UniFLoW: Universal Multi-Modal Federated LoRA Fine-Tuning Framework with Analytical Aggregation

Anonymous authors
Paper under double-blind review

Abstract

As Multimodal Large Language Models (MLLMs) continue to be trained, the availability of public data diminishes, limiting the possibility for further training and adaptation. However, private data remains an underutilized yet valuable resource. Federated Learning (FL) enables decentralized training on private data, yet extending it to MLLMs is challenging: heterogeneous client modalities induce architectural incompatibility, and full-parameter fine-tuning of billion-scale models incurs prohibitive communication costs. Parameter-efficient methods like LoRA alleviate these issues but introduce aggregation inconsistency, as averaged low-rank updates fail to recover the true global update faithfully. To address these issues, we propose **UniFLoW** (**Uni**versal multi-modal **F**ederated **Lo**RA fine-tuning framework **W**ith Analytical Aggregation), a unified federated framework that leverages pre-trained large language models and multi-modal Encoder architecture, and our proposed **Fed**erated **A**ggregating **A**nalytical **Lo**w-**R**ank **A**daption ($FedA^2$-$LoRA$). **UniFLoW** effectively utilizes fragmented client-side multi-modal data while $FedA^2$-$LoRA$ ensuring consistent aggregation. And modality-specific encoders and a II stage training strategy ensure effective integration of diverse modalities without overfitting. Experiments on text, image, and speech demonstrate that **UniFLoW** enables scalable, communication-efficient, and aggregation-consistent federated fine-tuning, with $FedA^2$-$LoRA$ achieving state-of-the-art performance compared to existing FedLoRA approaches. We envision **UniFLoW** as a promising solution to the growing scarcity of public data.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated remarkable performance across a broad spectrum of tasks and have garnered widespread recognition. These advancements are primarily driven by Pre-trained Language Models (Min et al., 2023; Li et al., 2024) trained on massive multimodal datasets. However, as publicly available data sources become increasingly saturated, it is becoming progressively harder to obtain new public multimodal data. Consequently, ① ***alleviating the multimodal data acquisition bottleneck by unlocking previously inaccessible data sources*** has emerged as a critical challenge for sustaining the progress of MLLMs.

Benefiting from the rapid development of mobile devices (Mairittha et al., 2020), personal data has become abundant, offering a vast resource for MLLMs pre-training and fine-tuning. Yet, such data often contains sensitive private information, which clients are reluctant to share. For example, a street-view image may inadvertently reveal a personal license plate number. To address these privacy concerns, Federated Learning (FL) (McMahan et al., 2017; Liang et al., 2025) allows clients to train models locally on private data while sharing and aggregating model parameters to fully exploit these resources. Despite its effectiveness, traditional FL restricts parameter aggregation to unimodal models. In the real world, clients often possess heterogeneous multimodal data (as illustrated in Figure 1(a)). ② *This modality inconsistency induces architectural incompatibility, thereby rendering existing parameter aggregation approaches infeasible in multimodal scenarios.*

Another prominent challenge of applying FL to MLLMs lies in ③ ***communication overhead***. Given that MLLMs typically contain tens of billions of parameters, full-parameter fine-tuning (Lv et al.,

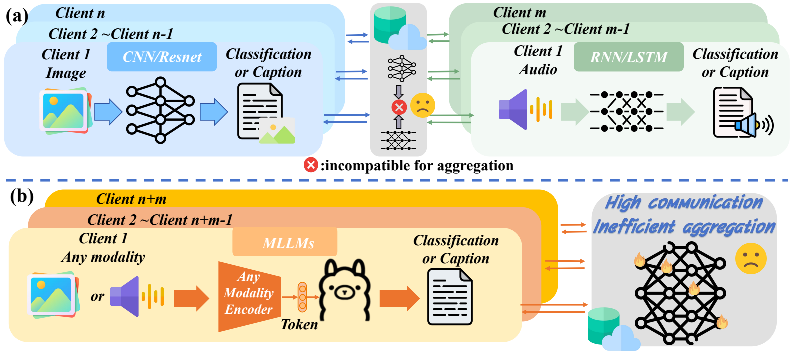


Figure 1: (a) Traditional FL faces architecture heterogeneity: clients handling different modalities (e.g., images with CNNs and audio with RNNs) employ incompatible model structures, making parameter aggregation infeasible. (b) A unified architecture based on MLLMs can process diverse modalities through modality encoders. However, full-parameter training and transmission introduce high communication costs and inefficient aggregation.

2023) not only incurs substantial communication costs—since an enormous number of parameters must be transmitted to the server—but also aggravates issues such as knowledge forgetting and overfitting, as shown in Figure 1 (b). To address this issue, incorporating Parameter-Efficient Fine-Tuning (PEFT) (Han et al., 2024) into Federated Large Language Models (FedLLMs) (Ye et al., 2024a) offers a promising and scalable solution. Among these PEFT methods, LoRA-based (Hu et al., 2022) methods have become increasingly popular. Although LoRA offers significant advantages in training efficiency and model generalization, it introduces aggregation inconsistency in federated settings. Specifically, when aggregating low-rank updates from multiple clients, the global weight update Δ_W computed by averaging the low-rank components and then performing matrix multiplication deviates from the true average of the local updates Δ_W^*, as shown in ④:

$$\textbf{In most cases,} \quad ④\ \Delta_W = \left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{B}_k\right)\left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{A}_k\right) \neq \Delta_W^* = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{B}_k\boldsymbol{A}_k$$

Recently, several FedLoRA (Wu et al., 2024) variants have been proposed to address this issue. Most of them focus on uploading a single matrix, either $\boldsymbol{A}$ or $\boldsymbol{B}$, as the optimal choice. For example, one line of work uploads only one of these matrices to the server, which offers a certain level of personalization but still cannot fully recover the complete global update Δ_W. Another line of work (Singhal et al., 2024) directly transmits the residual matrix $\Delta_W - \Delta_W^*$, but this significantly increases communication costs compared to exchanging low-rank updates, thereby weakening the efficiency advantage of LoRA. ⑤ *Current FedLoRA methods still fail to achieve an effective balance between recovery accuracy and communication efficiency.*

To address the challenges of ① **unlocking private multimodal data at scale**, ② **modality-induced architectural incompatibility**, ③ **communication overhead**, ④ **aggregation inconsistency**, and ⑤ **accuracy–efficiency trade-off** in exploiting private multimodal client data, we propose the **Uni**versal multi-modal **F**ederated **Lo**RA fine-tuning framework **W**ith Analytical Aggregation (**UniFLoW**). Building on the FL and using modality-specific encoders (ImageBind (Girdhar et al., 2023)) together with a shared base model (Vicuna-7B (Zheng et al., 2023)) resolves ① and ②,in **UniFLoW**. This design enables the system to process up to six modalities and seamlessly scale with additional encoders, while inserting LoRA adapters into both the encoders and the base model so that only a small number of trainable parameters need to be communicated across federated rounds.

Within the **UniFLoW** framework, we further develop **Federated Aggregating Analytical Low-Rank Adaptation** ($FedA^2$**-**$LoRA$) as the core analytical aggregation module to address ③, ④, and ⑤ for LoRA-based FedLLMs. Inspired by (Guo et al., 2024)'s analysis of the optimal solution and our gradient analysis (in the Appendix A), we observe that the A matrices tend to capture more global and stable directions than the B matrices. Based on this observation, $FedA^2$-$LoRA$ aggregates the $\boldsymbol{A}$ matrices across clients via averaging to obtain $\boldsymbol{A}^{t+1}$, and then analytically reconstructs the corresponding $\boldsymbol{B}^{t+1}$ by solving a regularized least-squares (Ridge Regression) problem (Zhang et al., 2010) so that the resulting low-rank update $\boldsymbol{B}^{t+1}\boldsymbol{A}^{t+1}$ closely approximates the desired full-rank update Δ_W^*. Compared to FedExLoRA (Singhal et al., 2024), which relies on transmitting residuals, $FedA^2$-$LoRA$ not only improves recovery accuracy but also reduces communication costs, thereby providing a more consistent and efficient federated aggregation scheme for LoRA parameters.

Furthermore, in real-world multimodal federated learning (MMFed) (Feng et al., 2023) scenarios, each client typically possesses only one modality, and different clients have different modalities. For instance, some hospitals may only have ultrasound data, while others may only have X-ray data. This diversity complicates modality alignment in FL for **UniFLoW**. To prevent LLMs from over-integrating modality-specific information and neglecting content information during training, we propose an adaptive II stage training, where we first fine-tune only the modality encoders while keeping the base model frozen, and then, after a warm-up period, switch to fine-tuning only the base model with the encoders frozen. This approach ensures a more balanced and effective learning process by gradually incorporating both modality-specific and global model updates.

In our multimodal experiments, we involved speech, images, and text. We demonstrated that **UniFLoW**'s visual modality training can enhance the LLM and facilitate audio QA tasks. Furthermore, multimodal client-side training continuously improves the LLM's understanding capabilities. To the best of our knowledge, **UniFLoW** is the first framework capable of fine-tuning the multi-task FedMLLM across three modalities. In text question-answering experiments, $FedA^2$-$LoRA$ demonstrated significant success in restoring Δ_W^* without increasing communication costs, as shown through comparative experiments.

Our main contributions are summarized as follows:

- To fully leverage distributed multi-modal private data, we design a federated fine-tuning framework **UniFLoW** for MLLMs with universal modality support. To the best of our knowledge, this is the first federated Universal MLLMs fine-tuning framework that enables general-purpose modality integration.
- To address the aggregation inconsistency in federated LoRA fine-tuning, we propose $FedA^2$-$LoRA$, which retains the parameter-efficient nature of LoRA without introducing additional communication costs. Furthermore, $FedA^2$-$LoRA$ enables faithful reconstruction of the global update Δ_W, thereby substantially mitigating the inconsistency issue.
- Through extensive experiments, we validate the effectiveness of our proposed multi-modal fine-tuning framework. In addition, our $FedA^2$-$LoRA$ consistently achieves state-of-the-art performance in both multi-modal and single-modal settings.

2 RELATED WORK

2.1 MULTIMODAL IN FEDERATED LEARNING

The rapid progress of multi-modal learning (Huang et al., 2021; Ye et al., 2025) has significantly broadened the application scope of artificial intelligence. Federated learning (Huang et al., 2024), initially explored in single-modality scenarios, has also recently begun to extend to multi-modal tasks. Existing methods can be grouped into two categories. The first category addresses **multiple independent single-modality tasks**. For example, QFL (Pokharel et al., 2025) performs classification across different modalities such as speech and images, yet each task remains modality-specific. The second category targets **a single multi-modal task**. For instance, FedCola (Sun et al., 2024a) combines complementary local training with collaborative global aggregation to enable cross-modal knowledge sharing, achieving strong performance on image-to-text generation. Similarly, MLLM-LLaVA-FL (Zhang et al., 2025) leverages large models to supervise and guide the training of smaller federated models, enhancing vision-to-language capabilities (Liu et al., 2024). Despite these advances, current FL approaches typically operate with relatively small models and remain confined to a single multi-modal task, limiting their ability to generalize across multiple tasks or scale seamlessly to new modalities. **To the best of our knowledge, a unified and extensible federated framework capable of supporting various multimodal tasks simultaneously has yet to be developed.**

2.2 LORA IN FEDERATED LEARNING

Fine-tuning large models has proven effective in adapting them to become task-specific experts (Panigrahi et al., 2023). However, in FL, full-parameter fine-tuning (FPFT) (Bian et al., 2025) incurs

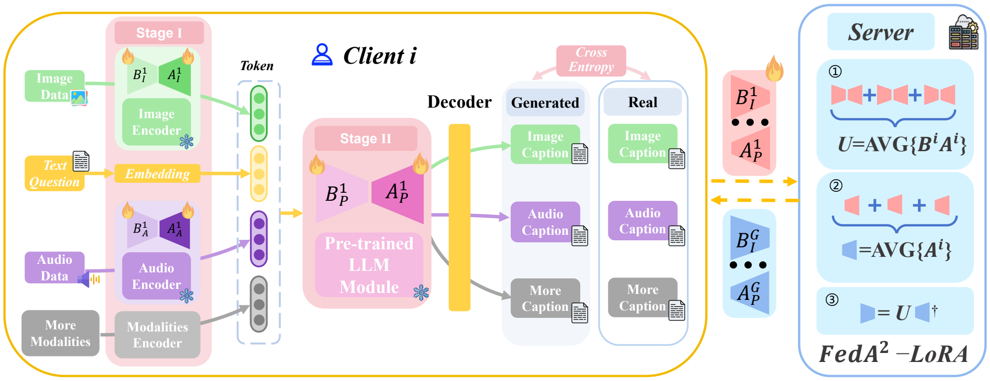


Figure 2: Overview of the UniFLoW. The inference process operates on each client, where local multi-modal data (e.g., text, images, or audio) is processed by modality-specific encoders (e.g., ImageBind), and the resulting features are then passed to a pretrained LLM (e.g., Vicuna-7B). The fine-tuning procedure is organized into two stages: in the I stage, we apply LoRA to the modality encoders, and in the II stage, we use LoRA to fine-tune the pretrained large language model. The server aggregates parameter updates from clients using the proposed $FedA^2$-$LoRA$ method, which ensures consistent and efficient updates while minimizing communication overhead. This framework is designed to handle heterogeneous data sources across clients, thereby enhancing scalability and model performance across different modalities.

prohibitive communication costs, and the limited data available on each client further increases the risk of overfitting. As a result, recent research on FedLLM (Ye et al., 2024b) has increasingly shifted toward PEFT (Chua et al., 2023; Rong et al., 2025) to overcome these challenges.

Due to LoRA's (Hu et al., 2022) low training cost and strong generalization ability, it has become the predominant fine-tuning strategy for FedLLMs. FedLoRA (Yi et al., 2023) first introduced LoRA into FL, substantially reducing both communication and computation costs while mitigating overfitting. Building on this, FDLoRA (Qi et al., 2024) employed two parallel LoRA modules—one for capturing personalized information and the other for modeling global knowledge—but this design nearly doubled the computation overhead. To further reduce costs, FFA-LoRA (Sun et al., 2024b) proposed freezing the A matrix and uploading only B, which alleviates the inconsistency of Δ_W and Δ_W^*. However, this approach decreases the number of trainable parameters and places excessive reliance on the initialization of A. To address this, FedSA-LoRA (Guo et al., 2024) suggested uploading only A for aggregation while keeping both A and B trainable locally, partially alleviating the dependence on the initial A. While this strategy enables the training of personalized models, it remains insufficient for producing a stronger global model. Moreover, although FedSA-LoRA claims that B is client-data-specific, it does not conclusively demonstrate that B depends solely on local data. Therefore, a more effective LoRA aggregation strategy remains necessary.

3 METHOD

To exploit the fragmented and diverse modal information available on clients, we propose a **Uni**versal multi-modal **F**ederated **Lo**RA fine-tuning framework **W**ith Analytical Aggregation (**UniFLoW**), as shown in Figure 2. We describe our **UniFLoW** from two perspectives: *client training* and *server aggregation*.

3.1 TRAINING PROCESS ON THE CLIENT SIDE

Unlike previous FedMLLMs approaches, **UniFLoW** focuses on heterogeneous modalities, where client modalities are inconsistent. For instance, in our experiments, some clients only have audio-based Q&A or image-based Q&A, and some have both. Therefore, for client k, there is a dataset $D_k = \{(\mathcal{M}_{n_k}^m, Q_{n_k}, \mathcal{A}_{n_k})\}_{n=1}^N$, where $m \in M_{avail} = \{\text{image}, \text{audio}, \dots\}$ represents the modality.

For the encoder (Imagebind) and the base model (Vicuna), there are also some differences. We not only apply LoRA to the base model, but also to the modal encoder. Any layer ℓ satisfies:

$$\tilde{W}^{(m,\ell)} \;=\; W_0^{(m,\ell)} \;+\; \alpha \boldsymbol{B}^{(m,\ell)} \boldsymbol{A}^{(m,\ell)}, \quad h^{(m,\ell+1)} = \phi\left(\tilde{W}^{(m,\ell)} h^{(m,\ell)} + b^{(m,\ell)}\right) \tag{1}$$

Where W_0 represents the frozen parameters of the model, α denotes the scaling parameter indicating the degree of injection, and $\boldsymbol{B}$ and $\boldsymbol{A}$ represent the trainable parameters. h is the input feature.

For a client data sample $(\mathcal{M}^m_{n_k}, Q_{n_k}, \mathcal{A}_{n_k})$, the modal data $\mathcal{M}^m_{n_k}$ will enter Imagebind f_m to process and produce the feature $Z^m_{n_k}$. Here, $\Theta_0^{(m)}$ represents the set of frozen parameters $W_0^{(m,\ell)}$, and $\Phi^{(m)}_{E,k}$ represents the set of trainable parameters $\boldsymbol{B}^{(m,\ell)}_{E,k}$ and $\boldsymbol{A}^{(m,\ell)}_{E,k}$. As shown in Eq. 2:

$$Z^m_{n_k} \;=\; f_m(\mathcal{M}^m_{n_k}; \Theta_0^{(m)}, \Phi^{(m)}_{E,k}), \quad \Phi^{(m)}_{E,k} \triangleq \{\boldsymbol{B}^{(m,\ell)}_{E,k}, \boldsymbol{A}^{(m,\ell)}_{E,k}\}_{\ell=1}^{L_m} \tag{2}$$

Then, the modal feature $Z^m_{n_k}$ is mapped to a unified low-dimensional space through a pooling layer, and then projected to the dimension required by the LLM via a trainable fully connected layer $\boldsymbol{P} \in \mathbb{R}^{d_{\text{LLM}} \times d_{\text{IB}}}$. The layer $\boldsymbol{P}$ learns to bridge the gap between different modal pooling outputs and the LLM, thereby producing the feature $\mathcal{X}^m_{n_k}$ required by the LLM. Specifically, as shown in Eq. 3:

$$\mathcal{X}^m_{n_k} = \boldsymbol{P}(\text{Pool}(Z^m_{n_k})) \tag{3}$$

Before feeding into the LLMs, we combine the question Q_{n_k} with the feature $\mathcal{X}^m_{n_k}$ extracted from modality $\mathcal{M}^m_{n_k}$ to construct the prompt. Specifically, multiple modal features are concatenated and then joined with the common prompt tokens $e_{\text{BOS}}, P_{\text{before}}, P_{\text{gap}}, P_{\text{after}}$ (as shown in Appendix) to form the final prompt X_{n_k}, as shown in Eq. 4:

$$X_{n_k} = \left[e_{\text{BOS}}, P_{\text{before}}, \bigoplus_{m \in M_{\text{avail}}} \mathcal{X}^m_{n_k}, P_{\text{gap}}, Q_{n_k}, P_{\text{after}} \right] \tag{4}$$

The LLMs processes the prompt X_{n_k} like Eq. 2. In this formulation, $\mathcal{F}(\cdot)$ denotes the Vicuna model, v is used to distinguish between the ImageBind parameters and the Vicuna parameters, and ℓ indicates the layer where LoRA is inserted. Here, $\Theta_0^{(v)}$ represents the set of frozen parameters, while $\Phi^{(v)}_{LLM}$ denotes the set of trainable parameters $\boldsymbol{B}$ and $\boldsymbol{A}$. As shown in Eq. 5:

$$H_{n_k} \;=\; \mathcal{F}(X_{n_k}; \Theta_0^{(v)}, \Phi^{(v)}_{LLM,k}), \quad \Phi^{(v)}_{LLM,k} \triangleq \{\boldsymbol{B}^{(v,\ell)}_{LLM,k}, \boldsymbol{A}^{(v,\ell)}_{LLM,k}\}_{\ell=1}^{L_v} \tag{5}$$

The hidden state H_{n_k} generated from the $(\mathcal{M}^m_{n_k}, Q_{n_k}, \mathcal{A}_{n_k})$ is mapped to a probability distribution p_t through the projection matrix w_o and the softmax$(\cdot)$ operation, representing the likelihood of each token in the vocabulary. The true token y_t is provided by the answer $\mathcal{A}_{n_k} = \{y_1, y_2, \dots\}$, and the loss between the predicted probability distribution p_t and the y_t is computed as shown in Eq. 6:

$$\mathcal{L} = -\sum_{t=1}^{T} \log p_t(y_t \mid y_{<t}, \{X_{n_k}\}_{m \in M_{\text{avail}}}), \quad p_t = \text{softmax}(w_o(H^t_{n_k})) \tag{6}$$

Previously, training of MLLMs was centralized, where data from different modalities were jointly shuffled and optimized in a single pipeline, naturally aligning their representations and yielding balanced performance (e.g., ImageBind). In FL, however, each client may hold only a single or highly biased modality, so directly fine-tuning the base LLM on such inputs makes it overfit modality-specific patterns rather than content-level semantics, thus harming cross-modal generalization.

To address this, we adopt a *two-stage training process*, as shown in Eq.7. When the number of local iterations *step* is less than τ, we update only the encoder parameters so that they absorb modality-specific biases and adapt to the local input distribution. When $step \geq \tau$, we update only the LLM, allowing the LLM to focus on content-related modeling based on already-normalized multimodal features. This staged schedule decouples modality adaptation from content learning and mitigates the risk of the LLM collapsing to a single dominant modality in federated settings.

$$\Phi_k^{(t)} = \{\Phi^{(t)}_{E,k}, \Phi^{(t)}_{LLM,k}\} = \begin{cases} \Phi^{(t)}_{E,k} = \Phi^{(t-1)}_{E,k} - \eta \cdot \nabla \Phi_{E,k} \mathcal{L} & , if\ step < \tau \\ \Phi^{(t)}_{LLM,k} = \Phi^{(t-1)}_{LLM,k} - \eta \cdot \nabla \Phi_{LLM,k} \mathcal{L} & , if\ step \geq \tau \end{cases} \tag{7}$$

Table 1: In our **UniFLoW** system, only 1% of the parameters require updating during fine-tuning, ensuring high efficiency and reduced communication overhead.

	Encoder		**Encoder LoRA**		**LLM**		**LLM LoRA**	
	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—	$Vicuna$	7B❄	$FedA^2$-$LoRA$	33M🔥
Image	ImageBind	1.2B❄	$FedA^2$-$LoRA$	6M🔥				
Audio								

3.2 Aggregation Process on the Server Side

In FL, as shown in Eq.8, direct aggregation can introduce bias because computing the global update as the product of the averaged low-rank matrices, i.e., $\Delta_W = (\frac{1}{K}\sum \boldsymbol{B}_k)(\frac{1}{K}\sum \boldsymbol{A}_k)$, is not equivalent to averaging the full-rank updates $\Delta_W^* = \frac{1}{K}\sum \boldsymbol{B}_k \boldsymbol{A}_k$. Specifically, the product of the averaged low-rank matrices $\boldsymbol{B}_k$ and $\boldsymbol{A}_k$ does not yield the same result as the average of the original weight matrices $\boldsymbol{B}_k\boldsymbol{A}_k$. Consequently, directly aggregating low-rank matrices $\boldsymbol{B}_k$ and $\boldsymbol{A}_k$ may introduce inconsistencies, resulting in suboptimal global updates and slower convergence in FL.

$$\Delta_W = \left(\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{B}_k\right)\left(\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{A}_k\right), \Delta_W^* = \frac{1}{K}\sum_{k=1}^{K} \Delta w_k = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{B}_k \boldsymbol{A}_k. \tag{8}$$

Existing FedLoRA variants (Qi et al., 2024) mainly focus on uploading only one of the two matrices ($\boldsymbol{A}$ or $\boldsymbol{B}$) to resolve the mismatch between Δ_W and Δ_W^* in Eq.8, but the reduction in uploaded parameters also limits the amount of shared information that can be aggregated. This hinders the aggregation of some shared information. Inspired by FedSA-LoRA (Guo et al., 2024) and our gradient analysis (see Appendix), we observe that the matrices $\boldsymbol{A}_k$ tend to encode more global and stable directions, whereas $\boldsymbol{B}_k$ are more sensitive to client-specific data. To better extract this global information and reduce bias without increasing communication overhead, we propose $FedA^2$-$LoRA$.

Since $\boldsymbol{A}_k \in \Phi_k$ primarily captures such global directions, we directly aggregate $\boldsymbol{A}_k$ collected from the clients to obtain the global $\boldsymbol{A}$, as shown in Eq.9. This approach follows the same processing method as in the previous FedLoRA approach, where each client uploads its local $\boldsymbol{A}_k$ parameters, which are then aggregated on the server to form the global $\boldsymbol{A}$. This method ensures that the global $\boldsymbol{A}$ reflects the shared global information across clients, without being influenced by the local data distributions. Consequently, the aggregation process is more consistent and unbiased, enabling the model to capture the generalizable patterns better while avoiding overfitting to client-specific data.

$$\boldsymbol{A} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{A}_k \tag{9}$$

To effectively integrate the aggregated information from the parameters $\boldsymbol{B}_k$ and $\boldsymbol{A}_k \in \Phi_k$, while preserving global information and still reflecting the unique contribution of each client k, we perform aggregation in $\boldsymbol{B}_k\boldsymbol{A}_k$, which is also our goal of optimization. This approach ensures that the federated learning process captures both shared global knowledge and client-specific data without overfitting to individual data distributions, as shown in Eq. 10:

$$U = \Delta_W^* = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{B}_k \boldsymbol{A}_k \tag{10}$$

By the definitions of Eq. 9 and Eq. 10, we have transformed the multi-objective optimization into a single-objective optimization. The optimization objective can be written as:

$$\min_{B} \|\boldsymbol{B}\boldsymbol{A} - U\|_F^2 \tag{11}$$

The optimal solution $\boldsymbol{B}^* = U\boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{A}^T]^{-1}$ can be obtained by taking the derivative.When $\boldsymbol{A}\boldsymbol{A}^T$ is invertible, the pseudoinverse is equal to the standard inverse ($(\boldsymbol{A}\boldsymbol{A}^T)^\dagger = (\boldsymbol{A}\boldsymbol{A}^T)^{-1}$) and is unique, allowing for a straightforward solution. However, when $\boldsymbol{A}\boldsymbol{A}^T$ is non-invertible (or singular), the solution (via the pseudoinverse) is non-unique, and direct (brute-force) solving often leads to an ill-conditioned matrix and numerically unstable results. Therefore, when $\boldsymbol{A}\boldsymbol{A}^T$ is non-invertible, we

Table 2: Performance Comparison of Different LoRA Methods on the Heysquad and PandaGPT's Datasets, with Only the **Image** Modality Client Participating in Training.'- -' indicates methods where no training was conducted.

Test Modality	**Image**				**Audio**			
Method	Acc	P_{Bert}	R_{Bert}	F_{Bert}	Acc	P_{Bert}	R_{Bert}	F_{Bert}
- -	61.23	78.59	81.74	80.09	50.19	78.96	81.71	80.02
LoRA	64.68	81.55	81.58	81.55	52.27	81.86	79.40	80.61
LoRA (II stage)	66.67	84.67	82.24	83.43	55.15	83.02	80.09	81.53
FFA-LoRA	64.70	82.41	82.31	82.36	52.07	82.82	81.14	81.96
FFA-LoRA (II stage)	70.58	85.04	80.76	82.84	55.29	81.71	82.28	81.99
FedSA-LoRA	68.18	**87.28**	76.64	81.59	54.54	83.09	78.93	80.96
FedSA-LoRA (II stage)	69.23	82.82	81.14	81.96	56.27	82.91	80.64	81.71
FedEx-LoRA	67.02	79.15	85.19	81.99	53.88	80.74	80.64	80.68
FedEx-LoRA (II stage)	69.40	80.75	**86.91**	83.72	58.33	81.28	**83.02**	82.13
$FedA^2$-$LoRA$	69.02	80.66	86.85	83.63	55.56	80.75	83.00	81.86
$FedA^2$-$LoRA$ (II stage)	**72.22**	82.27	86.36	**84.27**	**58.44**	**85.89**	80.31	**83.00**

need to constrain the solution $\boldsymbol{B}$ through regularization (Tikhonov regularization (Groetsch, 1984)), ensuring a unique and stable.Therefore, we can rewrite the optimization objective as follows:

$$\min_{B} \|\boldsymbol{B}\boldsymbol{A} - U\|_F^2 + \lambda\|\boldsymbol{B}\|_F^2 \tag{12}$$

In the same manner, by taking the derivative, we can obtain the optimal solution for B^* as $U\boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{A}^T + \lambda I]^{-1}$ when $\boldsymbol{A}\boldsymbol{A}^T$ is non-invertible. By summarizing the results of the above results, we obtain the optimal method for solving $\boldsymbol{B}$, as follows Eq. 13:

$$\boldsymbol{B} = \begin{cases} U\boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{A}^T + \lambda I]^{-1} & \text{, if } (\boldsymbol{A}\boldsymbol{A}^T)^{-1} \text{ does not exist} \\ U\boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{A}^T]^{-1} & \text{, if } (\boldsymbol{A}\boldsymbol{A}^T)^{-1} \text{ exists} \end{cases} \tag{13}$$

Where λ is a hyperparameter. When the rank r is small, $\boldsymbol{A}\boldsymbol{A}^\top$ is typically well-conditioned and the influence of λ is minor. As r increases, a smaller λ can better restore $\boldsymbol{B}$, but too small a λ may lead to numerical instability when inverting $\boldsymbol{A}\boldsymbol{A}^\top + \lambda I$ over many communication rounds. In practice, we find that setting $\lambda = 1$ strikes a good balance between stability and reconstruction accuracy, and ensures that $FedA^2$-$LoRA$ converges reliably in multi-round FL training.

4 Experiments

In this section, we evaluate the effectiveness of **UniFLoW** on multimodal question answering using the open Audio QA from HeySQuAD (Wu et al., 2023) and the open image QA from PandaGPT (Su et al., 2023). Our **UniFLoW** implementation adopts Vicuna (Chiang et al., 2023) as the base LLM and ImageBind (Girdhar et al., 2023) as the modality encoder, with detailed configurations summarized in Table1. Beyond multimodal understanding and generation, we further assess the proposed $FedA^2$-LoRA on natural language understanding to verify its generality as a federated LoRA aggregation scheme. For these experiments, we use RoBERTa (Liu et al., 2019) on the GLUE benchmark (Wang et al., 2018), including MNLI, SST-2, and RTE. Our $FedA^2$-LoRA implementation is built on the FederatedScope-LLM library (Kuang et al., 2023). LoRA-based experiments on GLUE are conducted with half-precision for improved efficiency on NVIDIA GeForce RTX 4090 GPUs, while multimodal experiments for **UniFLoW**, rsLoRA, and VeRA are run on NVIDIA A800 GPUs. Unless otherwise specified, the main results reported in tables are averaged over multiple runs with mean and standard deviation, and others are obtained from a single run.

4.1 UniFlow Multimodal Performance Evaluation

Since datasets Heysquad and PandGPT are both open-domain QA, traditional evaluation schemes are not well-suited. Therefore, we adopt BERTScore (P_{Bert}, R_{Bert}, F_{Bert}) (Devlin et al., 2019) and token accuracy (Acc) (Jiang et al., 2021) as evaluation metrics.We selected recent works, Fed-LoRA (Wu et al., 2024), FFA-LoRA (Sun et al., 2024b), and FedSA-LoRA (Guo et al., 2024), for

Table 3: Performance Comparison of Different LoRA Methods on the Heysquad and PandaGPT's Datasets, with Only the **Audio** Modality Client Participating in Training.'- -' indicates methods where no training was conducted.

Test Modality	**Image**				**Audio**			
Method	Acc	P_{Bert}	R_{Bert}	F_{Bert}	Acc	P_{Bert}	R_{Bert}	F_{Bert}
- -	61.23	78.59	81.74	80.09	50.19	78.96	81.71	80.02
LoRA	57.14	77.10	79.74	78.38	54.87	76.78	83.19	79.85
LoRA (II stage)	62.89	79.20	81.14	80.16	54.54	80.25	82.64	81.42
FFA-LoRA	62.94	77.33	81.84	79.52	57.14	80.87	84.12	82.40
FFA-LoRA (II stage)	63.29	81.72	81.64	81.68	55.55	83.06	83.20	83.21
FedSA-LoRA	63.19	78.72	83.30	80.93	53.84	80.37	80.21	80.25
FedSA-LoRA(II stage)	63.15	82.82	81.14	81.96	53.84	84.77	84.43	84.60
FedEx-LoRA	62.71	80.73	80.76	80.71	55.52	79.55	82.68	81.08
FedEx-LoRA (II stage)	62.96	81.39	81.72	81.44	56.80	81.31	83.89	82.58
$FedA^2$-$LoRA$	63.00	**83.62**	80.30	81.92	56.63	80.32	82.89	81.57
$FedA^2$-$LoRA$ (II stage)	**63.63**	81.01	**83.33**	**82.14**	**58.15**	**85.51**	**85.61**	**85.56**

Table 4: Performance Comparison of Different LoRA Methods on the Heysquad and PandaGPT's Datasets, with the **Image** and **Audio** Modality Client Participating in Training.'- -' indicates methods where no training was conducted.

Test Modality	**Image**				**Audio**			
Method	Acc	P_{Bert}	R_{Bert}	F_{Bert}	Acc	P_{Bert}	R_{Bert}	F_{Bert}
- -	61.23	78.59	81.74	80.09	50.19	78.96	81.71	80.02
LoRA	64.89	**86.13**	82.22	84.11	58.33	82.42	**87.75**	84.97
LoRA (II stage)	66.21	85.75	81.57	83.60	59.63	88.64	82.44	85.42
FFA-LoRA	63.15	85.18	79.27	82.08	57.87	**91.20**	77.04	83.26
FFA-LoRA (II stage)	70.58	85.61	82.07	83.80	58.43	86.18	85.86	85.82
FedSA-LoRA	66.29	84.22	83.15	83.68	58.33	82.32	78.93	83.67
FedSA-LoRA (II stage)	68.42	84.79	83.71	84.22	60.23	84.79	83.71	84.22
FedEx-LoRA	66.67	81.21	85.81	83.45	58.33	83.92	83.82	83.79
FedEx-LoRA (II stage)	69.40	81.89	88.06	84.86	59.76	82.78	86.50	84.60
$FedA^2$-$LoRA$	75.00	84.51	84.34	84.43	60.17	84.97	85.83	85.39
$FedA^2$-$LoRA$ (II stage)	**76.19**	83.44	**88.08**	**85.69**	**60.90**	87.15	85.49	**86.31**

comparison. Since the model is a pre-trained large model, only a limited communication rounds are required. In these experiments, we set the communication rounds to 10, with 10 participating clients per modality. Each client is provided with 2,000 samples for training and 200 samples for testing. The detailed descriptions of the datasets and evaluation metrics are provided in Appendix G.

As shown in Tables 3 and 2, training on data from a single modality enhances the base model's QA ability. Moreover, Table 4 demonstrates that the performance of clients improves even when modality inconsistencies exist across clients. However, as indicated in Table 3, directly applying LoRA in FL may lead to performance degradation. From Tables 2, 3, and 4, we observe that our II stage training strategy provides an effective approach for mitigating modality differences in FL.

4.2 $FedA^2$-$LoRA$ Performance Evaluation

As shown in Table 5, the proposed $FedA^2$-LoRA aggregation strategy consistently improves the performance of different LoRA variants on the GLUE benchmark. In particular, $FedA^2$-rsLoRA achieves the highest average accuracy of 61.19, a clear improvement over the original rsLoRA (53.50). Similar gains are observed for both LoRA and VeRA, indicating that $FedA^2$-LoRA can generally enhance the effectiveness of existing federated LoRA methods. These results suggest that $FedA^2$-LoRA not only alleviates the aggregation inconsistency typically encountered in federated LoRA, but also improves model stability and generalization across diverse tasks, without introduc-

Table 5: Performance of different methods on the GLUE benchmark. For all tasks, we report accuracy evaluated across 3 runs with mean and standard deviation.

	Method	MNLI	RTE	SST2	QQP	Avg.
LoRA	LoRA	53.64$_{\pm 0.14}$	58.13$_{\pm 0.17}$	54.36$_{\pm 0.08}$	51.23$_{\pm 0.07}$	54.28
	FFA-LoRA	53.65$_{\pm 0.16}$	58.51$_{\pm 0.08}$	65.23$_{\pm 0.05}$	60.40$_{\pm 0.12}$	59.44
	FedDPA-LoRA	41.16$_{\pm 0.11}$	52.70$_{\pm 0.22}$	50.92$_{\pm 0.06}$	47.17$_{\pm 0.15}$	47.98
	FedSA-LoRA	53.65$_{\pm 0.32}$	52.84$_{\pm 0.05}$	53.58$_{\pm 0.19}$	63.18$_{\pm 0.05}$	55.81
	$FedA^2\text{-}LoRA$	53.66$_{\pm 0.32}$	58.56$_{\pm 0.22}$	69.52$_{\pm 0.25}$	63.89$_{\pm 0.16}$	61.43
rsLoRA	rsLoRA	53.64$_{\pm 0.22}$	55.90$_{\pm 0.03}$	69.17$_{\pm 0.02}$	46.69$_{\pm 0.18}$	56.35
	FFA-rsLoRA	52.09$_{\pm 0.12}$	52.42$_{\pm 0.17}$	53.97$_{\pm 0.15}$	52.20$_{\pm 0.32}$	52.67
	FedDPA-rsLoRA	53.64$_{\pm 0.18}$	52.70$_{\pm 0.12}$	51.95$_{\pm 0.21}$	54.25$_{\pm 0.17}$	53.13
	FedSA-rsLoRA	53.64$_{\pm 0.28}$	58.47$_{\pm 0.17}$	60.19$_{\pm 0.28}$	48.94$_{\pm 0.00}$	55.31
	$FedA^2\text{-}rsLoRA$	53.66$_{\pm 0.10}$	59.59$_{\pm 0.33}$	70.33$_{\pm 0.19}$	53.30$_{\pm 0.28}$	59.22
VeRA	VeRA	53.64$_{\pm 0.34}$	52.41$_{\pm 0.32}$	53.58$_{\pm 0.21}$	59.38$_{\pm 0.11}$	54.75
	FFA-VeRA	53.66$_{\pm 0.21}$	52.05$_{\pm 0.10}$	53.64$_{\pm 0.05}$	63.18$_{\pm 0.20}$	55.63
	FedDPA-VeRA	53.44$_{\pm 0.09}$	52.70$_{\pm 0.12}$	51.94$_{\pm 0.11}$	53.36$_{\pm 0.14}$	52.86
	FedSA-VeRA	53.64$_{\pm 0.21}$	52.05$_{\pm 0.13}$	53.84$_{\pm 0.12}$	56.65$_{\pm 0.24}$	54.04
	$FedA^2\text{-}VeLoRA$	53.66$_{\pm 0.25}$	53.46$_{\pm 0.00}$	54.60$_{\pm 0.30}$	61.99$_{\pm 0.06}$	55.92

ing additional communication overhead. In these experiments, we use 3 participating clients and run 500 communication rounds to evaluate the aggregation strategy under a practical FL setup.

4.3 Ablation Study

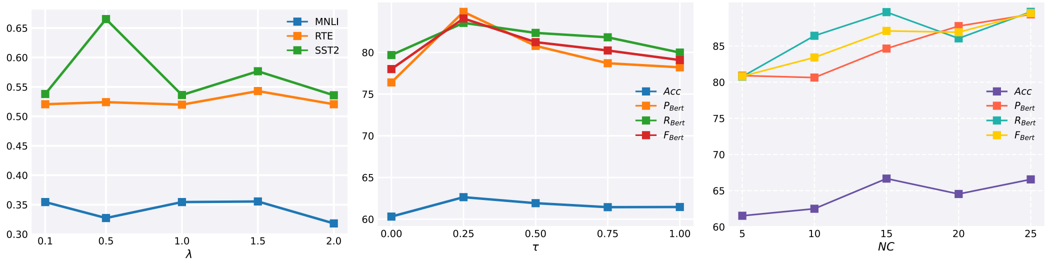
Sensitivity to λ. As shown in Figure 3a, when $FedA^2$-$LoRA$ restores $\boldsymbol{B}$, it does so by inverting it. A hyperparameter λ is introduced during **regularization**. Although mathematically, smaller values of λ are preferable, it grows exponentially during federated learning iterations. As a result, an excessively small λ can lead to memory overflows. To address this, we set $\lambda = 1$ and conducted a sensitivity analysis. Our results show that $FedA^2$-$LoRA$ is not significantly affected by λ, likely because the number of regularizations decreases when the matrix r of $\boldsymbol{B}$ and $\boldsymbol{A}$ is small. This was verified through testing on MNLI, RTE, and SST2 tasks.

Sensitivity to τ. As shown in Figure 3b, performance reaches its peak at $\tau = 0.25$. This may be attributed to the parameter ratio between the ImageBind and Vicuna models. At $\tau = 0$, only the base model is fine-tuned, with the encoder left unchanged. Conversely, at $\tau = 1$, only the encoder is fine-tuned, while the base model remains fixed. This observation indirectly highlights the effectiveness of two-stage training, where both components contribute to achieving optimal performance.

Sensitivity to NC. As shown in Figure 3c, the performance of the model varies with the number of clients (denoted as NC). The results indicate that as the number of clients increases from 5 to 25, the accuracy (Acc) and other metrics (such as Precision, Recall, and F1 score) improve significantly, demonstrating the positive impact of adding more clients in FL. This suggests that increasing the number of participating clients helps enhance the **UniFLoW** to generalize, as it learns from a more diverse set of data sources. However, the improvement in performance starts to plateau beyond a certain point, indicating diminishing returns with a higher number of clients. These observations highlight the importance of optimizing the number of clients to balance performance gains with computational and communication efficiency.

5 Conclusion

In this work, we introduced **UniFLoW**, a unified federated framework for multimodal large language model fine-tuning. Our framework effectively addresses critical challenges in federated learning, including modality heterogeneity, the risk of overfitting with limited private data, and aggregation bias in LoRA-based fine-tuning. Through our novel and efficient scheme $FedA^2$-$LoRA$ and our stage training strategy, **UniFLoW** demonstrated its ability to effectively utilize fragmented multimodal data without incurring additional communication costs. Our extensive experiments across



(a) Sensitivity to λ

(b) Sensitivity to τ

(c) Sensitivity to NC

Figure 3: Sensitivity analysis of the performance with respect to different hyperparameters. (a) Sensitivity to λ, showing the performance variation across different tasks (MNLI, RTE, and SST2). (b) and (c) Sensitivity to τ and NC (the Number of Clients), demonstrating the impact on Acc, P_{Bert},R_{Bert}, and F_{Bert}.

speech, image, and text modalities have shown that **UniFLoW** not only achieves state-of-the-art performance but also maintains consistency and efficiency in aggregation. The II stage training approach we implemented proved crucial in mitigating overfitting and ensuring the proper integration of both modality-specific and content-specific information. We believe **UniFLoW** offers a promising direction for unlocking the full potential of private multimodal data. This is especially relevant in an era where public resources for large-model training are becoming increasingly scarce. Looking ahead, future research could explore the application of **UniFLoW** to a wider range of modalities and investigate its scalability with a larger number of clients.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Jieming Bian, Yuanzhe Peng, Lei Wang, Yin Huang, and Jie Xu. A survey on parameter-efficient fine-tuning for foundation models in federated learning. *arXiv preprint arXiv:2504.21099*, 2025.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Terence Jie Chua, Wenhan Yu, Jun Zhao, and Kwok-Yan Lam. Fedpeat: Convergence of federated learning, parameter-efficient fine tuning, and emulator assisted tuning for artificial intelligence foundation models with mobile edge computing. *arXiv preprint arXiv:2310.17491*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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. 4171–4186, 2019.

Moming Duan, Duo Liu, Xianzhang Chen, Renping Liu, Yujuan Tan, and Liang Liang. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Transactions on Parallel and Distributed Systems*, 32(1):59–71, 2020.

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 ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4035–4045, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Charles Groetsch. The theory of tikhonov regularization for fredholm equations. *Boston Pitman Publication*, 104, 1984.

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.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

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*, 1(2):3, 2022.

Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16312–16322. IEEE, 2023.

Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. Federated learning for generalization, robustness, fairness: A survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9387–9406, 2024.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. *Advances in neural information processing systems*, 34:18590–18602, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

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. *arXiv preprint arXiv:2309.00363*, 2023.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39, 2024.

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

Haoyuan Liang, Xinyu Zhang, Shilei Cao, Guowen Li, and Juepeng Zheng. Tta-feddg: Leveraging test-time adaptation to address federated domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18658–18666, 2025.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*, 2024.

Xuezheng Liu, Zhicong Zhong, Yipeng Zhou, Di Wu, Xu Chen, Min Chen, and Quan Z Sheng. Accelerating federated learning via parallel servers: A theoretically guaranteed approach. *IEEE/ACM Transactions on Networking*, 30(5):2201–2215, 2022.

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.

Kai Lv, Yuqing Yang, Tengxiao Liu, Qinghui Gao, Qipeng Guo, and Xipeng Qiu. Full parameter fine-tuning for large language models with limited resources. *arXiv preprint arXiv:2306.09782*, 2023.

Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26752–26762, 2024.

Nattaya Mairittha, Tittaya Mairittha, and Sozo Inoue. Improving activity data collection with on-device personalization using fine-tuning. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pp. 255–260, 2020.

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

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models. In *International Conference on Machine Learning*, pp. 27011–27033. PMLR, 2023.

Atit Pokharel, Ratun Rahman, Thomas Morris, and Dinh C Nguyen. Quantum federated learning for multimodal data: A modality-agnostic approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 545–554, 2025.

Jiaxing Qi, Zhongzhi Luan, Shaohan Huang, Carol Fung, Hailong Yang, and Depei Qian. Fdlora: Personalized federated learning of large language model via dual lora tuning. *arXiv preprint arXiv:2406.07925*, 2024.

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International conference on machine learning*, pp. 18250–18280. PMLR, 2022.

Xuankun Rong, Wenke Huang, Jian Liang, Jinhe Bi, Xun Xiao, Yiming Li, Bo Du, and Mang Ye. Backdoor cleaning without external guidance in mllm fine-tuning. *arXiv preprint arXiv:2505.16916*, 2025.

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

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Guangyu Sun, Matias Mendieta, Aritra Dutta, Xin Li, and Chen Chen. Towards multi-modal transformers in federated learning. In *European Conference on Computer Vision*, pp. 229–246. Springer, 2024a.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

Zhen Wang, Daniyal M Alghazzawi, Li Cheng, Gaoyang Liu, Chen Wang, Zeng Cheng, and Yang Yang. Fedcsa: Boosting the convergence speed of federated unlearning under data heterogeneity. In *2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pp. 388–393. IEEE, 2023.

Xinghao Wu, Xuefeng Liu, Jianwei Niu, Haolin Wang, Shaojie Tang, and Guogang Zhu. Fedlora: When personalized federated learning meets low-rank adaptation. 2024.

Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. Heysquad: A spoken question answering dataset. *arXiv preprint arXiv:2304.13689*, 2023.

Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*, 2025.

Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, and Yanfeng Wang. Feddisco: Federated learning with discrepancy-aware collaboration. In *International Conference on Machine Learning*, pp. 39879–39902. PMLR, 2023.

Rui Ye, Rui Ge, Xinyu Zhu, Jingyi Chai, Du Yaxin, Yang Liu, Yanfeng Wang, and Siheng Chen. Fedllm-bench: Realistic benchmarks for federated learning of large language models. *Advances in Neural Information Processing Systems*, 37:111106–111130, 2024a.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6137–6147, 2024b.

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

Edvin Listo Zec, Adam Breitholtz, and Fredrik D Johansson. Overcoming label shift in targeted federated learning. *arXiv preprint arXiv:2411.03799*, 2024.

Jianyi Zhang, Hao Yang, Ang Li, Xin Guo, Pu Wang, Haiming Wang, Yiran Chen, and Hai Li. Mllm-llava-fl: Multimodal large language model assisted federated learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4066–4076. IEEE, 2025.

Zhihua Zhang, Guang Dai, Congfu Xu, and Michael I Jordan. Regularized discriminant analysis, ridge regression and beyond. *The Journal of Machine Learning Research*, 11:2199–2228, 2010.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

UniFLoW: Universal Multi-Modal Federated LoRA Fine-Tuning Framework with Analytical Aggregation (Supplementary material)

Table of Contents in Appendix

A Proof of Lemma 1

Lemma 1. *the A matrices tend to capture more global and stable directions than the B matrices.*

Proof. Inspired by (Guo et al., 2024), we provide the following proof: since the computations of B and A are independent, we can fix one to optimize the other. First, we consider fine-tuning A with fixed $B = \mathbb{B}$. The loss function becomes:

$$\mathcal{L} = \mathbb{E}_{(X_{n_k},\mathcal{A}_{y_{n_k}})}[\|\mathcal{A}_{y_{n_k}} - (W_0 + \mathbb{B}A)X_{n_k}\|_2^2]. \tag{14}$$

Then, the gradients of Eq. (17) w.r.t. A is:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial A} &= \frac{\partial \mathbb{E}_{(X_{n_k},\mathcal{A}_{y_{n_k}})}[\|\mathcal{A}_{y_{n_k}} - (W_0 + \mathbb{B}A)X_{n_k}\|_2^2]}{\partial A} \\ &= \frac{\partial \mathbb{E}[\|W_t X_{n_k} - (W_0 + \mathbb{B}A)X_{n_k}\|_2^2]}{\partial A} \\ &= \frac{\partial \mathbb{E}[\|(W_0 + \Delta_W)X_{n_k} - (W_0 + B\mathbb{A})X_{n_k}\|_2^2]}{\partial A} \\ &= \frac{\partial \mathbb{E}[\|(\Delta_W - \mathbb{B}A)X_{n_k}\|_2^2]}{\partial A} \\ &= \mathbb{E}[2\mathbb{B}^T[(\Delta_W - \mathbb{B}A)X_{n_k}]X_{n_k}^T]\end{aligned} \tag{15}$$

To obtain the optimal A^*, we set Eq. (15) to zero, which means:

$$\begin{aligned}\mathbb{E}[2\mathbb{B}^T[(\Delta_W - \mathbb{B}A)X_{n_k}]X_{n_k}^T] &= 0 \\ 2\mathbb{B}^T\Delta_W\mathbb{E}[X_{n_k}X_{n_k}^T] - 2\mathbb{B}^T\mathbb{B}A\mathbb{E}[X_{n_k}X_{n_k}^T] &= 0 \\ \mathbb{B}^T\mathbb{B}A\mathbb{E}[X_{n_k}X_{n_k}^T] &= \mathbb{B}^T\Delta_W\mathbb{E}[X_{n_k}X_{n_k}^T] \\ A &= \mathbb{B}^\dagger\Delta_W.\end{aligned} \tag{16}$$

Thus, we obtain $A^* = \mathbb{B}^\dagger\Delta_W$.

Then, we consider fine-tuning B while freezing $A = \mathbb{A}$. We abstract the loss function as:

$$\mathcal{L} = \mathbb{E}_{(X_{n_k},\mathcal{A}_{y_{n_k}})}[\|\mathcal{A}_{y_{n_k}} - (W_0 + BA)X_{n_k}\|_2^2]. \tag{17}$$

Then, the gradient of Eq. (17) w.r.t. B is:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial B} &= \frac{\partial \mathbb{E}_{(X_{n_k},\mathcal{A}_{y_{n_k}})}[\|\mathcal{A}_{y_{n_k}} - (W_0 + B\mathbb{A})X_{n_k}\|_2^2]}{\partial B} \\ &= \frac{\partial \mathbb{E}[\|\mathcal{A}_{y_{n_k}} - (W_0 + B\mathbb{A})X_{n_k}\|_2^2]}{\partial B} \\ &= \frac{\partial \mathbb{E}[\|(W_0 + \Delta_W)X_{n_k} - (W_0 + B\mathbb{A})X_{n_k}\|_2^2]}{\partial B} \\ &= \frac{\partial \mathbb{E}[\|(\Delta_W - \mathbb{B}A)X_{n_k}\|_2^2]}{\partial B} \\ &= \mathbb{E}[2[(\Delta_W - B\mathbb{A})X_{n_k}](-X_{n_k}^T\mathbb{A}^T)] \\ &= \mathbb{E}[2(B\mathbb{A} - \Delta_W)X_{n_k}X_{n_k}^T\mathbb{A}^T].\end{aligned} \tag{18}$$

To obtain the optimal B^*, we set Eq. (18) to zero, which means:

$$\begin{aligned}\mathbb{E}[2(B\mathbb{A} - \Delta_W)X_{n_k}X_{n_k}^T\mathbb{A}^T] &= 0 \\ 2B\mathbb{A}\mathbb{E}[X_{n_k}X_{n_k}^T]\mathbb{A}^T - 2\Delta_W\mathbb{E}[X_{n_k}X_{n_k}^T]\mathbb{A}^T &= 0 \\ 2B\mathbb{A}\mathbb{E}[X_{n_k}X_{n_k}^T]\mathbb{A}^T - 2\Delta_W\mathbb{E}[X_{n_k}X_{n_k}^T]\mathbb{A}^T &= 0 \\ B\mathbb{A}\mathbb{E}[X_{n_k}X_{n_k}^T]\mathbb{A}^T &= \Delta_W\mathbb{E}[X_{n_k}X_{n_k}^T]\mathbb{A}^T \\ B &= \Delta_W\mathbb{E}[X_{n_k}X_{n_k}^T]\mathbb{A}^T(\mathbb{A}\mathbb{E}[X_{n_k}X_{n_k}^T]\mathbb{A}^T)^{-1}.\end{aligned} \tag{19}$$

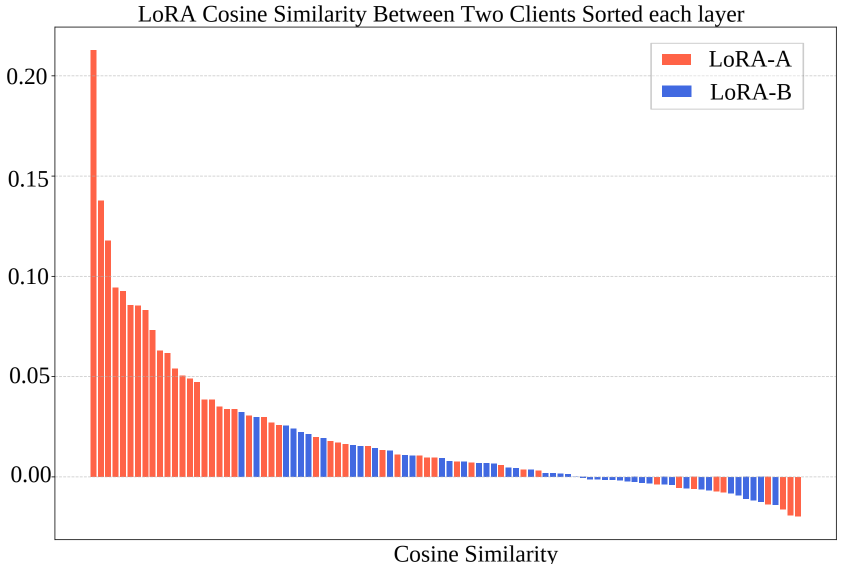


Figure 4: Similarity between LoRA-A and LoRA-B at each layer of the two clients

Therefore, we obtain $B^* = \Delta_W \mathbb{E}[X_{n_k} X_{n_k}^T]\mathbb{A}^T(\mathbb{A}\mathbb{E}[X_{n_k} X_{n_k}^T]\mathbb{A}^T)^{-1}$.

Therefore, we can derive the following relationship between A^* and B^*:

$$B^* = \mathbb{B}A^*\mathbb{E}[X_{n_k} X_{n_k}^T]\mathbb{A}^T(\mathbb{A}\mathbb{E}[X_{n_k} X_{n_k}^T]\mathbb{A}^T)^{-1} \tag{20}$$

Where $\mathbb{A}$ and $\mathbb{B}$ are Gaussian random initialization matrices.

From Eq. 20, we know that A^* only contains information unrelated to the data, which we consider to be general information. B^* can be written as a function of A^* and X_{n_k}, meaning it can capture both general and data information. Since A and B usually have the same rank, A^* contains more general information when acquiring the same amount of information.

Meanwhile, since B^* can be written as a function of A^* and X_{n_k}, it indicates that B also contains general information, making the aggregation of B necessary. Therefore, we propose $FedA^2$-$LoRA$ to fully capture global information by recovering $\Delta_W^* = \frac{1}{K}\sum_{k=1}^K \boldsymbol{B}_k\boldsymbol{A}_k$. □

B PROOF OF Eq. 13

$FedA^2$-$LoRA$ primarily aims to mitigate the problem of $\Delta_W = \left(\frac{1}{K}\sum_{k=1}^K \boldsymbol{B}_k\right)\left(\frac{1}{K}\sum_{k=1}^K \boldsymbol{A}_k\right) \neq \Delta_W^* = \frac{1}{K}\sum_{k=1}^K \boldsymbol{B}_k\boldsymbol{A}_k$. Therefore, we need to design B and A to restore u as much as possible, and thus we can express the optimization objective as Eq. 21.

$$\min_B \|\boldsymbol{BA} - U\|_F^2 \tag{21}$$

As shown in Appendix A, A more easily captures general information, while B contains both general and personalized information. Therefore, we consider first solving for $\boldsymbol{A}$ using a weighted average method, as shown in Eq. 9. Then, we only need to design the optimization function to solve for $\boldsymbol{B}$. The optimization function is as follows:

$$f(\boldsymbol{B}) = \|\boldsymbol{BA} - U\|_F^2 \tag{22}$$

We compute the derivative of the objective function with respect to $\boldsymbol{B}$, we have

$$
\begin{aligned}
\frac{\partial f}{\partial B} &= \frac{\partial \|\boldsymbol{B}\boldsymbol{A} - U\|_F^2}{\partial B} \\
&= \frac{\partial}{\partial B}\text{tr}((\boldsymbol{B}\boldsymbol{A} - U)^T(\boldsymbol{B}\boldsymbol{A} - U)) \\
&= \frac{\partial}{\partial B}\text{tr}(\boldsymbol{A}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{A} - 2U^T\boldsymbol{B}\boldsymbol{A} + U^TU) \\
&= 2\boldsymbol{B}\boldsymbol{A}\boldsymbol{A}^T - 2U\boldsymbol{A}^T
\end{aligned}
\tag{23}
$$

To obtain the optimal $\boldsymbol{B}$, we set Eq. 23 to zero, which means:

$$
\begin{aligned}
2\boldsymbol{B}\boldsymbol{A}\boldsymbol{A}^T - 2U\boldsymbol{A}^T &= 0 \\
\boldsymbol{B}\boldsymbol{A}\boldsymbol{A}^T &= U\boldsymbol{A}^T \\
\boldsymbol{B}^* &= U\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^\dagger
\end{aligned}
\tag{24}
$$

When $\boldsymbol{A}\boldsymbol{A}^T$ is invertible, the pseudoinverse is equal to the standard inverse ($(\boldsymbol{A}\boldsymbol{A}^T)^\dagger = (\boldsymbol{A}\boldsymbol{A}^T)^{-1}$) and is unique, allowing for a straightforward solution. However, when $\boldsymbol{A}\boldsymbol{A}^T$ is non-invertible (or singular), the solution (via the pseudoinverse) is non-unique, and direct (brute-force) solving often leads to an ill-conditioned matrix and numerically unstable results. Therefore, when $\boldsymbol{A}\boldsymbol{A}^T$ is non-invertible, we need to constrain the solution $\boldsymbol{B}$ through regularization (Tikhonov regularization), ensuring a unique and stable.Therefore, we can rewrite the optimization objective as follows:

$$
\min_{B} \|\boldsymbol{B}\boldsymbol{A} - U\|_F^2 + \lambda\|\boldsymbol{B}\|_F^2
\tag{25}
$$

By rewriting the optimization function we can obtain :

$$
F(\boldsymbol{B}) = |\boldsymbol{B}\boldsymbol{A} - U\|_F^2 + \lambda\|\boldsymbol{B}\|_F^2
\tag{26}
$$

We compute the derivative of the objective function with respect to $\boldsymbol{B}$, we have

$$
\begin{aligned}
\frac{\partial F}{\partial B} &= \frac{\partial[|\boldsymbol{B}\boldsymbol{A} - U\|_F^2 + \lambda\|\boldsymbol{B}\|_F^2]}{\partial B} \\
&= \frac{\partial}{\partial B}\text{tr}((\boldsymbol{B}\boldsymbol{A} - U)^T(\boldsymbol{B}\boldsymbol{A} - U)) + \lambda\frac{\partial}{\partial B}\text{tr}(\boldsymbol{B}^T\boldsymbol{B}) \\
&= \frac{\partial}{\partial B}\text{tr}(\boldsymbol{A}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{A} - 2U^T\boldsymbol{B}\boldsymbol{A} + U^TU + \lambda\boldsymbol{B}^T\boldsymbol{B}) \\
&= 2\boldsymbol{B}\boldsymbol{A}\boldsymbol{A}^T - 2U\boldsymbol{A}^T + 2\lambda\boldsymbol{B}
\end{aligned}
\tag{27}
$$

To obtain the optimal $\boldsymbol{B}$, we set Eq. 27 to zero, which means:

$$
\begin{aligned}
2\boldsymbol{B}\boldsymbol{A}\boldsymbol{A}^T - 2U\boldsymbol{A}^T + 2\lambda\boldsymbol{B} &= 0 \\
\boldsymbol{B}\boldsymbol{A}\boldsymbol{A}^T + \lambda\boldsymbol{B} &= U\boldsymbol{A}^T \\
\boldsymbol{B}(\boldsymbol{A}\boldsymbol{A}^T + \lambda I) &= U\boldsymbol{A}^T
\end{aligned}
\tag{28}
$$

The $\boldsymbol{A}\boldsymbol{A}^T$ is Symmetric Positive Semi-definite (PSD), because the matrix $\boldsymbol{A}\boldsymbol{A}^T$ is a Gram matrix formed by the product of a matrix and its transpose. For any non-zero vector v, the quadratic form $v^T(\boldsymbol{A}\boldsymbol{A}^T)v$ can be rewritten as:

$$
v^T(\boldsymbol{A}\boldsymbol{A}^T)v = (\boldsymbol{A}^Tv)^T(\boldsymbol{A}^Tv) = \|\boldsymbol{A}^Tv\|_2^2 \geq 0
$$

Since the squared L_2 norm is always non-negative, $\boldsymbol{A}\boldsymbol{A}^T$ is PSD.

And $\boldsymbol{A}\boldsymbol{A}^T + \lambda I$ is Positive Definite (PD) For any non-zero vector v, we examine the quadratic form v^TMv:

$$
\begin{aligned}
v^T(\boldsymbol{A}\boldsymbol{A}^T + \lambda I)v &= v^T(\boldsymbol{A}\boldsymbol{A}^T)v + v^T(\lambda I)v \\
&= \underbrace{\|\boldsymbol{A}^Tv\|_2^2}_{\geq 0} + \underbrace{\lambda\|v\|_2^2}_{>0 \text{ (since } \lambda>0)}
\end{aligned}
$$

Since $v^T(\boldsymbol{A}\boldsymbol{A}^T)v \geq 0$ and, critically, $\lambda\|v\|_2^2 > 0$ (for $v \neq 0$ and $\lambda > 0$), the sum is strictly positive:

$$
v^T(\boldsymbol{A}\boldsymbol{A}^T + \lambda I)v > 0
$$

Since the quadratic form $v^T(\boldsymbol{A}\boldsymbol{A}^T + \lambda I)v$ is strictly positive for all non-zero vectors v, the matrix $\boldsymbol{A}\boldsymbol{A}^T + \lambda I$ is Positive Definite (PD). All positive definite matrices are non-singular and therefore invertible.

Therefore, when $\boldsymbol{A}\boldsymbol{A}^T$ is irreversible, the optimal solution for B^* is as follows:

$$\boldsymbol{B}^* = U\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T + \lambda I)^{-1} \tag{29}$$

Therefore, when $\boldsymbol{A}\boldsymbol{A}^T$ is both reversible and irreversible, the optimal solution for $\boldsymbol{B}$, which is the $\boldsymbol{B}$ returned by our server, is as follows:

$$\boldsymbol{B} = \begin{cases} U\boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{A}^T + \lambda I]^{-1} & \text{, if } (\boldsymbol{A}\boldsymbol{A}^T)^{-1} \text{ does not exist} \\ U\boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{A}^T]^{-1} & \text{, if } (\boldsymbol{A}\boldsymbol{A}^T)^{-1} \text{ exists} \end{cases} \tag{30}$$

C FURTHER ANALYTICAL EXPERIMENTS

C.1 COST ANALYSIS

From Table 6, we can see that $FedA^2 - LoRA$ remains highly cost-effective in terms of overhead: the number of trainable parameters and the per-round communicated parameters are the same as standard LoRA (1.83M / 0.78M) and are much lower than FedEx-LoRA, which requires uploading the full model (25.74M). Therefore, it does not introduce additional communication or memory pressure. Although we introduce a closed-form aggregation step with computational complexity $O(Krd^2)$, this operation is executed only on the server side and remains manageable under a small LoRA rank r. In practice, this small extra computational cost yields more stable consistent aggregation and better federated performance.

Table 6: Time and space costs for each method on the RTE and QNLI tasks. # Communication round denotes the number of communication rounds to reach the predefined target performance.

	# Trainable Parm.	# Per-round Communicated Parm.	# Computational Complexity.
LoRA	1.83M	0.78M	$O(Krd)$
FFA-LoRA	1.44M	0.39M	$O(Krd)$
FedDPA-LoRA	2.62M	0.78M	$O(Krd)$
FedSA-LoRA	1.83M	0.39M	$O(Krd)$
FedEx-LoRA	1.83M	25.74M	$O(Krd^2)$
$FedA^2 - LoRA$	1.83M	0.78M	$O(Krd^2)$

C.2 OTHER INDICATORS

Table 7 shows that $FedA^2-LoRA$ consistently improves performance across all metrics—including BLEU, ROUGE-L, and METEOR—while maintaining strong results on both image and audio quality assurance. Compared with standard LoRA and other federated LoRA variants, our method achieves the highest or near-highest scores across both modalities. Notably, under the Stage-II setting, $FedA^2 - LoRA$ achieves:

These results demonstrate that our analytical aggregation method not only improves overall accuracy but also enhances output fluency (BLEU), semantic coverage (ROUGE-L), and paraphrase robustness (METEOR). The improvements remain consistent across different modalities, highlighting the effectiveness of FedA²-LoRA in multimodal federated learning scenarios.

C.3 PERFORMANCE UNDER THE SAME COMMUNICATION COST

As shown in Table 8, under the same communication cost (i.e., comparable numbers of communication rounds and per-round transmitted parameters), our proposed $FedA^2 - LoRA$ achieves the best overall performance across both image and audio modalities. For the *image* modality, $FedA^2 - LoRA$ attains 76.92 Acc and 85.93 F_{BERT}, which is competitive with or superior to all baselines while using only half the communication rounds of FFA-LoRA and FedSA-LoRA. For the *audio* modality, $FedA^2 - LoRA$ yields the highest Acc (78.84) and the best F_{BERT} (85.94), outperforming all competing methods by a clear margin. These results demonstrate that the proposed

Table 7: Performance Comparison of Different LoRA Methods on the Heysquad and PandaGPT's Datasets, with the **Image** and **Audio** Modality Client Participating in Training.'- -' indicates methods where no training was conducted.

Test Modality	Image			Audio		
Method	BLEU	ROUGE−L	METEOR	BLEU	ROUGE−L	METEOR
- -	31.62	50.14	50.42	29.94	53.33	45.39
LoRA	32.10	65.00	57.14	33.06	62.50	51.34
LoRA (II stage)	35.45	64.29	**57.74**	**36.86**	64.29	53.29
FFA-LoRA	33.49	62.50	50.76	32.06	65.00	53.57
FFA-LoRA (II stage)	35.88	65.38	52.24	35.53	62.50	55.38
FedSA-LoRA	31.13	62.50	50.20	31.58	61.90	50.26
FedSA-LoRA (II stage)	35.45	63.33	51.88	35.00	64.29	52.24
$FedA^2$-$LoRA$	36.85	64.29	54.59	33.86	65.00	54.17
$FedA^2$-$LoRA$ (II stage)	**38.96**	**66.67**	55.21	35.87	**67.86**	**55.88**

Table 8: Performance comparison across methods for Image and Audio modalities Under the same communication cost.

Method	Rounds	Image				Audio			
		Acc	P_{BERT}	R_{BERT}	F_{BERT}	Acc	P_{BERT}	R_{BERT}	F_{BERT}
LoRA	33	75.00	80.14	86.12	83.03	72.22	79.88	**87.46**	83.46
FFA-LoRA	66	**78.57**	86.74	**86.85**	**86.63**	76.92	85.18	86.54	85.72
FedSA-LoRA	66	72.72	**86.88**	84.35	85.48	78.26	**86.33**	84.43	85.37
FedEx-LoRA	1	61.53	79.78	80.91	80.31	55.56	79.32	84.18	81.68
$FedA^2$-$LoRA$	33	76.92	85.58	86.16	85.93	**78.84**	84.94	86.97	**85.94**

analytical aggregation not only preserves the communication efficiency of standard LoRA, but also delivers consistently stronger multimodal performance under realistic federated constraints.

C.4 HETEROGENEITY-AWARE MECHANISMS

To better reflect realistic heterogeneous-client scenarios, we further evaluate a heterogeneity-aware variant of **UniFLoW**. In this setting, different clients adopt different LoRA ranks according to their local data scale and computational budget we simulate five clients whose encoders use ranks r=4 or r=8, and we employ zero-padding to handle heterogeneous sequence lengths. As reported in Table 9, the proposed $FedA^2 - LoRA(zero - padding)$ consistently outperforms standard $LoRA(zero - padding)$ on both modalities: for images, it improves Acc and F_{BERT} from 61.64/82.34 to 61.92/84.46, and for audio from 56.27/82.22 to 57.93/83.46. These results demonstrate that our analytical aggregation remains effective when LoRA ranks and data sizes vary across clients, and that UniFLoW naturally extends to heterogeneity-aware configurations without sacrificing performance.

C.5 THE PERFORMANCE OF $FedA^2 - LoRA$ ON IMAGE AND AUDIO MODALITIES

As summarized in Table 10, $FedA^2 - LoRA$ consistently achieves the best overall performance among all federated LoRA baselines on both *image* and *audio* modalities. In particular, it substantially improves accuracy and F_{BERT} over standard LoRA, FFA-LoRA, FedSA-LoRA, and FedEx-LoRA, while maintaining the same communication budget. These results demonstrate that $FedA^2 - LoRA$ serves as a state-of-the-art federated LoRA approach for multimodal settings, effectively handling heterogeneous modality clients without sacrificing performance.

C.6 ABLATION STUDY ON USING TIKHONOV REGULARIZATION

Without regularization, the solved matrix $\boldsymbol{B}$ can become ill-conditioned, which may accumulate numerical errors across communication rounds and eventually degrade model performance. This

Table 9: Performance comparison of zero-padding variants on Image and Audio modalities.

Method	Image				Audio			
	Acc	P_{BERT}	R_{BERT}	F_{BERT}	Acc	P_{BERT}	R_{BERT}	F_{BERT}
–	61.23	78.59	81.74	80.09	50.19	78.96	81.71	80.02
LoRA (zero-padding)	61.64	84.44	80.38	82.34	56.27	84.18	80.37	82.22
FedA^2-LoRA (zero-padding)	**61.92**	**86.87**	**82.21**	**84.46**	**57.93**	**86.94**	**80.26**	**83.46**

Table 10: Performance comparison across methods on Image and Audio modalities.

Method	Image				Audio			
	Acc	P_{BERT}	R_{BERT}	F_{BERT}	Acc	P_{BERT}	R_{BERT}	F_{BERT}
–	61.23	78.59	81.74	80.09	50.19	78.96	81.71	80.02
LoRA	64.89	86.13	82.22	84.11	58.33	82.42	87.75	84.97
FFA-LoRA	63.15	85.18	79.27	82.08	57.87	91.20	77.04	83.26
FedSA-LoRA	66.29	84.22	83.15	84.22	58.33	82.32	78.93	83.67
FedEx-LoRA	66.67	81.21	85.81	83.45	58.33	83.92	83.82	83.79
$FedA^2 - LoRA$	**75.00**	84.51	84.34	84.43	60.17	84.97	85.83	**85.39**

effect is clearly reflected in Table 11, $FedA^2 - LoRA(\lambda = 0, w/oTik.)$ only yields marginal gains over vanilla LoRA and even harms some Image metrics (e.g., F_{BERT} drops from 84.11 to 80.96), indicating that the unregularized solution is unstable. In contrast, $FedA^2 - LoRA$ with Tikhonov regularization ($\lambda = 0.1$) achieves a substantial improvement, boosting Image Acc from 64.89 to 76.19 and Audio F_{BERT} from 84.97 to 86.31. These results corroborate the ridge-regression view of Eq.13: a mild Tikhonov term stabilizes the inversion of $AA^\top$, leading to more reliable global aggregation. Consistently, the trend observed here on SST2 matches Figure 3(a) in the main text, where a moderate λ yields the best overall performance.

C.7 Evaluation of alternative modality encoders and a different LLM

The results demonstrate that **UniFLoW** is model-agnostic and can be seamlessly applied to different multimodal encoder–LLM combinations.

To further assess the generality and modularity of UniFLoW, we evaluate the framework using *different encoder–LLM pairs*, replacing ImageBind with UniBindLyu et al. (2024) and substituting Vicuna-7B with LLaMA-1BZhao et al. (2024). As shown in Table 12, UniFLoW consistently improves performance across both image and audio tasks under these alternative backbones, confirming that the proposed analytical aggregation and two-stage training strategy are model-agnostic. Notably, $FedA^2 - LoRA$ and its II stage variant achieve the best overall results, demonstrating that UniFLoW remains effective even when the encoder and LLM architectures differ substantially from the default configuration. This highlights the flexibility and robustness of our framework in real-world multimodal federated settings.

C.8 Effectiveness of the II Stage Training Strategy

To further understand the role of the proposed II stage training strategy, we compare three scheduling variants: (i) simultaneous training of the encoder and LLM LoRA ($FedA^2 - LoRA$), (ii) a *reverse* schedule that first updates the LLM and then fine-tunes the encoder $FedA^2 - LoRA(Reverse)$, and (iii) our II stage strategy, which first calibrates the modality encoders and then updates the LLM. Intuitively, aligning the modality space in the first stage and only then letting the LLM learn semantics on top of these stabilized representations should be more favorable; reversing this order breaks this dependency and exposes the LLM to highly inconsistent, modality-biased features. The results in Table13 confirm this intuition: while the reverse variant brings only marginal improvements over the vanilla $FedA^2 - LoRA$ baseline, our II stage strategy consistently achieves the best performance on both image and audio metrics (e.g., Image Acc 76.19 and Audio F_{BERT} 86.31), demonstrating that encoder-first then LLM-second is a more effective training schedule in heterogeneous multimodal FL.

Table 11: Ablation study on Tikhonov regularization for recovering matrix $\boldsymbol{B}$.

Method	Image				Audio			
	Acc	P_{BERT}	R_{BERT}	F_{BERT}	Acc	P_{BERT}	R_{BERT}	F_{BERT}
LoRA	64.89	86.13	82.22	84.11	58.33	82.42	87.75	84.97
FedA^2-LoRA ($\lambda = 0$, w/o Tik.)	65.09	83.09	78.93	80.96	62.64	86.53	82.16	84.29
FedA^2-LoRA (Tikhonov, $\lambda = 0.1$)	**76.19**	83.44	**88.08**	85.69	60.90	**87.15**	85.49	**86.31**

Table 12: Generalization of UniFLoW with alternative encoders (UniBind) and LLMs (LLaMA-1B).

Method	Image				Audio			
	Acc	P_{BERT}	R_{BERT}	F_{BERT}	Acc	P_{BERT}	R_{BERT}	F_{BERT}
–	52.17	80.85	82.47	81.61	52.94	81.92	79.36	80.61
LoRA	63.15	83.80	**83.55**	83.67	61.90	85.16	77.49	81.14
FedA^2-LoRA	65.00	**88.41**	81.06	**84.58**	62.50	84.29	82.57	83.41
FedA^2-LoRA (II stage)	**66.67**	83.77	83.39	83.58	**63.63**	**86.01**	**84.01**	**85.00**

C.9 Impact of Training Order on Multimodal Federated Optimization

In heterogeneous multimodal FL, different clients hold different modalities (e.g., image, audio, and text), which induces highly inconsistent hidden representations across clients. If we update the encoders and the LLM LoRA simultaneously, these modality-specific representations drive the LLM in conflicting directions, causing it to overfit modality-specific noise and harming both convergence and generalization. To disentangle modality calibration and semantic modeling, we adopt a II stage training schedule: Stage I first fine-tunes the encoders to calibrate the modality space across clients, and Stage II then updates the LLM on top of these stabilized representations. We further compare this design with a coarse-grained schedule that trains the encoder for the first $T = 0.5$CR communication rounds(CR)and the LLM for the remaining rounds. As shown in Table 14 this T-based variant $FedA^2 - LoRA(T = 0.5CR)$ yields only limited improvements and is consistently worse than our proposed II-stage strategy, whereas $FedA^2 - LoRA$ (uppercaseii stage) achieves the best performance on both image and audio metrics (e.g., Image Acc 76.19 and Audio F_{BERT} 86.31). These results confirm that explicitly separating encoder alignment and LLM adaptation at the stage level is more effective than joint or coarse-grained training in heterogeneous multimodal federated settings.

D Use of Large Language Models (LLMs)

In the preparation of this manuscript, Large Language Models (LLMs) are employed as a general-purpose assistive tool aimed at enhancing the quality, clarity, and presentation of the writing. While the LLMs provide valuable support in specific areas, the core research, experimental design, data analysis, and intellectual contributions remain entirely the work of the authors.

The specific applications of LLMs in this work include:

- **Text Polishing and Refinement**: The LLM is used to review the manuscript for grammatical accuracy, enhance sentence structure, and ensure consistency in phrasing and tone throughout the paper. This process is similar to employing an advanced grammar and style checker, aimed at improving the readability, fluency, and overall quality of the manuscript. The model aids in refining language, ensuring it meets academic writing standards, while maintaining the integrity and originality of the authors' ideas.
- **Coherence and Logical Flow**: We use the LLM to help organize and structure our arguments more effectively. By presenting drafts of sections to the model, we receive suggestions on improving the logical transitions between paragraphs, identifying gaps in the narrative, and strengthening the overall flow. This assistance helps ensure that the document presents a coherent, well-structured, and compelling argument that enhances the readability for our audience.

Table 13: Comparison of different training strategies for FedA^2-LoRA.

Method	Image				Audio			
	Acc	P_{BERT}	R_{BERT}	F_{BERT}	Acc	P_{BERT}	R_{BERT}	F_{BERT}
–	61.23	78.59	81.74	80.09	50.19	78.96	81.71	80.02
LoRA	64.89	86.13	82.22	84.11	58.33	82.42	**87.75**	84.97
FedA^2-LoRA	75.00	84.51	84.34	84.43	60.17	84.97	85.83	85.39
FedA^2-LoRA (Reverse)	69.63	**86.71**	81.09	83.81	60.61	**87.33**	82.94	85.07
FedA^2-LoRA (II stage)	**76.19**	83.44	**88.08**	**85.69**	**60.90**	87.15	85.49	**86.31**

Table 14: Effect of different training schedules: simultaneous training, encoder-first (T = 0.5CR), and our II stage strategy.

Method	Image				Audio			
	Acc	P_{BERT}	R_{BERT}	F_{BERT}	Acc	P_{BERT}	R_{BERT}	F_{BERT}
–	61.23	78.59	81.74	80.09	50.19	78.96	81.71	80.02
LoRA	64.89	**86.13**	82.22	84.11	58.33	82.42	**87.75**	84.97
FedA^2-LoRA	75.00	84.51	84.34	84.43	60.17	84.97	85.83	85.39
FedA^2-LoRA (T = 0.5CR)	66.15	84.03	85.78	84.90	60.35	**88.98**	82.72	85.73
FedA^2-LoRA (II stage)	**76.19**	83.44	**88.08**	**85.69**	**60.90**	87.15	85.49	**86.31**

- **Supplementing and Articulating Ideas**: At various stages of the manuscript preparation, the LLM serves as a sounding board to supplement and articulate the authors' ideas. It assists in expressing complex thoughts more clearly, offering alternative ways to present concepts that are already formulated by the authors. The LLM does not contribute to the generation of new ideas or novel research findings but instead supports the authors in refining their expression, ensuring that their original insights are communicated effectively.

All suggestions and modifications proposed by the LLM are thoroughly reviewed, edited, and approved by the authors to ensure they accurately reflect the research's intent and underlying meaning. The final responsibility for the content, interpretation, and presentation of this paper rests solely with the authors.

E More related work

As privacy concerns grow, companies are increasingly reluctant to upload sensitive data to the cloud for centralized model training. This issue becomes more pronounced in the era of large language models (LLMs) and multi-modal learning, where private data spans diverse modalities such as text, speech, and images. To effectively utilize these distributed data, FedAvg (McMahan et al., 2017) pioneered the federated learning (FL) community and laid the foundation for Parallel Federated Learning (PFL) (Liu et al., 2022). Building upon this paradigm, most federated optimization algorithms improve FedAvg through various enhancements (Li et al., 2020; Duan et al., 2020; Karimireddy et al., 2020; Acar et al., 2021; Qu et al., 2022). For instance, SCAFFOLD (Karimireddy et al., 2020) mitigates client drift through control variates, while FedSAM (Qu et al., 2022) enhances model generalization under heterogeneous client distributions using the Sharpness Aware Minimization optimizer.

Although these methods substantially advance the handling of traditional non-IID issues, they remain limited when applied to complex tasks involving large models and multimodal data. Specifically, challenges such as domain shift (Huang et al., 2023) and category shift (Zec et al., 2024) become more severe as model scales and modality heterogeneity increase. To address domain shift, FedCSA (Wang et al., 2023) employs model bias-based clustering to improve global model consistency, while FedDisco (Ye et al., 2023) mitigates poor convergence under category shift. However, these PFL approaches still fall short in fully exploiting client data in large-model scenarios, since each client model is restricted to its local modality-specific data, limiting cross-modal knowledge sharing and undermining the potential of LLM-based multi-modal learning.

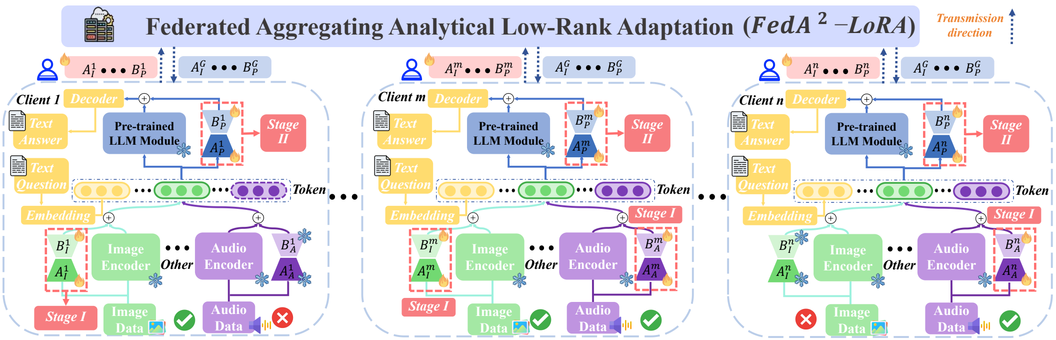


Figure 5: The Overview of the UniFLoW.

F Detailed flowchart of UniFlow

The inference process operates on each client, where local multi-modal data (e.g., text, images, or audio) is processed by modality-specific encoders (e.g., ImageBind), and the resulting features are then passed to a pretrained LLM (e.g., Vicuna-7B). The fine-tuning procedure is organized into two stages: in the I stage, we apply LoRA to the modality encoders, and in the II stage, we use LoRA to fine-tune the pretrained large language model. The server aggregates parameter updates from clients using the proposed $FedA^2$-$LoRA$ method, which ensures consistent and efficient updates while minimizing communication overhead. This framework is designed to handle heterogeneous data sources across clients, thereby enhancing scalability and model performance across different modalities.

G Dataset detail

$BERTScore$ Devlin et al. (2019) is a metric designed to assess the quality of text generation by comparing the contextual embeddings of predicted and reference tokens using the BERT model. It calculates three components: precision (P), recall (R), and F1 score (F1), all based on the cosine similarity between the token embeddings of the predicted and reference sentences. Precision measures how similar the predicted tokens are to the reference tokens, while recall assesses how well the reference tokens are captured by the prediction. The F1 score provides a balanced measure by combining both precision and recall. By using embeddings that capture deeper semantic meaning, $BERTScore$ is particularly useful for tasks like text generation, machine translation, and summarization, where understanding the meaning behind words is more important than the exact word matches. This metric has been shown to align better with human judgment in evaluating the quality of generated text, making it a robust alternative to traditional word-overlap based metrics.

Token accuracy (Acc) (Jiang et al., 2021) is a straightforward metric used to evaluate the performance of a model at the token level. It measures the percentage of tokens (words or subwords) in the model's predicted output that match the corresponding tokens in the reference or ground truth. This metric is particularly useful for tasks such as machine translation, text generation, and token-level classification, where precision in individual token prediction is crucial. Unlike other metrics like sentence-level accuracy, token accuracy provides a more granular view of model performance, allowing for better insight into how well a model captures the structure and meaning of language at the token level.

$HeySQuAD$ (Wu et al., 2023) is to evaluate a model's ability to understand noisy spoken queries and provide accurate responses. By including both human-spoken and machine-generated questions, it helps assess how models handle variations in spoken language, making it a valuable resource for improving SQA systems. This dataset is particularly useful for training models that deal with real-world conversational scenarios, improving their robustness in noisy environments.

$PandaGPT's$ (Su et al., 2023) visual instruction dataset is designed to improve multimodal instruction-following models by providing a collection of image-language instruction pairs. This dataset enables models to learn how to process and respond to multimodal inputs, combining images with textual instructions and responses. It plays a crucial role in training models like PandaGPT to handle complex tasks across different modalities, enhancing their ability to generate responses based on visual and textual inputs.