pFedMMA

Personalized Federated Fine-Tuning with Multi-Modal Adapter for Vision-Language Models

Sajjad Ghiasvand, Mahnoosh Alizadeh, & Ramtin Pedarsani

Department of Electrical and Computer Engineering
UC Santa Barbara

ICLR 2026

The Challenge: Federated VLM Adaptation

Problem

CLIP achieves impressive zero/few-shot generalization

Privacy constraints prevent centralized training

Fine-tuning on decentralized, heterogeneous data is challenging

Gap in Existing Methods

Prompt tuning methods sacrifice generalization for personalization

FedOTP: 97% local acc but only 31% HM acc

Poor performance on unseen classes/domains

Goal: Achieve strong personalization AND generalization with communication efficiency

pFedMMA: Our Approach

1

Multi-Modal Adapter

Insert lightweight adapters in upper layers of both image & text encoders. Each has down-projection, shared projection, and up-projection.

2

Asymmetric Updates

Clients update ALL adapter params locally. Only the shared projection is uploaded & aggregated globally via FedAvg.

3

Communication Efficient

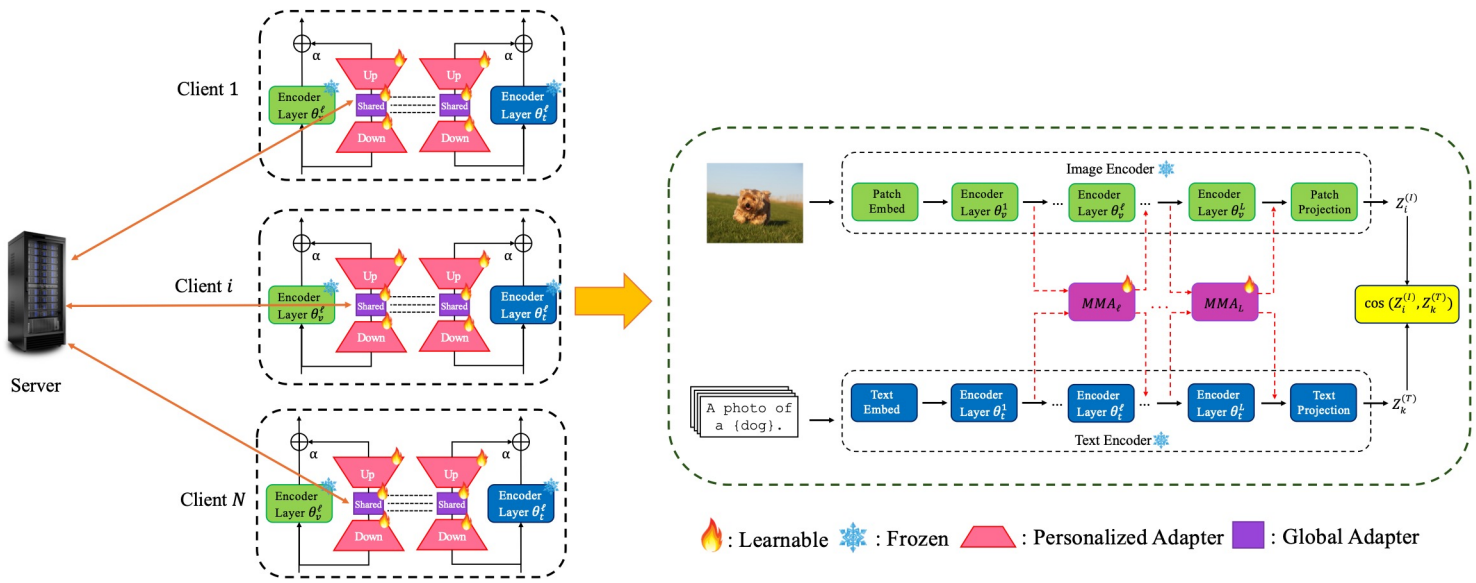
Only 3,072 params/round communicated (vs 8,192+ for baselines). Local adapters stay on-device for personalization.

$$\text{Output}(x) = f(x) + \alpha \cdot A(x), \quad \text{where } A(x) = \text{Up}(\delta(\text{Shared}(\text{Down}(x))))$$

pFedMMA Framework Overview

ICLR 2026

Each client trains local (up/down) + shared adapters. Only the shared projection is uploaded to the server for global aggregation.



1 Local Adapters

Up/down projections stay on-device. Each client personalizes to its own data distribution.

2 Shared Projection

Global cross-modal alignment layer aggregated by the server via FedAvg each round.

3 Communication-Efficient

Only 3,072 params/round transmitted — the shared projection — 60% less than baselines.

Results: Personalization & Generalization

ICLR 2026

84.15%
HM Accuracy
Best across all methods

97.17%
Local Accuracy
Strong personalization

+6.4%
vs Best Baseline
Over FedPGP

16-shot, 7 Datasets (ViT-B/16) — HM Accuracy

Method	Local	Base	Novel	HM ↑	Δ HM
pFedMMA (Ours)	**97.17**	**77.40**	**81.49**	**84.15**	—
FedPGP	95.38	76.49	71.68	79.09	-5.1
pFedMoAP	97.89	61.82	66.60	71.05	-13.1
FedOTP	97.34	18.00	36.69	31.08	-53.1
PromptFL	88.93	88.95	75.36	83.09	-1.1
CLIP (zero-shot)	76.36	76.81	81.21	78.03	-6.1

Results: Domain Generalization & Personalization

Feature Shift: DomainNet (β=0.5)

Method	Avg Acc ↑
pFedMMA (Ours)	**47.17%**
pFedMoAP	24.65%
FedPGP	24.90%
PromptFL	11.24%
CoCoOp+FedAvg	18.25%
CLIP (zero-shot)	10.11%

+22.3% over best competitor on DomainNet ✓

Label Shift: CIFAR-100 (100 clients, Dirichlet β)

Method	β=0.1	β=0.5	β=1
pFedMMA (Ours)	**81.82**	**76.92**	**75.70**
pFedMoAP	80.29	75.68	74.53
FedOTP	77.53	72.21	70.99
FedPGP	74.72	74.85	74.18
PromptFL	75.34	72.85	72.83
CLIP (zero-shot)	73.14	69.53	64.95

Best accuracy across all β settings on CIFAR-100 ✓

Why pFedMMA? — Communication & Computational Efficiency

Params Communicated (up+down per round)

3,072 / 3,072

vs 8,192 for baselines

Best HM Accuracy

84.15%

+6.4% vs next best

Local Trainable Params

248,832

all kept on-device

GPU Memory

4,634 MiB

4th lowest overall

Full Cost Analysis (ViT-B/16)

Method	Local Params	Comm. (↑/↓)	Train Time (s)	GPU (MiB)	Local Acc	HM Acc ↑
PromptFL	8,192	8K / 8K	1,645	5,116	88.93	83.09
FedPGP	12,416	8K / 8K	3,980	13,374	95.38	79.09
FedOTP	16,384	8K / 8K	1,328	3,014	97.34	31.08
pFedMoAP	74,240	8K / 74K	902	3,108	97.89	71.05
pFedMMA (Ours)	**248,832**	**3K / 3K ★**	**2,175**	**4,634**	**97.17**	**84.15**

★ pFedMMA sends only 3K params/round (60% less than baselines), yet achieves the highest HM accuracy — the best accuracy-efficiency trade-off.

Key Takeaways

1 Multi-Modal Adapter (MMA)

Lightweight visual/textual up/down projection adapters in upper CLIP layers. Effective fine-tuning without prompt engineering.

2 Asymmetric Personalization

Local adapters personalize to each client's distribution. Shared adapter aligns text-image features globally via federation.

3 Best Accuracy Trade-off

84.15% HM accuracy — +6.4% over nearest competitor across 7 datasets.

4 Communication Efficient

Only 3,072 params per round — 60% less than prompt-tuning baselines. Scales to heterogeneous federated settings.

Code: github.com/sajjad-ucsb/pFedMMA

ICLR 2026