Federated Cross-Modal Style-Aware Prompt Generation (Student Abstract)

Suraj Prasad, Navyansh Mahla, Sunny Gupta, Amit Sethi

Indian Institute of Technology Bombay, Maharashtra 400076, India
{suraj.prasad, 210040106, 22d1631, asethi}@iitb.ac.in

Abstract

Existing federated prompt learning methods for vision-language models like CLIP rely solely on text-based prompts and final-layer visual features, missing crucial multiscale visual details and client-specific style variations. This limits its generalization across non-IID distributions and novel classes. We introduce FedCSAP (Federated Cross-Modal Style-Aware Prompt Generation), which harnesses multiscale features from CLIP's vision encoder alongside domain-aware style statistics from client data. By fusing these visual representations with textual context, FedCSAP generates adaptive, context-aware prompts that enhance robustness across seen and unseen classes. Our privacy-preserving approach operates through local training and global aggregation, effectively handling heterogeneous client distributions. Experiments on multiple image classification datasets demonstrate that FedCSAP significantly outperforms existing federated prompt learning methods in both accuracy and generalization.

Code — https://github.com/Jarus77/FedCSAP.git

Introduction

The integration of vision-language models (VLMs) such as CLIP into federated learning (FL) settings promises powerful cross-modal understanding, yet is often bottlenecked by computational and communication constraints that make end-to-end training impractical. Prompt-based learning has emerged as a compelling alternative, enabling the adaptation of large pretrained VLMs to downstream tasks through minimal, efficient modifications. Techniques like Context Optimization (CoOp) replace hand-engineered prompts with soft learnable vectors, demonstrating remarkable efficiency and task-specific performance within few-shot scenarios (Zhou et al. 2022).

Recently, Federated Text-driven Prompt Generation (FedTPG) (Qiu et al. 2024) advanced this paradigm by introducing a text-conditioned unified prompt generation network collaboratively learned across clients. FedTPG leverages task-specific textual inputs to create context-aware prompts, improving generalization beyond the fixed prompt vectors of earlier methods. However, FedTPG primarily focuses on textual context without incorporating multi-scale visual features or client-specific style variations. Similarly, Federated Context Optimization (FedCoOp) extended prompt learning to FL by learning a unified prompt vector set collaboratively from decentralized, heterogeneous clients but relied solely on textual information and features from only the final layer of the vision encoder. This myopic focus results in a fundamental limitation: the rich, multi-scale visual cues and domain-dependent style variations inherently present in diverse client data remain underutilized. As a consequence, these models often struggle to generalize to new domains or unseen classes, especially in the presence of non-IID distributions and context shifts, ultimately limiting their robustness in real-world federated environments.

To address this crucial gap, we introduce FedCSAP (Federated Cross-Modal Style-Aware Prompt Generation), a framework that transcends prior boundaries by weaving together multi-scale visual features from various depths of CLIP's vision encoder with client-specific style statistics derived from batch activations. By fusing these richer visual signals with task-relevant textual cues, our system dynamically generates context-aware prompts finely tuned to each client's domain, thereby improving generalization and robustness across heterogeneous federated data.

Methodology

Our framework, FedCSAP, learns context-aware prompt tokens by fusing multi-scale visual and domain-specific style features with textual embeddings across federated clients as shown in Fig. 1.

Problem Setup

Each federated client has local, non-IID data $\mathcal{D}_i$ with unique class subsets and style variations. Standard prompt learning with CLIP ignores valuable intermediate visual signals and domain statistics. We aim to generate prompt tokens that:

(i) integrate multi-scale CLIP visual features, (ii) encode client-specific style via batch normalization, and (iii) aintain diversity via redundancy penalization.

Prompt Generation

Given class names $\{c_j\}$, textual context is embedded as:

$$\mathcal{T} = \{E_{\text{text}}(c_j)\}_{j=1}^{n} \in \mathbb{R}^{n \times d}$$

A learnable query matrix $Q \in \mathbb{R}^{m \times d}$ interacts via cross-attention:

$$K_{\mathcal{T}} = \mathcal{T} W_K, \quad V_{\mathcal{T}} = \mathcal{T} W_V$$

$$\mathcal{P} = h_\phi\big(\text{CrossAttention}(Q, K_{\mathcal{T}}, V_{\mathcal{T}})\big)$$

To capture visual details at different levels, we extract features from multiple layers as shown in Fig. 1 of CLIP's vision encoder:

$$\mathbf{f}_v^l(x) \in \mathbb{R}^{W_l \times H_l \times C_l}$$

$$\widehat{\mathbf{f}}_v^l(x) = \text{GAP}(\mathbf{f}_v^l(x)) \in \mathbb{R}^{C_l}$$

Combined, with style mean $\boldsymbol{\mu}_i$:

$$\mathbf{F}(x) = [\widehat{\mathbf{F}}(x); \boldsymbol{\mu}_i]$$

Full multi-modal vector:

$$\mathbf{M}(x) = [\mathcal{T}; \mathbf{F}(x)]$$

Attention-Based Fusion

The injection block B_ϕ fuses content via channel-wise attention:

$$\mathbf{O}_1 = \mathbf{F}(x) \otimes A_1(\mathbf{F}(x)) + \mathbf{F}(x)$$

$$\mathbf{O}_q = \mathbf{O}_{q-1} \otimes A_q(\mathbf{O}_{q-1}) + \mathbf{O}_{q-1}$$

$$A_q(\mathbf{v}) = \sigma\big(W_2 \delta(W_1 \mathbf{v})\big)$$

Projecting to visual tokens:

$$v_m = h_m(\mathbf{O}_Q), \quad m = 1, \ldots, M$$

Prompt tokens and visual tokens are merged:

$$c'_m = c_m + v_m$$

Composite prompt for class j:

$$t_j = \big\{[v_1 + c_1], \ldots, [v_M + c_M], [\text{CLS}_y]\big\}$$

Federated Training

Each client optimizes via total loss:

$$L_{\text{total}} = L_{\text{ce}} + \lambda L_{\text{CRP}}$$

where

$$L_i(\theta^r, \phi^r; \mathcal{T}_i) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\big[y \log p_{\theta^r,\phi^r}(y|x, \mathcal{T}_i)\big]$$

$$p_{\theta,\phi}(y = j|x, \mathcal{T}) = \frac{\exp(\cos(E_{image}(x), E_{text}(t_j))/\tau)}{\sum_i^n \exp(\cos(E_{image}(x), E_{text}(t_i))/\tau)}$$

and context redundancy penalization:

$$L_{\text{CRP}} = \sum_{j \neq l} \big|c'_j \cdot c'_l - I\big|$$

Server aggregates client updates via FedAvg:

$$\theta^{r+1} = \frac{1}{|S_r|} \sum_{i \in S_r} \theta_i^{r+1}, \quad \phi^{r+1} = \frac{1}{|S_r|} \sum_{i \in S_r} \phi_i^{r+1}$$

Experiments

We evaluate FedCSAP on nine diverse image classification datasets (Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVC Aircraft, SUN397, UCF101, and DTD) following standard federated learning protocols. Each dataset is split into base classes for training and new classes for generalization evaluation. Base classes are distributed non-IID across clients, with each client receiving 20 distinct classes and 8 labeled images per class.

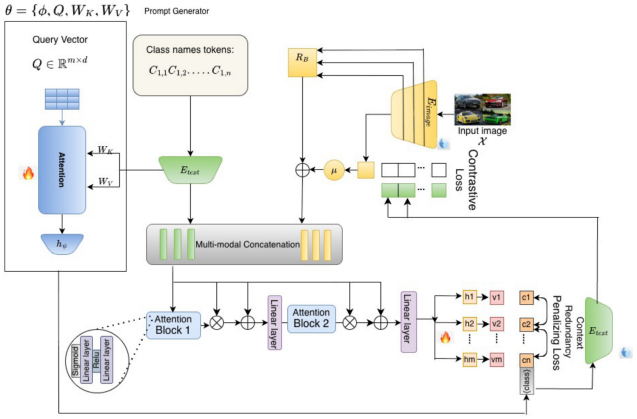

Figure 1: **FedCSAP:** A federated prompt generator that fuses multi-scale visual features and local style cues with CRP loss to create distinct, context-aware prompt tokens.

We measure performance using local accuracy (client-specific tasks), base accuracy (aggregated base classes), and new accuracy (unseen classes), combined via harmonic mean. All results are averaged over three independent runs.

Results and Discussion

FedCSAP is compared against five baselines: CLIP, CoOp, FedCoOp, FedCoCoOp,FedMaple and FedTPG. While CoOp achieves strong local performance (83.41%), it struggles with base classes (72.05%) and new classes (71.27%). Federated methods improve base-class collaboration but often sacrifice generalization. FedCSAP achieves the highest harmonic mean (76.06%), outperforming the best baseline by 0.84%. Notably, FedCSAP excels on new classes (75.61%), demonstrating superior generalization. This validates our core hypothesis: multi-scale visual features and style-aware prompt generation enable better adaptation to domain shifts and unseen categories. The integration of batch-level style statistics and context redundancy penalization loss effectively balances local specialization with global generalization in heterogeneous federated environments.

Conclusion

We presented a federated learning framework for multi-modal vision-language tasks using attention-based fusion and adaptive prompt generation. Our approach combines cross-attention mechanisms with multi-layer visual features from CLIP, while context redundancy penalization prevents prompt collapse across federated clients. The framework effectively handles heterogeneous data distribution while preserving privacy. Future work could explore adaptive penalty weighting and extensions to other vision-language tasks.

References

Qiu, C.; Li, X.; Mummadi, C. K.; Ganesh, M. R.; Li, Z.; Peng, L.; and Lin, W.-Y. 2024. Federated text-driven prompt generation for vision-language models. In *ICLR 2024*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to Prompt for Vision-Language Models. *Int. J. of Computer Vision*.