南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

FedMVP: Federated Multimodal Visual Prompt Tuning for Vision-Language Models

Mainak Singha[1] Subhankar Roy[2] Sarthak Mehrotra[3] Ankit Jha[4]
Moloud Abdar[5] Biplab Banerjee[3] Elisa Ricci[1,6]

[1] University of Trento, Italy [2] University of Bergamo, Italy [3] Indian Institute of Technology Bombay, India
[4] LNMIIT Jaipur, India [5] The University of Queensland, Australia [6] Fondazione Bruno Kessler, Italy

ICCV 2025

汇报人：白楚榆

Background

CLIP：用类别文本提示做分类

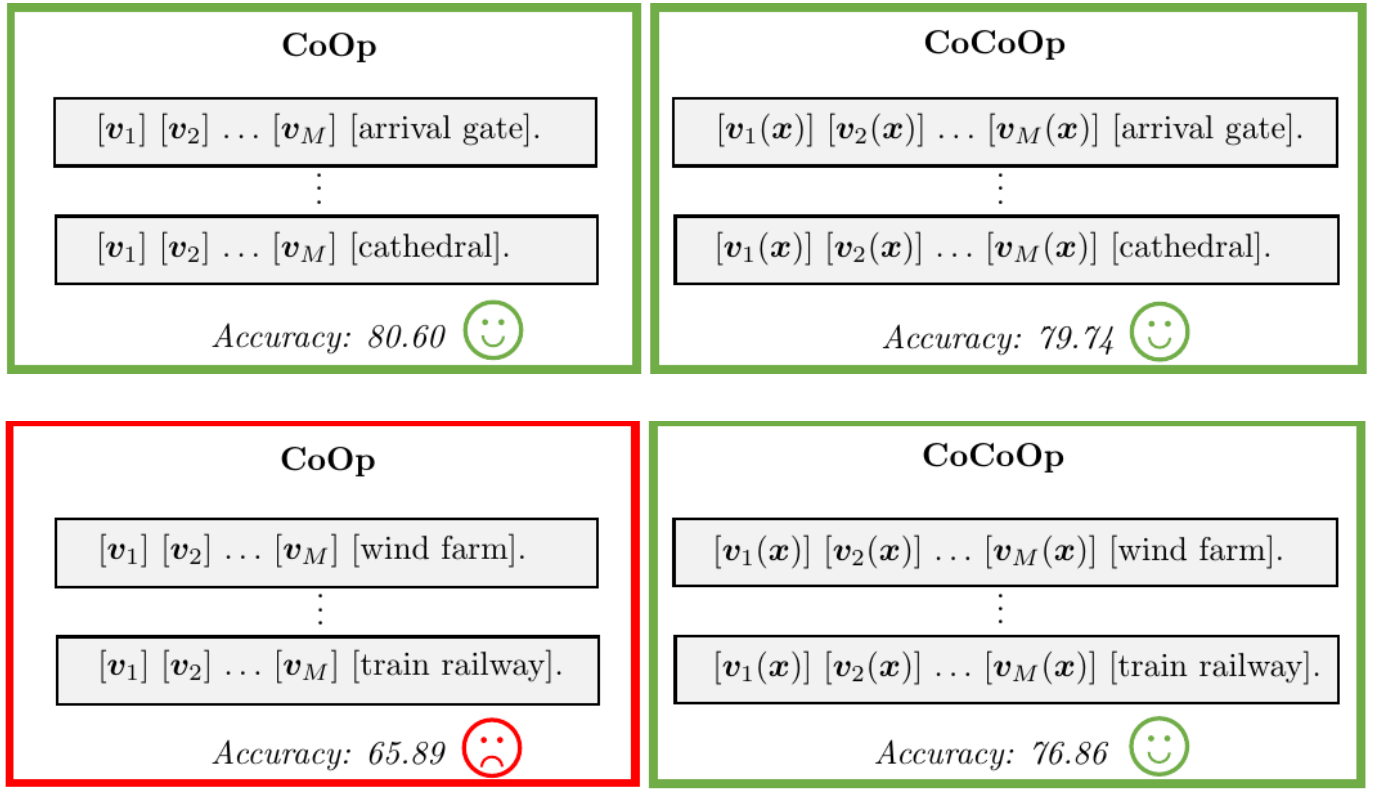
例如 “a photo of a [class]”

CoOp：学习固定文本提示

训练图片共享同一组提示

问题：训练类 Base 好，未见类 New 呢?

上： Base （训练类） 下： New （未见类）



CoOp 在 New 上明显下降；下一页看 CoCoOp 怎样让提示随图像变化。

CoCoOp

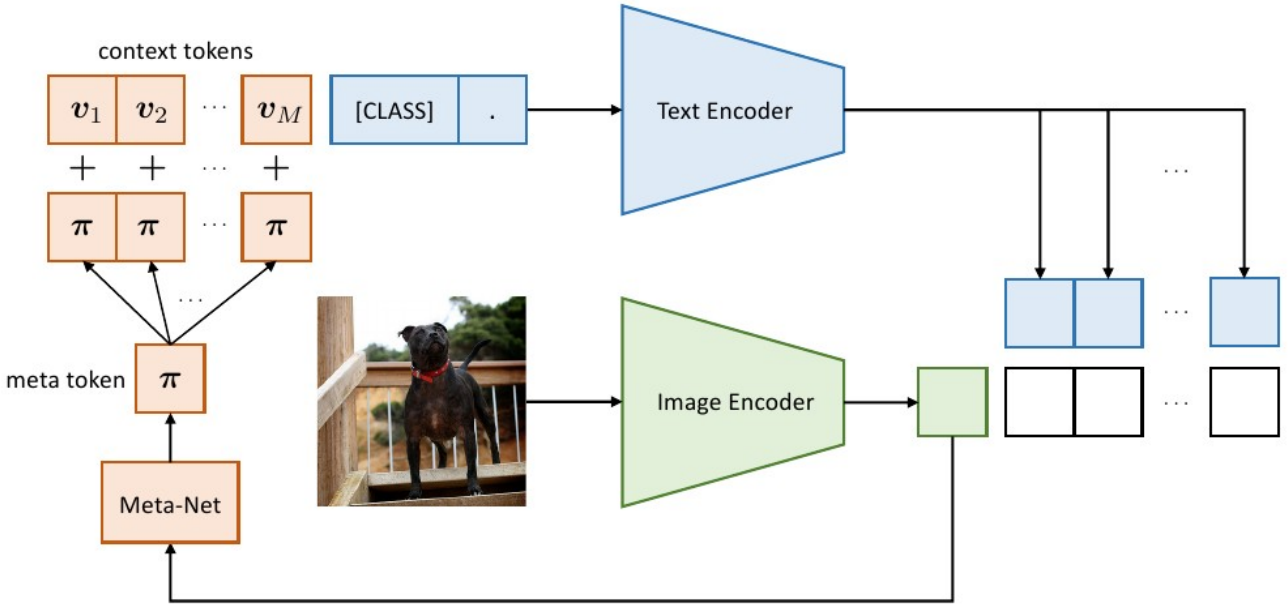


Figure 2. Our approach, Conditional Context Optimization (CoCoOp), consists of two learnable components: a set of context vectors and a lightweight neural network (Meta-Net) that generates for each image an input-conditional token.

条件来自当前图像；动态结果进入文本侧。

FedTPG

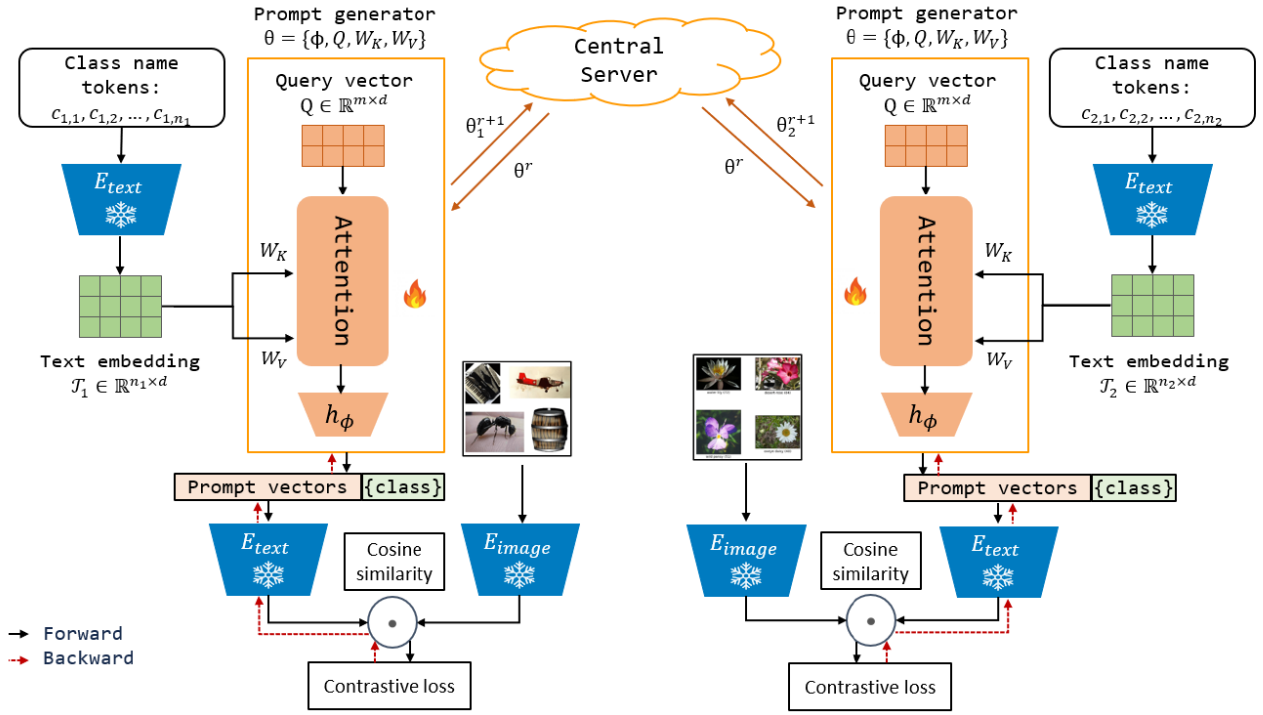
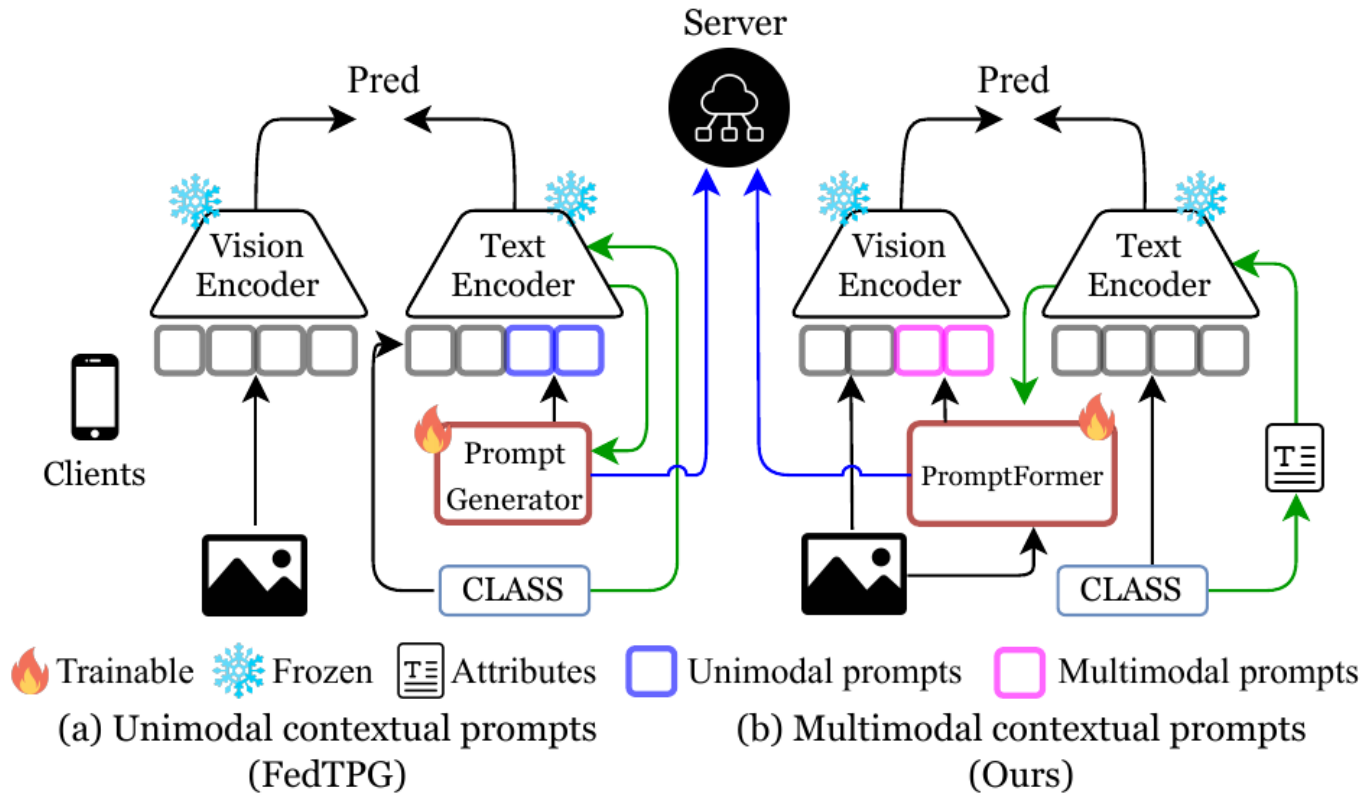


Figure 2: Our proposed FedTPG learns a unified prompt generator over the frozen CLIP model for converting task-related text input $\mathcal{T}$ to context-aware prompt vectors. The prompt generator is learned across multiple clients with different classification datasets collaboratively.

任务相关文本生成提示，服务器聚合提示生成器。

Motivation



FedTPG： 类别文本条件，文本侧提示。
FedMVP： 图像 + 类别属性条件，视觉侧提示。

FedMVP

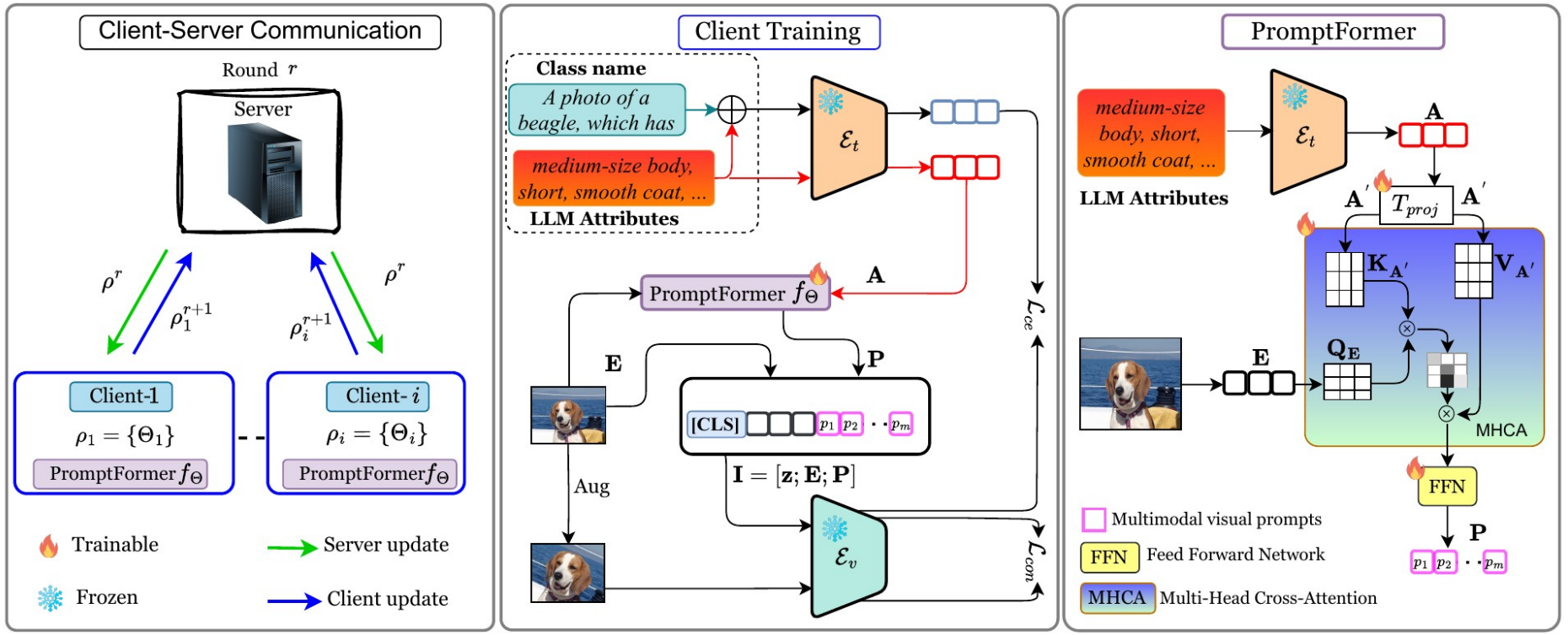
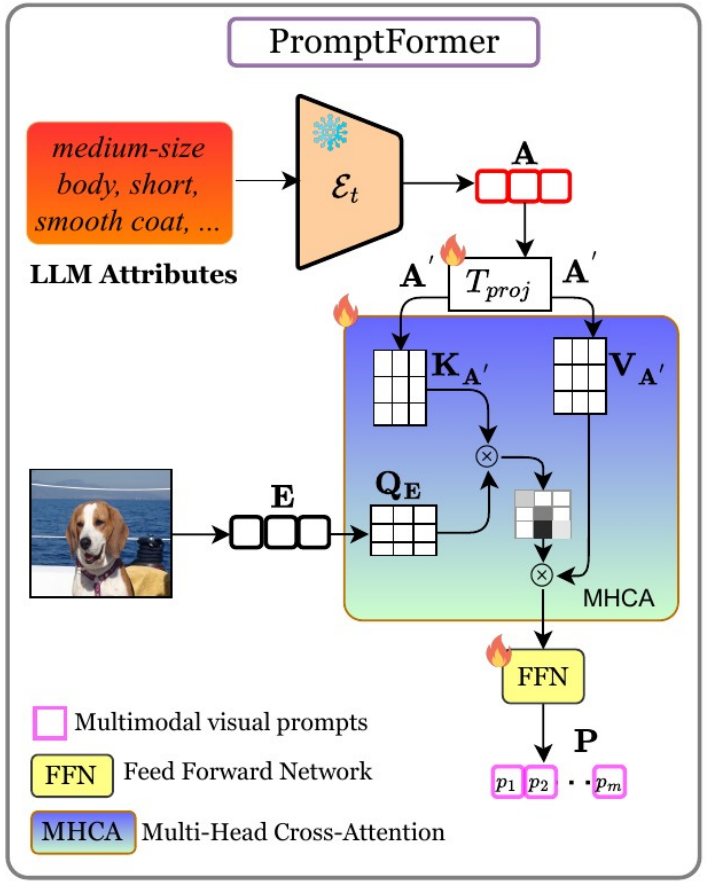


Figure 2. **Overview of Federated Multimodal Visual Prompt Tuning (FedMVP)**. Each client i trains the parameters of the Former network f_Θ and multi-head self-attention on its local data. At each round, the server aggregates the trainable parameters clients and sends the updated parameters back to the clients. The PromptFormer uses cross-attention mechanism to fuse the v textual attribute feature to generate multimodal visual prompts $\mathbf{P}$, which are then used for visual prompt tuning.

共享 PromptFormer； CLIP 主干冻结；本地图像不上传。

PromptFormer



$$\mathbf{P} = \{p_j\}_{j=1}^{m} = f_\Theta(\mathbf{A}, \mathbf{E}). \tag{1}$$

$$\mathbf{P}(\mathbf{A}', \mathbf{E}) = \mathrm{FFN}(\mathrm{CrossAttention}(\mathbf{Q}_\mathbf{E}, \mathbf{K}_{\mathbf{A}'}, \mathbf{V}_{\mathbf{A}'}));$$
$$\text{with } \mathbf{Q}_\mathbf{E} = \mathbf{E}W_\mathbf{Q}, \mathbf{K}_{\mathbf{A}'} = \mathbf{A}'W_\mathbf{K}, \mathbf{V}_{\mathbf{A}'} = \mathbf{A}'W_\mathbf{V}, \tag{2}$$

E： 当前图像 patch； A： 候选类别属性； P： 动态视觉提示（ m=4 ）。

Loss

$$\mathcal{L}_{\text{ce}} = -\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} y \log p(y|\mathbf{I}), \tag{3}$$

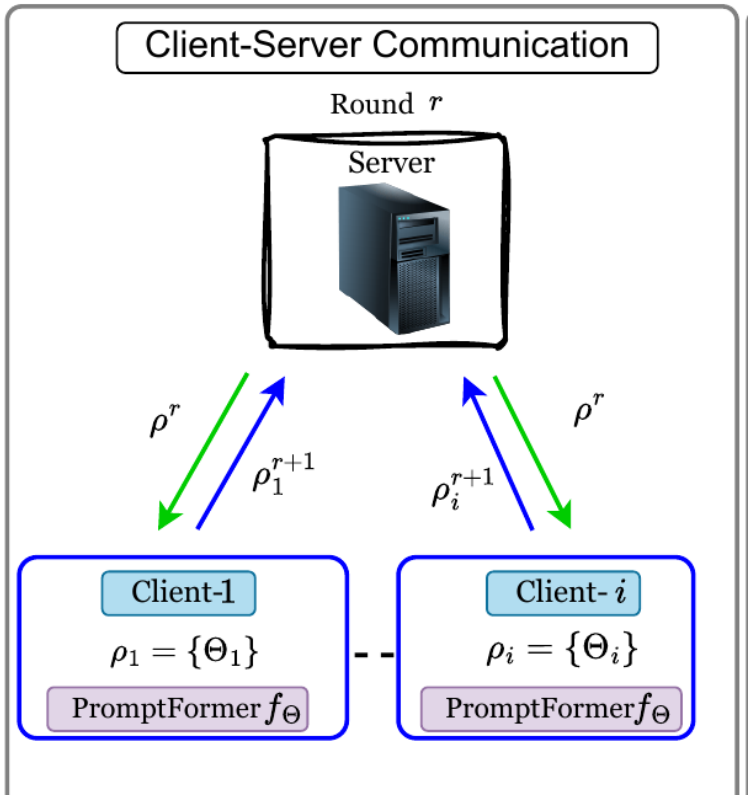
$$p(y=k|\mathbf{I}) = \frac{\exp(\cos(\mathbf{v},\mathbf{t}_k))/\tau)}{\sum_{k'=1}^{K}\exp(\cos(\mathbf{v},\mathbf{t}_{k'}))/\tau)}, \tag{4}$$

$$\mathcal{L}_{\text{con}} = 1 - \cos(\mathcal{E}_v(\mathbf{I}), \mathcal{E}_v(\mathbf{x}')), \tag{5}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \alpha \cdot \mathcal{L}_{\text{con}}, \tag{6}$$

分类损失约束正确类别；一致性损失约束增强前后的图像表示。

FL Training



$$\rho_i^{r+1} = \mathrm{SGD}_J(\eta^r, \rho^r, \mathcal{C}_i, \mathcal{D}_i). \tag{7}$$

$$\rho^{r+1} = \frac{1}{|S_r|} \sum_{i \in S_r} \rho_i^{r+1}. \tag{8}$$

if training loss $\mathcal{L}_{\text{total}}$ in a client starts below a certain loss value σ, which we set to 0.5. In particular, only the LoRA weights corresponding to the query, key and value matrices of the attention layers are trained, while freezing the parameters ρ. In that case, the client shares only the LoRA matrices with the server. This not only helps in boosting generalization, but also brings down the communication overhead of clients by around $\times 267$.

选客户端 → 本地更新 → 等权聚合；低损失时改训 PromptFormer 的 LoRA 。

Base-to-New

Table 2. Base-to-New Generalization

Methods	Local	Base	New	HM
ZS-CLIP [22]	76.72	70.51	75.78	74.24
FedOTP [14]	74.82	65.22	57.04	64.89
FedCoCoOp [35]	81.46	73.76	66.00	73.20
FedVPT [11]	76.29	70.43	74.89	73.79
FedCLIP [15]	76.87	71.04	75.06	74.24
FedMaPLe [13]	81.63	74.44	70.62	75.29
FedKgCoOp [30]	78.38	72.18	75.87	75.39
PromptFL [7]	81.75	74.47	71.70	75.74
FedTPG [21]	80.75	73.68	76.02	76.70
FedMVP (Ours)	**81.89**	**75.37**	**77.82**	**78.27**
Improvement	**+0.14**	**+0.90**	**+1.79**	**+1.57**

9 个数据集； new 类图片不参与训练。 FedMVP New 77.82%， FedTPG 76.02%。

Domain Generalization

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 3. DomainBed averages: MSST, SSMT

Method	Average	Average
ZS-CLIP [22]	70.41	70.41
FedOTP [14]	59.49	57.07
FedCoCoOp [35]	61.79	55.61
FedTPG [21]	65.43	64.36
PromptFL [7]	66.77	65.39
FedKgCoOp [30]	66.82	65.50
FedMaPLe [13]	68.70	67.80
FedVPT [11]	69.98	69.22
FedCLIP [15]	71.03	70.72
FedMVP (Ours)	**73.02**	**72.63**
Improvement	**+1.99**	**+1.91**

Table 4. ImageNet DG

Method	Source	Target				
	IN	INV2	IN-S	IN-A	IN-R	Average
ZS-CLIP [22]	66.75	60.79	46.12	47.79	74.00	57.18
FedOTP [14]	51.68	45.53	34.73	16.64	63.98	40.22
FedMaPLe [13]	66.96	60.65	44.69	46.24	74.62	56.55
FedVPT [11]	66.92	60.34	46.43	48.03	74.56	57.34
PromptFL [7]	67.80	61.59	45.61	48.78	74.49	57.62
FedCLIP [15]	67.26	61.35	46.66	48.24	74.36	57.65
FedKgCoOp [30]	67.53	61.60	46.69	48.37	74.71	57.84
FedCoCoOp [35]	68.51	62.29	46.90	50.33	76.49	59.00
FedTPG [21]	69.51	62.90	47.65	49.97	76.35	59.22
FedMVP (Ours)	**70.87**	**63.72**	**50.93**	**51.76**	**77.23**	**60.91**
Improvement	**+1.36**	**+0.82**	**+3.28**	**+1.43**	**+0.74**	**+1.69**

MSST： 多源训练，留一个目标域。 SSMT： 单源训练，测其余域。

Cross-Dataset

Table 5. **Cross-Dataset Generalization setting.** Accuracies (%) on the seen classes of the source dataset, *i.e.*, ImageNet, and the unseen classes of the target datasets. Best number in each column is in bold and second-best underlined. Improvement row reports the absolute difference between the best and second-best numbers.

Method	Source	Target										
	ImageNet	Caltech101	Flowers102	FGVCAircraft	UCF101	OxfordPets	Food101	DTD	StanfordCars	SUN397	EuroSAT	Average
ZS-CLIP [22]	66.75	92.90	71.29	24.72	66.75	89.15	86.09	44.33	65.29	62.59	47.68	65.08
FedOTP [14]	51.68	88.60	48.36	10.95	46.10	73.43	61.01	41.55	29.61	48.34	50.26	49.82
FedMaPLe [13]	66.96	92.49	68.25	23.52	60.32	89.67	83.52	44.68	60.16	61.85	45.38	62.98
PromptFL [7]	67.80	91.87	68.13	21.44	64.13	88.70	85.85	42.43	63.59	62.77	43.26	63.22
FedCLIP [15]	67.26	93.39	67.19	24.03	65.64	88.28	85.14	44.44	65.50	63.42	41.67	63.87
FedKgCoOp [30]	67.53	93.63	69.31	23.06	64.46	88.55	85.37	44.74	64.99	63.85	43.29	64.13
FedVPT [11]	68.56	91.42	69.24	21.09	65.70	87.79	85.45	44.57	65.00	64.68	47.82	64.28
FedCoCoOp [35]	68.51	94.11	66.34	20.79	62.75	89.04	85.40	43.20	63.98	64.02	**55.40**	64.50
FedTPG [21]	69.51	93.75	70.04	23.22	64.72	**90.60**	85.91	46.25	63.98	66.78	47.86	65.31
FedMVP (Ours)	**70.87**	**95.37**	**72.80**	**25.94**	**70.58**	89.27	**87.06**	**49.78**	**65.83**	**68.19**	50.84	**67.57**
Improvement	+1.36	+1.62	+1.51	+1.22	+3.83	-1.33	+0.97	+3.53	+0.33	+1.41	-4.56	+2.26

只在 ImageNet 联邦调优；目标平均 67.57%， FedTPG 为 65.31%。

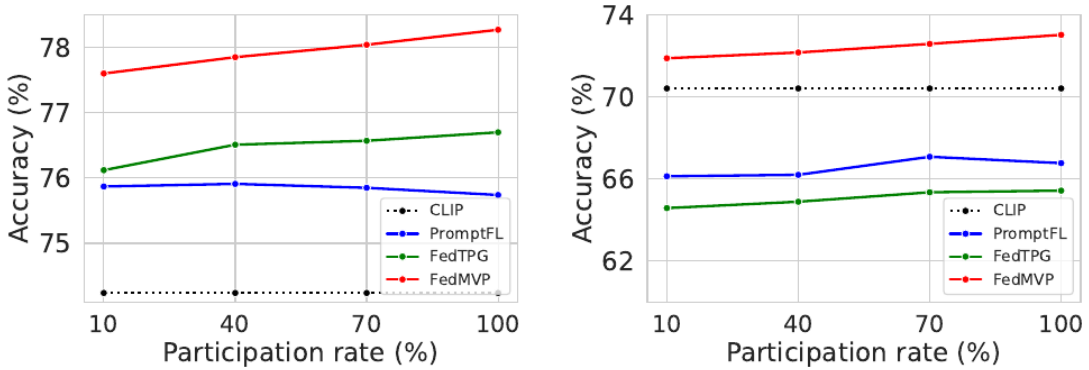
Ablation

Table 6. **Ablating FedMVP components.** We report numbers for the Base-to-New and MMST DG (on DomainBed) settings.

Components	Base-to-New	MSST DG
ZS-CLIP	74.24	70.41
f_Θ *only*	75.94	71.85
$f_\Theta + \mathcal{L}_{\text{con}}$	76.27	72.14
w/o LoRA	77.41	72.58
FedMVP (Full)	**78.27**	**73.02**

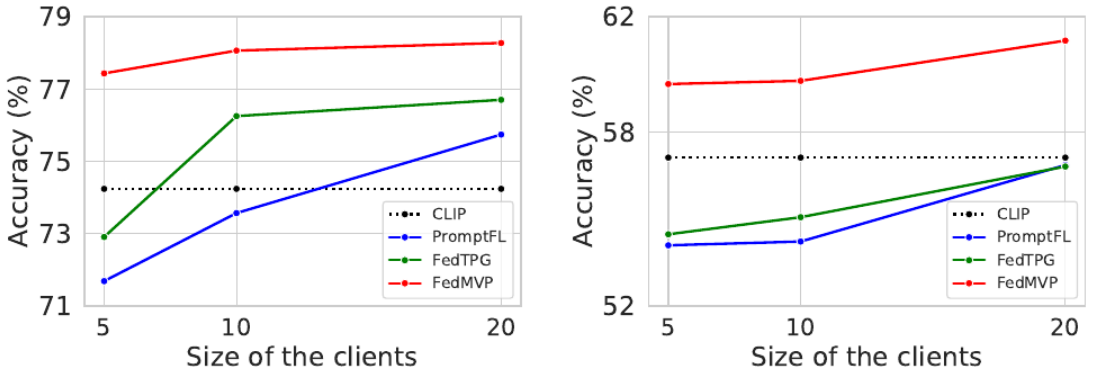
PromptFormer 是主增益；一致性损失小幅提升；完整方法高于 w/o LoRA 。

Clients



(a) Base-to-New (b) MSST DG

Figure 3. **Sensitivity to participation rate** of the clients, on (a) Base-to-New and (b) MSST DG (DomainBed) settings.

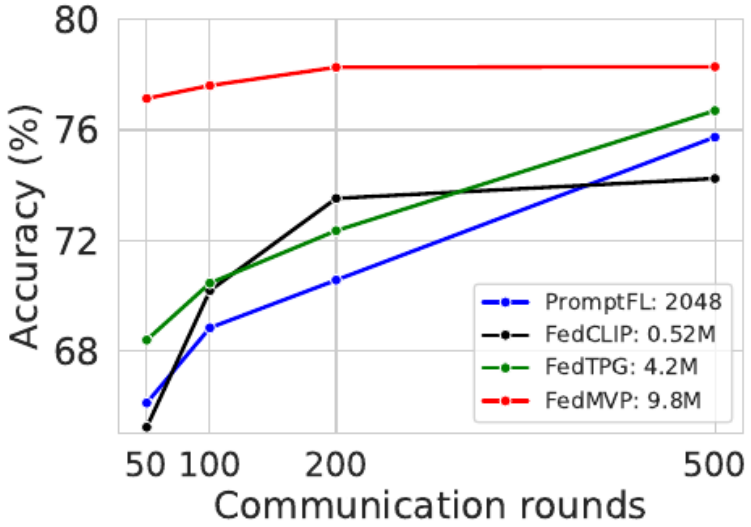


(a) Base-to-New (b) MSST DG

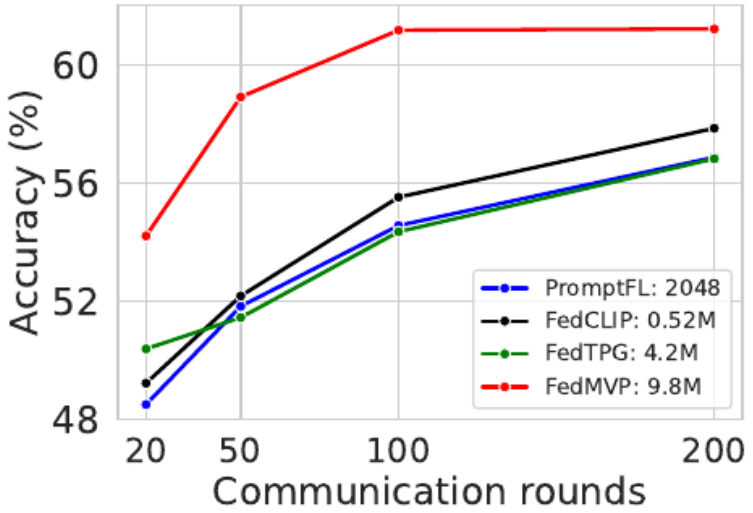
Figure 5. **Sensitivity to number of classes per client** on (a) Base-to-New and (b) MSST DG (DomainNet) setting.

参与率 10%–100%；每端类别数 K=5/10/20 。

Efficiency



(a) Base-to-New



(b) MSST DG

Figure 4. **Computational analysis of the methods** on (a) Base-to-New and (b) MSST DG (DomainNet) setting.

参数规模
FedMVP 9.8M
FedTPG 4.2M

曲线上， FedMVP 较早达到更高准确率。

Discussion

主要结论

图像 patch 与类别属性共同生成逐图变化的视觉提示。
三类泛化设置下， FedMVP 的平均结果占优。

讨论问题

属性是否可靠?

当 New 类没有训练图片时，仅靠类别名生成的属性，能否让 PromptFormer 正确区分它与相似类别？如果 New 类与 Base 类共享的属性很少，或者属性写得过于笼统，泛化效果会怎样?

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Thanks