}_i}\left[\left\|\nabla f_i^{\mathbf{H}}(\mathbf{X})-\nabla\phi_i(\mathbf{X})\right\|^2\right] \leq \tau_i^2, \tag{29}$$

$$\sup_{\mathbf{X}\in\mathcal{C},\, \mathbf{H}\in\mathrm{supp}(\mathcal{H}_i)} \mathbb{E}_{\xi\sim\mathcal{D}_i}\left[\left\|\nabla_{\mathbf{X}}\ell(\mathbf{W}_0+\mathbf{BHA};\xi)\right\|^2\right] \leq M. \tag{30}$$

Assumption A.1 is standard in optimization literature (Bottou et al., 2018; Fang et al., 2026). Assumption A.2 is commonly adopted in FL to bound sampling randomness and stochastic-gradient moments (Fang et al., 2026).

A.4 Lemma A.1 and Its Proof

Lemma A.1. *Under Assumption A.2, we have*

$$\mathbb{E}[\mathbf{G}_B^t] = \nabla_{\mathbf{B}}\Phi(\mathbf{X}^t), \qquad \mathbb{E}[\mathbf{G}_A^t] = \nabla_{\mathbf{A}}\Phi(\mathbf{X}^t), \tag{31}$$

where $\mathbf{G}_B^t = \frac{1}{p}\mathbf{S}_B^t\mathbf{Z}^t$ and $\mathbf{G}_A^t = \frac{1}{1-p}(\mathbf{I}_R - \mathbf{Z}^t)\mathbf{S}_A^t$ denote the scaled aggregated gradients, and $\mathbf{S}_Y^t = \sum_{i\in\mathcal{S}_t}\frac{w_i}{q}\nabla_{\mathbf{Y}}\ell(\mathbf{W}_0+\mathbf{BH}_i^t\mathbf{A};\xi_i^{t,0})\Big|_{\mathbf{X}=\mathbf{X}^t}$ are the pre-selection aggregated gradients at the common round-start iterate $\mathbf{X}^t$ for $Y \in \{B, A\}$, with local minibatch sample $\xi_i^{t,0} \sim \mathcal{D}_i$.

Proof. Let $I_i^t = \mathbf{1}\{i \in \mathcal{S}_t\}$, $q = m/N$, and $\alpha_i^t = w_i I_i^t/q$. Uniform client sampling gives $\mathbb{E}[I_i^t] = q$. By the independence of client sampling and activation masks, we have

$$\mathbb{E}[\mathbf{S}_B^t] = \sum_i \frac{w_i}{q}\mathbb{E}[I_i^t]\mathbb{E}_{\mathbf{H}\sim\mathcal{H}_i}\left[\nabla_{\mathbf{B}} f_i^{\mathbf{H}}(\mathbf{X}^t)\right] = \nabla_{\mathbf{B}}\Phi(\mathbf{X}^t). \tag{32}$$

Similarly, $\mathbb{E}[\mathbf{S}_A^t] = \nabla_{\mathbf{A}}\Phi(\mathbf{X}^t)$ holds. Since $\mathbf{Z}^t$ is conditionally independent of $\mathbf{S}_B^t, \mathbf{S}_A^t$ and $\mathbb{E}[\mathbf{Z}^t] = p\mathbf{I}_R$, iterated expectations yield

$$\begin{aligned}
\mathbb{E}[\mathbf{G}_B^t] &= \frac{1}{p}\mathbb{E}\left[\mathbf{S}_B^t\mathbb{E}[\mathbf{Z}^t \mid \mathbf{S}_B^t]\right] = \nabla_{\mathbf{B}}\Phi(\mathbf{X}^t),\\
\mathbb{E}[\mathbf{G}_A^t] &= \frac{1}{1-p}\mathbb{E}\left[\mathbb{E}[\mathbf{I}_R - \mathbf{Z}^t \mid \mathbf{S}_A^t]\mathbf{S}_A^t\right] = \nabla_{\mathbf{A}}\Phi(\mathbf{X}^t).
\end{aligned} \tag{33}$$

This leads to equation 31, which concludes the proof. □

A.5 Proof of Theorem 3

Proof. Let $p_{\min} = \min\{p, 1-p\}, \kappa = p_{\min}^{-1}, \gamma = \eta K, \alpha_i^t = \frac{w_i}{q}\mathbf{1}\{i \in \mathcal{S}_t\}, I_i^t = \mathbf{1}\{i \in \mathcal{S}_t\}$, where $p \in (0,1)$ is fixed and $q = m/N$.

Step 1: Selection and client-sampling identities. To formalize the complementary channel selection mechanism in CFLoRA, let $\mathbf{u} = (\mathbf{u}_B, \mathbf{u}_A)$ denote an arbitrary partitioned variable in the parameter or gradient space. We define the random scaling operator $\mathcal{R}_t$ as: $\mathcal{R}_t(\mathbf{u}) = \left(\frac{\mathbf{u}_B\mathbf{Z}^t}{p}, \frac{(\mathbf{I}_R-\mathbf{Z}^t)\mathbf{u}_A}{1-p}\right)$. By the definition of $\kappa = p_{\min}^{-1}$, for any $\mathbf{u}$, the operator's norm is bounded:

$$\|\mathcal{R}_t(\mathbf{u})\| \leq \kappa\|\mathbf{u}\|. \tag{34}$$

Furthermore, for any $\mathcal{F}_t$-measurable $\mathbf{u}$, taking the expectation over the random mask yields unbiasedness and a bounded second moment:

$$\mathbb{E}[\mathcal{R}_t(\mathbf{u})] = \mathbf{u}, \qquad \mathbb{E}\|\mathcal{R}_t(\mathbf{u})\|^2 = \frac{\|\mathbf{u}_B\|_F^2}{p} + \frac{\|\mathbf{u}_A\|_F^2}{1-p} \leq \kappa\|\mathbf{u}\|^2. \tag{35}$$

Uniform client sampling gives $\mathbb{E}[I_i^t] = q$ and $\mathbb{E}[I_i^t I_j^t] = \frac{m(m-1)}{N(N-1)}$, and hence for vectors $\mathbf{y}_i$, we have

$$\begin{aligned}\mathbb{E}\left[\sum_i \alpha_i^t \mathbf{y}_i\right] &= \sum_i w_i \mathbf{y}_i, \\ \mathbb{E}\left\|\sum_i \alpha_i^t \mathbf{y}_i\right\|^2 &= \|\sum_i w_i \mathbf{y}_i\|^2 + \frac{N-m}{m(N-1)}\left(N\sum_i w_i^2\|\mathbf{y}_i\|^2 - \|\sum_i w_i \mathbf{y}_i\|^2\right).\end{aligned} \tag{36}$$

In particular, define $\omega_t = \sum_i \alpha_i^t$, we have

$$\mathbb{E}[\omega_t] = 1, \qquad \mathbb{E}[\omega_t^2] = 1 + \frac{N-m}{m(N-1)}\left(N\sum_i w_i^2 - 1\right) = \chi_m. \tag{37}$$

Step 2: Endpoint decomposition and sampling noise. The lifted local updates satisfy $\mathbf{X}_i^{t,0} = \mathbf{X}^t$ and $\mathbf{X}_i^{t,k+1} = \mathbf{X}_i^{t,k} - \eta\mathcal{R}_t(\mathbf{g}_i^{t,k})$, where $\mathbf{g}_i^{t,k}$ is the stochastic gradient of $f_i^{\mathbf{H}_i}$ at $\mathbf{X}_i^{t,k}$. The server update becomes $\mathbf{X}^{t+1} = \mathbf{X}^t - \gamma\mathbf{v}^t$, with $\mathbf{v}^t$ defined as

$$\begin{aligned}\mathbf{v}^t = \mathcal{R}_t\left(\sum_i \alpha_i^t \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0})\right) + \frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t(\nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k}) - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0})), \\ + \frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t((\mathbf{g}_i^{t,k} - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k}))).\end{aligned} \tag{38}$$

Independent activation masks and equation 36 yield

$$\begin{aligned}\mathbb{E}\left\|\sum_i \alpha_i^t \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0})\right\|^2 &= \mathbb{E}\left\|\sum_i \alpha_i^t \nabla \phi_i(\mathbf{X}^t)\right\|^2 + \frac{1}{q}\sum_i w_i^2 \mathbb{E}\|\nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0}) - \nabla\phi_i(\mathbf{X}^t)\|^2 \\ &\le \|\nabla\Phi(\mathbf{X}^t)\|^2 + \frac{N-m}{m(N-1)}\zeta_w^2 + \frac{\tau_w^2}{q}.\end{aligned} \tag{39}$$

where $\zeta_w^2 = \sup_{\mathbf{X}\in\mathcal{C}}\left\{N\sum_i w_i^2\|\nabla\phi_i(\mathbf{X})\|^2 - \|\nabla\Phi(\mathbf{X})\|^2\right\}$ and $\tau_w^2 = \sum_{i=1}^N w_i^2\tau_i^2$. Since the round-start gradients are independent of $\mathbf{Z}^t$,

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_t\left(\sum_i \alpha_i^t \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0})\right)\right] &= \nabla\Phi(\mathbf{X}^t), \\ \mathbb{E}\left\|\mathcal{R}_t\left(\sum_i \alpha_i^t \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0})\right)\right\|^2 &\le \kappa\left(\|\nabla\Phi(\mathbf{X}^t)\|^2 + \frac{N-m}{m(N-1)}\zeta_w^2 + \frac{\tau_w^2}{q}\right).\end{aligned} \tag{40}$$

Let $\mathcal{F}_{t,k}$ be the filtration extending $\mathcal{F}_t$ with all randomness up to local step k. Fresh minibatches give $\mathbb{E}[(\mathbf{g}_i^{t,k} - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k})) \mid \mathcal{F}_{t,k}] = 0$ and $\mathbb{E}[\|(\mathbf{g}_i^{t,k} - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k}))\|^2 \mid \mathcal{F}_{t,k}] \le \sigma_i^2$.

Noise cross terms vanish across steps by conditional centering and across clients by conditional independence. Since $\mathbb{E}[(\alpha_i^t)^2] = w_i^2/q$,

$$\begin{aligned}\mathbb{E}[\frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t((\mathbf{g}_i^{t,k} - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k})))] = 0, \\ \mathbb{E}\|\frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t((\mathbf{g}_i^{t,k} - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k})))\|^2 \le \frac{\kappa^2\sigma_w^2}{qK}, \\ \mathbb{E}\left\langle \mathcal{R}_t\left(\sum_i \alpha_i^t \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0})\right), \frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t((\mathbf{g}_i^{t,k} - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k})))\right\rangle = 0.\end{aligned} \tag{41}$$

The last identity uses that $\mathcal{R}_t\left(\sum_i \alpha_i^t \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0})\right)$ is fixed given the selected set and round masks.

Step 3: Local drift. Write $\mathbb{E}_{t,S}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t, \mathcal{S}_t]$. Assumption A.2 and equation 34 imply

$$\begin{aligned}\left(\mathbb{E}_{t,S}\|\mathbf{X}_i^{t,k} - \mathbf{X}^t\|^2\right)^{1/2} &\le \eta \sum_{s=0}^{k-1} \left(\mathbb{E}_{t,S}\|\mathcal{R}_t(\mathbf{g}_i^{t,s})\|^2\right)^{1/2} \\ &\le \eta k \kappa \sqrt{M}.\end{aligned} \tag{42}$$

Smoothness and the conditional L^2 triangle inequality give

$$\begin{aligned}&\left(\mathbb{E}_{t,S}\|\frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t(\nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k}) - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0}))\|^2\right)^{1/2} \\ &\le \frac{\kappa L}{K}\sum_{k,i}\alpha_i^t \left(\mathbb{E}_{t,S}\|\mathbf{X}_i^{t,k} - \mathbf{X}^t\|^2\right)^{1/2} \\ &\le \frac{\omega_t}{2} L\eta\kappa^2\sqrt{M}(K-1).\end{aligned} \tag{43}$$

For any $\mathcal{F}_t$-measurable unit vector $\mathbf{z}$, self-adjointness and Cauchy–Schwarz yield

$$\begin{aligned}&\left|\langle \mathbf{z}, \mathbb{E}_{t,S}[\frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t(\nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k}) - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0}))]\rangle\right| \\ &\le \frac{1}{K}\sum_{k,i}\alpha_i^t \left(\mathbb{E}_{t,S}\|\mathcal{R}_t(\mathbf{z})\|^2\right)^{1/2}\left(\mathbb{E}_{t,S}\|\nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k}) - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0})\|^2\right)^{1/2} \\ &\le \frac{\omega_t}{2} L\eta\kappa^{3/2}\sqrt{M}(K-1).\end{aligned} \tag{44}$$

Define $d = \frac{1}{4}L^2\eta^2\kappa^3 M(K-1)^2 = \frac{1}{4}L^2\gamma^2\kappa^3 M\left(1-\frac{1}{K}\right)^2$, averaging over $\mathcal{S}_t$ and using equation 37, then taking the supremum over unit $\mathbf{z}$, gives

$$\begin{aligned}\|\mathbb{E}[\frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t(\nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k}) - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0}))]\|^2 \le d, \\ \mathbb{E}\|\frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t(\nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k}) - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0}))\|^2 \le \chi_m \kappa d.\end{aligned} \tag{45}$$

Step 4: Descent and telescoping. By equation 40, equation 41, and equation 45,

$$\begin{aligned}\mathbb{E}\|\mathbf{v}^t\|^2 \le& 2\mathbb{E}\|\mathcal{R}_t\left(\sum_i \alpha_i^t \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0})\right) + \frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t((\mathbf{g}_i^{t,k} - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k})))\|^2 \\ &+ 2\mathbb{E}\|\frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t(\nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k}) - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0}))\|^2 \\ \le& 2\kappa\|\nabla\Phi(\mathbf{X}^t)\|^2 + 2\mathcal{V}_K + 2\chi_m\kappa d.\end{aligned} \tag{46}$$

Moreover, $\mathbb{E}[\mathbf{v}^t] = \nabla\Phi(\mathbf{X}^t) + \mathbb{E}[\frac{1}{K}\sum_{k=0}^{K-1}\sum_i \alpha_i^t \mathcal{R}_t(\nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,k}) - \nabla f_i^{\mathbf{H}}(\mathbf{X}_i^{t,0}))]$ and $ab \le a^2/4 + b^2$ imply

$$-\langle\nabla\Phi(\mathbf{X}^t), \mathbb{E}[\mathbf{v}^t]\rangle \le -\frac{3}{4}\|\nabla\Phi(\mathbf{X}^t)\|^2 + d. \tag{47}$$

By L-smoothness and $\gamma \le (4L\kappa)^{-1}$,

$$\begin{aligned}\mathbb{E}[\Phi(\mathbf{X}^{t+1})] &\le \Phi(\mathbf{X}^t) - \gamma\langle\nabla\Phi(\mathbf{X}^t), \mathbb{E}[\mathbf{v}^t]\rangle + \frac{L\gamma^2}{2}\mathbb{E}\|\mathbf{v}^t\|^2 \\ &\le \Phi(\mathbf{X}^t) - \gamma\left(\frac{3}{4} - L\gamma\kappa\right)\|\nabla\Phi(\mathbf{X}^t)\|^2 + L\gamma^2\mathcal{V}_K + \gamma(1 + L\gamma\kappa\chi_m)d \\ &\le \Phi(\mathbf{X}^t) - \frac{\gamma}{2}\|\nabla\Phi(\mathbf{X}^t)\|^2 + L\gamma^2\mathcal{V}_K + \gamma\left(1 + \frac{\chi_m}{4}\right)d.\end{aligned}$$

Taking total expectations, summing over t, and using $\Phi \geq \Phi^\star$ yields

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\Phi(\mathbf{X}^t)\|^2 &\leq \frac{2\Delta_0}{\gamma T} + 2L\gamma\mathcal{V}_K + \left(2+\frac{\chi_m}{2}\right)d \\ &= \frac{2\Delta_0}{\gamma T} + 2L\gamma\mathcal{V}_K + \mathcal{D}_K\gamma^2,\end{aligned} \tag{48}$$

where

$$\mathcal{V}_K = \kappa\left(\frac{\tau_w^2}{q} + \frac{N-m}{m(N-1)}\zeta_w^2\right) + \frac{\kappa^2\sigma_w^2}{qK}, \qquad \mathcal{D}_K = \frac{4+\chi_m}{8}L^2\kappa^3 M\left(1-\frac{1}{K}\right)^2.$$

Step 5: Learning-rate choice. For $\gamma_T = K\eta_T$ in equation 19, omitting candidates with zero denominators,

$$\frac{2\Delta_0}{\gamma_T T} = \max\left\{\frac{8L\kappa\Delta_0}{T}, 2\sqrt{\frac{L\Delta_0\mathcal{V}_K}{T}}, \left(\frac{4\Delta_0^2\mathcal{D}_K}{T^2}\right)^{1/3}\right\},$$

and

$$2L\gamma_T\mathcal{V}_K \leq 2\sqrt{\frac{L\Delta_0\mathcal{V}_K}{T}}, \qquad \mathcal{D}_K\gamma_T^2 \leq \left(\frac{4\Delta_0^2\mathcal{D}_K}{T^2}\right)^{1/3}.$$

Substituting into equation 48 and bounding the maximum by the sum establishes equation 20, which completes the proof. □

A.6 Proof of Corollary 4

Proof. When $r_i = R$ for every client, $\Phi = F$ and $\tau_w^2 = 0$. Substituting $\gamma = (4L\kappa\sqrt{T})^{-1}$ into equation 48 gives

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{X}^t)\|^2 \leq{}& \frac{8L\kappa\Delta_0 + \frac{N-m}{m(N-1)}\zeta_w^2/2 + \kappa\sigma_w^2/(2qK)}{\sqrt{T}} \\ &+ \frac{(4+\chi_m)\kappa M}{128T}\left(1-\frac{1}{K}\right)^2,\end{aligned}$$

which is equation 21. □

A.7 Insights from Theorem 3.

Asymptotic Convergence and Exact Aggregation. Theorem 3 establishes that, under an appropriately tuned learning rate, CFLoRA achieves a sublinear convergence rate of $\mathcal{O}(T^{-1/2})$ to a stationary point of the surrogate objective. Crucially, this bound isolates the effects of the learning rate and statistical noise without suffering from any non-vanishing error floor. By employing complementary-channel selection, CFLoRA achieves exact aggregation, thereby completely eliminating this bilinear aggregation error from the asymptotic rate. The parameter $\kappa = 1/\min\{p, 1-p\}$ explicitly captures the variance expansion cost of alternating between the factor matrices $\mathbf{A}$ and $\mathbf{B}$.

Components in Variance ($\mathcal{V}_K$). The term $\mathcal{V}_K$ encapsulates three distinct sources of statistical variance in the federated system, each scaled by the protocol's alternating selection factor κ:

- **Activation-mask variability (τ_w^2):** This variance arises from the heterogeneity of client rank budgets. Because clients randomly sample sub-matrices corresponding to their local capacity, their updates fluctuate around the true gradient of Φ. If all clients possess the capacity to use the full server rank (i.e., $r_i = R$ and $\mathbf{H}_i^t = \mathbf{I}_R$), this term vanishes.
- **Client-sampling variance (ζ_w^2):** This term accounts for the data heterogeneity across different clients under partial participation ($m < N$). The coefficient $\frac{N-m}{m(N-1)}$ reflects standard sampling without replacement; thus, if the server aggregates from all clients in every round ($m = N$), this inter-client variance vanishes.
- **Minibatch noise (σ_w^2):** The standard stochastic gradient noise from local data sampling is effectively suppressed by the number of local gradient steps executed across participating clients (K).

Tradeoffs of Local Training Steps (K). Theorem 3 reveals a fundamental theoretical tradeoff governed by the number of local steps K, explicitly controlled by the balance between $\mathcal{V}_K$ and the local drift term $\mathcal{D}_K$. Executing more local steps reduces the impact of minibatch noise (via the $\frac{\kappa^2\sigma_w^2}{qK}$ term in $\mathcal{V}_K$). However, this variance reduction comes at the cost of increased local drift across clients ($\mathcal{D}_K$), which scales with $(1-1/K)^2$. Because this structural drift vanishes entirely when $K=1$, the choice of K should balance statistical minibatch variance and client local drifts. Note that the activation-mask variability (τ_w^2) remains unaffected by K, as the masks are drawn only once per communication round.

Heterogeneous Ranks and the Surrogate Objective (Φ). Since clients with low rank budgets evaluate gradients using masked sub-matrices, their updates are structurally biased with respect to the full-adapter objective F. Consequently, the theoretical guarantee concerns the surrogate objective Φ, which incorporates these heterogeneous client ranks through the activation distributions $\{\mathcal{H}_i\}_{i=1}^N$. When every client uses the full server rank ($\mathbf{H}_i^t = \mathbf{I}_R$), Φ reduces to F.

B Algorithm of CFLoRA

Algorithm B.1: CFLoRA with Homogeneous Client Ranks

Input: $\mathbf{W}_0$, $(\mathbf{B}_{\text{init}}, \mathbf{A}_{\text{init}})$, R, $\{w_i\}_{i=1}^N$, m, T, K, η, $p \in (0,1)$
Output: $(\mathbf{B}^T, \mathbf{A}^T)$

$\mathbf{B}^0 \leftarrow \mathbf{B}_{\text{init}}$, $\mathbf{A}^0 \leftarrow \mathbf{A}_{\text{init}}$;
$q \leftarrow m/N$, $\eta_B \leftarrow \eta/p$, $\eta_A \leftarrow \eta/(1-p)$;
for $t \leftarrow 0$ **to** $T-1$ **do**
  Sample m clients from $\{1,\dots,N\}$ to form $\mathcal{S}^t$;
  Draw $z_j^t \sim \text{Bernoulli}(p)$ independently for $j = 1,\dots,R$;
  $\mathbf{Z}^t \leftarrow \text{diag}(z_1^t,\dots,z_R^t)$;
  Broadcast $(\mathbf{B}^t, \mathbf{A}^t, \mathbf{Z}^t)$ to clients in $\mathcal{S}^t$;
  for $i \in \mathcal{S}^t$ in parallel **do**
    $\mathbf{B}_i^{t,0} \leftarrow \mathbf{B}^t$, $\mathbf{A}_i^{t,0} \leftarrow \mathbf{A}^t$;
    for $k \leftarrow 0$ **to** $K-1$ **do**
      Compute gradients $(\mathbf{g}_{B,i}^{t,k}, \mathbf{g}_{A,i}^{t,k})$ of $\ell\Big(\mathbf{W}_0 + \mathbf{B}_i^{t,k}\mathbf{A}_i^{t,k}; \xi_i^{t,k}\Big)$ with respect to the local factors;
      $\mathbf{B}_i^{t,k+1} \leftarrow \mathbf{B}_i^{t,k} - \eta_B \mathbf{g}_{B,i}^{t,k}\mathbf{Z}^t$;
      $\mathbf{A}_i^{t,k+1} \leftarrow \mathbf{A}_i^{t,k} - \eta_A(\mathbf{I}_R - \mathbf{Z}^t)\mathbf{g}_{A,i}^{t,k}$;
    end
    $\Delta\mathbf{B}_i^t \leftarrow \mathbf{B}_i^{t,K} - \mathbf{B}_i^{t,0}$;
    $\Delta\mathbf{A}_i^t \leftarrow \mathbf{A}_i^{t,K} - \mathbf{A}_i^{t,0}$;
    Upload the active entries of $\Delta\mathbf{B}_i^t$ and $\Delta\mathbf{A}_i^t$;
  end
  // Factor-wise aggregation
  $\mathbf{B}^{t+1} \leftarrow \mathbf{B}^t + \sum_{i\in\mathcal{S}^t} \frac{w_i}{q}\Delta\mathbf{B}_i^t$;
  $\mathbf{A}^{t+1} \leftarrow \mathbf{A}^t + \sum_{i\in\mathcal{S}^t} \frac{w_i}{q}\Delta\mathbf{A}_i^t$;
end
return $(\mathbf{B}^T, \mathbf{A}^T)$;

C Implementation and Complexity Analysis

Efficient Implementation. CFLoRA introduces *negligible* additional communication overhead beyond exchanging compact LoRA factors and updates. The additional control information only involves R-bit complementary mask and $\mathcal{O}(r_i \log R)$ bits specifying client i's selected channel indices. Even more, this information can alternatively be reproduced from shared random seeds. The

Algorithm B.2: `CFLoRA` with Heterogeneous Client Ranks

Input: $\mathbf{W}_0$, $(\mathbf{B}_{\text{init}}, \mathbf{A}_{\text{init}})$, R, $\{r_i\}_{i=1}^N$, $\{w_i\}_{i=1}^N$, m, T, K, η, $p \in (0,1)$
Output: $(\mathbf{B}^T, \mathbf{A}^T)$

$\mathbf{B}^0 \leftarrow \mathbf{B}_{\text{init}}$, $\mathbf{A}^0 \leftarrow \mathbf{A}_{\text{init}}$;
$q \leftarrow m/N$, $\eta_B \leftarrow \eta/p$, $\eta_A \leftarrow \eta/(1-p)$;
for $t \leftarrow 0$ **to** $T-1$ **do**
  Sample m clients from $\{1, \dots, N\}$ to form $\mathcal{S}^t$;
  `// Shared complementary mask`
  Draw $z_j^t \sim \text{Bernoulli}(p)$ independently for $j = 1, \dots, R$;
  $\mathbf{Z}^t \leftarrow \text{diag}(z_1^t, \dots, z_R^t)$;
  `// Client-specific rank selections`
  for $i \in \mathcal{S}^t$ **do**
    Uniformly sample r_i distinct indices from $\{1, \dots, R\}$;
    Construct $\mathbf{E}_i^t$ from the standard basis vectors corresponding to the sampled indices;
  end
  for $i \in \mathcal{S}^t$ in parallel **do**
    $\widetilde{\mathbf{B}}_i^{t,0} \leftarrow \mathbf{B}^t \mathbf{E}_i^t$, $\widetilde{\mathbf{A}}_i^{t,0} \leftarrow (\mathbf{E}_i^t)^\top \mathbf{A}^t$;
    $\widetilde{\mathbf{Z}}_i^t \leftarrow (\mathbf{E}_i^t)^\top \mathbf{Z}^t \mathbf{E}_i^t$;
    for $k \leftarrow 0$ **to** $K-1$ **do**
      Compute gradients $(\mathbf{g}_{B,i}^{t,k}, \mathbf{g}_{A,i}^{t,k})$ of $\ell\Big(\mathbf{W}_0 + \widetilde{\mathbf{B}}_i^{t,k} \widetilde{\mathbf{A}}_i^{t,k}; \xi_i^{t,k}\Big)$ with respect to the compact factors;
      $\widetilde{\mathbf{B}}_i^{t,k+1} \leftarrow \widetilde{\mathbf{B}}_i^{t,k} - \eta_B \mathbf{g}_{B,i}^{t,k} \widetilde{\mathbf{Z}}_i^t$;
      $\widetilde{\mathbf{A}}_i^{t,k+1} \leftarrow \widetilde{\mathbf{A}}_i^{t,k} - \eta_A (\mathbf{I}_{r_i} - \widetilde{\mathbf{Z}}_i^t) \mathbf{g}_{A,i}^{t,k}$;
    end
    $\Delta\widetilde{\mathbf{B}}_i^t \leftarrow \widetilde{\mathbf{B}}_i^{t,K} - \widetilde{\mathbf{B}}_i^{t,0}$;
    $\Delta\widetilde{\mathbf{A}}_i^t \leftarrow \widetilde{\mathbf{A}}_i^{t,K} - \widetilde{\mathbf{A}}_i^{t,0}$;
    Upload the active entries of $\Delta\widetilde{\mathbf{B}}_i^t$ and $\Delta\widetilde{\mathbf{A}}_i^t$;
  end
  `// Factor-wise aggregation`
  $\mathbf{B}^{t+1} \leftarrow \mathbf{B}^t + \sum_{i \in \mathcal{S}^t} \frac{w_i}{q} \Delta\widetilde{\mathbf{B}}_i^t (\mathbf{E}_i^t)^\top$;
  $\mathbf{A}^{t+1} \leftarrow \mathbf{A}^t + \sum_{i \in \mathcal{S}^t} \frac{w_i}{q} \mathbf{E}_i^t \Delta\widetilde{\mathbf{A}}_i^t$;
end
return $(\mathbf{B}^T, \mathbf{A}^T)$;

main additional operation is channel indexing: each participating client receives the r_i channels specified by $\mathbf{E}_i^t$, and the server maps the returned increments back to their corresponding global positions. These operations are implemented directly through indexing, without explicitly constructing the selection matrices. Local training uses two complementary branches to compute only the selected factor gradients, while the server aggregates the compact increments with weights w_i/q. Thus, the additional computation is limited to channel indexing and masking, which are negligible.

Trainable Parameters. At the same local rank r_i, `CFLoRA` retains the same number of adapter parameters as standard LoRA but updates only one factor per rank channel: either a column of $\widetilde{\mathbf{B}}_i$ or the corresponding row of $\widetilde{\mathbf{A}}_i$. For square projections (i.e., $d_{\text{in}} = d_{\text{out}}$), these two vectors have equal size, so `CFLoRA` trains exactly half of the adapter parameters, regardless of the realized complementary mask.

The computational saving comes from omitting parameter-gradient calculations for the frozen factor entries. Both compact factors remain necessary for the forward pass.

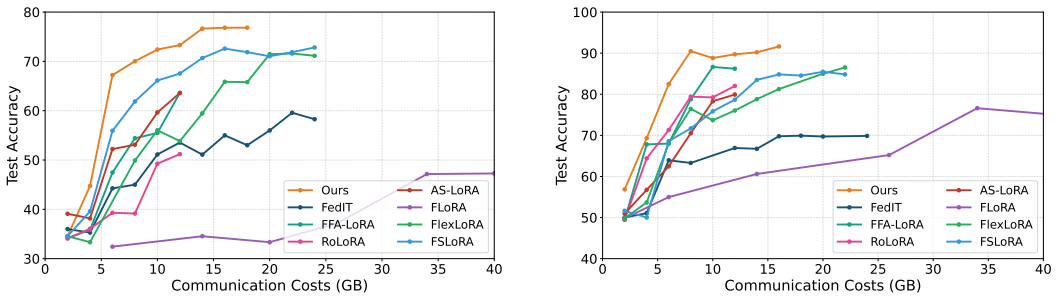
Communication Costs. Each participating client downloads both compact adapter factors and uploads only the updated columns of $\widetilde{\mathbf{B}}_i$ and rows of $\widetilde{\mathbf{A}}_i$ in each round. At the same local rank,

Table C.1: Normalized cost comparison at the same local rank for square projections between standard federated LoRA and our CFLoRA. MAC counts assume that input gradients are required and only selected factor gradients are computed. Communication assumes compact transmission at equal precision. Control information is negligible, which is excluded.

Quantity	Standard federated LoRA	CFLoRA
Active adapter parameters	1	1
Trainable adapter parameters	1	$1/2$
Forward MACs	1	1
Input-gradient MACs	1	1
Factor-gradient MACs	1	$1/2$
Total adapter MACs per step	1	$5/6$
Download payload per round	1	1
Upload payload per round	1	$1/2$
Combined communication per round	1	$3/4$

the adapter download is unchanged compared with standard federated LoRA, because both trainable and frozen entries are required for the local forward pass. For square projections, only half of the adapter entries are updated and uploaded. Therefore, assuming the same transmission precision and excluding the negligible control metadata, CFLoRA reduces the adapter upload by 50% and the combined adapter download and upload by 25%.

Table C.1 summarizes the comparison between standard federated LoRA and CFLoRA. Moreover, Figure C.1 illustrates accuracy versus communication costs under heterogeneous client ranks.



(a) Test accuracy (matched) versus communication costs on the MNLI dataset.

(b) Test accuracy versus communication costs on the QNLI dataset.

Figure C.1: Test accuracy versus communication costs of CFLoRA (Ours) and federated LoRA baselines on the MNLI and QNLI datasets.

D Experiments

D.1 Implementation Details

In the homogeneous setting, all clients share a uniform rank of $R = 8$. In the heterogeneous setting, the server maintains rank $R = 32$, and each client is assigned a fixed rank $r_i \in \{4, 8, 16, 32\}$ throughout training. To model heterogeneous client capabilities, we draw rank values from a normal distribution with mean $\mu = (a+b)/2$ and standard deviation $\sigma = (b-a)/6$, where $a = 4$ and $b = 32$. The sampled values are then clipped to $[4, 32]$ and mapped to the nearest allowed rank. The LoRA factors $\mathbf{A}$ and $\mathbf{B}$ are initialized randomly and with zeros, respectively. For the shared global adapter, the LoRA scaling multiplier is fixed at 2. In each round, all clients use the same complementary channel selection matrix $\mathbf{Z}$. For the commonsense reasoning benchmark, the training corpus is a mixture of the ARC-c/e (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), OBQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), and SIQA (Sap et al., 2019) datasets. Unless otherwise stated, the default hyperparameters used in our experiments are summarized in Table D.2.

Table D.2: parameter settings.

Setting	RoBERTa / GLUE	LLaMA-3.2-3B-Instruct / commonsense reasoning
Total clients	25	25
Client participation ratio q	0.4	0.4
Communication rounds T	200	100
Local steps K	50	40
Batch size	8	8
Learning rate	1×10^{-4}	2×10^{-4}
LoRA dropout rate	0.1	0.1
Channel sampling probability p	0.9	0.9

D.2 Experiments of LLaMA-3.2-3B-Instruct

In this section, we present an extended evaluation of fine-tuning LLaMA-3.2-3B-Instruct on the commonsense reasoning benchmark under both homogeneous and heterogeneous client ranks. The complete empirical results are summarized in Table D.3 and Table D.4.

Discussions on Homogeneous Client Ranks. As shown in Table D.3, CFLoRA achieves an overall average accuracy of 74.28%, outperforming all baseline methods including the strongest baseline FLoRA (72.83%) and RoLoRA (72.27%). Crucially, comparison with single-factor adaptation schemes reveals a fundamental empirical insight for large language models: methods that freeze one matrix factor (such as FFA-LoRA at 69.62% and AS-LoRA at 69.71%) suffer a severe performance drop of over 4.5 percentage points compared to CFLoRA.

Discussions on Heterogeneous Client Ranks. Table D.4 demonstrates that the advantages of CFLoRA are pronounced in the presence of client rank heterogeneity ($r_i \in \{4, 8, 16, 32\}$). CFLoRA delivers an average accuracy of 75.36%, outperforming FSLoRA by 2.16 points, FLoRA by 2.76 points, FlexLoRA by 3.77 points, and standard FedIT by 5.45 points.

Table D.3: Testing accuracy for fine-tuning the LLaMA-3.2-3B-Instruct on the commonsense reasoning tasks under *homogeneous* client ranks.

Method	**ARC-c**	**ARC-e**	**BoolQ**	**OBQA**	**PIQA**	**SIQA**	**Avg.**
FedIT	67.72 ± 0.45	84.48 ± 1.10	62.44 ± 0.27	74.53 ± 0.47	70.51 ± 2.39	66.47 ± 1.89	71.03
FFA-LoRA	67.45 ± 0.10	80.62 ± 0.29	64.17 ± 0.04	63.37 ± 1.93	70.39 ± 3.89	71.69 ± 0.54	69.62
RoLoRA	69.60 ± 1.93	**86.78** ± 0.12	64.18 ± 0.07	73.73 ± 2.45	69.61 ± 3.88	69.73 ± 2.27	72.27
AS-LoRA	67.53 ± 0.06	80.75 ± 0.26	64.12 ± 0.01	63.60 ± 1.70	70.30 ± 3.03	71.95 ± 0.69	69.71
FLoRA	70.65 ± 0.16	83.67 ± 0.03	**65.26** ± 1.10	72.07 ± 0.75	72.67 ± 1.34	72.67 ± 1.41	72.83
CFLoRA	**70.67** ± 1.39	85.55 ± 1.05	64.24 ± 0.11	**75.60** ± 0.90	**75.73** ± 1.27	**73.88** ± 0.58	**74.28**

Table D.4: Testing accuracy for fine-tuning the LLaMA-3.2-3B-Instruct on the commonsense reasoning benchmark under *heterogeneous* client ranks.

Method	**ARC-c**	**ARC-e**	**BoolQ**	**OBQA**	**PIQA**	**SIQA**	**Avg.**
FedIT	67.04 ± 0.32	80.13 ± 0.97	62.37 ± 0.81	71.33 ± 1.40	70.24 ± 2.13	68.32 ± 1.37	69.91
FLoRA	66.98 ± 1.20	84.18 ± 0.42	63.55 ± 1.94	**76.00** ± 0.57	74.27 ± 1.63	70.63 ± 1.88	72.60
FlexLoRA	67.45 ± 0.74	82.84 ± 1.03	62.35 ± 0.97	74.02 ± 0.55	73.31 ± 0.86	69.59 ± 1.64	71.59
FSLoRA	**69.88** ± 0.27	84.51 ± 0.49	64.21 ± 0.75	73.80 ± 0.92	74.96 ± 1.50	71.83 ± 1.69	73.20
CFLoRA	69.51 ± 0.12	**85.82** ± 0.77	**66.51** ± 0.35	75.40 ± 0.39	**78.73** ± 1.45	**76.20** ± 1.52	**75.36**

D.3 Further Experiments on Ablation Study on p

In this section, we provide additional ablation studies to further analyze the impact of the complementary channel selection probability p. In addition to Tables 3, we evaluate the performance of CFLoRA across varying values of p for RoBERTa on the GLUE benchmark under heterogeneous client ranks (Table 4), as well as for LLaMA-3.2-3B-Instruct on the commonsense reasoning benchmark under both homogeneous and heterogeneous client ranks (Table D.5 and D.6). Both sets of experiments are conducted under a partial-participation setting with $q = 0.4$ to simulate practical resource-constrained federated learning environments.

Table D.5: Ablation study on selection probability p for LLaMA-3.2-3B-Instruct on the commonsense reasoning benchmark under homogeneous client ranks.

Method	ARC-c	ARC-e	BoolQ	OBQA	PIQA	SIQA	Avg.
$p = 0.3$	65.98 ± 1.67	84.71 ± 0.44	62.54 ± 0.50	70.40 ± 0.57	72.33 ± 1.79	72.25 ± 2.07	71.37
$p = 0.5$	69.51 ± 0.96	84.87 ± 0.15	63.62 ± 0.67	74.63 ± 0.65	74.53 ± 1.48	73.26 ± 0.62	73.40
$p = 0.7$	68.65 ± 0.14	**85.64** ± 1.24	64.17 ± 0.13	75.09 ± 0.41	**75.74** ± 2.00	73.63 ± 1.31	73.82
$p = 0.9$	**70.67** ± 1.39	85.55 ± 1.05	64.24 ± 0.11	**75.60** ± 0.90	75.73 ± 1.27	**73.88** ± 0.58	**74.28**
$p = 1.0$	67.56 ± 0.13	80.54 ± 0.24	**64.31** ± 0.08	64.25 ± 1.39	70.15 ± 2.44	71.63 ± 0.73	69.74

Table D.6: Ablation study on selection probability p for LLaMA-3.2-3B-Instruct on the commonsense reasoning benchmark under heterogeneous client ranks.

Method	ARC-c	ARC-e	BoolQ	OBQA	PIQA	SIQA	Avg.
$p = 0.3$	68.81 ± 0.16	82.78 ± 0.76	62.25 ± 0.11	74.97 ± 0.61	71.47 ± 3.24	73.44 ± 0.20	72.29
$p = 0.5$	68.89 ± 0.45	83.76 ± 0.71	65.25 ± 0.44	75.50 ± 1.13	76.93 ± 0.46	75.61 ± 0.07	74.32
$p = 0.7$	69.10 ± 0.60	84.53 ± 0.66	66.18 ± 0.50	**75.82** ± 0.57	**78.88** ± 4.90	76.01 ± 0.17	75.09
$p = 0.9$	**69.51** ± 0.12	**85.82** ± 0.77	**66.51** ± 0.35	75.40 ± 0.39	78.73 ± 1.45	**76.20** ± 1.52	**75.36**
$p = 1.0$	69.13 ± 0.38	83.65 ± 0.73	62.34 ± 0.23	67.73 ± 0.94	73.33 ± 0.63	74.12 ± 0.34	71.72

Asymmetric Allocation. Theoretically, the convergence upper bounds established in Theorem 3 and Corollary 4 depend inversely on $p_{\min} = \min\{p, 1-p\}$ via the parameter $\kappa = p_{\min}^{-1}$. Under worst-case analysis where both factors are treated symmetrically, the theoretical bound is minimized at $p = 0.5$. However, our empirical findings reveal a distinct asymmetric behavior during actual federated optimization:

- **Asymmetric Importance of Factor $\mathbf{B}$:** Increasing p from 0.3 to 0.9 consistently improves downstream performance across all benchmark tasks. For instance, on RoBERTa under homogeneous ranks (Table 3), the average score rises from 74.72 at $p = 0.3$ to 78.70 at $p = 0.5$, peaking at 79.73 when $p = 0.9$. A similar steady gain is observed in the heterogeneous setting (Table 4), rising from 73.37 ($p = 0.3$) to 79.57 ($p = 0.9$). This behavior aligns closely with findings in centralized PEFT literature (Hayou et al., 2024), which demonstrate that the two low-rank factor matrices contribute asymmetrically to model representation updates: $\mathbf{B}$ typically requires larger effective updates or faster gradient movement to steer model representations, whereas $\mathbf{A}$ acts primarily as a random subspace projection. Prioritizing updates to $\mathbf{B}$ ($p = 0.9$) thus accelerates optimization while preserving sufficient exploration along the latent channels.
- **Penalty for Starving $\mathbf{B}$ ($p = 0.3$):** When p is decreased to 0.3 (meaning the majority of active channels update $\mathbf{A}$ instead of $\mathbf{B}$), we observe a pronounced drop of 5.01 points on RoBERTa homogeneous and 6.20 points on heterogeneous ranks. This confirms that allocating insufficient updates to $\mathbf{B}$ acts as a bottleneck for parameter adaptation.

Hyperparameter Robustness Across $p \in [0.5, 0.9]$. Despite the empirical preference for higher p, CFLoRA exhibits robustness across the wide range. On RoBERTa (both homogeneous and heterogeneous), variations across $p \in \{0.5, 0.7, 0.9\}$ remain within a tight margin of approximately 1.0 point on average (e.g., $78.70 \to 78.84 \to 79.73$ in Table 3, and $78.74 \to 79.45 \to 79.57$ in Table 4). A similar plateau is evident on LLaMA-3.2-3B-Instruct in Table D.5, where tasks such as OBQA,

PIQA, and SIQA perform consistently well from $p = 0.5$ onward. This hyperparameter insensitivity is of substantial practical value in federated learning: practitioners do not need exhaustive tuning to identify a value of p.

Catastrophic Degradation at Single-Factor Boundary ($p = 1.0$). Crucially, when p reaches 1.0, which entirely eliminates updates to $\mathbf{A}$, performance experiences a sharp drop across different settings. On RoBERTa under homogeneous ranks, the overall average drops sharply by 4.39 points (from 79.73 at $p = 0.9$ down to 75.34 at $p = 1.0$), with severe regressions across individual tasks such as CoLA (-5.06) and MNLI-M (-6.80). The degradation is even more pronounced when fine-tuning LLaMA-3.2-3B-Instruct on reasoning tasks (Table D.5). Comparing $p = 0.9$ to $p = 1.0$, accuracy drops 4.54% on average. These empirical observations directly corroborate our core theoretical insights: freezing one low-rank projection matrix permanently severely curtails the adapter's expressive capacity.

D.4 Experiments with Varying Client Participation

We evaluate the effect of client participation by varying the participation rate q from 0.2 to 1.0. Tables D.7 and D.8 report test accuracy on the QNLI dataset under homogeneous and heterogeneous client ranks, respectively. Under homogeneous ranks, `CFLoRA` achieves the highest average accuracy, exceeding the strongest baseline by 1.38 points. Under heterogeneous ranks, it improves the average over the strongest baseline by 1.52 points. We further evaluate full participation across six GLUE tasks under homogeneous ranks in Table D.9, `CFLoRA` obtains the highest average score of 84.82, surpassing the strongest baseline (FLoRA) by 2.26 points. These results indicate that its performance advantage extends beyond the partial-participation setting used in the main experiments.

Table D.7: Testing accuracy for fine-tuning the RoBERTa model on the QNLI dataset under homogeneous client ranks with varying client participation.

Method	$q = 0.2$	$q = 0.4$	$q = 0.6$	$q = 0.8$	$q = 1.0$	**Avg.**
FedIT	85.01 $\pm$ 0.23	86.61 $\pm$ 0.26	88.90 $\pm$ 0.26	93.06 $\pm$ 0.90	90.02 $\pm$ 0.46	88.72
FFA-LoRA	86.13 $\pm$ 0.35	88.49 $\pm$ 0.23	90.94 $\pm$ 0.33	**94.09** $\pm$ 0.32	91.42 $\pm$ 0.19	90.21
RoLoRA	86.52 $\pm$ 0.27	87.60 $\pm$ 2.21	92.60 $\pm$ 0.46	93.23 $\pm$ 0.14	93.10 $\pm$ 0.46	90.61
AS-LoRA	87.13 $\pm$ 0.35	88.26 $\pm$ 0.67	91.28 $\pm$ 0.38	92.71 $\pm$ 0.17	92.46 $\pm$ 1.03	90.37
FLoRA	88.83 $\pm$ 0.24	89.27 $\pm$ 0.20	90.69 $\pm$ 0.41	92.93 $\pm$ 0.44	91.41 $\pm$ 0.35	90.63
CFLoRA	**88.95** $\pm$ 0.06	**90.61** $\pm$ 0.25	**92.98** $\pm$ 0.31	93.59 $\pm$ 0.06	**93.93** $\pm$ 0.34	**92.01**

Table D.8: Testing accuracy for fine-tuning the RoBERTa model on the QNLI dataset under heterogeneous client ranks with varying client participation.

Method	$q = 0.2$	$q = 0.4$	$q = 0.6$	$q = 0.8$	$q = 1.0$	**Avg.**
FedIT	68.01 $\pm$ 4.58	70.34 $\pm$ 2.01	85.89 $\pm$ 0.37	90.03 $\pm$ 3.26	88.22 $\pm$ 2.36	80.50
FLoRA	70.97 $\pm$ 4.32	87.35 $\pm$ 1.84	87.75 $\pm$ 3.21	92.67 $\pm$ 0.72	91.77 $\pm$ 2.35	86.10
FlexLoRA	**78.41** $\pm$ 5.10	86.66 $\pm$ 1.55	86.81 $\pm$ 1.36	93.11 $\pm$ 1.25	**94.02** $\pm$ 0.95	87.80
FSLoRA	73.68 $\pm$ 4.01	86.68 $\pm$ 2.27	86.92 $\pm$ 1.48	89.40 $\pm$ 2.45	88.54 $\pm$ 1.60	85.04
CFLoRA	76.84 $\pm$ 3.73	**91.29** $\pm$ 0.45	**91.46** $\pm$ 2.10	**93.60** $\pm$ 0.38	93.39 $\pm$ 0.91	**89.32**

Table D.9: Testing accuracy for fine-tuning the RoBERTa model on the GLUE benchmark under homogeneous client ranks with full participation.

Method	QNLI	MRPC	CoLA	MNLI-M	MNLI-MM	QQP	Avg.
FedIT	90.02 ± 0.46	70.30 ± 0.39	75.16 ± 0.15	82.40 ± 0.44	82.19 ± 0.33	80.98 ± 1.05	80.18
FFA-LoRA	91.42 ± 0.19	73.20 ± 0.42	74.16 ± 0.45	84.79 ± 0.16	84.78 ± 0.16	81.89 ± 1.05	81.71
RoLoRA	93.10 ± 0.46	73.86 ± 1.26	73.93 ± 0.29	85.76 ± 0.43	85.84 ± 0.53	81.25 ± 0.38	82.29
AS-LoRA	92.46 ± 1.03	72.44 ± 0.67	75.04 ± 0.94	82.79 ± 0.33	**86.76** ± 0.35	80.95 ± 1.67	81.74
FLoRA	91.41 ± 0.35	76.34 ± 0.80	**77.26** ± 0.25	83.66 ± 0.77	83.88 ± 0.66	82.79 ± 0.32	82.56
CFLoRA	**93.93** ± 0.34	**80.83** ± 0.85	76.74 ± 0.46	**86.54** ± 0.40	86.71 ± 0.26	**84.15** ± 0.53	**84.82**