Harmonizing Generalization and Personalization in Federated Prompt Learning

Tianyu Cui [1] Hongxia Li [1] Jingya Wang [1] Ye Shi [1 *]

Abstract

Federated Prompt Learning (FPL) incorporates large pre-trained Vision-Language models (VLM) into federated learning through prompt tuning. The transferable representations and remarkable generalization capacity of VLM make them highly compatible with the integration of federated learning. Addressing data heterogeneity in federated learning requires personalization, but excessive focus on it across clients could compromise the model's ability to generalize effectively. To preserve the impressive generalization capability of VLM, it is crucial to strike a balance between personalization and generalization in FPL. To tackle this challenge, we proposed **Fed**erated **P**rompt Learning with CLIP **G**eneralization and low-rank **P**ersonalization (FedPGP), which employs pre-trained CLIP to provide knowledge-guidance on the global prompt for improved generalization and incorporates a low-rank adaptation term to personalize the global prompt. Further, FedPGP integrates a prompt-wise contrastive loss to achieve knowledge guidance and personalized adaptation simultaneously, enabling a harmonious balance between personalization and generalization in FPL. We conduct extensive experiments on various datasets to explore base-to-novel generalization in both category-level and domain-level scenarios with heterogeneous data, showing the superiority of FedPGP in balancing generalization and personalization.

1. Introduction

Federated Learning (McMahan et al., 2017) has been proposed as an efficient collaborative learning strategy, enabling clients to jointly train a global model while preserving data privacy. In this context, the ability of large pre-trained Vision-Language models (VLM) like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) to learn transferable representations across downstream tasks makes them a natural fit for integration with federated learning. This collaborative approach not only harnesses the outstanding performance and generalization capabilities of pre-trained models but also ensures efficient and privacy-preserving global model training across multiple clients. However, due to the millions of parameters in VLM, fine-tuning the entire model in federated learning leads to high communication costs and memory footprint issues. Prompt tuning addresses these challenges by adapting pre-trained models to diverse downstream tasks with a reduced parameter count, and its integration of federated learning has been explored in previous research (Zhao et al., 2022; Guo et al., 2023b). We term the combination of prompt-tuning and federated learning as Federated Prompt Learning (FPL) for simplicity. Currently, studies in FPL have not been thoroughly explored in terms of personalization and generalization. Methods derived from traditional federated learning studies fail to capture the multi-modality of VLM, which hinders a direct transfer of these methods into FPL.

In federated learning, it is essential to account for the generalization capability to unseen domains or categories. However, existing studies in FL, like (Nguyen et al., 2022; Liu et al., 2021; Zhang et al., 2021), have struggled to achieve satisfactory results in evaluating generalization on target datasets in unseen domains. With the help of VLM which has strong generalization performance, this problem may be solved. Unfortunately, the generalization issues in prompt-based VLM have been revealed in recent research (Khattak et al., 2023a;b). For instance, CoOp (Zhou et al., 2022b) struggles with generalizing to unseen categories within the same dataset due to overfitting, resulting in lower test accuracy on novel categories compared to the zero-shot CLIP baseline with a handcrafted prompt. PromptSRC (Khattak et al., 2023b) addresses this issue with self-regularization constraints, maximizing the mutual agreement between prompted and frozen VLM features. However, the generalization of FPL is still an open challenge. FedTPG (Qiu et al., 2023) takes a step toward exploring generalization by learning a unified prompt generation network among

[1] ShanghaiTech University. Correspondence to: Ye Shi <shiye@shanghaitech.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

multiple clients but disregards the data heterogeneity.

Data heterogeneity results in another challenge in federated learning, where the data distributions among clients are not independently and identically distributed (Non-IID). This leads to discrepancies between local and global optimization objectives, making it difficult for a single global prompt to adapt to the varied local distributions. In the endeavor to learn personalized models, pFedPrompt (Guo et al., 2023a) incorporates personalized attention modules into FPL while learning a consensus among users through shared text prompts. Nevertheless, if we employ strong personalized techniques to fully adapt the prompts to local distributions, it may lead to the loss of inherent generalization in VLM. This raises the question we aim to explore:

How can we strike a balance between generalization and personalization in FPL?

To overcome the problems outlined above, we proposed **Fed**erated **P**rompt Learning with CLIP **G**eneralization and low-rank **P**ersonalization (FedPGP), an effective method that reaches a balance between personalization and generalization. In FedPGP, each client learns a personalized prompt, combining a global prompt and an adaptation term to accommodate heterogeneous local distributions. The incorporation of the adaptation term allows fine-tuning of the global prompt for specific client needs. To enhance prompt generalization, we incorporate category-agnostic knowledge from CLIP, aligning the global prompt in each client towards a unified direction.

To balance generalization and personalization, we utilize a low-rank decomposition for the adaptation term, ensuring robust generalization capabilities in comparison to a full-rank term. To enable personalized prompts better access to client-specific knowledge, we aim to separate representations of global and personalized prompts for better personalization. Considering this and the knowledge-guidance from CLIP to global prompt, we introduce an additional contrastive loss into the optimization objective to further balance personalization and generalization of our FedPGP. This involves treating global prompt representations with personalized ones as negative pairs for personalization, while simultaneously treating them as positive pairs with the representation of handcrafted prompt of CLIP for generalization.

Our main contributions are summarized as follows:

- We are the first to consider both personalization and generalization in federated prompt learning. We aim to learn a personalized prompt for each client in heterogeneous federated scenarios while preserving the remarkable generalization capacity in Visual-Language models, leading to the balance of generalization and personalization.

- We propose FedPGP, utilizing low-rank decomposition adaptation to flexibly adjust the global prompt to heterogeneous local distributions, which prevents overfitting on local datasets. Additionally, we integrate an extra contrastive loss, treating representations of global and personalized prompts as negative pairs and representations of global and handcrafted prompts as positive pairs.

- We conduct extensive experiments on widely adopted datasets to investigate the base-to-novel generalization of FedPGP on both category-level and domain-level in the case of heterogeneous data. Our comparative experimental results demonstrate FedPGP's superiority in harmonizing generalization and personalization.

2. Related Work

Federated Learning This subsection mainly introduces the research on personalization and generalization in federated learning. Personalized federated learning (PFL) algorithms, rather than creating a universal model for all clients, tackle data heterogeneity by learning customized models for each client. Various strategies for achieving PFL have been suggested in previous studies. Some existing methods combine global model optimization with additional local model customization involving local fine-tuning (Wang et al., 2019; Mansour et al., 2020; Tan et al., 2022a), regularization (Li et al., 2020; 2021b; T Dinh et al., 2020), parameter decomposition (Jeong & Hwang, 2022; Hyeon-Woo et al., 2021; Arivazhagan et al., 2019), parameter generation (Shamsian et al., 2021; Ma et al., 2022; Li et al., 2023), and clustering methods for client grouping (Huang et al., 2021; Zhang et al., 2020; Sattler et al., 2020; Zhang et al., 2022; Cao et al., 2023; Cai et al., 2024). The theoretical significance of PFL was pointed out by (Huang et al., 2023). To be specific, FedPer (Arivazhagan et al., 2019), FedBABU (Oh et al., 2021), and FedRep (Collins et al., 2021) share the base layers while learning personalized classifier heads locally. FedBN (Li et al., 2021c) uses local batch normalization to alleviate feature shift before averaging models. FedRod (Chen & Chao, 2021) proposed learning a global predictor for generic FL and a local predictor for personalized FL.

Many existing studies in federated learning commonly assume the test dataset is a subset of the client dataset. However, a research gap exists for scenarios where the target dataset (i.e., the test dataset) is not included in the training process. This scenario is also referred to as domain generalization in centralized machine learning. FedSR (Nguyen et al., 2022) employs regularization techniques for simplified data representation, intending to achieve improved generalization capabilities. ELCFS (Liu et al., 2021) tackled federated domain generalization with continuous frequency space interpolation and the boundary-oriented episodic learning scheme. FedADG (Zhang et al., 2021) utilizes federated adversarial learning for dynamic universal feature representation. Unlike traditional federated learning approaches, we are the first to consider both personalization and generalization in the context of FPL.

Federated Prompt Learning Prompt tuning is a technique employed to adapt pre-trained models to diverse downstream tasks. For instance, CoOp (Zhou et al., 2022b) uses tunable text prompts to replace the fixed template in CLIP, and CoCoOp (Zhou et al., 2022a) utilizes image feature to instruct the optimization of the soft text prompt. ProGrad (Zhu et al., 2023) selectively updates prompts based on aligned gradients with general knowledge to prevent forgetting essential information from VLMs. Some works also incorporate prompt tuning into federated learning. PromptFL (Guo et al., 2023b) introduced prompt learning into Federated Learning. FedPR (Feng et al., 2023) focuses on learning federated visual prompts within the null space of the global prompt for MRI reconstruction. pFedPG (Yang et al., 2023) employs a client-specific prompt generator on the server side for personalized prompts, while FedTPG (Qiu et al., 2023) also trains a global prompt generation network to enhance generalization. pFedprompt (Guo et al., 2023a) maintains a non-parametric personalized attention module for each client to generate local personalized spatial visual features. FedOTP (Li et al., 2024) employs unbalanced Optimal Transport to promote the collaboration between global and local prompts across heterogeneous clients. Designed for domain discrepancy, FedAPT (Wei et al., 2023) unlocks specific domain knowledge for each test sample to provide personalized prompts, and Fed-DPT (Su et al., 2022) applies both visual and textual prompt tuning to facilitate domain adaptation over decentralized data. However, these methods overlook the aspect of generalization. We propose FedPGP, a framework that effectively balances personalization and generalization in FPL.

Contrastive learning Contrastive learning methodologies have gained significant attention by consistently attaining state-of-the-art outcomes results in the field of visual representation learning (Chen et al., 2020a;b; Xie et al., 2022). The fundamental principle behind contrastive learning is to minimize the distance between representations generated from diverse augmentations of the same image (positive pairs) while simultaneously maximizing the distance between representations obtained from augmented views of different images (negative pairs). A proportion of research (Tan et al., 2022b; Li et al., 2021a; Mu et al., 2023) combines contrastive learning with federated learning, which improves the local training process and achieves higher model effectiveness. FedPCL (Tan et al., 2022b) employs prototype-wise contrastive learning for client-specific representations, promoting alignment with global and local prototypes to enhance knowledge sharing. MOON (Li et al., 2021a) minimizes the distance between local and global model representations while increasing the distance from the previous local model's representation. Different from model-wise of MOON and prototype-wise of FedPCL, our FedPGP introduces a novel prompt-wise contrastive methodology. In addition, unlike previous work focusing on aligning the global and local components, FedPGP treats representations of global and personalized prompts as negative pairs and representations of global and handcrafted prompts as positive pairs.

3. Proposed Method

In this section, we delve into the details of our proposed FedPGP, illustrated in Figure 1. FedPGP leverages CLIP knowledge-guidance and low-rank adaptation with an additional contrastive loss to balance generalization and personalization.

3.1. Preliminaries of Prompt Learning

Prompt learning methods (Zhou et al., 2022b;a) offer an efficient approach to adapting pre-trained models like CLIP to downstream tasks by training a part of the parameters in the prompt. Unlike the zero-shot transfer that utilizes a fixed word embedding $W = \{w_1, w_2, ..., w_l\}$ mapped from a hand-crafted prompt (e.g., "a photo of a $\langle$label$\rangle$"), prompt learning replaces a set of M continuous context vectors $p = \{p_1, ..., p_M\} \in \mathbb{R}^{d\times k}$ as the learnable prompt. Specifically, we use $p = \{p_1, ..., p_M\}$ to replace $\{w_2, ..., w_{M+1}\}$ to be consistent with previous methods. Then the textual prompt of k class can be reformulated as $P_k = \{w_1, p_1, ..., p_M, w_{M+2}, ..., w_l\}$ and is fed into pre-trained text encoder $g(\cdot)$. Denote image encoder as $f(\cdot)$, the prediction probability for each category of input image x is computed through matching scores:

$$p(\hat{y} = k|x) = \frac{\exp(\mathrm{sim}(f(x), g(P_k))/\tau)}{\sum_{c=1}^{K} \exp(\mathrm{sim}(f(x), g(P_c))/\tau)}, \quad (1)$$

where sim$(\cdot,\cdot)$ denotes a metric function (e.g., cosine similarity), $\hat{y}$ denotes the predicted label, K denotes the number of classes, and τ denotes the temperature of Softmax. Then we optimize the learnable prompt by cross-entropy loss:

$$\mathcal{L}_{ce} = -\frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \sum_{k} y \log p(\hat{y} = k|x), \quad (2)$$

where y denotes the one-hot ground-truth annotation.

3.2. Federated Prompt Learning

Suppose there are N clients and a central server. Each client i holds local dataset D_i with n_i samples and $D = \{D_1, D_2, ..., D_N\}$ represents the total dataset where each

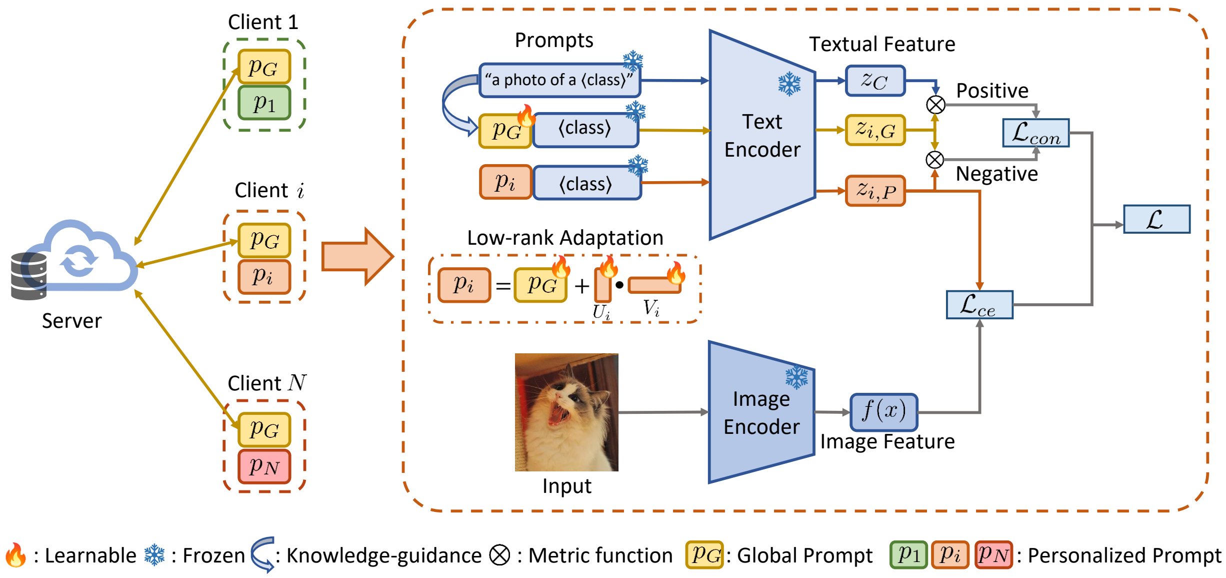


Figure 1: Pipeline of FedPGP. On the left, clients send global prompts to the server for aggregation while retaining adaptation term locally. The right shows the workflow of CLIP knowledge-guidance and low-rank adaptation with an additional contrastive loss to balance generalization and personalization.

dataset is derived from a distinct data distribution $\mathcal{D}_i$. Each client is equipped with a pre-trained CLIP model and a prompt learner in our federated learning setup. Let C_t represent the set of selected clients participating in communication round t. For each communication round t, the selected clients initialize the global prompt with p_G^{t-1} and perform local training p_i^t through cross-entropy loss $\mathcal{L}_{ce}$ for E local epoch, at e local epoch the update of global prompt is:

$$p_{G,i}^{t,e} = p_{G,i}^{t,e-1} - \eta \nabla \mathcal{L}_{ce}(p_{G,i}^{t,e-1}). \tag{3}$$

After E local epoch training, each client in C_t uploads the global prompt $p_{G,i}^{t,E}$ to the server for aggregation:

$$p_G^t = \sum_{i\in C_t} \frac{n_i}{\sum_{j\in C_t} n_j} p_{G,i}^{t,E}. \tag{4}$$

The global prompt is aggregated in the context of federated learning, carrying the unique characteristics learned from other clients. The optimization objective of FPL can be formulated as:

$$\min_{p_G} \sum_{i=1}^{N} \frac{n_i}{\sum_j n_j} \mathcal{L}_{ce}^{D_i}(p_G), \tag{5}$$

where $\mathcal{L}_{ce}^{D_i}(p_G)$ represents the cross-entropy loss on dataset D_i of client i.

3.3. Generalization and Personalization for FPL

Previous research has discovered that prompted visual-language models, such as CoOp, overfit to the base classes and cannot generalize to the unseen class observed during training (Zhou et al., 2022a; Zhu et al., 2023; Ma et al., 2023). This phenomenon of overfitting to base classes implies that the prompt fails to capture more generalized elements that are crucial for recognizing a wider range of scenarios. On the contrary, the manually designed prompts adopted by the zero-shot CLIP are relatively generalizable. The problem of generalization in prompt vision-language models remains unresolved in FPL. The objective of the client's local training can be formulated as:

$$\mathcal{L}_{ce}^{D_i}(p_{G,i}^{t,e}) = -\frac{1}{|\mathcal{D}_i|} \sum_{(x,y)\in\mathcal{D}_i} \sum_k y \log p(\hat{y}=k|x). \tag{6}$$

$\mathcal{L}_{ce}^{D_i}$ is to optimize the prompts for the client-specific task. Despite aggregating prompts in federated learning, overfitting to client-specific tasks remains a challenge. Leveraging the remarkable generalization capabilities of CLIP, FedPGP utilizes knowledge from CLIP to guide the global prompt to enhance generalization. Specifically, we obtain the representations of the handcrafted prompt $g(p_C)$ of CLIP and the global prompt $g(p_G)$ and align them through a metric function, such as cosine similarity. This knowledge-guidance from CLIP promotes the preservation of category-agnostic information within learnable global prompts, contributing to improve model generalization.

Due to data heterogeneity, it is difficult for a single global prompt to adapt to diverse local distributions. Different from tuning model parameters in traditional federated learning, FPL involves frozen client models and learnable prompts, leading to a distinct approach in adapting global prompt to local distributions. In FedPGP, the adaptation of global prompt p_G to the client-specific prompt p_i is achieved by introducing an additional adaptation term Δp_i:

$$p_i = p_G + \Delta p_i, \tag{7}$$

where $\Delta p_i \in \mathbb{R}^{d\times k}$ owns the same dimension of p_G. Then the objective of federated learning can be formulated as:

$$\min_{p_G, \{\Delta p_i\}_{i=1}^N} \sum_{i=1}^{N} \frac{n_i}{\sum_j n_j} \mathcal{L}_{ce}^{D_i}(p_G + \Delta p_i). \tag{8}$$

3.4. Balance FPL's Generalization and Personalization

Previous research (Aghajanyan et al., 2020) has shown that pre-trained language models with lower intrinsic dimensions tend to exhibit better evaluation accuracy and lower relative generalization gaps across various tasks. Inspired by this, we propose that the prompt may also possess a low "intrinsic rank" during the adaptation process. To retain information derived from aggregation and knowledge-guidance of CLIP, our adaptation term is designed in a low-rank form instead of adding a full-rank term to overwrite the global prompt entirely. Specifically, the additional term is decomposed as:

$$\Delta p_i = U_i V_i. \tag{9}$$

We decompose Δp_i into multiplication between two low-rank matrices $U_i \in \mathbb{R}^{d\times b}$ and $V_i \in \mathbb{R}^{b\times k}$, where b denotes the bottleneck dimension of low-rank decomposition. Consequently, each client's personalized learnable prompt $p_i \in \mathbb{R}^{d\times k}$ can be reformulated as:

$$p_i = p_G + \Delta p_i = p_G + U_i V_i, \tag{10}$$

where $p_G \in \mathbb{R}^{d\times k}$ is the full-rank matrix and Δp_i is the low-rank component of personalized p_i. Consequently, the objective of FedPGP can be reformulated as:

$$\min_{p_G, \{U_i, V_i\}_{i=1}^N} \sum_{i=1}^{N} \frac{n_i}{\sum_j n_j} \mathcal{L}_{ce}^{D_i}(p_G + U_i V_i). \tag{11}$$

Employing the low-rank adaptation, FedPGP introduces personalization while preserving generalizability, striking a balance between the model's ability to generalize and personalize. Moreover, our objective is to go beyond the consensus knowledge communicated by clients through the global prompt and instead offer them personalized knowledge. To enable personalized prompts better access to client-specific knowledge, we aim to increase dissimilarity between representations of global and personalized prompts for better personalization.

Summarizing the target we mentioned, our objective is 1) to bring close the representations of the handcrafted prompt $g(p_C)$ of CLIP and the global prompt $g(p_G)$, and 2) to create a clear distinction between the representations of the global prompt $g(P_G)$ and personalized prompt $g(P_i)$. Building upon the above analysis, We consider the global prompt representations z_G with handcrafted prompt representation z_C as positive pairs, while simultaneously treating them as negative pairs with personalized prompt representations z_i. Consequently, we design an additional contrastive loss $\mathcal{L}_{con}$ for FedPGP to balance generalization and personalization:

$$\mathcal{L}_{con} = -\log \frac{\exp(\mathrm{sim}(z_G, z_C)/\tau)}{\exp(\mathrm{sim}(z_G, z_C)/\tau) + \exp(\mathrm{sim}(z_G, z_i)/\tau)}, \tag{12}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes a metric function (e.g., cosine similarity). The contrastive loss can guide the global prompt to gain complementary knowledge from pre-trained CLIP representation and enable Δp_i to learn personalized knowledge distinct from the global prompt. Consequently, our overall training objective thus becomes:

$$\mathcal{L} = \mathcal{L}_{ce} + \mu \mathcal{L}_{con}, \tag{13}$$

where $\mu \geq 0$ is a hyper-parameter. We offer comprehensive algorithmic details for FedPGP in Algorithm 1. For every communication round t, the selected clients locally train both the global prompt $p_{G,i}$ and the low-rank adaptation term Δp_i. After local training, the updated global prompt $p_{G,i}^{t,E}$ are sent to the server for aggregation, while the low-rank adaptation term Δp_i is retained locally.

4. Experiments

In this section, we conduct extensive experiments aiming at evaluating the generalization and personalization capability of FedPGP in scenarios of heterogeneous data distribution.

4.1. Experimental Setup

Datasets and Data Heterogeneity. Following previous research (Guo et al., 2023a;b), we selected five datasets to investigate base-to-novel class generalization ability: OxfordPets (Parkhi et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), DTD (Cimpoi et al., 2014), Caltech101 (Fei-Fei, 2004), Food101 (Bossard et al., 2014). We equally split the datasets into base and novel classes and utilized the pathological setting by assigning a specific number of non-overlapping base classes to each client. Each client model is trained on their local classes and evaluated on both local classes, base classes (classes seen on other clients), and

Algorithm 1 FedPGP

Input: Communication rounds T, local epochs R, client number N, local dataset D_i , sample numbers m_i, pre-trained CLIP model text encoder $g(\cdot)$ and image encoder $f(\cdot)$, class number K, learning rate η, the temperature of Softmax τ, hyper-parameterμ, and bottleneck number b.

```
Initialize parameters $p_i^0 = p_G^0 + \Delta p_i^0$
for each communication rounds $t \in \{1, ..., T\}$ do
  Sample client $C^t \in \{1, ..., N\}$
  for each client $i \in C^t$ do
    Initialize $p_G^{t,0} = p_G^{t-1}$,$p_i^{t,0} = p_G^{t,0} + \Delta p_i^{t-1}$
    for each local epoch $e \in \{1, ..., E\}$ do
      Sample a mini-batch $B_i \in D_i$
      Obtain the image feature $f(x)(x \in B_i)$ through image encoder $f(\cdot)$
      Obtain the global text feature $g(P_G^{t,e})$ ,the personalized text feature $g(P_i^{t,e})$, the CLIP general text feature $g(P_C)$ through text encoder $g(\cdot)$
      Calculate the cross-entropy loss $\mathcal{L}_{ce}$ ,the contrastive loss $\mathcal{L}_{con}$ according to (12) and the optimization objective $\mathcal{L} = \mathcal{L}_{ce} + \mu\mathcal{L}_{con}$
      Update prompts $p_{G,i}^{t,e} \leftarrow p_G^{t,e-1} - \eta\nabla\mathcal{L}_{D_i}$
    end for
  end for
  Aggregate and calculate the global prompt $p_G^t = \sum_{i\in C_t} \frac{n_i}{\sum_{j\in C_t} n_j} p_{G,i}^{t,E}$
end for
return $p_i = p_G + \Delta p_i$
```

novel classes (unseen in the whole training process). For domain generalization, we evaluate FedPGP on two datasets with multi-domains: DomainNet (Peng et al., 2019) with six domains and Office-Caltech10 (Gong et al., 2012) with four domains. Similar to previous research (Nguyen et al., 2022; Zhang et al., 2021), we utilize the leave-one-domain-out validation strategy. Each client participating in the federated learning system is assigned data from one of the distinct domains. We pick one domain to serve as the target domain and use the rest as source domains. Each client possesses a distinct source domain for training and then tests its model generalization ability on the whole target domain.

For evaluation of personalization, beyond the datasets used in base-to-novel class generalization, we employed two additional benchmark datasets: CIFAR-10 (Krizhevsky et al., 2010) and CIFAR-100 (Krizhevsky et al., 2009). We applied the Dirichlet Distribution, as in previous work where the datasets were partitioned randomly among clients using a symmetric Dirichlet distribution. Besides, we employ the Pathlogicacl setting the same as in base-to-novel class generalization with non-overlapping classes across clients. The Appendix Section A.1 contains comprehensive information regarding each dataset and provides additional details about the Non-IID settings.

Baselines. For generalization, we compare FedPGP with (i) Zero-shot CLIP (Radford et al., 2021) with hand-crafted text prompt template, e.g., "a photo of a [class]" (ii) CoOp (Zhou et al., 2022b) with learnable prompt vectors replacing hand-crafted text prompts trained on each client locally. (iii) PromptFL (Guo et al., 2023b) with unified prompt vectors learned across clients via FedAvg (McMahan et al., 2017) collectively. For personalization, we consider (iv) pFedPrompt (Guo et al., 2023a) which learns a unified prompt with personalized attention modules for each client and four baselines introduced in (Guo et al., 2023a), which are derived from traditional personalized federated learning techniques: (v) PromptFL+FT (Cheng et al., 2021), (vi) Prompt+Per (Arivazhagan et al., 2019), (vii) Prompt+Prox (Li et al., 2020) and Prompt+AMP (Huang et al., 2021).

Implementation Details. All methods presented in this paper are based on a frozen CLIP using two backbones, ResNet50 (He et al., 2016) and ViT-B16 (Dosovitskiy et al., 2020), defaulting to ViT-B16 if not explicitly specified. In federated learning, we set the client's local training epoch $E = 1$ and communication round $T = 150$ with $N = 100$ clients and partition rate $r = 10\%$ for CIFAR-10/CIFAR-100 datasets. Besides, we consider training epoch $E = 2$ and communication round $T = 25$ with client numbers $N = 10$ and a full partition rate, i.e., $r = 100\%$ for other datasets. The low-rank decomposition bottleneck is set to $b = 8$, and the hyperparameter μ for the contrastive loss is set to 1. We employ cosine similarity as the metric function in contrastive loss. For the setting of learnable prompts, the length of prompt vectors p is 16 with a dimension of 512, token position is "end" with "random" initialization. Apart from the few-shot learning, batch sizes are set to 32 during training and 100 during testing. Additional implementation details can be found in the Appendix Section A.2.

4.2. Performance Evaluation

Base-to-Novel Class Generalization. We evaluated the performance of FedPGP against baselines on their local classes, base classes, and novel classes respectively. We present the harmonic mean (HM) of these three accuracies to demonstrate the overall performance. The experiment results are summarized in Table 1. As indicated in Table 1(a), FedPGP achieves the best performances in local classes, highlighting its exceptional personalization capability. Moreover, FedPGP outperforms other methods in both base classes and the harmonic meanwhile also exhibiting the second-best performance in novel classes. These results show its exceptional capacity for balancing personalization and generalization. CoOp achieves the second-best performance in local classes owing to the pathological setting. However, it

Table 1: Accuracy comparison (%) on clients' local classes and Base-to-novel generalization.

(a) Average over 5 datasets.

Methods	Local	Base	Novel	HM
CLIP (Radford et al., 2021)	79.18	79.83	**83.25**	80.72
CoOp (Zhou et al., 2022b)	94.28	69.40	73.16	77.55
PromptFL (Guo et al., 2023b)	90.00	85.65	78.53	84.46
Prompt+Prox (Li et al., 2020)	89.84	85.04	77.40	83.78
FedMaPLe	90.81	84.90	81.49	85.56
FedCoCoOp	90.01	85.08	81.4	85.35
FedPGP	**95.67**	**85.69**	81.75	**87.33**

(b) OxfordPets.

Methods	Local	Base	Novel	HM
CLIP (Radford et al., 2021)	89.34	89.31	96.86	91.70
CoOp (Zhou et al., 2022b)	95.33	82.51	92.92	89.90
PromptFL (Guo et al., 2023b)	95.12	95.16	91.89	94.03
Prompt+Prox (Li et al., 2020)	95.95	95.24	91.25	94.10
FedMaPLe	93.75	95.53	**97.45**	95.55
FedCoCoOp	96.02	96.01	97.25	96.42
FedPGP	**96.65**	**95.87**	97.33	**96.61**

(c) Flowers102.

Methods	Local	Base	Novel	HM
CLIP (Radford et al., 2021)	67.69	68.85	**77.23**	71.01
CoOp (Zhou et al., 2022b)	96.39	55.91	64.47	68.54
PromptFL (Guo et al., 2023b)	94.32	76.19	70.1	78.96
Prompt+Prox (Li et al., 2020)	92.73	73.06	66.09	75.75
FedMaPLe	94.89	77.49	70.46	79.71
FedCoCoOp	94.57	77.88	74.39	81.4
FedPGP	**99.68**	**78.48**	75.11	**83.13**

(d) DTD.

Methods	Local	Base	Novel	HM
CLIP (Radford et al., 2021)	53.79	54.62	**58.20**	55.47
CoOp (Zhou et al., 2022b)	86.38	39.2	37.65	47.13
PromptFL (Guo et al., 2023b)	72.71	71.41	49.28	62.44
Prompt+Prox (Li et al., 2020)	74.07	**71.84**	50.20	63.37
FedMaPLe	78.37	65.35	55.85	65.26
FedCoCoOp	72.61	68.20	54.4	64.08
FedPGP	**89.07**	69.65	55.25	**68.15**

(e) Caltech101.

Methods	Local	Base	Novel	HM
CLIP (Radford et al., 2021)	95.72	96.96	93.99	95.54
CoOp (Zhou et al., 2022b)	99.39	86.37	86.12	90.22
PromptFL (Guo et al., 2023b)	97.04	97.27	92.79	95.65
Prompt+Prox (Li et al., 2020)	96.76	**97.34**	91.99	95.30
FedMaPLe	96.47	96.7	**94.32**	95.82
FedCoCoOp	96.65	95.45	92.46	94.82
FedPGP	**99.46**	96.09	93.62	**96.33**

(f) Food101.

Methods	Local	Base	Novel	HM
CLIP (Radford et al., 2021)	89.38	89.39	**89.98**	89.58
CoOp (Zhou et al., 2022b)	**93.92**	82.99	84.62	86.92
PromptFL (Guo et al., 2023b)	90.79	88.22	88.6	89.19
Prompt+Prox (Li et al., 2020)	89.68	87.72	87.49	88.29
FedMaPLe	90.59	**89.43**	89.38	89.80
FedCoCoOp	90.18	87.86	88.51	88.84
FedPGP	93.51	88.37	88.44	**90.04**

Table 2: The average classification accuracy using leave-one-domain-out validation on Offica-Caltech10 and DomainNet.

Datasets	Office-Caltech10					DomainNet						
Domains	A	C	D	W	Avg	C	I	P	Q	R	S	Avg.
CLIP (Radford et al., 2021)	19.40	18.32	21.87	18.59	19.55	49.89	47.23	53.61	32.10	48.19	50.79	46.96
CoOp (Zhou et al., 2022b)	41.54	15.55	56.04	43.60	39.18	83.42	53.28	80.80	49.41	75.18	82.88	70.83
PromptFL (Guo et al., 2023b)	96.34	91.57	97.96	98.30	96.04	95.28	73.72	94.50	61.60	95.72	95.43	86.04
Prompt+Prox (Li et al., 2020)	96.13	**92.52**	97.57	97.96	96.05	95.47	69.44	94.95	61.24	75.18	95.41	81.95
FedPGP	**96.55**	91.92	**98.93**	**98.75**	**96.54**	**96.45**	**74.46**	**95.43**	**62.12**	**96.06**	**96.05**	**86.76**

cannot effectively generalize its performance to other base classes and new classes. When it comes to FPL, PromptFL and PromptProx sacrifice the personalization capability in local classes to gain better generalization ability.

Leave-One-Domain-Out Generalization. Table 2 shows the average classification accuracy with leave-one-domain-out validation on Office-Caltech10 and DomainNet. As we can see, FedPGP achieves the highest average accuracy and outperforms all baselines across nearly all target domains. Through local prompt tuning, CoOp's domain generalization capabilities generally surpass those of CLIP. We notice that FPL enhances the model's domain generalization power, marking a significant improvement compared

Table 3: Accuracy comparison (%) on the Pathological Non-IID setting over 10 clients.

Methods	OxfordPets	Flowers102	DTD	Caltech101	Food101
CoOp (Zhou et al., 2022b)	83.21±1.30	70.14±0.76	44.23±0.63	87.37±0.44	70.43±2.42
PromptFL (Zhou et al., 2022b)	90.79±0.61	72.80±1.14	54.11±0.22	89.70±1.99	77.31±1.64
PromptFL+FT (Cheng et al., 2021)	91.23±0.50	72.31±0.91	53.74±1.36	89.70±0.25	77.16±1.56
Prompt+PER (Arivazhagan et al., 2019)	89.50±1.62	72.11±1.35	50.23±0.82	86.72±1.45	71.29±1.87
Prompt+Prox (Li et al., 2020)	89.24±0.41	66.40±0.29	44.26±1.11	89.41±0.55	76.24±1.94
Prompt+AMP (Huang et al., 2021)	80.21±0.44	69.10±0.13	47.16±0.92	87.31±1.60	74.48±1.71
pFedPrompt (Guo et al., 2023a)	91.84±0.41	86.46±0.15	77.14±0.09	96.54±1.31	92.26±1.34
FedPGP	**98.96±0.42**	**99.29±0.03**	**91.52±0.41**	**98.90±0.19**	**95.52±0.15**

Table 4: The detailed classification accuracy using leave-one-domain-out validation on Offica-Caltech10 dataset.

Datasets	Office-Caltech10					
Source Domains		Amazon	Caltech	DSLR	Webcam	Avg.
CoOp (Zhou et al., 2022b)	Amazon	——	89.03	16.49	19.1	41.54±41.15
	Caltech	26.89	——	5.87	13.89	15.55±10.61
	DSLR	64.96	86.62	——	16.56	56.04±35.87
	Webcam	50.16	76.94	3.72	——	43.6±37.05
FedPGP	Amazon	——	96.45	96.03	97.18	**96.55±0.58**
	Caltech	94.66	——	86.92	93.59	**91.92±4.19**
	DSLR	98.08	99.36	——	99.36	**98.93±0.74**
	Webcam	98.98	98.98	98.3	——	**98.75±0.39**

to the local approach. Moreover, FedPGP enhances the model's ability to generalize while accomplishing personalization, which proves the effectiveness of our framework's design. We provide the detailed classification accuracy on each source domain within the Office-Caltech10 dataset in Table 4. Additional experiment results on specific client domain generalization are available in the Appendix Section B.1.

Evaluation on Personalization. We report the performance of FedPGP against baselines in Table 3, 5. To facilitate comparison, we present the results in Table 3 utilizing ResNet50 as the backbone, aligning with the setting in (Guo et al., 2023a). As shown in Table 3, our FedPGP demonstrates significantly superior performance compared to baseline methods across all datasets. This confirms that our framework's ability to personalize effectively is successful in addressing extreme non-IID scenarios. Table 5 shows the results of FedPGP and baseline methods on CIFAR-10 and CIFAR100 datasets with Dirichlet Non-IID setting over 100 clients with 10% partition. In the scenario with Dirichlet settings and the substantial number of clients, our approach FedPGP consistently demonstrates superior performance compared to the baseline methods. This further emphasizes the effectiveness of our approach.

Table 5: Accuracy comparison (%) on the Dirichlet Non-IID setting in CIFAR-10 and CIFAR-100 over 100 clients.

Methods	CIFAR-10	CIFAR-100
CLIP (Radford et al., 2021)	87.52±0.56	64.83±0.49
CoOp (Zhou et al., 2022b)	93.13±0.34	74.78±0.41
PromptFL (Zhou et al., 2022b)	92.32±0.79	73.72±0.61
Prompt+Prox (Li et al., 2020)	91.79±0.46	71.08±0.89
FedPGP	**94.82±0.37**	**77.44±0.15**

4.3. Ablation Study

Effect of Parameter μ of Contrastive Loss In this subsection, we investigated the impact of the contrastive loss parameter μ in a Pathological Non-IID setting across four datasets with varying shot numbers. The results are presented in Figure 2, which shows an improvement in test accuracy with an increase in the number of shots. Upon observation, optimal results are mostly achieved with μ set to 1 in experiments, leading to our adoption of $\mu = 1$ for other experiments.

Effective of Low-rank Adaption In this subsection, we explored the effectiveness of low-rank adaptation by comparing it with full-rank adaptation. The results of the two

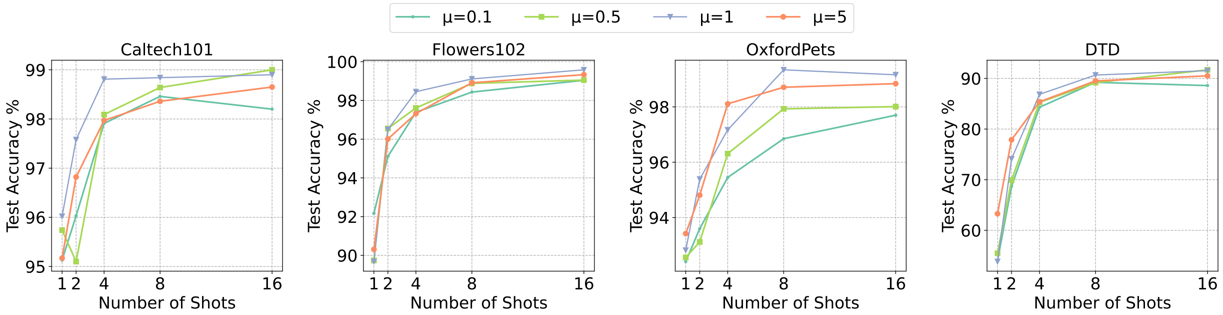


Figure 2: Quantitative comparisons on four datasets across varying shot numbers and parameter μ of contrastive loss in FedPGP over 10 clients.

Table 6: Accuracy (%) of ablation study on adaption and additional loss for clients' local classes and Base-to-novel generalization.

Methods	Local	Base	Novel	HM
FedPGP w/o Positive	94.63	84.68	77.75	85.13
FedPGP w/ Full-rank Adaption	**98.57**	48.00	63.40	64.17
FedPGP	95.67	**85.69**	**81.75**	**87.33**

methods are shown in Table 6. As we can see, full-rank adaptation achieves the best performance on local classes but it completely overwrites the global prompt, resulting in a loss of category- agnostic knowledge and the generalization capacity. Although low-rank adaptation performance on the local class is below the full-rank adaptation, it significantly outperforms the full-rank adaptation in terms of base-to-novel generalization.

Effective of Contrastive Loss We demonstrate the efficacy of contrastive loss in achieving balance by separately testing the generalization ability of the model without positive pairs and the personalization ability of the model with negative pairs. Table 6 shows the performance of the model without knowledge-guidance from CLIP (positive pairs), compared to the performance when the contrastive loss is employed. FedPGP outperforms the model without knowledge-guidance across all three accuracies and the harmonic mean. Table 7 shows the performance of the model without pushing the representation of global prompt and personalized prompt (negative pairs), compared to the performance when the contrastive loss is employed. The results show the negative pairs enhance the model's ability to personalize for Non-IID data distribution in federated prompt learning. In general, our ablation study on contrastive loss demonstrates its ability to balance personalization and generalization in federated prompt learning.

Table 7: Accuary (%) of ablation study on additional loss for personalization.

Methods	OxfordPets	Flowers102	DTD	Caltech101	Food101
FedPGP w/o Negative	97.65±0.20	98.63±0.11	90.78±0.31	98.48±0.17	94.72±0.18
FedPGP	**98.96±0.42**	**99.29±0.03**	**91.52±0.41**	**98.90±0.19**	**95.52±0.15**

5. Conclusion

In this paper, we propose a novel approach named FedPGP, which represents a pioneering effort to harmonize personalization and generalization in federated prompt learning. In our approach, each client gains generalization capabilities through knowledge-guidance of CLIP and acquires personalization abilities by adapting the global prompt to a personalized prompt. Further, with low-rank decomposition adaptation and an extra contrastive loss, FedPGP learns a personalized prompt for each client in heterogeneous federated scenarios while preserving the remarkable generalization capacity in pre-trained Vision-Language models. Extensive experiments on various datasets explored base-to-novel generalization in both unseen categories and domains, showing the superiority of FedPGP in balancing generalization and personalization. In future work, we aim to explore the theoretical foundations of low-rank adaptation in federated prompt learning.

Acknowledgement

This work was supported by NSFC (No.62303319), Shanghai Sailing Program (22YF1428800, 21YF1429400), Shanghai Local College Capacity Building Program (23010503100), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human Machine Collaboration (ShanghaiTech University), and Shanghai Engineering Research Center of Intelligent Vision and Imaging.

Impact Statement

There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 446–461. Springer, 2014.

Cai, Z., Shi, Y., Huang, W., and Wang, J. Fed-CO$_2$: Cooperation of online and offline models for severe data heterogeneity in federated learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Cao, Y.-T., Shi, Y., Yu, B., Wang, J., and Tao, D. Knowledge-aware federated active learning with non-iid data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22279–22289, 2023.

Chen, H.-Y. and Chao, W.-L. On bridging generic and personalized federated learning for image classification. *arXiv preprint arXiv:2107.00778*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Cheng, G., Chadha, K., and Duchi, J. Fine-tuning is fine in federated learning. *arXiv preprint arXiv:2108.07313*, 3, 2021.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fei-Fei, L. Learning generative visual models from few training examples. In *Workshop on Generative-Model Based Vision, IEEE Proc. CVPR, 2004*, 2004.

Feng, C.-M., Li, B., Xu, X., Liu, Y., Fu, H., and Zuo, W. Learning Federated Visual Prompt in Null Space for MRI Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8064–8073, 2023.

Gong, B., Shi, Y., Sha, F., and Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2066–2073. IEEE, 2012.

Guo, T., Guo, S., and Wang, J. pFedPrompt: Learning Personalized Prompt for Vision-Language Models in Federated Learning. In *Proceedings of the ACM Web Conference 2023*, pp. 1364–1374, 2023a.

Guo, T., Guo, S., Wang, J., Tang, X., and Xu, W. PromptFL: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 2023b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Huang, W., Shi, Y., Cai, Z., and Suzuki, T. Understanding convergence and generalization in federated learning through feature learning theory. In *The Twelfth International Conference on Learning Representations*, 2023.

Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., and Zhang, Y. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 7865–7873, 2021.

Hyeon-Woo, N., Ye-Bin, M., and Oh, T.-H. Fedpara: Low-rank hadamard product for communication-efficient federated learning. *arXiv preprint arXiv:2108.06098*, 2021.

Jeong, W. and Hwang, S. J. Factorized-FL: Personalized Federated Learning with Parameter Factorization & Similarity Matching. *Advances in Neural Information Processing Systems*, 35:35684–35695, 2022.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023a.

Khattak, M. U., Wasim, S. T., Naseer, M., Khan, S., Yang, M.-H., and Khan, F. S. Self-regulating Prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15190–15200, 2023b.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5(4):1, 2010.

Li, H., Cai, Z., Wang, J., Tang, J., Ding, W., Lin, C.-T., and Shi, Y. FedTP: Federated Learning by Transformer Personalization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Li, H., Huang, W., Wang, J., and Shi, Y. Global and local prompts cooperation via optimal transport for federated learning. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Li, Q., He, B., and Song, D. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021a.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021b.

Li, X., Jiang, M., Zhang, X., Kamp, M., and Dou, Q. FedBN: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021c.

Liu, Q., Chen, C., Qin, J., Dou, Q., and Heng, P.-A. FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1013–1023, 2021.

Ma, C., Liu, Y., Deng, J., Xie, L., Dong, W., and Xu, C. Understanding and mitigating overfitting in prompt tuning for vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Ma, X., Zhang, J., Guo, S., and Xu, W. Layer-wised model aggregation for personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10092–10101, 2022.

Mansour, Y., Mohri, M., Ro, J., and Suresh, A. T. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mu, X., Shen, Y., Cheng, K., Geng, X., Fu, J., Zhang, T., and Zhang, Z. FedProc: Prototypical contrastive federated learning on non-iid data. *Future Generation Computer Systems*, 143:93–104, 2023.

Nguyen, A. T., Torr, P., and Lim, S. N. FedSR: A simple and effective domain generalization method for federated learning. *Advances in Neural Information Processing Systems*, 35:38831–38843, 2022.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Oh, J., Kim, S., and Yun, S.-Y. FedBABU: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*, 2021.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Qiu, C., Li, X., Mummadi, C. K., Ganesh, M. R., Li, Z., Peng, L., and Lin, W.-Y. Text-driven Prompt Generation for Vision-Language Models in Federated Learning. *arXiv preprint arXiv:2310.06123*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Sattler, F., Müller, K.-R., and Samek, W. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

Shamsian, A., Navon, A., Fetaya, E., and Chechik, G. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pp. 9489–9502. PMLR, 2021.

Su, S., Yang, M., Li, B., and Xue, X. Cross-domain federated adaptive prompt tuning for clip. *arXiv preprint arXiv:2211.07864*, 2022.

T Dinh, C., Tran, N., and Nguyen, J. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Tan, A. Z., Yu, H., Cui, L., and Yang, Q. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022a.

Tan, Y., Long, G., Ma, J., Liu, L., Zhou, T., and Jiang, J. Federated learning from pre-trained models: A contrastive learning approach. *Advances in Neural Information Processing Systems*, 35:19332–19344, 2022b.

Wang, K., Mathews, R., Kiddon, C., Eichner, H., Beaufays, F., and Ramage, D. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019.

Wei, G., Wang, F., Shah, A., and Chellappa, R. Dual prompt tuning for domain-aware federated learning. *arXiv preprint arXiv:2310.03103*, 2023.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Yang, F.-E., Wang, C.-Y., and Wang, Y.-C. F. Efficient model personalization in federated learning via client-specific prompt generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19159–19168, 2023.

Zhang, L., Lei, X., Shi, Y., Huang, H., and Chen, C. Federated learning with domain generalization. *arXiv preprint arXiv:2111.10487*, 2021.

Zhang, L., Shi, Y., Chang, Y.-C., and Lin, C.-T. Federated fuzzy neural network with evolutionary rule learning. *IEEE Transactions on Fuzzy Systems*, 2022.

Zhang, M., Sapra, K., Fidler, S., Yeung, S., and Alvarez, J. M. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.

Zhao, H., Du, W., Li, F., Li, P., and Liu, G. Reduce communication costs and preserve privacy: Prompt tuning method in federated learning. *arXiv preprint arXiv:2208.12268*, 2022.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15659–15669, 2023.

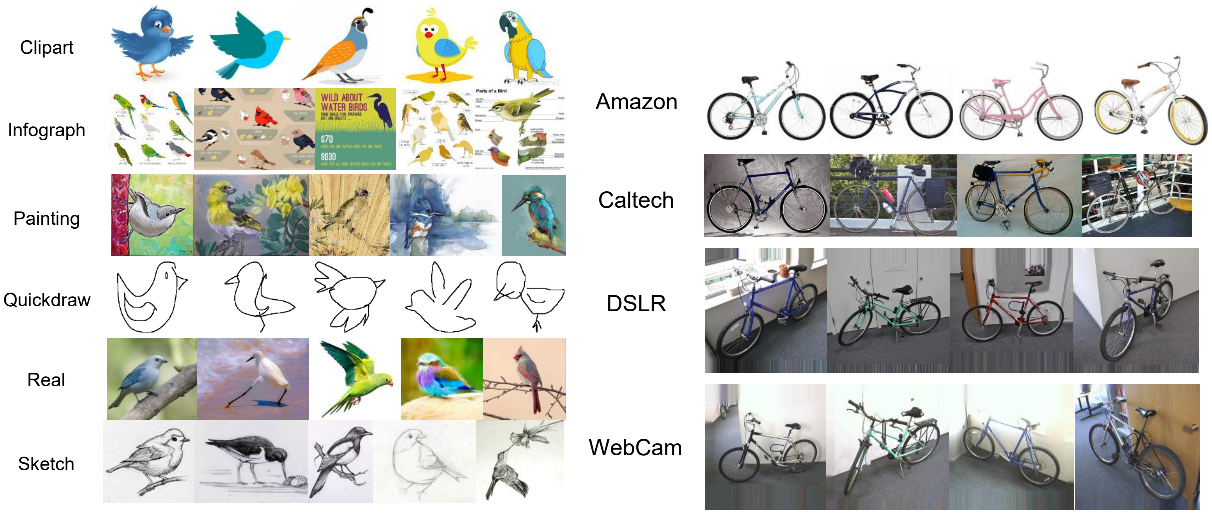
A. Experimental Details

A.1. Dataset Setup

For our evaluation, we've chosen nine diverse visual classification datasets as our benchmark. Table 8 provides a detailed overview, including information on original tasks, class numbers, training and testing sample sizes, and domain counts. In datasets with multiple domains, we utilize the well-established Office-Caltech10 benchmark, featuring four domains: Amazon, Caltech, DSLR, and WebCam. These domains capture variations arising from different camera devices and real-world environments. Additionally, we leverage DomainNet, a large-scale dataset comprising six domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. We focus on training with 10 selected classes from each dataset. Visual examples of raw instances from these two multi-domain datasets can be found in Figure 3.

Table 8: Statistical details of datasets used in experiments.

Dataset	Classes	Train	Test	Domains	Task
OxfordPets (Parkhi et al., 2012)	37	2,944	3,669	1	Fine-grained pets recognition
Flowers102 (Nilsback & Zisserman, 2008)	102	4,093	2,463	1	Fine-grained flowers recognition
DTD (Cimpoi et al., 2014)	47	2,820	1,692	1	Texture recognition
Caltech101 (Fei-Fei, 2004)	100	4,128	2,465	1	Object recognition
Food101 (Bossard et al., 2014)	101	50,500	30,300	1	Fine-grained food recognition
CIFAR10 (Krizhevsky et al., 2009)	10	50,000	10,000	1	Image Classification
CIFAR100 (Krizhevsky et al., 2009)	100	50,000	10,000	1	Image Classification
DomainNet (Peng et al., 2019)	10	18278	4573	6	Image recognition
Office-Caltech10 (Gong et al., 2012)	10	2025	508	4	Image recognition



(a) DomainNet

(b) Office-Caltech10

Figure 3: Visual examples of raw instances from two datasets with multiple domains: "Bird" in DomainNet (left) and "Bike" in Office-Caltech10 (right).

A.2. Experimental Setup

We employ SGD optimizer with learning rate $\eta = 0.001$. The experiments were conducted three times using different seeds. We calculated the average performance and the final result in federated prompt learning is obtained by averaging the performance across all clients. All experiments are conducted with Pytorch on NVIDIA A40 GPUs.

Base-to-Novel Class Generalization. For Base-to-Novel generalization, we separate each dataset into base and novel classes equally and distribute the base classes to each client without overlapping. Each client trains their local model on their local classes, and we evaluate their personalized prompt on both local classes, base classes (classes seen on other clients but unseen during local training), and novel classes (unseen in the whole training process). The accuracy is the average overall 10 clients.

Leave-One-Domain-Out Generalization. For Leave-One-Domain-Out generalization, each client participating in the federated learning system is assigned data from one of the distinct domains. We pick one domain to serve as the target domain and use the rest of the domains as source domains. Each client possesses a distinct source domain for training and then tests its model generalization ability on the whole target domain. The accuracy is the average overall 3 clients in Office-Caltech10 and 5 clients in DomainNet.

Personalization. For evaluation of personalization, we apply Dirichlet distribution on CIFAR-10 and CIFAR-100 over 100 clients. Specifically, the datasets are partitioned randomly among clients using a symmetric Dirichlet distribution with hyperparameter α = 0.3. Besides, we employ the Pathlogicacl setting the same as in base-to-novel class generalization with non-overlapping classes across 10 clients for other datasets.

Effect of Individual Components. For the ablation study, we employ $\mathcal{L}_{neg} = 1 - \text{sim}(z_G, z_i)$ as the additional loss for FedPGP without positive pairs, and $\mathcal{L}_{pos} = \text{sim}(z_G, z_C)$ as the additional loss for FedPGP without negative pairs. The hyperparameter is set as $\mu = 1$ for the additional loss.

B. Additional Experimental Results

B.1. Detailed Results of Leave-One-Domain-Out Generalization

In Table 4 and Table 9, we provide the detailed classification accuracy on each source domain within Office-Caltech10 and DomainNet datasets, respectively. Notably, as PromptFL and PromptProx utilize a single global prompt, their results remain consistent across different source domains. Therefore, the presented results specifically focus on CoOp and our FedPGP, both employing distinct local models for each client. To be specific, the values shown in the table indicate the testing results on the target domain across clients with different source domains. Comparing the results of CoOp and our FedPGP, we observe that FedPGP consistently outperforms CoOp in all cases with a significantly smaller standard deviation, showcasing the robust generalization capability of our proposed method.

Table 9: The detailed classification accuracy using leave-one-domain-out validation on DomainNet dataset.

Datasets	DomainNet							
Source Domains		Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg.
CoOp (Zhou et al., 2022b)	Clipart	——	81.17	82.80	69.98	90.71	92.46	83.42±8.96
	Infograph	57.41	——	60.00	30.64	65.09	53.24	53.28±15.20
	Painting	86.80	70.88	——	67.06	92.62	83.04	80.80±11.67
	Quickdraw	48.50	44.02	50.00	——	53.02	51.50	49.41±3.94
	Real	89.95	79.27	93.70	27.29	——	85.69	75.18±30.05
	Sketch	86.55	84.27	86.91	64.51	92.15	——	82.88±12.09
FedPGP	Clipart	——	96.23	96.80	95.96	96.27	96.99	**96.45±0.43**
	Infograph	76.45	——	74.44	69.29	76.24	75.90	**74.46±3.21**
	Painting	96.18	95.89	——	93.01	96.02	96.08	**95.43±1.50**
	Quickdraw	66.22	52.30	63.42	——	62.30	66.36	**62.12±6.11**
	Real	97.12	95.75	97.03	93.17	——	97.25	**96.06±1.87**
	Sketch	96.71	95.88	96.56	94.75	96.38	——	**96.05±0.81**

B.2. Detailed Results of Individual Components in Base-to-Novel Generalization

Table 10 presents the per-dataset results for each component of our FedPGP framework in the Base-to-novel generalization setting. The results demonstrate the effectiveness of CLIP knowledge-guidance in enhancing performance for both base and novel classes. Additionally, even though full-rank adaptation outperforms our low-rank adaptation on local classes, its generalization on both base and novel classes significantly diminishes due to overwriting the global prompt. These findings emphasize the efficacy of FedPGP in enhancing model generalization across diverse datasets.

Table 10: Accuracy (%) of ablation study on adaption and additional loss for clients' local classes and Base-to-novel generalization.

Dataset	Methods	Local	Base	Novel	HM
Average over 5 datasets	FedPGP w/o Positive	94.63	84.68	77.75	85.13
	FedPGP w/ Full-rank Adaption	**98.57**	48.00	63.40	64.17
	FedPGP	95.67	**85.69**	**81.75**	**87.33**
Oxford Pets	FedPGP w/o Positive	95.88	95.44	96.77	96.03
	FedPGP w/ Full-rank Adaption	**99.94**	41.28	73.32	62.67
	FedPGP	96.65	**95.87**	**97.33**	**96.61**
Flowers102	FedPGP w/o Positive	98.73	77.18	62.22	76.61
	FedPGP w/ Full-rank Adaption	**99.91**	35.43	53.97	52.85
	FedPGP	99.68	**78.48**	**75.11**	**83.13**
DTD	FedPGP w/o Positive	87.34	67.33	49.83	64.70
	FedPGP w/ Full-rank Adaption	**96.29**	25.12	34.81	38.01
	FedPGP	89.07	**69.65**	**54.25**	**68.15**
Caltech101	FedPGP w/o Positive	98.34	96.02	92.57	95.58
	FedPGP w/ Full-rank Adaption	**99.89**	75.94	81.08	84.48
	FedPGP	99.46	**96.09**	**93.62**	**96.33**
Food101	FedPGP w/o Positive	92.85	87.42	87.36	89.14
	FedPGP w/ Full-rank Adaption	**96.84**	62.21	73.8	75.09
	FedPGP	93.51	**88.37**	**88.44**	**90.04**

B.3. Effect of Number of Bottleneck

In this subsection, we explore the impact of the number of bottleneck b in our low-rank decomposition of adaptation term Δp_i. We present the accuracy results considering the impact of both the bottleneck and shot number using a random seed. It can be observed that the classification accuracy improves as the bottleneck and shot number increase, showing the number of bottleneck determines the extent to which the knowledge in the global prompt is rewritten. We select the number of bottleneck $b = 8$ for the balance of generalization and personalization.

B.4. Learning Curves

To analyze the convergence pattern of our FedPGP, we visualized the test accuracy across 10 clients with a local training epoch $E = 2$ and communication round $T = 25$. The results are illustrated in Figure 4, revealing accelerated convergence and enhanced stability exhibited by FedPGP.

Table 11: Quantitative comparisons on 4 datasets across varying number of shots with different number of bottleneck in FedPGP over 10 clients.

Dataset	Bottleneck	1 shot	2 shots	4 shots	8 shots	16 shots
Oxford Pets	1	92.4	92.89	93.96	94.28	95.12
	2	92.39	93.04	94.93	95.91	96.39
	4	92.51	**93.62**	94.66	96.72	97.32
	8	**93.16**	93.12	**96.31**	**97.93**	**97.81**
Flowers102	1	86.89	91.92	96.26	98.56	98.75
	2	87.79	93.95	96.28	97.60	98.71
	4	87.77	94.86	97.61	**98.92**	**99.37**
	8	**89.74**	**96.55**	**97.64**	98.88	99.05
DTD	1	53.13	60.52	70.41	85.61	83.00
	2	52.63	58.77	73.97	87.75	91.05
	4	55.02	66.05	76.80	**89.27**	90.08
	8	**55.47**	**69.91**	**85.27**	89.16	**92.00**
Caltech101	1	93.46	93.93	96.06	97.62	98.40
	2	93.27	94.36	96.69	97.89	98.33
	4	94.44	**96.32**	97.02	98.20	98.30
	8	**95.74**	95.10	**98.09**	**98.28**	**99.00**

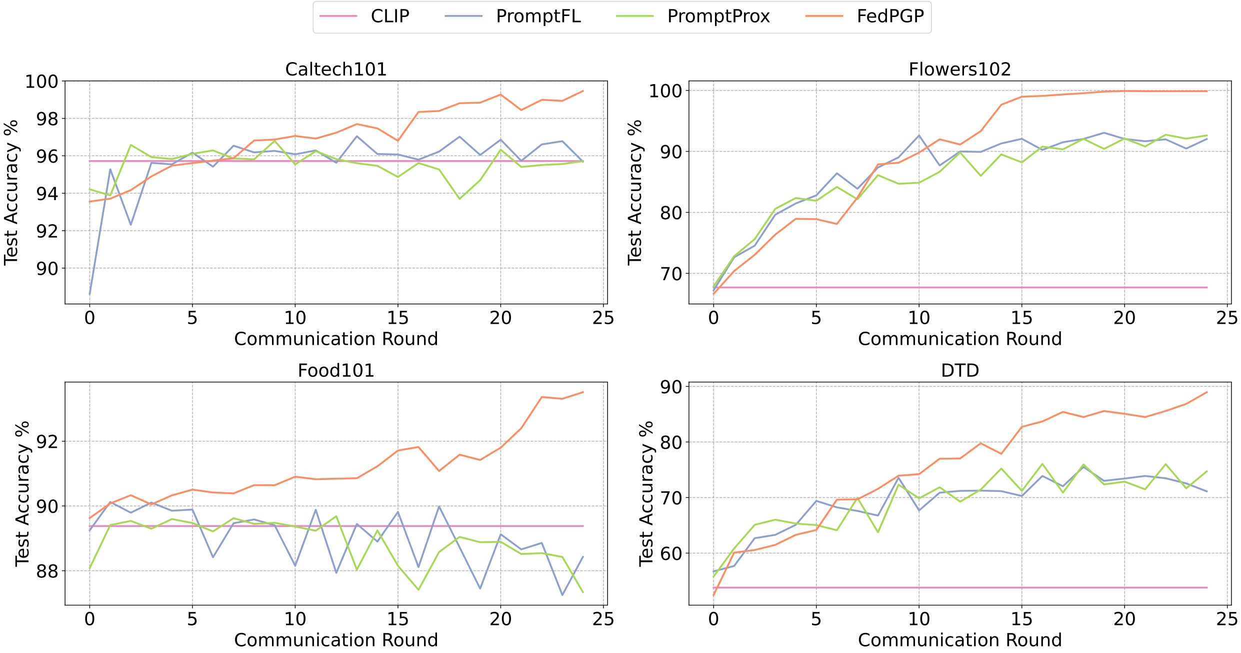


Figure 4: Accuracy learning curves of FedPGP and baselines on four datasets over 10 clients.