ICML International Conference On Machine Learning

2026
POSTER

国家超级计算深圳中心 National Supercomputing Center in Shenzhen 深圳云计算中心 Shenzhen Cloud Computing Center

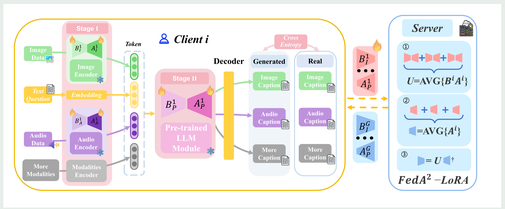
UniFLoW: Universal Multi-Modal Federated LoRA Fine-Tuning with Analytical Aggregation

Aggregation-consistent FedMLLM training over private image, audio, and text clients.

Haoyuan Liang · Zhiyu Ye · Jielong Tang · Yang Yang · Shilei Cao · Guowen Li · Fei Hu · Zhiwei Zhang · Haohuan Fu · Juepeng Zheng✉

Sun Yat-sen University · Tsinghua SIGS · National Supercomputing Center in Shenzhen · South China University of Technology

SYSU · Tsinghua · Wechat



UNIFLOW **Stage I aligns modality encoders, Stage II adapts the LLM, and FedA²-LoRA reconstructs a consistent global update.**

3 modalities image/audio/text	**10** clients per modality
0.78M params per round	**II** stage training

1 Why Federated MLLMs Are Hard

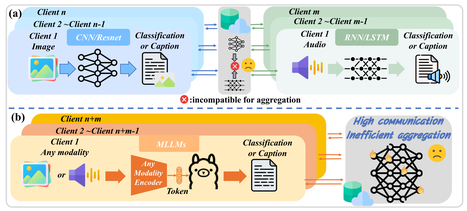
Private multimodal data is valuable but cannot simply be uploaded for centralized MLLM adaptation. Federated learning protects data locality, yet heterogeneous client modalities break classical parameter aggregation and billion-scale full tuning is too expensive to communicate.

- **Architecture mismatch:** image, audio, and text clients may use incompatible local models.
- **LoRA mismatch:** averaging low-rank factors does not recover the average full update.
- **Communication pressure:** full-parameter tuning sends far more than low-rank adapters.

Question. Can a single federated framework learn from fragmented multimodal clients while preserving LoRA efficiency?

2 Unify Modalities Before Aggregation

The paper motivates a shift from client-specific unimodal networks to a shared MLLM architecture with modality encoders. This makes federated training possible across clients that hold different private modalities.



Left: traditional FL cannot aggregate incompatible modality-specific models. **Right:** a unified MLLM accepts any modality but still needs communication-efficient, consistent LoRA aggregation.

3 Client-Side Training

Each client maps its local modality through ImageBind-style encoders, pools projected features into the LLM token space, and applies LoRA to both encoder and Vicuna modules.

LORA-INJECTED LAYER

$$\tilde{W}^{(m,l)} = W_0^{(m,l)} + \alpha B^{(m,l)} A^{(m,l)}$$

STAGE I	BRIDGE	STAGE II
freeze LLM adapt encoders	pool + project to LLM space	freeze encoders adapt LLM

This schedule first calibrates modality-specific representations, then lets the LLM learn on stabilized features rather than noisy modality drift.

4 FedA²-LoRA ★ KEY

Direct FedLoRA averaging is biased because multiplication happens after averaging the low-rank factors. UniFLoW instead averages the stable A direction and analytically solves for the matching B.

INCONSISTENCY

$$(1/K\,\Sigma_k B_k)(1/K\,\Sigma_k A_k) \neq 1/K\,\Sigma_k B_k A_k = \Delta W^*$$

ANALYTICAL RECONSTRUCTION

$$A = 1/K\,\Sigma_k A_k,\ U = \Delta W^*,\ B = UA^T(AA^T + \lambda I)^{-1}$$

Why it works. The analysis finds A captures more global directions, while B carries more client-specific signal; ridge recovery keeps the global update stable.

Server Aggregation Recipe

1. Average client A factors to keep the shared direction basis.
2. Average full LoRA updates as the target global update.
3. Solve the ridge system for B so BA matches that target.

5 Multimodal QA Results ★ Headline

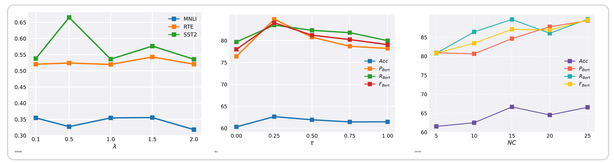
HeySQuAD audio QA + PandaGPT image QA; 10 communication rounds, 10 clients per modality, 2,000 train and 200 test samples per client.

Method	Img Acc	Img F_{BERT}	Aud Acc	Aud F_{BERT}
No train	61.23	80.09	50.19	80.02
LoRA	64.89	84.11	58.33	84.97
FedA²-LoRA	75.00	84.43	60.17	85.39
FedA²-LoRA (II)	**76.19**	**85.69**	**60.90**	**86.31**

+11.30 image Acc vs LoRA	**+1.34** audio F_{BERT} vs LoRA	**76.19** best image accuracy	**86.31** best audio F_{BERT}

6 Ablations & Sensitivity

Tikhonov regularization stabilizes the inverse in the B solve: the paper reports Image Acc rising from 64.89 for LoRA to 76.19 with $\lambda = 0.1$, while Audio F_{BERT} rises from 84.97 to 86.31.



Sensitivity analysis shows the two-stage split peaks around $\tau = 0.25$, and adding clients improves multimodal QA before diminishing returns.

Regularization is not cosmetic: without it, the recovered B becomes ill-conditioned and image F_{BERT} drops to 80.96 in the ablation.

10 rounds communication	**10 clients** per modality	**2k / 200** train/test

7 General Federated LoRA Gains

RoBERTa on GLUE; accuracy averaged across three runs. FedA² improves LoRA-family aggregation beyond multimodal QA.

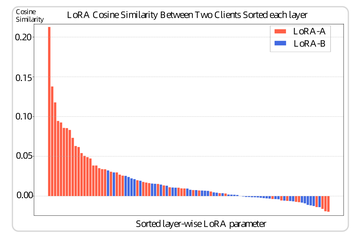
Family	Base Avg	FedA² Avg	Gain
LoRA	89.17	90.50	+1.33
rsLoRA	**89.16**	**90.68**	+1.52
VeRA	86.86	88.02	+1.16

LoRA	**0.78M** communicated params / round
FedEx	**25.74M** communicated params / round
FedA²	**0.78M** communicated params / round with analytical recovery

FedA²-LoRA keeps LoRA-level communication while avoiding FedEx-style residual transmission.

8 Heterogeneity & Evidence

The appendix tests heterogeneous ranks ($r = 4$ or $r = 8$) with zero-padding. FedA²-LoRA improves image F_{BERT} from 82.34 to 84.46 and audio F_{BERT} from 82.22 to 83.46.



Layer-wise similarity evidence supports the design choice: aggregate the more global A matrices, then recover B analytically.

The same aggregation idea works across multimodal QA, GLUE LoRA variants, and heterogeneous-client settings.

9 Takeaways

IDEA. Unified MLLM clients. | **METHOD.** Average A; solve B. | **RESULT.** Best image/audio QA. | **USE.** Private multimodal adaptation.

UniFLoW