FedNano: Toward Lightweight Federated Tuning for Pretrained Multimodal Large Language Models

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) excel in tasks like multimodal reasoning and cross-modal retrieval but face deployment challenges in real-world scenarios due to distributed multimodal data and strict privacy requirements. Federated Learning (FL) offers a solution by enabling collaborative model training without centralizing data. However, integrating MLLMs into FL introduces challenges such as high computational demands, limited client capacity, substantial communication costs, and heterogeneous client data. Existing FL methods, which require deploying full models on clients, are impractical in these settings. To address these limitations, we propose ***FedNano***, a novel FL framework that centralizes the LLM on the server while introducing *NanoEdge*, a lightweight module for client-specific adaptation. *NanoEdge* employs modality-specific encoders, connectors, and trainable *NanoAdapters* with low-rank adaptation, achieving a **95% reduction** in client-side model storage and a transmission overhead of just **0.01%** of model parameters. By transmitting compact updates of *NanoAdapters*, *FedNano* effectively handles client heterogeneity and resource constraints, providing a scalable, privacy-preserving solution for MLLM deployment. Experiments show that *FedNano* outperforms existing methods, bridging the gap between MLLM complexity and FL constraints and enabling efficient, decentralized multimodal AI systems.

1 Introduction

Multimodal Large Language Models (MLLMs) (Zhu et al., 2023; Liu et al., 2024b; Peng et al., 2023b; Alayrac et al., 2022; Li et al., 2023) excel in tasks like multimodal reasoning and cross-modal retrieval (Yin et al., 2024), making them indispensable for applications such as visual question answering (VQA) (Antol et al., 2015). However, deploying MLLMs in real-world scenarios poses significant challenges, particularly in decentralized environments where sensitive multimodal data resides on resource-constrained edge devices. The computational and storage demands of MLLMs make deploying the full model on clients infeasible, while privacy regulations limit centralized data aggregation. These constraints highlight the need for innovative solutions to enable MLLM deployment in distributed, privacy-sensitive systems.

Federated Learning (FL) (McMahan et al., 2017) offers a promising solution for privacy-preserving collaborative model training, enabling decentralized learning without requiring data centralization. However, existing FL methods (Yang et al., 2024; Yi et al., 2023; Zhang et al., 2024a; Chen et al., 2023; Che et al., 2024; Chen and Zhang, 2022) are poorly suited for MLLMs as they fail to address the following critical challenges. First, client data in FL is typically non-IID, with heterogeneous distributions that hinder global model alignment and degrade performance. Second, parameter-efficient fine-tuning (PEFT) techniques (Houlsby et al., 2019; Lester et al., 2021; Zaken et al., 2021; Hu et al., 2021), such as those used in FedDPA-F(Yang et al., 2024), pFedLoRA (Yi et al., 2023), and FedIT (Zhang et al., 2024a), reduce the number of trainable parameters but still require deploying the full MLLM—often exceeding 10 billion parameters—on clients, making them impractical for resource-constrained devices like mobile phones or IoT systems. Third, while PEFT reduces trainable parameters, transmitting substantial parameter updates to the server during each communication round imposes significant overhead, making it unsustainable in bandwidth-limited environments. Finally, existing methods rely on client devices for complex local optimizations, which are infeasible for resource-limited devices. Together, these limitations prevent the effective integration of MLLMs into FL frameworks.

To address these challenges, we propose ***Fed-***

Nano, a novel FL framework specifically designed for MLLMs. *FedNano* establishes a new design paradigm by centralizing the large language model (LLM) on the server, where it remains frozen to retain general-purpose capabilities, while deploying *NanoEdge*, a lightweight client-side adaptation module, for efficient local tuning. *NanoEdge* employs modality-specific encoders, connectors, and trainable *NanoAdapters* optimized using low-rank decomposition (Hu et al., 2021). This design eliminates the need to deploy the full MLLM on clients, reducing client-side storage requirements by over **95%**, as demonstrated in Tab. 1, and making it highly practical for mobile and IoT devices. Furthermore, *NanoEdge* transmits only compact updates of the *NanoAdapter* between clients and the server, achieving **an over 99% reduction** in the number of uploaded parameters compared to existing PEFT-based FL methods, e.g., FedDAT (Chen et al., 2023) and FedDPA-F (Yang et al., 2024) while retaining task-relevant information. By the transmission of compact updates of *NanoAdapter*, *FedNano* also enhances privacy protection, making it suitable for privacy-sensitive FL environments.

To address client heterogeneity, *FedNano* adapts Fisher Merging (Matena and Raffel, 2022) to align global updates with client-specific data distributions. This adaptation improves performance on non-IID datasets and outperforms traditional aggregation methods such as FedAvg (McMahan et al., 2017) and FedProx (Li et al., 2020). By integrating these innovations, *FedNano* effectively bridges the gap between the computational complexity of MLLMs and the constraints of FL, enabling efficient deployment in real-world scenarios.

Experiments across diverse MLLM and multimodal tasks demonstrate that *FedNano* not only outperforms existing methods but also significantly reduces resource and communication costs, enabling the scalable, efficient, and privacy-preserving deployment of MLLMs. This framework lays a strong foundation for advancing multimodal AI systems in decentralized real-world applications, including personalized healthcare, cross-device collaboration, and multimodal user interfaces.

The key contributions of this work are:

- Novel Federated Framework: We propose a new paradigm that centralizes the LLM on the server while enabling lightweight client-side adaptation through *NanoEdge*. This reduces client-side storage by over 95%, making it highly practical for resource-constrained devices.
- Efficient Communication with Low-Rank Updates: *FedNano* employs low-rank decomposition in *NanoAdapters*, achieving achieving an over 99% reduction in the number of transmitted parameters, allowing efficient deployment in bandwidth-constrained environments.
- Improved Generalization on Non-IID Data: We adapt Fisher Merging for federated learning, aligning global updates with client-specific distributions to significantly improve model performance on heterogeneous datasets.
- Comprehensive Validation: Extensive experiments on diverse multimodal tasks demonstrate the superior performance of *FedNano* and resource efficiency, establishing it as a scalable solution for deploying MLLMs in real-world decentralized applications.

Approach	Client Params	Server Uploads
FedNano	304.55M (4.30%)	1.05M (0.01%)
FedDPA-F	7222.81M (100%)	180.89M (2.50%)
Reduction	**↓ 95.8%**	**↓ 99.4%**

Table 1: Comparison of parameter distribution and communication efficiency between *FedNano* and FedDPA-F (Yang et al., 2024) on LLaVA-1.5-7B (Liu et al., 2024b). *Client Params* refers to parameters retained on client devices, while *Server Uploads* denotes parameter updates sent to the server per round. Both methods use adapters with rank 64. *FedNano* achieves a 95.8% reduction in client parameters and 99.4% in server uploads, highlighting its efficiency.

2 Related Work

2.1 Multimodal Large Language Models

MLLMs (Zhu et al., 2023; Liu et al., 2024b; Peng et al., 2023b; Alayrac et al., 2022; Li et al., 2023; Dai et al., 2023) extend LLMs (Touvron et al., 2023; Peng et al., 2023a; Bai et al., 2023) to process multimodal data by incorporating modality-specific encoders and connectors. Encoders map input modalities, such as images, into representations compatible with LLMs, while connectors align these representations with the embedding space of LLM. Recent advances focus on optimizing connectors for efficient alignments, such as the linear connector in MiniGPT-4 (Zhu et al.,

2023) and LLaVA (Liu et al., 2024b), as well as the lightweight MLP bridge employed in LLaVA-1.5 (Liu et al., 2024a). However, these centralized approaches are ill-suited for FL, where resource constraints and data heterogeneity demand lightweight and scalable solutions. *FedNano* addresses these challenges by centralizing the LLM and deploying *NanoAdapters* on clients, enabling efficient and scalable multimodal FL.

2.2 Parameter Efficient Fine-tuning

PEFT techniques (Houlsby et al., 2019; Lester et al., 2021; Zaken et al., 2021; Hu et al., 2021) adapt large pretrained models to downstream tasks by minimizing trainable parameters, significantly reducing computational costs compared to full-model fine-tuning. They can be categorized into additive methods, such as Adapters (Houlsby et al., 2019), which introduce trainable components like MLP layers, and Soft Prompts (Lester et al., 2021), which learn tunable embeddings prepended to the input; selective approaches, such as BitFit (Zaken et al., 2021), which update only specific parameters; and reparameterized methods, such as LoRA (Hu et al., 2021), which leverage low-dimensional spaces for efficient adaptations. While effective in centralized settings, PEFT faces challenges in FL due to the communication overhead of parameter updates, the heterogeneity of client datasets, and the impracticality of modifying model architectures on resource-constrained clients that cannot host full LLMs. *FedNano* addresses these challenges by introducing *NanoAdapters*, which enable efficient task-specific tuning while keeping the LLM frozen on the server, significantly reducing communication costs and ensuring scalability in FL environments.

2.3 Multimodal Federated Learning

The application of FL in multimodal models, particularly vision-language (VL) tasks, has gained significant attention for addressing data heterogeneity and client diversity (Yang et al., 2024; Yi et al., 2023; Zhang et al., 2024a; Chen et al., 2023; Che et al., 2024; Chen and Zhang, 2022). Early works, such as Yu et al. (Yu et al., 2023), focused on tackling modality and model disparities between servers and clients, while Chen et al. (Chen et al., 2023) introduced PEFT-based methods for efficiently fine-tuning VL models and addressing data heterogeneity issues in federated settings. To support these advancements, benchmarks from Feng et al. (Feng et al., 2023) and Xu et al. (Xu et al., 2024) provided standardized frameworks for evaluating multimodal learning under heterogeneous conditions. Building on these efforts in VL tasks, recent FL research has begun exploring the adaptation of MLLMs to decentralized and heterogeneous environments. Che et al. (Che et al., 2024) addressed the challenge of incomplete modalities in client-local data, while frameworks like FedMSplit (Chen and Zhang, 2022) and DisentAFL (Chen and Zhang, 2024) focused on resolving modality incongruities and asymmetrical knowledge sharing. Zhang et al. (Zhang et al., 2024b) introduced FedMLLM, targeting data heterogeneity and long-tailed distributions to enable effective fine-tuning of MLLMs on diverse datasets. However, existing approaches fail to address the resource-intensive nature of MLLMs in FL. Even with PEFT methods, deploying MLLMs on clients remains impractical due to their substantial computational and memory demands. Moreover, PEFT still requires transmitting a significant number of updated parameters, leading to considerable communication overhead. To overcome these limitations, *FedNano* centralizes the LLM on the server, drastically reducing client storage requirements and minimizing communication costs by restricting updates to lightweight parameters. This design enables efficient and scalable adaptation, extending the applicability of MLLMs to federated multimodal learning, even in resource-constrained environments.

3 Methodology

3.1 Problem Definition

This work addresses federated fine-tuning for multimodal large language models (MLLMs) in decentralized, data-heterogeneous environments. Each client k holds a private multimodal dataset $D^k = \{(v_i^k, q_i^k, a_i^k)\}$, comprising image-question-answer triplets. Due to data heterogeneity, the marginal distributions of v_i^k, q_i^k, and a_i^k vary across clients, leading to significant differences in both visual and textual feature spaces. Such heterogeneity introduces challenges for achieving consistent generalization across clients, as traditional aggregation strategies fail to reconcile diverse local updates.

Our objective is to collaboratively fine-tune a shared global foundation model f_θ for VQA (Antol et al., 2015). Following (Liu et al., 2024a), we formulate this as an open-ended generation problem, where the model generates free-form answers given

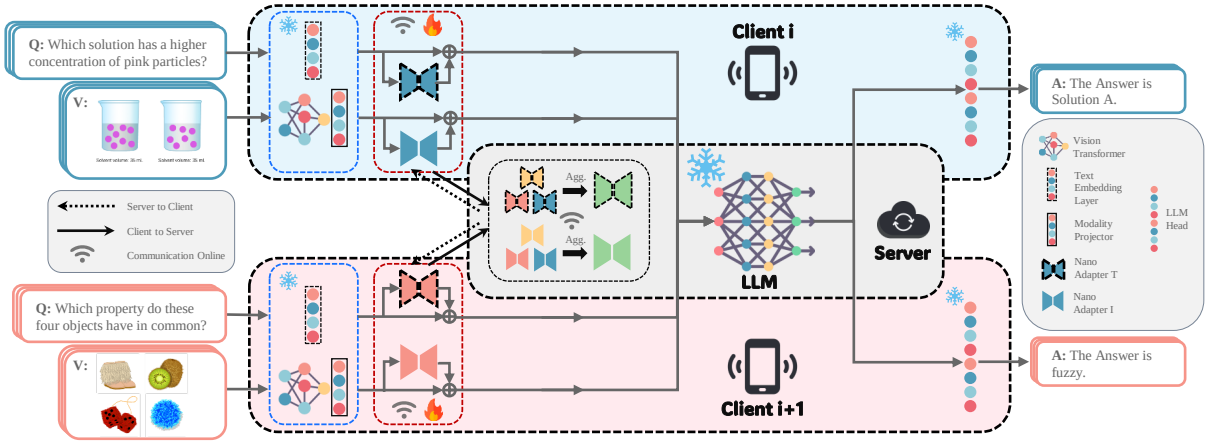


Figure 1: Overview of the *FedNano* framework. The server hosts the frozen LLM backbone, while clients handle lightweight modality-specific adapters. Each client adapts local tasks using compact updates, which are aggregated on the server to improve global performance. This design reduces computational overhead and communication costs, enabling scalable federated learning for multimodal tasks.

image-question pairs. Existing approaches struggle under the resource constraints of client devices, which include limited computational power and high communication costs. Furthermore, privacy concerns prevent raw data sharing, necessitating efficient mechanisms for client-server collaboration while preserving data privacy.

To address these challenges, we propose ***FedNano***, a parameter-efficient framework that centralizes the computationally intensive LLM on the server while enabling lightweight, client-specific tuning. In the following sections, we detail the design of *FedNano*, focusing on how it minimizes computational and communication overhead, and addresses data heterogeneity.

3.2 Overview of *FedNano* Architecture

FedNano is designed to address the inherent challenges of deploying MLLMs in FL environments. As illustrated in Fig.1, the computationally intensive LLM backbone, constituting the majority of the model parameters, is centralized on the server, while clients manage lightweight modules, including *NanoAdapters*, for task-specific updates. This design eliminates the need to deploy the full model on resource-constrained devices, alleviating the computational limitations of edge environments like mobile and IoT systems. By centralizing the LLM, *FedNano* significantly reduces the client-side burden while ensuring that raw data and sensitive computations remain local, preserving data privacy.

The architecture tackles four critical challenges in FL: high computational demands, communication overhead, data heterogeneity, and privacy concerns. Centralizing the LLM mitigates the computational burden on clients, requiring them to handle only *NanoEdge*, which involves less than 5% of the total model parameters. During model aggregation, clients upload only the *NanoAdapter*, which accounts for just 0.01% of the total model size. This lightweight communication design ensures reduced bandwidth usage, enabling efficient operation even in bandwidth-constrained settings. Privacy is further safeguarded as no raw data or labels are shared, and only task-specific updates are exchanged between clients and the server.

To address data heterogeneity, *FedNano* adapts Fisher Merging (Matena and Raffel, 2022) into FL as an advanced aggregation method, leveraging client-specific posterior distributions to balance global generalization with local adaptability. This approach ensures robust performance across diverse tasks and datasets, even in non-IID scenarios. By effectively integrating these design principles, *FedNano* bridges the gap between the computational challenge of deploying MLLMs and the practical constraints of FL, providing a scalable, efficient, and privacy-preserving framework for deploying MLLMs in decentralized systems.

3.3 *NanoEdge*: Client-Side Tuning Module

MLLMs are composed of three key components: modality encoders, a connector, and a pretrained LLM backbone. The modality encoders extract embeddings from raw inputs, such as images and text, while the connector aligns these embeddings into a unified representation compatible with the LLM. Together, these components enable MLLMs

to effectively handle diverse multimodal tasks by leveraging their pretrained capabilities.

Building on this structure, *NanoEdge* introduces *NanoAdapters* at the interface between the connector and the LLM to facilitate efficient task-specific tuning while preserving the pretrained alignment across modalities. By freezing the modality encoders and the connector, *NanoEdge* maintains their alignment with the LLM, ensuring the foundational structure of the pretrained model remains intact. This design allows *NanoAdapters* to focus solely on learning task-specific patterns from local client data and integrating federated knowledge updates, avoiding any disruption to the pretrained alignment. By restricting training to the lightweight *NanoAdapter* parameters, *NanoEdge* minimizes client-side computational demands while enabling efficient and privacy-preserving adaptation.

The *NanoAdapters* employ a low-rank decomposition mechanism, inspired by LoRA (Hu et al., 2021), consisting of a down-projection to reduce embedding dimensionality and an up-projection to restore it. This design balances parameter efficiency and adaptation capability, enabling *NanoEdge* to perform localized tuning and transmit updates efficiently. Each modality is equipped with a dedicated *NanoAdapter*—$\mathcal{A}_I$ for images and $\mathcal{A}_T$ for text—capturing modality-specific patterns essential for multimodal tasks.

Extensive ablation studies demonstrate the effectiveness of this design. Configurations combining both $\mathcal{A}_I$ and $\mathcal{A}_T$ consistently outperform those using a single adapter, highlighting the importance of modality-specific tuning. By transmitting only highly compact *NanoAdapter* updates, which account for just 0.01% of the total model size per round, *NanoEdge* achieves a 98% reduction in the number of uploaded parameters compared to traditional PEFT-based FL methods, while retaining task-relevant information. This compact communication strategy not only ensures scalability in bandwidth-constrained environments but also inherently minimizes the exposure of sensitive client-specific information.

3.4 Fisher-Guided Adaptive Aggregation

In FL, the aggregation process can be interpreted as maximizing the joint likelihood of the posterior distributions of all clients. The traditional method, FedAvg, can be viewed as equivalent to this process under the assumption of isotropic Gaussian posterior distributions for client models. (Matena and Raffel, 2022), which oversimplifies the process and struggles with data heterogeneity. *FedNano* overcomes this limitation by employing Fisher Merging (Matena and Raffel, 2022), leveraging the Laplace approximation for a more accurate estimation of posterior distributions. The aggregation formula is defined as:

$$\theta_{global}^{(i)} = \frac{\sum_{k=1}^{K} \frac{|D_k|}{\sum_{k=1}^{K}|D_k|} F_k^{(i)} \theta_k^{(i)}}{\sum_{k=1}^{K} \frac{|D_k|}{\sum_{k=1}^{K}|D_k|} F_k^{(i)}}, \quad (1)$$

where F is the Fisher Information Matrix (FIM), which serves as the precision matrix of the Laplace approximation, k denotes the k-th client, and i indexes the parameters. This approach improves the precision of maximizing the joint likelihood of the posterior distributions of all clients by providing a more accurate posterior distribution estimation for each client. As a result, it achieves superior FL performance, particularly in scenarios with non-IID data distributions.

To enhance scalability, *FedNano* introduces an efficient approach for FIM computation. We approximate the full FIM with its diagonal form (Kirkpatrick et al., 2017), reducing computational complexity from $O(|\theta|^2)$ to $O(|\theta|)$. Then, the diagonal FIM is computed directly from squared gradients during backpropagation (Wu et al., 2023), minimizing additional overhead while maintaining aggregation accuracy. By combining Fisher Merging with this efficient computation strategy, *FedNano* effectively addresses the dual challenges of heterogeneity and scalability in FL. Compared to uniform averaging in FedAvg, which fails to adapt to diverse client data, *FedNano* dynamically prioritizes impactful updates, achieving better generalization across non-IID datasets.

4 Experiment

4.1 Experimental Setup

Datasets and Partitioning We evaluate our approach on the Visual Question Answering (VQA) task using two established benchmarks: ScienceQA (Lu et al., 2022) and IconQA (Lu et al., 2021). These datasets were selected for their well-defined categorical structures and multimodal complexities, making them particularly suitable for assessing the performance of FL in non-IID settings. To simulate FL, we partitioned the datasets using Dirichlet distributions following (Che et al., 2023; Lai et al.,

Algorithm 1 *FedNano*

Server Update:

Randomly initialize θ^0, and distribute to clients
for $r = 1$ to R **do**
  for $k = 1$ to K **in parallel do**
    $\theta_k^r \leftarrow$ Client Update(θ^{r-1}, D_k)
    Compute FIM F_k
  end for
  $\theta^r \leftarrow$ ServerAgg$(\{\theta_k^r, F_k^r\})$ ▷ Eq. 1
end for

Client Update (θ^{r-1}, D_k):

$\theta_k^{r-1} \leftarrow \theta^{r-1}$
for local step $t = 1$ to T **do**
  Sample $\{\boldsymbol{X}, \boldsymbol{y}\}$ from D_k
  $\theta_k^{r(t)} \leftarrow$ Optimization$(\theta_k^{r(t-1)}, \boldsymbol{X}, \boldsymbol{y})$
end for
return θ_k^r

2022; Zhang et al., 2024a) with a concentration parameter $\alpha = 1$ to create strongly non-IID splits. Partitioning was guided by topic annotations in ScienceQA and skill annotations in IconQA, ensuring heterogeneous yet meaningful distributions across five simulated clients. Each partition, representing an individual client dataset, maintains consistent train-validation-test splits for evaluation.

Metrics Performance is evaluated using accuracy, computed following the official methodology specified for each dataset.

4.2 Implementation Details

Backbones We evaluate our approach on two state-of-the-art pretrained MLLMs, MiniGPT-4 (Zhu et al., 2023) and LLaVA-1.5 (Liu et al., 2024b). Both integrate a vision encoder with a pretrained LLM via lightweight connectors, enabling effective multimodal alignment and reasoning.

Baselines To the best of our knowledge, *FedNano* is the first framework designed to deploy MLLMs in FL by centralizing the LLM on the server. Given the absence of prior work addressing this setting, we evaluate *FedNano* against three widely recognized FL baselines: FedAvg (McMahan et al., 2017), a foundational aggregation method with limited handling of data heterogeneity; FedProx (Li et al., 2020), which mitigates client drift through a proximal term but lacks parameter-specific adaptation; and FedDPA-F (Yang et al., 2024), which integrates advanced alignment strategies but incurs high computational and communication overheads. We further include comparisons with a centralized model, representing the performance upper bound achieved with access to all data, and locally fine-tuned models, which operate in isolation without collaboration.

Training Configurations Each dataset partition is assigned to a single client, resulting in $K = 5$ clients. The training process includes 10 communication rounds ($R = 10$), with each client performing one local epoch per round using a batch size of 8. All experiments were conducted on NVIDIA A100 80G GPUs.

4.3 Main Results

The main results in Tab. 2 demonstrate that FL methods consistently outperform locally fine-tuned models, i.e., LocFT, highlighting the advantages of leveraging global knowledge to enhance client-specific performance in distributed and heterogeneous environments.

FedNano achieves the highest average performance across all FL methods, narrowing the gap with centralized training more effectively than baselines. While FedAvg performs competitively by employing simple weighted averaging, its lack of adaptability to non-IID data results in lower performance on heterogeneous clients. FedProx mitigates client drift by constraining local updates to remain close to the global model, but this rigid approach limits its ability to adapt to diverse local patterns, making it insufficient for complex multimodal tasks requiring flexibility. FedDPA-F, although designed for personalization, depends on careful tuning of the number of global training epochs. However, it carries the risk of forgetting the knowledge encoded in the global adapter, as the parameters of the global adapter are overwritten during subsequent local adapter training, leading to potential decreases in model performance.

In contrast, the superior results of *FedNano* are driven by its innovative design. The use of *NanoAdapters* for lightweight, modality-specific tuning significantly reduces client-side computational and storage burdens, enabling efficient deployment on resource-constrained devices. Additionally, the Fisher Merging mechanism leverages diagonal approximations of the FIM to prioritize critical updates, aligning client contributions with global objectives. These innovations allow *FedNano* to outperform baselines consistently across datasets, ad-

Backbone	Approach	ScienceQA (Clients)						IconQA (Clients)					
		C1	C2	C3	C4	C5	Avg	C1	C2	C3	C4	C5	Avg
MiniGPT-4	Centralized	73.70	88.34	89.83	84.52	87.41	84.76	80.76	86.62	81.16	82.74	85.36	83.33
	LocFT	67.74	74.69	77.42	72.46	74.07	73.28	67.70	73.48	70.63	70.86	77.53	72.04
	FedAvg	70.22	79.65	79.65	75.19	75.56	76.05	70.31	75.61	74.98	72.76	81.25	74.98
	FedProx	70.97	80.40	80.15	75.19	75.80	76.50	70.94	77.36	74.58	71.50	80.70	75.01
	FedDPA-F	**71.96**	78.41	**81.14**	76.42	75.80	76.75	70.94	**77.91**	74.51	73.08	80.30	75.35
	FedNano	68.98	**81.89**	80.89	**76.43**	**77.04**	**77.05**	**72.21**	77.28	**75.85**	**74.27**	**82.52**	**76.42**
LLaVA-1.5	Centralized	83.87	91.07	89.33	90.57	89.38	88.84	86.62	88.92	84.88	87.25	88.45	87.22
	LocFT	71.96	80.89	76.92	79.65	75.80	77.04	**75.93**	78.94	72.53	74.35	76.50	75.65
	FedAvg	73.20	**84.37**	83.62	82.13	**80.49**	80.76	71.18	79.89	76.80	**77.51**	**83.23**	77.72
	FedProx	73.95	**84.37**	83.87	81.39	80.00	80.71	70.23	80.13	76.72	**77.51**	82.36	77.39
	FedDPA-F	73.70	84.12	84.12	81.89	79.51	80.67	72.12	79.65	76.80	77.43	82.36	77.68
	FedNano	**74.94**	84.12	**84.86**	**82.88**	80.25	**81.41**	72.13	**80.44**	**77.36**	77.43	82.83	**78.04**

Table 2: Performance comparison of MiniGPT-4-7B and LLaVA-1.5-7B on ScienceQA and IconQA. Results include centralized training, local fine-tuning (LocFT), and various federated approaches. Metrics include individual client accuracies (C1–C5) and their average performance (Avg). *FedNano* achieves superior average performance on both datasets compared to other federated approaches, demonstrating its effectiveness in handling client heterogeneity.

dressing the challenges of data heterogeneity while maintaining scalability and communication efficiency. The results validate the effectiveness of *FedNano* in federated learning environments, highlighting its ability to balance generalization and personalization while overcoming limitations inherent in existing FL methods.

4.4 Ablation Studies

To illurstrate the importance of different decisions we made for *FedNano*, we conduct a series of ablation experiments.

The Necessity of Combining Both $\mathcal{A}_T$ and $\mathcal{A}_I$ To assess the necessity of textual $\mathcal{A}_T$ and visual $\mathcal{A}_I$ adapters, we conducted ablation experiments with configurations using only $\mathcal{A}_T$, only $\mathcal{A}_I$, or both. As shown in Tab. 3, combining both adapters consistently achieves the best performance, highlighting their complementary roles. The results show that $\mathcal{A}_I$ outperforms $\mathcal{A}_T$ individually, emphasizing its critical role in addressing the modality gap for visual embeddings. However, $\mathcal{A}_T + \mathcal{A}_I$ outperforms $\mathcal{A}_I$ alone, validating the importance of $\mathcal{A}_T$ in enhancing cross-client generalization through federated updates. In summary, $\mathcal{A}_I$ is essential for visual adaptation, while $\mathcal{A}_T$ refines textual updates, and their combination effectively addresses modality-specific challenges, achieving the best overall performance.

Trade-offs in Fisher-Guided Adaptive Aggregation FIM is specific to a particular set of model parameters and plays a key role in the ability of *FedNano* to achieve superior global alignment by capturing parameter importance. To compute the FIM precisely, *FedNano* employs additional forward and backward passes per communication round, ensuring accurate parameter estimation. While this enhances accuracy, it introduces modest computational overhead. To explore the trade-offs between precision and efficiency, we conduct an ablation study with *FedNano-EF*, a variant that approximates the FIM during standard training, eliminating the need for additional computation steps. This modification reduces computational overhead to the level of FedAvg. Despite this simplification, *FedNano-EF* incurs only a slight accuracy trade-off and consistently outperforms baselines, as shown in Tab. 4. These results demonstrate the adaptability of *FedNano*: the standard version excels in accuracy-critical tasks by leveraging

Backbone	Variants	ScienceQA	IconQA
MiniGPT-4	$\mathcal{A}_T$	45.91	57.77
	$\mathcal{A}_I$	74.57	75.17
	$\mathcal{A}_T + \mathcal{A}_I$	**76.42**	**76.04**
LLaVA-1.5	$\mathcal{A}_T$	50.08	48.15
	$\mathcal{A}_I$	77.03	77.12
	$\mathcal{A}_T + \mathcal{A}_I$	**78.04**	**77.83**

Table 3: Performance comparison of adapter configurations on ScienceQA and IconQA. Combining both textual adapter $\mathcal{A}_T$ and visual adapter $\mathcal{A}_I$ yields the best results, highlighting the importance of modality-specific adapters for improved multimodal performance.

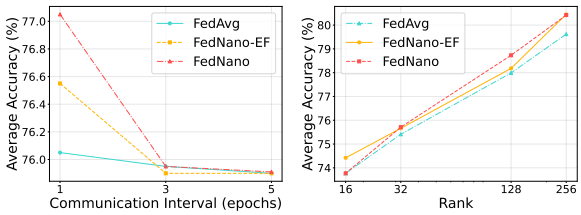
Dataset	Variants	MiniGPT-4	LLaVA-1.5
ScienceQA	*FedNano*	**77.05**	**81.41**
	FedNano-EF	76.55	80.81
	FedAvg	76.05	80.76
	FedProx	76.50	80.71
	FedDPA-F	76.75	80.67
IconQA	*FedNano*	**76.42**	**78.04**
	FedNano-EF	76.04	77.83
	FedAvg	74.98	77.72
	FedProx	75.01	77.39
	FedDPA-F	75.35	77.68

Table 4: Performance comparison of *FedNano* and *FedNano-EF* on ScienceQA and IconQA. *FedNano* achieves the highest accuracy, while *FedNano-EF* offers a trade-off with reduced computational overhead, demonstrating strong performance across both datasets.

precise FIM computation to optimize alignment, while *FedNano-EF* provides a practical alternative for resource-constrained environments, achieving strong performance with reduced overhead.

Higher Adapter Ranks Enhance *FedNano* Performance Fig. 2b illustrates the impact of adapter rank on model performance, comparing *FedNano* with FedAvg on the ScienceQA dataset. As the adapter rank increases, accuracy improves due to the enhanced capacity to encode task-specific and client-specific information, which is particularly important in non-IID settings. However, higher ranks also incur greater communication costs, necessitating a trade-off between performance and resource efficiency in FL. *FedNano* consistently outperforms FedAvg across all ranks, with the performance gap widening at higher ranks. This improvement is driven by the FIM aggregation, which leverages richer client-specific updates at higher ranks to achieve better alignment between local contributions and the global model. In contrast, at lower ranks, the limited adapter capacity constrains the quality of updates, reducing the effectiveness of FIM aggregation.

Frequent Communication Amplifies the Advantages of *FedNano* This study on the MiniGPT-4 backbone with the ScienceQA dataset evaluates the impact of communication frequency. As shown in Fig. 2a, reduced communication frequency leads to a general decline in global model performance across all methods due to increased parameter divergence, which hinders effective aggregation. Importantly, the results highlight that *FedNano* outperforms FedAvg by a larger margin when communication is more frequent. With shorter intervals, FIM mechanism of *FedNano* can better leverage aligned client parameters to prioritize impactful updates, amplifying its advantages in handling data heterogeneity. In contrast, FedAvg struggles with parameter divergence regardless of communication frequency, showing minimal improvement with more frequent updates. These findings underscore that while frequent communication benefits all methods, it significantly enhances the effectiveness of *FedNano*, reinforcing its superior ability to integrate client-specific updates and maintain robust performance in federated learning environments.



(a) Communication Freq. (b) Adapter Rank

Figure 2: (a) Impact of communication frequency on average accuracy. *FedNano* outperforms FedAvg, with more frequent communication amplifying its advantages; (b) Effect of adapter rank on accuracy. *FedNano* consistently achieves superior performance across all ranks, demonstrating its ability to capture task-specific and client-specific information effectively.

5 Conclusion

This work introduced *FedNano*, an FL framework designed to address the unique challenges of deploying MLLMs in decentralized settings. By centralizing the LLM on the server and employing lightweight *NanoAdapters* on clients, *FedNano* achieves significant efficiency in resource utilization and communication, while effectively handling data heterogeneity in non-IID environments. Comprehensive evaluations on ScienceQA and IconQA benchmarks demonstrate that *FedNano* consistently outperforms SOTA FL baselines, narrowing the gap between federated and centralized training paradigms. By combining scalable design with robust performance, *FedNano* provides a practical and privacy-preserving solution, paving the way for advancing MLLM deployment in real-world applications.

Limitation and Future Work

While FedNano demonstrates robust performance and efficiency, certain areas warrant further exploration to enhance its applicability and effectiveness. One limitation lies in the assumption that all clients possess similar hardware capabilities for managing NanoAdapters. This assumption may not hold in real-world scenarios characterized by highly heterogeneous devices. Future research could investigate adaptive mechanisms that dynamically tailor NanoAdapter configurations to match the computational resources and capabilities of individual clients, broadening FedNano usability across diverse environments.

Although FedNano effectively mitigates data heterogeneity, federated learning in real-world settings often involves extreme client disparities in data size, quality, and distribution. Addressing such scenarios may require dynamic strategies to adapt aggregation weights or incorporate more sophisticated representations of client-specific characteristics. These enhancements could further strengthen FedNano resilience and generalization capabilities in highly non-IID environments. Moreover, while the current framework supports vision and language modalities, extending it to incorporate audio, sensor data, or other modalities could unlock applications in areas such as autonomous systems, multimodal healthcare, and industrial IoT.

Deploying FedNano in noisy or incomplete federated datasets presents another promising avenue for research. Benchmarking its performance under these challenging conditions would not only provide valuable insights but also identify additional opportunities for optimization. Furthermore, integrating FedNano into federated multi-agent systems—where distinct agents collaborate to learn and share knowledge—could enable groundbreaking applications in fields like logistics and autonomous vehicles, highlighting the framework versatility.

Finally, while FedNano achieves strong privacy guarantees by transmitting only NanoAdapter updates, integrating advanced privacy-preserving methods such as differential privacy or secure multi-party computation could provide even stronger safeguards for sensitive client data. A critical future direction lies in achieving these enhanced privacy measures without compromising the computational and communication efficiency that underpins FedNano practicality.

In summary, while FedNano addresses many critical challenges in federated learning for MLLMs, these future directions highlight its potential for further innovation. By extending its capabilities to tackle more diverse environments, extreme heterogeneity, and advanced privacy requirements, FedNano can serve as a foundational framework that inspires continued advancements in federated learning research and applications.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Liwei Che, Jiaqi Wang, Xinyue Liu, and Fenglong Ma. 2024. Leveraging foundation models for multimodal federated learning with incomplete modality. *Preprint*, arXiv:2406.11048.

Tianshi Che, Ji Liu, Yang Zhou, Jiaxiang Ren, Jiwen Zhou, Victor S Sheng, Huaiyu Dai, and Dejing Dou. 2023. Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization. *arXiv preprint arXiv:2310.15080*.

Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. 2023. Feddat: An approach for foundation model finetuning in multimodal heterogeneous federated learning. *Preprint*, arXiv:2308.12305.

Jiayi Chen and Aidong Zhang. 2022. Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 87–96.

Jiayi Chen and Aidong Zhang. 2024. On disentanglement of asymmetrical knowledge transfer for modality-task agnostic federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11311–11319.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi.

2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. 2023. Fedmultimodal: A benchmark for multimodal federated learning. *Preprint*, arXiv:2306.09486.

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

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Fan Lai, Yinwei Dai, Sanjay Singapuram, Jiachen Liu, Xiangfeng Zhu, Harsha Madhyastha, and Mosharaf Chowdhury. 2022. Fedscale: Benchmarking model and system performance of federated learning at scale. In *International conference on machine learning*, pages 11814–11827. PMLR.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716.

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

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023a. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023b. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chengyue Wu, Teng Wang, Yixiao Ge, Zeyu Lu, Ruisong Zhou, Ying Shan, and Ping Luo. 2023. π-tuning: transferring multimodal foundation models with optimal multi-task interpolation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 37713–37727.

Binqian Xu, Xiangbo Shu, Haiyang Mei, Guosen Xie, Basura Fernando, Mike Zheng Shou, and Jinhui Tang. 2024. Fedmllm: Federated fine-tuning mllm on multimodal heterogeneity data. *arXiv preprint arXiv:2411.14717*.

Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. 2024. Dual-personalizing adapter for federated foundation models. *arXiv preprint arXiv:2403.19211*.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, page nwae403.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. 2023. Multimodal federated learning via contrastive representation ensemble. *Preprint*, arXiv:2302.08888.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. 2024a. Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6915–6919. IEEE.

Jianyi Zhang, Hao Frank Yang, Ang Li, Xin Guo, Pu Wang, Haiming Wang, Yiran Chen, and Hai Li. 2024b. Mllm-fl: Multimodal large language model assisted federated learning on heterogeneous and long-tailed data. *arXiv preprint arXiv:2409.06067*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.