Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

FediLoRA: Practical Federated Fine-Tuning of Foundation Models Under Missing-Modality Constraints

Lishan Yang[1], **Wei Emma Zhang**[1], **Nam Kha Nguyen**[1], **Po Hu**[2], **Yanjun Shu**[3], **Weitong Chen**[1] and **Sim Mong Yuan**[1]
[1]Adelaide University
[2]Central China Normal University
[3]Harbin Institute of Technology
{lishan.yang, wei.e.zhang}@adelaide.edu.au

Abstract

Federated Learning with LoRA fine-tuning offers an efficient and privacy-aware solution for institutions to collaboratively leverage their large datasets to train VLLMs. However, participating institutions often possess heterogeneous computational resources, resulting in imbalanced LoRA ranks, which pose a major challenge for effective collaboration. In addition, real-world applications in domains such as healthcare and transportation frequently suffer from missing modalities due to user mistakes or device failures, which significantly degrade global model performance in federated settings. To the best of our knowledge, no prior work has addressed these two challenges simultaneously in federated VLLMs. To tackle these issues, we propose **FediLoRA**, a lightweight federated LoRA aggregation framework that effectively mitigates the impact of missing modalities in heterogeneous environment. FediLoRA is explicitly motivated by the observation that simple averaging and structured editing can jointly benefit both global and personalized models. Our approach achieves strong performance across multiple general-domain and medical-domain benchmark datasets. Additional experiments on healthcare data further demonstrate that FediLoRA is well-suited for practical, real-world deployment scenarios. Our code is released at https://github.com/gotobcn8/FediLoRA.

1 Introduction

Foundation Models (FMs) have demonstrated remarkable capacity to absorb such large-scale, heterogeneous data and exhibit extraordinary generalization abilities. Consequently, FMs have been deployed in various applications, including education [Wen *et al.*, 2024], finance [De Zarzà *et al.*, 2023], healthcare [Wang *et al.*, 2022; Chu *et al.*, 2025] and so on. Reliable decision-making systems generally depend on multimodal inputs, for instance, medical applications may combine diagnostic text reports with CT or X-ray images. However, constructing powerful real-world multimodal models faces two major challenges due to strict privacy regulations.

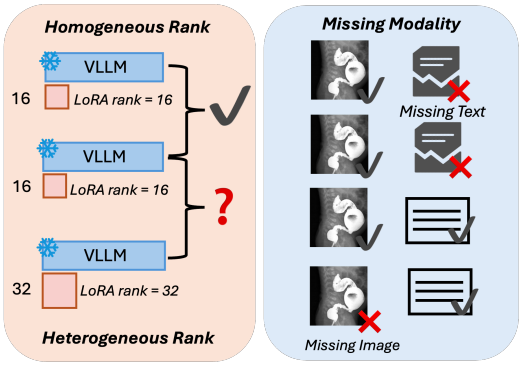


Figure 1: Two challenges for current Federated VLLMs. (1) Left: Inconsistencies in clients' local model configuration (heterogeneous rank), causing unstable and suboptimal aggregation. (2) Right: Missing modalities in different clients leads to poor multimodal representation learning and generalization.

First, institution data are inherently isolated across hospitals, which possess distinct patient populations and private data distributions. Federated Learning (FL) [McMahan *et al.*, 2017; Yang *et al.*, 2024a] provides a promising solution by enabling collaborative training across institutions without sharing raw data. Due to differences in local training environments, model inconsistency is commonly present in federated learning architectures. Second, multimodal data in the real world are often incomplete; certain modalities may be missing due to equipment constraints, inconsistent collection protocols, or human oversight [Cismondi *et al.*, 2013]. Such modality incompleteness can substantially degrade the performance and reliability of multimodal models, posing serious risks to the users, especially in clinical applications.

The heterogeneity of LoRA in VLLMs. Federated Learning consists of a global model and multiple clients (e.g., hospitals). Fine-tuning personalized Vision–Large Language Models becomes increasingly popular and stable in the specific domains [Savage *et al.*, 2025; Wang *et al.*, 2023], clients often have many fine-tuning options to choose from when adapting models to their own data and computing resources. Parameter-Efficient Fine-Tuning (PEFT) methods such as adapters [Houlsby *et al.*, 2019; Sun *et al.*, 2024a; Wu *et al.*, 2024], prefix-tuning [Li and Liang, 2021; Liu *et al.*,

2022], and Low-Rank Adaptation (LoRA) [Hu *et al.*, 2022a] have received growing attention. LoRA introduces trainable low-rank matrices to approximate gradient updates while keeping pretrained weights frozen. In practice, clients may choose different LoRA configurations to better fit their personalized data, training resources, or efficiency requirements. This naturally leads to heterogeneous LoRA settings, making it difficult for Multimodal FL to aggregate these structurally different client models. Existing works [Zhang *et al.*, 2024; Liu *et al.*, 2025; Kuang *et al.*, 2024; Sun *et al.*, 2024b; Yang *et al.*, 2024b] do not consider heterogeneous ranks. Although recent methods such as HetLoRA [Cho *et al.*, 2024], FLoRA [Wang *et al.*, 2024a], and FlexLoRA [Bai *et al.*, 2024] explore heterogeneous LoRA, they still lack effective mechanisms for handling our next challenge in multimodal data: missing modalities.

Missing Modality in Federated Learning. In real-world settings, missing modalities arise easily and are often unavoidable. For example, hospitals possess multimodal patient data, yet certain modalities are often missing due to privacy constraints, equipment limitations, or data collection issues. Traditional multimodal FL methods mainly focus on model aggregation or label distribution [Zong *et al.*, 2021; Chen and Li, 2022; Zeng *et al.*, 2024], but they do not address challenges for distributed missing modality. Prior approaches such as modality reconstruction [Xiong *et al.*, 2023], contrastive learning [Yu *et al.*, 2023], and prototype exchange [Bao *et al.*, 2024] attempt to mitigate missing modalities. However, when applied to foundation models, these methods often incur substantial computational overhead. As a result, they are impractical for real-world application. Therefore, there is a clear need for a lightweight yet effective solution to handle missing modalities in federated multimodal settings.

In summary, *the research effort of addressing missing modality in heterogeneous Federated LoRA, a common scenario in the era of large foundation models, is missing.* To this end, we propose **FediLoRA** to fill this gap. Our method is motivated by the observation that the global model achieves comparable performance when trained on either full-modality or missing-modality settings (Section 2). We attribute this robustness to the averaging effect of FedAvg aggregation [Collins *et al.*, 2022], the de facto standard aggregation method in federated learning, which aggregates client updates using a weighted average based on local data size. To extend its benefit to heterogeneous settings, we retain the core idea of FedAvg and adapt it to fit the heterogeneous LoRA and propose a dimension-wise reweighting mechanism.

Despite the global model's stability, we observe that client models suffer from a notable performance gap, particularly under missing-modality scenarios. This is due to both local data sparsity and FedAvg's emphasis on optimizing global performance. Given our success in handling missing modalities on the global model, we further explore whether aligning local models to the global model could improve both global and clients' performance. This leads us to adopt a simple yet effective solution: repairing local LoRA layers using their global counterparts. This layer-wise model editing strategy efficiently transfers useful global knowledge to clients without introducing additional training overhead (*Appendix D*). Our contributions are summarized as follows:

- We introduce a layer-wise model editing technique that leverages global LoRA parameters to correct local components, thereby enhancing client model performance without incurring additional computational cost, and we provide theoretical guarantees for its effectiveness.
- We propose a dimension-wise weighted aggregation and strategy that enables effective heterogeneous federated LoRA while preserving model robustness to missing modalities.
- We conduct comprehensive experiments on three VLLM benchmarks to validate our approach. In addition, we further demonstrate its capability on one medical-domain dataset.

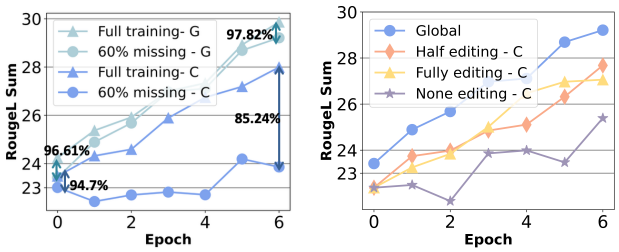
2 Motivation

Averaging aggregation mitigates the effect of missing modality. While the global model and client models often exhibit significant performance gaps when using FedAvg (or its variants), due to FedAvg's focus on global optimization, this observation motivates us to investigate whether averaging may help mitigate the adverse effects of missing modalities on the global server. Specifically, we hypothesize that the noise introduced by missing modalities can be diluted through the averaging nature of FedAvg aggregation.

To test this hypothesis, we conduct a preliminary experiment in the Federated LoRA setting. The evaluation is carried out under a **homogeneous-rank** configuration to align with the standard FedAvg aggregation strategy, following the setup of FedIT [Zhang *et al.*, 2024]. Experiments are performed on the image captioning dataset SAM-LLaVA [Chen *et al.*, 2023], considering both full-modality and missing-modality (60% missing) training scenarios. Figure 2a presents the global model's performance under this setting. We observe that while the global model initially shows instability in the missing-modality setting, it gradually recovers and approaches the full-modality performance after sufficient communication rounds. However, under identical conditions, the evaluation of client performance with missing versus complete modalities revealed pronounced percentage discrepancies (97% vs 85%), while this stands in contrast to their initial performance differences were relatively small and both exceeded 95%. The averaging effect of FedAvg probably contributes to this stabilization - the averaging process inherently reinforces common patterns shared across clients while smoothing out individual client-specific noise. [Collins *et al.*, 2022]. Consequently, even if some clients lack specific modalities or produce noisier updates, the aggregation process effectively pulls the global model back toward a stable, generalizable representation.

Simple model editing is effective. Inspired by PFedEdit [Yuan *et al.*, 2024][1], we test the effectiveness of the simple model editing to preserve the model's robustness.

[1]PFedEdit aims at personalization (i.e., focus on client performances) and its iterative method is on small deep learning models, which is computationally expensive for large foundation models.



(a) Overall performance between full and 60% missing modality training.

(b) Client (C) performance comparison among different editing strategies.

Figure 2: (a) The performance gap among global models and personalized models. The bidirectional arrow represents the percentage of the smaller value relative to the larger one. (b) The experiment investigates the impact of different LoRA editing strategies on local client performance. The blue curve represents the global performance used as a reference while there is a huge gap compared with None editing. G: Global, C: Client.

We first identify the layer with the lowest cosine similarity between the global and local models. Based on this, we apply two simple editing strategies: ***Half-editing***, where the selected client layer parameters are updated as a weighted average of the last round global layer parameter of the current local layer, and ***Full-editing***, where the entire local layer is replaced by the corresponding global layer. Figure 2b shows the edited client model performances under 60% missing modality training compared to the global model performance. We computed the weighted average of each client's local performance to obtain the client's performance in this evaluation.

It is evident that missing modalities significantly degrade the performance of client models and the editing could help improve the client performance. This result demonstrates that direct model editing, i.e., replacing the degraded client parameters with their stable global counterparts, can effectively compensate for modality sparsity at the client level. By selectively overriding parts of local models (e.g., LoRA adapters) with globally aggregated parameters, shared capabilities can be recovered without full retraining.

Motivated by these observations, we propose **FediLoRA**, a federated LoRA editing framework that combines dimension-wise averaging aggregation. Our approach aims to retain the inherent strength of FedAvg in mitigating the adverse effects of missing modalities and extends its applicability to heterogeneous LoRA ranks by reorganizing rank-specific dimension weights during aggregation. In addition, we introduce a lightweight editing strategy that selectively replaces degraded local parameters with their global counterparts at the layer level. This correction enables clients to better recover from modality sparsity without full retraining.

3 FediLoRA

In federated LoRA learning, the central server aggregates the LoRA updates submitted by participating clients. For a given LoRA layer with update matrix $\Delta W = B^T A$, collected from K clients, the standard homogeneous aggregation strategy adopted in FedIT [Zhang *et al.*, 2024] computes the global low-rank matrix as $A_{global} = \sum_{k=0}^{K} p_k A_k$, where p_k denotes the proportion of data owned by client k relative to the total data across all clients.

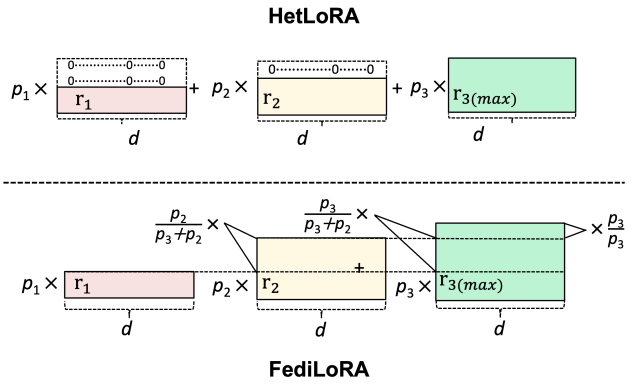


Figure 3: Visualization of the Aggregation Process for HetLoRA and Dimension-wise re-weighting of FediLoRA. $p_i = \frac{|\mathcal{D}_k|}{\sum_{j=1}^{K} |\mathcal{D}_j|}$ denotes the weights in FedAvg.

3.1 Dimension-wise Reweighting

To handle heterogeneous LoRA ranks, HetLoRA [Cho *et al.*, 2024] zero-pads each client's LoRA matrices to match the highest rank across all clients (also the same as the global rank). However, this zero-padding can significantly dilute the contributions of clients with higher ranks during the averaging step in FedAvg-based aggregation. Figure 3 illustrates the difference between our aggregation method and that of HetLoRA. In the following section, we present the detailed formulation of our aggregation process.

To mitigate information dilution in the FedAvg setting, we propose a *dimension-wise weighted aggregation* which excludes zero-padded dimensions and reweights non-zero ones proportionally based on the client's rank and data size.

$A_k \in \mathbb{R}^{r_k \times n}$ denotes the local LoRA-A matrix of client k. The global rank is r_g=max$\{r_1, r_2, \ldots, r_K\}$. A_k is zero-padded to size $r_g \times n$. Specifically, for each row index $d \in \{1, 2, \ldots, r_g\}$, we define a binary mask:

$$\text{mask}_k^{(d)} = \begin{cases} 1 & \text{if } d \le r_k \\ 0 & \text{if } r_k < d \le r_g \end{cases} \tag{1}$$

Given each client's aggregation weight p_k (e.g., based on local data size as introduced in Section FedAvg [McMahan *et al.*, 2017]), the normalized dimension-wise weight for client k at dimension d is:

$$\tilde{p}_k^{(d)} = \frac{\text{mask}_k^{(d)} \cdot p_k}{\sum_{j=1}^{K} \text{mask}_j^{(d)} \cdot p_j} \tag{2}$$

Using this reweighting, the aggregated global LoRA-A matrix $A_g \in \mathbb{R}^{r_g \times n}$ is computed row-wise as:

$$A_g^{(d,:)} = \sum_{k=1}^{K} \tilde{p}_k^{(d)} \cdot A_k^{(d,:)} \tag{3}$$

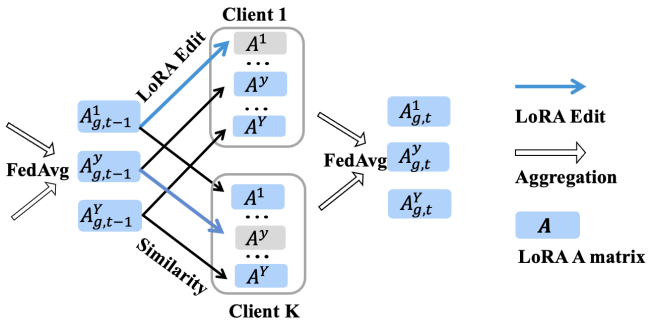


Figure 4: Layer-wise model editing. $A^y_{g,t-1}$ represents the $t-1$-th epoch, y-th global LoRA A matrix. LoRA-Editing happens at each ending of local fine-tuning and before aggregation.

The same strategy applies to the aggregation of the LoRA-B matrices. This method ensures that each dimension is aggregated only across the clients that contribute meaningful (non-zero) information to that dimension, while still respecting their overall weight in the federated system.

3.2 Layer-wise LoRA Editing

As shown in Figure 2, local clients are more adversely affected by missing modalities compared to the global model. Existing federated learning approaches typically address this issue by either reconstructing the missing modalities or employing contrastive learning techniques (Section 5.2). However, these solutions can be computationally expensive, particularly when applied to large foundation models (Refer to *Appendix D.2* for computational analysis). This motivates us to explore alternative strategies that are both computationally efficient and capable of preserving the performance of both global and local client models under missing modality conditions.

Building on our preliminary findings in Section 2, which demonstrate the effectiveness of simple model editing in addressing missing modalities, we propose a lightweight editing strategy that identifies and incorporates informative parameters from the global model into local models based on parameter similarity. Specifically, for a local model θ_t that has been fine-tuned, there are Y LoRA layers in total that have been updated, each comprising a pair of low-rank matrices A_k and B_k. To assess which layer is most affected by the missing modality, we compute the cosine similarity γ between the t-th round LoRA parameters of the local model and the corresponding global LoRA parameters from round $t-1$. We take the A matrix as an illustrative example. For each LoRA layer $y \in \{1, 2, \ldots, Y\}$ in the local model, we compute the cosine similarity between the local and global A matrices as follows:

$$\gamma_y = \frac{\langle A^y_{k,t}, A^y_{g,t-1} \rangle}{\left\| A^y_{k,t} \right\| \left\| A^y_{g,t-1} \right\|} \tag{4}$$

We then identify the layer with the lowest similarity score:

$$y^* = \arg\min_y \{\gamma_1, \gamma_2, \ldots, \gamma_Y\} \tag{5}$$

To edit this least-similar LoRA layer, we apply an interpolation that softly blends the local and global updates:

$$A^{y^*}_{k,t} \leftarrow \gamma_{y^*} A^{y^*}_{k,t} + (1 - \gamma_{y^*}) A^{y^*}_{g,t-1} \tag{6}$$

The whole federated editing process can be shown in Figure 4. As suggested by prior work [Guo *et al.*, 2024] and further evidenced by our empirical analysis in Section 4.2, the local LoRA matrix A tends to retain more information from the global model, whereas the matrix B captures more client-specific, personalized features. Based on this observation, to align the overall LoRA module with the global model while reducing computational complexity, we propose to compute similarity based solely on the A matrix. Under the consideration of minimal resource consumption, editing the fewest number of layers can achieve excellent, or even the best, performance (refer to *Appendix C* for details).

3.3 Theoretical Analysis

In this section, we conduct an analysis for the layer-wise LoRA Editing in FediLoRA. The whole proof details are in the *Appendix A*. According to the common **Assumptions (1-3)** from [Li *et al.*, 2019; Karimireddy *et al.*, 2020; Reddi *et al.*, 2021], shown in *Appendix A*.

Assumption 4 (Client Missing Modality Heterogeneity)**.** *For all* w, $\frac{1}{m}\sum_{i=1}^m \|\nabla F_i(w) - \nabla f(w)\|^2 \le \sigma_g^2$.

Assumption 5 (Block Hessian (layer-coupling) bound)**.** *There exists a nonnegative matrix* $H \in \mathbb{R}_+^{L\times L}$ *such that for all* w *and all layer indices* ℓ, r, $\left\|\nabla^2_{\ell r} f(w)\right\|_{\text{op}} \le H_{\ell r}$.

Lemma 1 (Exact layer-wise shrinkage induced by FediLoRA Editing)**.** *Fix client* i *and round* t*, and let* $\ell^* := \ell_i^t$, $s := \gamma_i^t \in [0,1]$ *as the finetuning parameters remain relatively stable within a limited range [Hu* et al.*, 2022b]:* $\|\Delta_i^t\|^2 - \left\|\tilde{\Delta}_i^t\right\|^2 = (1-s^2)\left\|\Delta^t_{i,\ell^\star}\right\|^2$.

Lemma 2 (Cross-layer smoothness inequality)**.** *Under Assumption 5, for any* w *and any block perturbation* $\Delta = (\Delta_1, ..., \Delta_L)$, $f(w+\Delta) \le f(w) + \sum_{\ell=1}^L \langle \nabla_\ell f(w), \Delta_\ell \rangle + \frac{1}{2}\sum_{\ell=1}^L \sum_{r=1}^L H_{\ell r} \|\Delta_\ell\| \|\Delta_r\|$.

Lemma 3 (Editing reduces coupling penalty)**.** *Define the quadratic coupling functional* $Q(\Delta) := \frac{1}{2}\sum_{\ell=1}^L \sum_{r=1}^L H_{\ell r} \|\Delta_\ell\| \|\Delta_r\|$*, then we have:*

$$Q(\tilde{\Delta}_i^t) \le Q(\Delta_i^t) - \frac{1}{2}(1-s^2) H_{\ell^\star\ell^\star} \left\|\Delta^t_{i,\ell^\star}\right\|^2 - (1-s)\sum_{r\neq\ell^\star} H_{\ell^\star r} \left\|\Delta^t_{i,\ell^\star}\right\| \left\|\Delta^t_{i,r}\right\|.$$

Theorem 1 (FediLoRA convergence with explicit coupling)**.** *Gather Assumptions 1–5 and let* $f^\star := \inf_w f(w) > -\infty$*. Choose client stepsize* η_l *sufficiently small (e.g.* $\eta_l = \Theta(1/(LK\sqrt{T}))$*). Then after* T *rounds,*

$$\min_{0\le t\le T-1} \mathbb{E}\left\|\nabla f(w^t)\right\|^2 \le \mathcal{O}\left(\frac{f(w^0) - f^\star}{K\eta_l T}\right) + \mathcal{O}\left(\eta_l(\sigma_l^2 + K\sigma_g^2)\right) - \mathcal{O}\left(\frac{1}{T}\sum_{t=0}^{T-1} \mathrm{CR}^t\right),$$

where the coupling reduction *term* $\mathrm{CR}^t \geq 0$ *aggregates the per-client decrease in coupling penalty induced by the FediLoRA edit:*

$$\mathrm{CR}^t := \frac{1}{m}\sum_{i=1}^{m}\Bigg[\frac{1}{2}(1-(\gamma_i^t)^2)\, H_{\ell_i^t\ell_i^t}\,\mathbb{E}\left\|\Delta_{i,\ell_i^t}^t\right\|^2 + (1-\gamma_i^t)\sum_{r\neq\ell_i^t} H_{\ell_i^t r}\,\mathbb{E}\left\|\Delta_{i,\ell_i^t}^t\right\|\,\mathbb{E}\left\|\Delta_{i,r}^t\right\|\Bigg].$$

In particular, since $\mathrm{CR}^t \geq 0$*, FedILoRA preserves the standard* $O(1/\sqrt{T})$ *stationarity rate, and the edit step can only improve the bound through explicit reduction of cross-layer coupling terms.*

4 Experiment

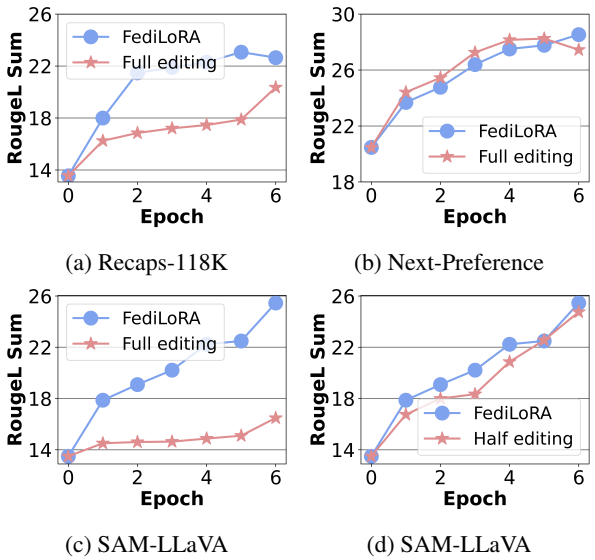
Dataset. Our evaluation spans three multimodal benchmarks: Recaps-118K [LMMs-Lab, 2024], a dialogue summarization dataset; SAM-LLaVA [Chen *et al.*, 2023], a multimodal image-text captioning benchmark; and Next-Preference [Mishra, 2024], a VQA task. We also conduct experiments on the Medical VQA dataset Slake [Liu *et al.*, 2021] using the same experimental setup in Section 4.5). The dataset contains over 14,000 pairs of knowledge-based and vision-grounded QA. Each dataset is partitioned into 11 subsets with randomly assigned sizes (one for global test set), and each subset is mutually exclusive and inaccessible to the others. We further split each subset into training and testing sets with ratio of 8:2 for each client. Following FedMultimodal [Feng *et al.*, 2023], we generate a certain sample of missing data for each dataset, simulating a certain missing modality scenario where text inputs are set to None or image inputs are zeros (corresponding input shape).

Evaluation Metrics. We evaluate our proposed framework using natural language generation evaluation metrics. We use ROUGE-LSum [Lin, 2004] (RSUM) to compute the longest common subsequence overlap between generated text and reference text, and BLEU score [Papineni *et al.*, 2002] to evaluate token-level generation accuracy.

Baselines. We employ the pre-trained LLaVA-1.5-7B model [Liu *et al.*, 2023] as the large multimodal foundation model in our benchmarks. We compare FediLoRA with two SOTA FL algorithms that support heterogeneous LoRA: (1) HetLoRA [Cho *et al.*, 2024]: a method that aggregates local updates through a sparsity-weighted scheme and demonstrates superior efficiency and performance. (2) FLoRA [Wang *et al.*, 2024b]: a method that aggregates LoRA parameters via a stacking strategy. We follow client and learning rate settings from [Wang *et al.*, 2024b]. All comparative baselines are trained under the same federated setting to ensure fair evaluation. All experiments are conducted on four NVIDIA RTX 4090 GPUs.

4.1 Main Results

We compare the global and personalized performances under the missing modality setting. As shown in Table 1, FediLoRA outperforms the baselines HetLoRA and FLoRA across the considered datasets and missing ratios.



(a) Recaps-118K (b) Next-Preference

(c) SAM-LLaVA (d) SAM-LLaVA

Figure 5: Comparing editing methods. (a), (b) and (c) present the performance comparison between full-editing and FediLoRA on three datasets respectively. (d) shows the comparison of half-editing and FediLoRA.

In the *global evaluation*, FediLoRA achieves the highest performance across all settings, except for a slight lag in the 30% missing SAM-LLaVA setting. For instance, on the Recaps-118K dataset with 30% **missing ratio (mr)**, FediLoRA achieves the highest BLEU and RSUM scores, outperforming FLoRA by +2.08 BLEU and +2.93 RSUM. Under the 60% mr setting, FediLoRA (BLEU 15.53, RSUM 25.47) still maintains the best performance among all methods. On the SAM-LLaVA dataset, where HetLoRA performance drops significantly under high mr, FediLoRA demonstrates strong robustness. It improves over FLoRA by at least +0.18 BLEU and +0.47 RSUM across the mrs. The largest absolute advantage is observed on the Next-Preference dataset, where FediLoRA surpasses FLoRA by +1.25 BLEU and +0.82 RSUM under 30% mr, and maintains a pass over 1.6 in two metrics lead under 60% mr.

In the *personalized evaluation* (i.e., client performance evaluation), FediLoRA remains competitive under both missing conditions. For example, on Recaps-118K with 30% mr, it achieves +2.28 BLEU and +0.8 RSUM over the best baseline. On SAM-LLaVA, it leads with up to +1.39 BLEU and +0.33 RSUM under 30% mr, and maintains a consistent advantage under 60% mr. Although FediLoRA performs best in most cases, we note that on the Next-Preference dataset with 60% mr, FLoRA achieves a slightly higher RSUM score (29.48 vs. 28.64), indicating its potential advantage in certain personalized settings. Across all datasets and conditions, FediLoRA achieves the most stable and highest overall performance.

4.2 Editing Different LoRA Matrices

This experiment investigates the impact of editing different LoRA components, specifically the A matrix, B matrix, both, or none, and the results are summarized in Table 2. The miss-

Global Evaluation	**Recaps-118K ↑**				**SAM-LLaVA ↑**				**Next-preference ↑**			
	Missing(30%)		Missing(60%)		Missing(30%)		Missing(60%)		Missing(30%)		Missing(60%)	
	BLEU	RSUM	BLEU	RSUM	BLEU	RSUM	BLEU	RSUM	BLEU	RSUM	BLEU	RSUM
HetLoRA	9.97	21.60	15.36	25.14	13.16	**25.39**	2.01	7.37	10.29	32.96	2.33	9.34
FLoRA	9.21	30.91	13.81	25.13	13.16	24.57	12.41	24.47	9.96	32.70	10.13	28.40
FediLoRA	**11.29**	**33.84**	**15.53**	**25.47**	**13.34**	25.15	**12.64**	**24.94**	**11.21**	**33.52**	**11.85**	**30.07**
Personalized Evaluation												
HetLoRA	8.74	19.98	9.22	15.08	12.91	24.91	2.69	6.31	11.15	29.94	2.02	7.86
FLoRA	12.67	**35.44**	14.33	24.79	13.24	25.18	9.89	22.10	10.30	24.69	9.92	**29.48**
FediLoRA	**14.95**	35.08	**14.86**	**25.59**	**13.74**	**25.51**	**11.28**	**22.33**	**13.86**	**33.65**	**10.52**	28.64

Table 1: Global and personalized performance comparisons. **Bold** font indicates the best performance. Higher values mean better results.

ing modality ratio is fixed at 60%.

Among all configurations, editing only the LoRA-A matrix consistently yields the best performance across all datasets, suggesting that the A matrix plays a more critical role in preserving global knowledge [Guo *et al.*, 2024]. Notably, editing both A and B matrices offers no additional benefit and, in some cases, even leads to performance degradation. This is likely due to the over-editing or disruption of client-specific adaptations. While FediLoRA modifies the LoRA B matrix, it still shows behavior similar to Both editing, and in the SAM-LLaVA dataset it even performs worse than None editing. As expected, models without any editing perform the worst in most scenarios, except on the SAM-LLaVA dataset. The experiment results highlight the effectiveness of the model editing strategy of FediLoRA in mitigating degradation caused by missing modalities. These results demonstrate that FediLoRA attains the strongest performance even with editing restricted to the LoRA-A module.

Edited Matrix	**Recaps-118K ↑**	**SAM-LLaVA ↑**	**Next-Preference ↑**
LoRA-A	**30.91**	**21.77**	**28.24**
LoRA-B	29.03	18.06	26.84
Both	30.22	16.94	26.11
None	28.45	19.87	24.68

Table 2: Comparison of editing different LoRA matrices in three datasets. We set the missing ratio as 60%. All results are evaluated at the global model using the RSUM metric.

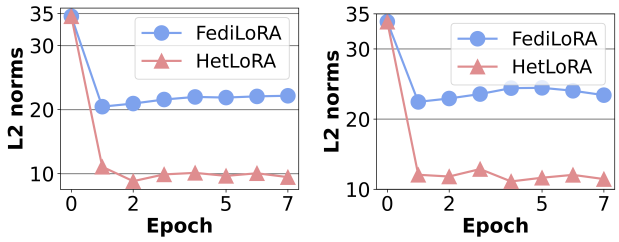
4.3 Full-Parameter Editing vs FediLoRA Editing

Figure 5 presents the personalized performance comparison between FediLoRA and full-layer editing under a 60% missing ratio on three datasets. In FediLoRA, let γ denote the similarity between the local LoRA and the global LoRA. Full editing corresponds to setting $\gamma = 0$, while half editing sets $\gamma = 0.5$. Unlike the experiments in the Section 2, this experiment evaluates the personalized performance under heterogeneous rank configurations.

Across all three datasets, Recaps-118K (Figure 5a), Next-Preference (Figure 5b), and SAM-LLaVA (Figure 5c), FediLoRA consistently outperforms full-layer editing. Notably, the gap is particularly evident in the SAM dataset, where the performance of full-layer editing stagnates while FediLoRA continues to improve steadily with each epoch. Specifically, Figure 5b witnesses degradation after the 6-th epoch while using full-layer editing. The results show that when the model edits only half of its layers, its personalized performance becomes somewhat close to FediLoRA's performance.

These results highlight a critical insight: increasing the proportion of edited parameters does not necessarily yield better performance. In contrast, FediLoRA's selective editing strategy demonstrates superior effectiveness and robustness under challenging learning conditions.

4.4 Comparative Study of Information Preservance



(a) 30% missing ratio (b) 60% missing ratio

Figure 6: Comparison of the Global Adapter L2 Norm between HetLoRA and FediLoRA under 30% and 60% missing ratios.

In this experiment, we examine the amount of information retained in LoRA parameters during training, comparing FediLoRA with HetLoRA. Specifically, we track the L_2 norm of the aggregated global LoRA for each epoch, where the 0-th round is the norm of the initialized parameters. This experiment is performed on SAM-LLaVA dataset. FediLoRA and HetLoRA are initialized with the same parameters. From Figure 6, after the first epoch of aggregation, a notable difference emerges in their L_2 norms: HetLoRA drops to around 10 under both missing modality ratios, while FediLoRA maintains a level above 20. This showcases that compared to HetLoRA, FediLoRA is able to keep the information away from being diluted. As mentioned in HetLoRA [Cho *et al.*,

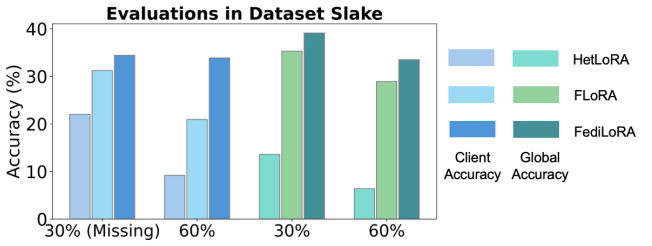


Figure 7: Performance of FediLoRA and baseline methods on the SLAKE dataset. The two bar clusters on the left show the global model accuracy, while the two clusters on the right present the client-side model accuracy.

2024], their aggregation strategy aims to balance high data complexity (abundant training data) with low-rank LoRA and low data complexity with high-rank LoRA. Important information from high-rank clients may be diluted during aggregation across heterogeneous ranks, especially when these clients possess highly complex training data.

4.5 Experiments in Medical Domain

This experiment compares FediLoRA, HetLoRA, and FloRA under two missing-modality settings (Figure 7), with all configurations consistent with the three benchmarks used in the main experiments. Under the 30% missing ratio. The global model accuracy of HetLoRA reaches approximately 20% but drops sharply when the missing ratio increases to 60%. Its client-side accuracy remains below 15%, indicating that it is not suitable as a medical foundation model for simple fine-tuning. FloRA exhibits similar behavior to HetLoRA in terms of global accuracy, although its client model reaches over 35% accuracy at the 30% missing ratio.

In contrast, FediLoRA consistently achieves the highest accuracy across all settings. Notably, it reaches 39.2% accuracy under the 30% missing-modality ratio. These results demonstrate that FediLoRA can be effectively applied to real-world domain specific scenarios: regardless of the extent of missing modality ratio, both small medical institutions and large hospitals can obtain strong diagnostic models through federated fine-tuning.

5 Related Work

5.1 Heterogeneous LoRA for FL

Given the homogeneous LoRA constraints the FL capability of capturing the full diversity of client contributions, a line of works using heterogeneous LoRA emerges. FlexLoRA [Bai *et al.*, 2024] aggregates the full LoRA matrix ΔW from clients on the global server followed by a Singular Value Decomposition (SVD) to each client to select their singular vectors. This method uploads the complete training parameters before performing aggregation and decomposition, which makes it difficult to apply in real-world scenarios. In HetLoRA [Cho *et al.*, 2024], during local training, the clients apply the rank self-pruning according to local resources. The global server aggregates the heterogeneous LoRA from clients by a simple zero-padding and weighted sparcity aggregation. The global LoRA is then truncated to distribute to the clients. The work in [Byun and Lee, 2025] replaces the zero-padding in HetLoRA with a strategy that replicates rows and columns from the high-quality clients' matrices to match the required dimensions, preserving the important information in high-priority clients and further improving the convergence speed. FLoRA [Wang *et al.*, 2024a] proposes a noise-free stacking-based aggregation method for heterogeneous LoRA without distributing the LoRA parameters back to the clients. LoRA-FAIR [Bian *et al.*, 2025] introduce additional module to keep fair in different clients LoRA. None of these works address multimodal or missing modalities.

5.2 Federated Learning for Missing Modality

CACMRN [Xiong *et al.*, 2023] is the first FL framework to tackle the missing modalities issue by performing modality **reconstruction**. Yu et al. [Yu *et al.*, 2023] aim to address the issue of missing modalities in cross-modal retrieval. Clients, each with distinct modalities, learn the representation information of specific modalities and upload it to the server for **contrastive learning**. Fed-Multimodal [Feng *et al.*, 2023] is a multimodal benchmark built upon the FL algorithm. The work presented a range of models and datasets and proposed a method to initialize the missing modal vector as **zero**. PmcmFL [Bao *et al.*, 2024] applied the common **prototype exchange** method against Non-IID to the missing modality, and they store prototypes of all modalities for all classes from each client on the server. When encountering the missing modalities, it utilizes the corresponding class-specific prototypes to fill in the representations. MMiC [Yang *et al.*, 2025] employs three techniques to mitigate the issue of missing modalities; they offer a potential solution without introducing additional modules; however, their reliance on frequent parameter exchanges limits their applicability to large-scale models. The aforementioned methods remain limited to traditional small-scale model architectures, rendering them impractical for medical or other VLLM-based application scenarios. In contrast, our approach does not impose restrictions on the types of multimodal tasks, meaning that we do not require the use of specific models.

6 Conclusion

In this paper, we presented FediLoRA, a simple yet effective framework for federated multimodal fine-tuning under heterogeneous LoRA ranks and missing modalities. By combining dimension-wise aggregation with lightweight layer-wise model editing, FediLoRA improves both global and personalized performance under missing-modality conditions. Experiments across multiple benchmark datasets, together with convergence analysis, confirm its robustness under severe modality incompleteness. Moreover, experiments on the medical dataset further demonstrate its effectiveness in domain-specific scenarios, highlighting its potential for real-world applications. In the future, we will focus on extending FediLoRA to more heterogeneous settings, such as scenarios involving different modality types and varying numbers of modalities.

Acknowledgments

The research is supported by Australian Research Council projects: DP240103070 and IE230100119.

References

[Bai *et al.*, 2024] 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 *Proceedings of the NeurIPS, 2024*, 2024.

[Bao *et al.*, 2024] Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Liang Hu, Ke Liu, Yang Liu, and Chongyang Shi. Multimodal federated learning with missing modality via prototype mask and contrast, 2024.

[Bian *et al.*, 2025] Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement. In *Proceedings of the CVPR*, 2025.

[Byun and Lee, 2025] Yuji Byun and Jaeho Lee. Towards federated low-rank adaptation of language models with rank heterogeneity. In *Proceedings of the 2025 NAACL Short Papers*, 2025.

[Chen and Li, 2022] Sijia Chen and Baochun Li. Towards optimal multi-modal federated learning on non-iid data with hierarchical gradient blending. In *IEEE INFOCOM, 2022*, 2022.

[Chen *et al.*, 2023] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

[Cho *et al.*, 2024] 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 EMNLP)*, 2024.

[Chu *et al.*, 2025] Xiaoli Chu, Yiheng Ye, Siqiao Tang, Miaoru Han, Guowei Wang, Shuai Lin, Bingzhen Sun, Qingchun Huang, Yan Zhang, Xiaodong Chu, et al. Personalized medication for chronic diseases using multimodal data-driven chain-of-decisions. *Advanced Science*, 2025.

[Cismondi *et al.*, 2013] Federico Cismondi, André S Fialho, Susana M Vieira, Shane R Reti, João MC Sousa, and Stan N Finkelstein. Missing data in medical databases: Impute, delete or classify? *Artificial intelligence in medicine*, 2013.

[Collins *et al.*, 2022] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Fedavg with fine tuning: Local updates lead to representation learning. *Proceedings of the NeurIPS, 2022*, 2022.

[De Zarzà *et al.*, 2023] I De Zarzà, J De Curtò, Gemma Roig, and Carlos T Calafate. Optimized financial planning: integrating individual and cooperative budgeting models with llm recommendations. *AI*, 2023.

[Feng *et al.*, 2023] Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. Fedmultimodal: A benchmark for multimodal federated learning. In *Proceedings of the 29th KDD, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, 2023.

[Guo *et al.*, 2024] Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. *arXiv preprint arXiv:2410.01463*, 2024.

[Houlsby *et al.*, 2019] 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 *ICML*, 2019.

[Hu *et al.*, 2022a] 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 *Proceedings of the 10th ICLR, 2022*, 2022.

[Hu *et al.*, 2022b] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

[Karimireddy *et al.*, 2020] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[Kuang *et al.*, 2024] 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 KDD, 2024*, 2024.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*. Association for Computational Linguistics, 2021.

[Li *et al.*, 2019] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, 2004.

[Liu *et al.*, 2021] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*. IEEE, 2021.

[Liu *et al.*, 2022] Xiao Liu, Heyan Huang, Ge Shi, and Bo Wang. Dynamic prefix-tuning for generative template-based event extraction. *arXiv preprint arXiv:2205.06166*, 2022.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[Liu *et al.*, 2025] Xiao-Yang Liu, Rongyi Zhu, Daochen Zha, Jiechao Gao, Shan Zhong, Matt White, and Meikang Qiu. Differentially Private Low-Rank Adaptation of Large Language Model Using Federated Learning. *ACM Trans. Manag. Inf. Syst.*, 2025.

[LMMs-Lab, 2024] LMMs-Lab. Llava-recap-118k. https://huggingface.co/datasets/lmms-lab/LLaVA-ReCap-118K, 2024. Accessed: 2026-05-11.

[McMahan *et al.*, 2017] 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 AISTATS*, 2017.

[Mishra, 2024] Sandeep Mishra. battlemaster/llava-next-preference-dataset-20k. https://huggingface.co/datasets/battleMaster/llava-next-Preference-dataset-20k, 2024. Accessed: 2026-05-11.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[Reddi *et al.*, 2021] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *Proceedings of the 9th ICLR, 2021*, 2021.

[Savage *et al.*, 2025] Thomas Savage, Stephen P Ma, Abdessalem Boukil, Ekanath Rangan, Vishwesh Patel, Ivan Lopez, and Jonathan Chen. Fine-tuning methods for large language models in clinical medicine by supervised fine-tuning and direct preference optimization: Comparative evaluation. *Journal of Medical Internet Research*, 2025.

[Sun *et al.*, 2024a] Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning. *arXiv preprint arXiv:2403.12313*, 2024.

[Sun *et al.*, 2024b] Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in Privacy-preserving Federated Learning. In *Proceedings of the 12th ICLR, 2024*, 2024.

[Wang *et al.*, 2022] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the EMNLP*, 2022.

[Wang *et al.*, 2023] Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. Finvis-gpt: A multimodal large language model for financial chart analysis. *arXiv preprint arXiv:2308.01430*, 2023.

[Wang *et al.*, 2024a] 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 *Proceedings of the NeurIPS, 2024*, 2024.

[Wang *et al.*, 2024b] 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. *arXiv preprint arXiv:2409.05976*, 2024.

[Wen *et al.*, 2024] Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and Jiliang Tang. Ai for education (ai4edu): Advancing personalized education with llm and adaptive learning. In *Proceedings of the 30th KDD*, 2024.

[Wu *et al.*, 2024] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024.

[Xiong *et al.*, 2023] Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. Client-adaptive cross-model reconstruction network for modality-incomplete multimodal federated learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM), MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 2023.

[Yang *et al.*, 2024a] Lishan Yang, Alireza Seyed Shakeri, Liangxi Pu, Weitong Chen, and Yanjun Shu. Efficient clustered federated learning by locality sensitive hashing. In *International Conference on Advanced Data Mining and Applications*. Springer, 2024.

[Yang *et al.*, 2024b] Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. Dual-Personalizing Adapter for Federated Foundation Models. In *Proceedings of the NeurIPS, 2024*, 2024.

[Yang *et al.*, 2025] Lishan Yang, Wei Emma Zhang, Quan Z Sheng, Lina Yao, Weitong Chen, and Ali Shakeri. Mmic: Mitigating modality incompleteness in clustered federated learning. *CIKM*, 2025.

[Yu *et al.*, 2023] Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. In *Proceedings of the 11th ICLR, 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[Yuan *et al.*, 2024] Haolin Yuan, William Paul, John Aucott, Philippe Burlina, and Yinzhi Cao. Pfededit: Personalized federated learning via automated model editing. In *European Conference on Computer Vision*, 2024.

[Zeng *et al.*, 2024] Huimin Zeng, Zhenrui Yue, and Dong Wang. Open-vocabulary federated learning with multimodal prototyping. In *Proceedings of the 2024 Conference of the NAACL: Human Language Technologies (Long Papers)*, 2024.

[Zhang *et al.*, 2024] 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 *Proceedings of the 2024 IEEE ICASSP*, 2024.

[Zong *et al.*, 2021] Linlin Zong, Qiujie Xie, Jiahui Zhou, Peiran Wu, Xianchao Zhang, and Bo Xu. Fedcmr: Federated cross-modal retrieval. In *Proceedings of the 44th SIGIR, Virtual Event, Canada, July 11-15, 2021*, 2021.