An Empirical Study of Federated Prompt Learning for Vision Language Model

Zhihao Wang[*1], **Wenke Huang**[*1], **Tian Chen**[*1], **Zekun Shi**[1], **Guancheng Wan**[1], **Yu Qiao**[1], **Bin Yang**[1], **Jian Wang**[†1,2], **Bing Li**[†1,2] and **Mang Ye**[†1]

[1]School of Computer Science, Wuhan University
[2]Zhongguancun Laboratory, China
{zhihao_wang, wenkehuang, tian.chen}@whu.edu.cn

Abstract

The Vision Language Model (VLM) excels in aligning vision and language representations, and prompt learning has emerged as a key technique for adapting such models to downstream tasks. However, the application of prompt learning with VLM in federated learning (*FL*) scenarios remains underexplored. This paper systematically investigates the behavioral differences between language prompt learning (LPT) and vision prompt learning (VPT) under data heterogeneity challenges, including label skew and domain shift. We conduct extensive experiments to evaluate the impact of various *FL* and prompt configurations, such as client scale, aggregation strategies, and prompt length, to assess the robustness of Federated Prompt Learning (FPL). Furthermore, we explore strategies for enhancing prompt learning in complex scenarios where label skew and domain shift coexist, including leveraging both prompt types when computational resources allow. Our findings offer practical insights into optimizing prompt learning in federated settings, contributing to the broader deployment of VLMs in privacy-preserving environments.

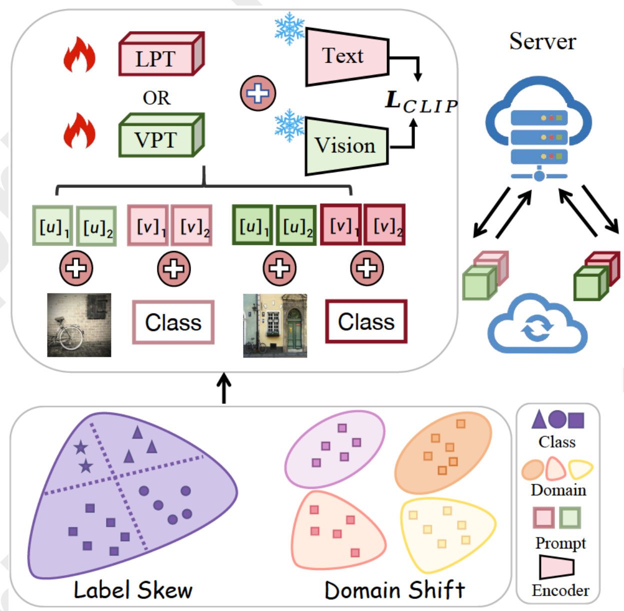


Figure 1: **Background** for Federated Prompt Learning. Participants learn shared visual or textual prompts. However, under data heterogeneity (Label Skew and Domain Shift), aggregated prompts struggle to capture consistency among clients. See details in Sec. 1.

1 Introduction

The advent of large-scale Vision Language Models, such as Contrastive Language-Image Pretraining (*CLIP*) [Radford *et al.*, 2021], has redefined multi-modal learning by enabling effective alignment between visual and textual representations. These models have demonstrated remarkable adaptability across various downstream tasks with prompt tuning [Shin *et al.*, 2020; Lester *et al.*, 2021], particularly in zero-shot and few-shot settings. However, due to the inconsistency of upstream and downstream data distribution and training costs, large-scale Vision Language Models such as *CLIP* still face many challenges in training and deployment.

Prompt learning [Bang *et al.*, 2024; Li *et al.*, 2024b], a lightweight yet effective paradigm for fine-tuning large pre-trained models, has emerged as a promising solution. By introducing task-specific prompts rather than retraining full model parameters, prompt learning offers an efficient means of adapting models to new tasks. While this approach has been extensively examined in centralized settings, its potential for optimizing VLM under federated conditions remains underexplored [Guo *et al.*, 2023b]. In these scenarios, the design and tuning of prompts are particularly critical, as they must account for the decentralized nature of data and the limited communication resources available in *FL*.

In Federated Learning scenarios, the challenges of data heterogeneity [Zhao *et al.*, 2018; Kairouz *et al.*, 2019; Li *et al.*, 2020a; Tan *et al.*, 2023] are further compounded by issues like **Label Skew and Domain Shift** as shown in Fig. 1. Label skew occurs when different clients have significantly imbalanced class distributions, leading to poor generalization on unseen data. Domain shift arises when the underlying data distributions vary across clients, making it difficult for the model to converge to a globally optimal solution. For VLMs enhanced with prompt learning, these issues manifest in distinct ways: language prompts may be less effective when the semantic representation of classes varies widely due to label skew, whereas visual prompts may struggle to capture consistent image features across diverse domains.

Motivated by these observations, our work conducts an in-depth empirical study of prompt learning for VLM in *FL* settings. We systematically analyze the distinct behaviors of language prompt learning and vision prompt learning in scenarios characterized by Label Skew and Domain Shift. Specifically, we explore how each type of prompt responds to imbalanced class distributions and varying domain characteristics, thereby uncovering the strengths and limitations inherent in each approach. We study the influence of different *FL* configurations alongside various prompt settings. This analysis aims to explore the robustness of federated prompt learning. Furthermore, recognizing that real-world applications often involve simultaneous challenges from both Label Skew and Domain Shift, we investigate strategies to boost the effectiveness of prompt learning in such complex environments.

Our investigation is structured around three main research objectives:

- **Q1:** *Do Language Prompt Tuning (LPT) and Vision Prompt Tuning (VPT) show* **Behavior Discrepancies** *under label skew and domain shift?*
- **Q2:** *Does federated prompt learning exhibits* **Robustness Patterns** *in various federated settings and prompt configurations?*
- **Q3:** *Do federated visual and textual prompts appear* **Collaboration Effect** *for complex heterogeneity?*

By addressing these three key areas, our work not only elucidates the differential impacts of language and vision prompts in federated settings but also provides practical guidelines for optimizing prompt learning in large-scale, distributed vision-language models. This study paves the way for more robust and resource-efficient deployments of VLM in real-world *FL* applications.

2 Related Work

2.1 Federated Learning

Federated Learning (*FL*) enables collaborative model training across decentralized clients while preserving data privacy [Konečný *et al.*, 2016; Kairouz *et al.*, 2019; Li *et al.*, 2020b; Huang *et al.*, 2022; Dai *et al.*, 2023; Huang *et al.*, 2023a; Huang *et al.*, 2024]. However, it faces significant challenges, particularly in scenarios involving label skew and domain shift, which substantially degrade model performance due to the presence of non-IID (non-independent and identically distributed) data [Luo *et al.*, 2021; Li *et al.*, 2022; Ma *et al.*, 2022; Ye *et al.*, 2023; Dai *et al.*, 2023; Hu *et al.*, 2024; Wang *et al.*, 2025]. Label skew occurs when the class distributions vary significantly across clients, leading to poor generalization on unseen data. Domain shift, on the other hand, arises when the input data distribution differs across clients, further complicating model convergence.

To address label skew, FedProx [Li *et al.*, 2020c]introduced a proximal term to the optimization objective, stabilizing training in non-IID settings. Similarly, MOON [Li *et al.*, 2021a] proposed contrastive learning at the client level to align model representations and mitigate discrepancies caused by skewed labels. For domain shift, FedBN [Li *et al.*, 2021b] employed client-specific batch normalization layers, allowing local domain-specific feature extraction without compromising global model updates. Other methods, such as MFL[Liu *et al.*, 2020] and CFL[Sattler *et al.*, 2021], attempted to balance global and local objectives by clustering clients with similar data distributions or introducing momentum-based updates to improve convergence under domain shifts. These methods focus on improving the communication efficiency and convergence of *FL* systems when clients experience heterogeneous data distributions. Despite these advancements in handling label skew and domain shift individually, there is a need for further research on how large-scale models like *CLIP* can be better adapted and applied in *FL* settings when facing these challenges. This paper aims to explore how *CLIP* can leverage prompt learning to maintain robust performance in federated settings, ensuring better cross-client adaptation and efficient model convergence despite the inherent challenges of data distribution.

2.2 Prompt Learning

Prompt learning is a powerful and flexible approach for adapting pre-trained models to new tasks with minimal adjustments. Initially developed for NLP and widely used in LLMs like GPT-3 [Brown *et al.*, 2020], it has since extended to vision-language models. Techniques like prompt tuning [Lester *et al.*, 2021] and prefix tuning show that large models can be efficiently adapted using lightweight prompt modifications without full retraining.

In Vision Language Models like *CLIP*, prompt learning has been particularly successful in enhancing the zero-shot and few-shot learning capabilities of the model [Bang *et al.*, 2024; Li *et al.*, 2024b]. Language prompts enable the model to better align textual and visual representations, improving performance on tasks like cross-modal retrieval. KgCoOp [Yao *et al.*, 2023] enhances language prompts by reducing the discrepancy between learned and handcrafted prompts, thereby improving performance in unseen classes. Vision prompts, on the other hand, modify the image encoding process, allowing the model to adapt to visual-specific tasks without fine-tuning the entire network[Chen *et al.*, 2023]. Besides, MaPLe [khattak *et al.*, 2023a] introduces multi-modal prompt learning, combining language and vision prompt learning. These methods have shown promising results when applied to vision-language models, improving the model's performance on tasks such as zero-shot image classification and visual question answering. However, the methods above did not provide a comprehensive study of prompt learning in *FL*. This paper explores both language and vision prompt learning of VLM in *FL*, where data heterogeneity shows a strong influence on model performance.

2.3 Federated Prompt Learning

Recent research has explored integrating prompt learning into federated systems to collaboratively train shared prompts, enhancing the generalization ability for specific tasks [Lu *et al.*, 2023; Guo *et al.*, 2023b; Su *et al.*, 2024; Wei *et al.*, 2023;

Gu *et al.*, 2023]. For instance, FedPR [Feng *et al.*, 2023] leverages federated visual prompts within the null space, while [Li *et al.*, 2023] highlights the role of visual prompts in balancing the trade-off between privacy and utility while minimizing privacy budget consumption. In addition to visual prompts, several studies have investigated textual modality adaptation. PromptFL [Guo *et al.*, 2023b] focuses on learning a shared text prompt, whereas FedTPG [Qiu *et al.*, 2024] introduces a unified prompt generation framework to coordinate prompt learning across clients. Moreover, personalized prompt learning has gained attention, with recent works exploring client-specific prompt optimization strategies to address data heterogeneity [Yang *et al.*, 2023; Guo *et al.*, 2023a; Su *et al.*, 2024]. FedOTP [Li *et al.*, 2024a] proposes an efficient collaborative prompt learning approach that enables each client to capture distinct category characteristics individually. Nevertheless, these methods do not comprehensively test the performance of different prompt learning under various federated settings and prompt configurations.

3 Method Details

We briefly review the foundations of *CLIP* and introduce the details of prompt learning of *CLIP* in federated settings.

3.1 Foundations of *CLIP*

Contrastive Language-Image Pretraining (*CLIP*) revolutionized multi-modal representation learning by aligning visual and textual data through contrastive learning. Given a dataset of paired image-text examples $D = \{(I_k, T_k)\}_{k=1}^{|D|}$, the goal of *CLIP* is to encourage semantic alignment between image and text embeddings while repelling unpaired samples. The architecture of *CLIP* includes a visual encoder f_v and a text encoder f_t,which transform an image I_k and a corresponding text T_k into normalized feature embeddings $\mathbf{z}_k = f_v(I_k)$ and $\mathbf{w}_k = f_t(T_k)$, respectively.

The training objective employs a symmetric contrastive loss based on the InfoNCE [Oord *et al.*, 2018] framework. For an image embedding $\mathbf{z}_k = f_v(I_k)$, the image-to-text contrastive loss is defined as:

$$\mathcal{L}_{I\to T}(\mathbf{z}, \mathbf{w}) = -\log \frac{\exp(\mathbf{z}_k \cdot \mathbf{w}_k/\tau)}{\sum_{b=1}^{|B|} \exp(\mathbf{z}_k \cdot \mathbf{w}_b/\tau)}, \quad (1)$$

where τ is a learnable temperature parameter controlling distribution scaling, $\cdot$ denotes the dot product, and B represents the batch size.

Similarly, the text-to-image contrastive loss, anchored on a text embedding $\mathbf{z}_k$, is formulated as:

$$\mathcal{L}_{T\to I}(\mathbf{z}, \mathbf{w}) = -\log \frac{\exp(\mathbf{w}_k \cdot \mathbf{z}_k/\tau)}{\sum_{b=1}^{|B|} \exp(\mathbf{w}_k \cdot \mathbf{z}_b/\tau)}. \quad (2)$$

The overall CLIP loss combines these two components symmetrically:

$$\mathcal{L}(\mathbf{z}, \mathbf{w}) = \frac{1}{2}(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}) \quad (3)$$

3.2 Prompt learning of *CLIP* in Federated Learning

In this section, we describe how both language and vision prompt learning techniques are adapted for *CLIP* in *FL*.

Language Prompt

In Language Prompt Learning (LPT), we utilize a learnable vector $\mathbf{P}_t$, which is combined with the class token for the text input, following the approach proposed in [Zhou *et al.*, 2022]. Specifically, the text input T_k is augmented with a series of prompt vectors $\{v_1, v_2, \cdots, v_L\}$ that are learned during training, where L is the number of tokens in the prompt. The augmented text input to the *CLIP* text encoder is:

$$\tilde{T}_k = [v_1, v_2, \cdots, v_L, \texttt{CLASS}], \quad (4)$$

where `CLASS` is a fixed token that represents the class information (or task-specific information). The input $\tilde{T}_k$ is then passed through the text encoder f_t, resulting in the modified text embedding $\tilde{\mathbf{w}}_t$:

$$\tilde{\mathbf{w}}_k = f_t(\tilde{T}_k). \quad (5)$$

The $\tilde{\mathbf{w}}_k$ is normalized before being used in downstream tasks. The use of learnable text prompts allows the model to adapt to domain-specific language and task-specific variations, improving performance in tasks where text data is heterogeneous across clients.

Vision Prompt

For Vision Prompt Learning (VPT), we follow the method in [Jia *et al.*, 2022], where a series of learnable vectors $\{u_1, u_2, \cdots, u_L\}$ are inserted between the class token `CLS` and the image patch embeddings $\mathbf{E}$ for the image encoder. The vision prompt $\mathbf{P}_v$ is thus represented as a sequence of these learnable vectors, which are concatenated with the image embeddings. The input to the image encoder is:

$$\tilde{I}_k = [\texttt{CLS}, u_1, u_2 \cdots, u_L, \mathbf{E}]. \quad (6)$$

The modified image $\tilde{I}_k$ is passed through the visual encoder f_v, yielding the image embedding $\tilde{\mathbf{z}}_k$:

$$\tilde{\mathbf{z}}_k = f_v(\tilde{I}_k). \quad (7)$$

Similar to the text embeddings, the resulting image embedding $\tilde{\mathbf{z}}_k$ is normalized before being used in downstream tasks. This modification helps the model focus on relevant features and adapt to the visual characteristics specific to each client's data distribution.

Federated Prompt Learning

In Federated Prompt Learning, the clients collaboratively learn shared prompt modules while maintaining the privacy of their local data. Following typical Federated Learning setup [McMahan *et al.*, 2017; Li *et al.*, 2020c; Li *et al.*, 2021a; Huang *et al.*, 2023b] , we assume that there are M clients (indexed by M). Each client m holds its own dataset D^m and optimizes its local prompts $\mathbf{P}^m$ ($\mathbf{P}_t$ or $\mathbf{P}_v$). The goal is to aggregate the local prompts across clients into a global prompt $\mathbf{P}^g$, which is then used to update the global model. The process can be divided into three steps: distribution, optimization, and aggregation.

$$\begin{aligned} \mathbf{P}^m &\leftarrow \mathbf{P}^g && \text{Distribution,} \\ \arg\min_{\mathbf{P}^m} \mathbb{E}_{(x,y)\sim D^m}\, \mathcal{L}(\mathbf{z},\tilde{\mathbf{w}})/\mathcal{L}(\tilde{\mathbf{z}},\mathbf{w}) & && \text{Optimization,} \\ \mathbf{P}^g &= \sum_m^M \frac{N^m}{N}\mathbf{P}^m && \text{Aggregation,} \end{aligned} \tag{8}$$

where N^m is the number of samples at client m, and $N = \sum_m^M N^m$ is the total number of samples across all clients. The global prompt $\mathbf{P}^g$ is then sent back to all clients, and the process is repeated iteratively. Through this federated learning setup, clients collaboratively adapt the shared prompts to their local data distributions, while the global model maintains privacy and benefits from the collective knowledge.

4 Experiments Setup

4.1 Datasets

Following [Li *et al.*, 2021a; Huang *et al.*, 2023b], we comprehensively evaluate the federated prompt learning with *CLIP* on the following four datasets.

- **Cifar-100** [Krizhevsky and Hinton, 2009] is a famous classification dataset, containing 32×32 images of 100 classes. Training and validating sets are composed of 50,000 and 10,000 images.
- **DomainNet** [Peng *et al.*, 2019] contains 6 domains: Sketch (S), real (R), painting (P), infograph (I), quickdraw (Q), and clipart (C) with 365 classes. It consists $48k - 172k$ images per domain.
- **Office-Home** [Venkateswara *et al.*, 2017] is a large classification dataset with 65 classes and includes four different domains: Art (Ar), Clipart (C), Product (P), and Real World (RW).
- **Office31** [Saenko *et al.*, 2010] has 31 classification number in three domains: Amazon (Am), DSLR (D), and Webcam (W). It consists of common objects in office scenarios, such as laptops, keyboards, and, file cabinets.

4.2 Implementation Details

We present the experimental setup from three key aspects as follows:

- Network Structure: We employ the pre-trained CLIP model [Radford *et al.*, 2021] with a ViT-B/16 image encoder backbone, using the official implementation from the repository[1]. For prompt construction, both for the text and visual encoders, we refer to the open-source implementations of CoOP[Zhou *et al.*, 2022], VPT[Jia *et al.*, 2022], and PromptSRC [Khattak *et al.*, 2023b] available at[2,3,4]. The prompt length for both visual and textual prompts is set to $L = 16$. Visual prompts $\mathbf{P}_v$ are inserted into the first 9 layers of the transformer. Both textual and visual prompts are randomly initialized with the normal distribution.

[1]https://github.com/openai/CLIP
[2]https://github.com/KaiyangZhou/CoOp
[3]https://github.com/KMnP/vpt
[4]https://github.com/muzairkhattak/PromptSRC

Methods	Cifar100 $\beta = 0.3$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 5.0$
ZS-CLIP	64.88			
VPT	78.74 $_{\uparrow 13.86}$	79.44 $_{\uparrow 14.56}$	79.38 $_{\uparrow 14.50}$	79.76 $_{\uparrow 14.88}$
LPT	75.24 $_{\uparrow 10.36}$	75.13 $_{\uparrow 10.25}$	76.05 $_{\uparrow 11.17}$	76.53 $_{\uparrow 11.65}$

Table 1: **Accuracy of different prompts learning in Label Skew** scenarios. Please refer to Sec. 5.1.

- Client Scale Construction: For datasets Office31 [Saenko *et al.*, 2010], Office-Home [Venkateswara *et al.*, 2017], and DomainNet [Peng *et al.*, 2019], which include multiple domains, we design the number of participants to be twice the number of domains. For example, we set the participant size to $K = 6$ for Office31. For Cifar-100, we set the client scale to $K = 10$.
- Training Settings: To ensure a fair comparison, we follow [Huang *et al.*, 2023b] and set the communication epoch $T = 50$ and the local update round $E = 1$ for all settings. The training batch size is 32, and we use SGD as the local update optimizer. The corresponding weight decay is $\eta = 1e - 5$ and momentum is set to 0.9. The local client learning rate is 0.001 for all scenarios. To mimic the label skew, we follow the common experiments setting and use Dirichlet sampling [Balakrishnan *et al.*, 2019]. We fix the random seed at 0 to ensure reproducibility and run the experiments on NVIDIA 3090.

5 Results and Analysis

Based on the above settings, we conduct extensive experiments to explore different types of prompt learning with *CLIP* under federated learning settings in detail.

5.1 Behavior Discrepancy between LPT and VPT in Label Skew and Domain Shift

To fully explore Language Prompt Learning (LPT) and Vision Prompt Learning (VPT) with *CLIP* in federated learning settings, we conduct both experiments in the Label Skew and Domain Shift scenarios.

Label Shift Scenarios. Applying Dirichlet sampling, we evaluate VPT and LPT in Cifar-100 with different degrees of Label Skew and the results are shown in Tab. 1. VPT outperforms LPT obviously in all settings.

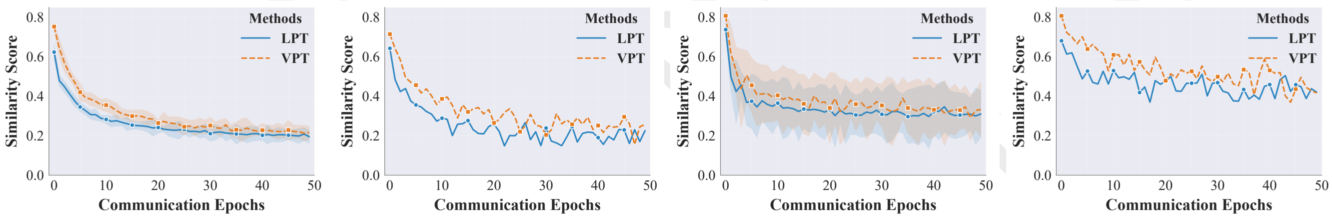
Domain Shift Scenarios. We conduct experiments on Office-Home, Office31 and DomainNet as for Domain Shift scenarios. The results in Tab. 2 indicate that LPT works better than VPT in most domains.

Since the results in the two scenarios are opposite, the key problem arises: *What makes the differences of results in Label Skew and Domain Shift?*

To address this problem, we further step in the training process in both scenarios. In concrete, we calculate the similarities of optimization directions of prompt learning between clients and the global server during training, and the visualization results are shown in Fig. 2. We can observe that in both Label Skew and Domain Shift scenarios, clients tend to optimize more similarly with the global server during the training of VPT. Besides, we also record the similarities

Methods	Office-Home					Office31				Domainnet						
	A	*C*	*PR*	*RW*	*AVG*	*AM*	*D*	*W*	*AVG*	*C*	*I*	*P*	*Q*	*R*	*S*	*AVG*
ZS-CLIP	84.30	66.17	89.18	89.66	82.32	80.96	72.45	74.05	75.82	87.69	69.66	79.77	28.11	91.91	84.82	73.66
VPT	84.13	75.23	93.84	92.14	86.34	87.26	90.00	92.53	89.93	88.93	75.34	82.57	48.83	93.45	87.44	79.43
LPT	86.20	75.87	94.57	93.70	**87.58**$_{\uparrow 5.26}$	89.40	92.25	95.32	**92.32**$_{\uparrow 16.50}$	90.30	75.84	84.05	44.59	93.24	87.58	**79.27**$_{\uparrow 5.61}$

Table 2: **Accuracy of different prompts learning in Domain Shift** scenarios. See details in Sec. 5.1



(a) Mean Value of Cifar-100 (b) Single Client of Cifar-100 (c) Mean Value of Office-Home (d) Single Client of Office-Home

Figure 2: **Similarities of Optimization Directions between clients and global server** in Cifar-100 and Office-Home in the training. Fig. 2a and Fig. 2c illustrate the mean value of the similarity of optimization direction between all clients and the global server, while the shadows indicate the standard deviation. Fig. 2b and Fig. 2d are the selected clients. Please refer to Sec. 5.1.

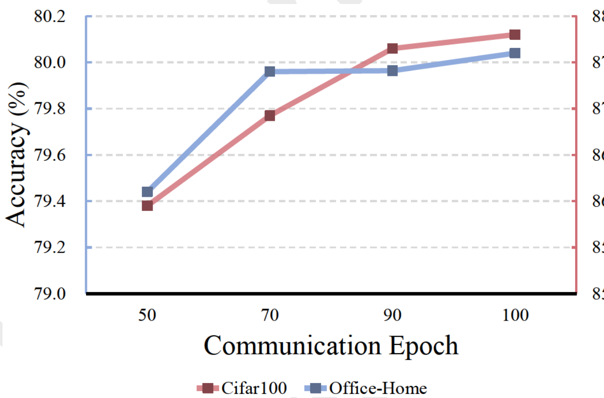


Figure 3: Effects of different **Communication Epochs**. Please refer to Sec. 5.2

among clients in Cifar-100 and Office-Home. We select the results of a single communication epoch and visualize them in Fig. 4. Likewise, the optimization directions among clients are more similar in VPT than in LPT.

Our experiments reveal that under label skew conditions, VPT exhibits significantly higher consistency in optimization directions among clients and between clients and the global server. This consistency helps align the visual features across disparate client data, effectively mapping the textual cues into a unified visual domain. In contrast, LPT shows more variability in its optimization paths, leading to inconsistent mapping and thus poorer performance when label distributions are imbalanced. Under domain shift scenarios, the optimization directions in LPT are less similar across clients, which appears to allow each client to better adapt to its own domain-specific data. This individualized adaptation enables LPT to map the textual features into diverse visual domains more effectively, yielding superior performance. Conversely, the high consistency of VPT in optimization directions tends to ignore the unique characteristics of each domain, thereby limiting its adaptability and resulting in lower performance in heterogeneous domain settings.

Regarding **Q1** in Sec. 1, we can obtain the answer: **A1**: *There exists* ***Behavior Discrepancy*** *between LPT and VPT under label skew and domain shift.* We restate this key phenomenon as follows:

> **Label Skew Prompt Consistency**:
> *For label skew, VPT exhibits consistent optimization across clients and with the global server, leading to unified visual mapping.*

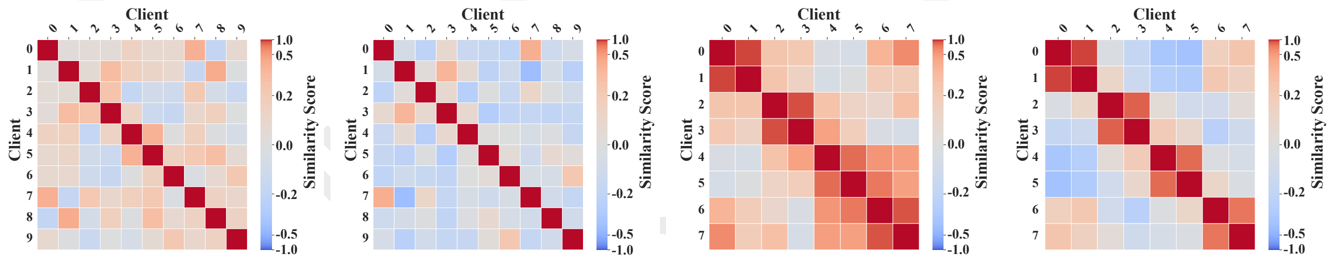
> **Domain Shift Prompt Discrepancy**:
> *For domain shift, diverse optimization in LPT enables better adaptation to distinct domains.*

5.2 Robustness of Federated Prompt Learning Under Various Settings

To comprehensively evaluate the robustness of FPL, we conduct extensive experiments under different federated learning settings and prompt configurations, analyzing their impact on model performance. The following subsections detail our findings across communication epochs, aggregation strategies, client scales, and prompt lengths.

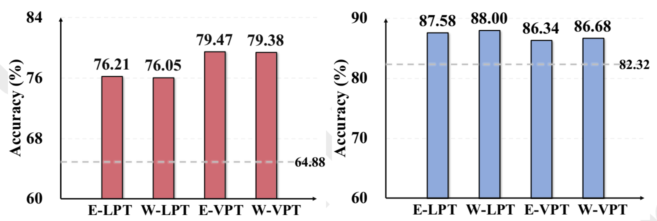
Communication Epochs. We assess the effect of varying the number of communication epochs on performance by conducting experiments on Cifar-100 and Office-Home. As shown in Fig. 3, increasing the number of communication rounds leads to only marginal improvements in accuracy, while still maintaining a performance advantage over direct zero-shot inference of *CLIP*. This suggests that Federated Prompt Learning is relatively robust to the number of training epochs, indicating that beyond a certain point, additional training contributes little to accuracy gains.

From a practical standpoint, this observation highlights an important trade-off: while increasing training rounds may slightly refine the model, it also incurs additional computa-



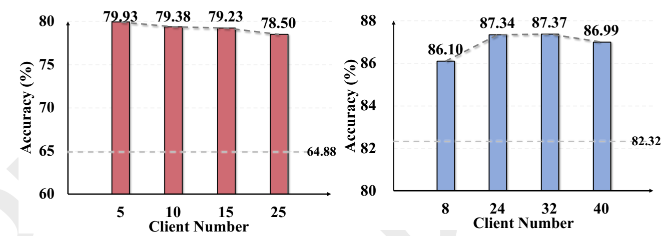
(a) VPT in Cifar-100 (b) LPT in Cifar-100 (c) VPT in Office-Home (d) LPT in Office-Home

Figure 4: **Similarities of Optimization Directions among clients** in Cifar-100 and Office-Home in the training epoch. Fig. 4a and Fig. 4b are the similarity between clients of VPT and LPT in Cifar-100 respectively, while Fig. 4c and Fig. 4d stand for Office-Home similarly. See details in Sec. 5.1.



(a) Cifar-100 (b) Office-Home

Figure 5: Effects of different **Aggregation Ways**. E is short for Equal, while W is short for Weighted. The dotted line represents the effect of *CLIP* with zero-shot inference. See details in Sec. 5.2.



(a) Cifar-100 (b) Office-Home

Figure 6: Effects of different **Scale of Clients**. The dotted line represents the effect of *CLIP* with zero-shot inference. Please refer to Sec. 5.2.

tional and communication costs. Therefore, in real-world applications, it is crucial to determine an optimal balance between training cost and performance improvement.

Aggregation Ways. To investigate the impact of aggregation strategies in FL settings, we compare two common methods: Weighted and Equal. While Weighted aggregates client models by assigning weights based on the size of each client's dataset, Equal assumes the weights of clients are equal.

Our experiments on Cifar-100 and Office-Home shown in Fig. 5 reveal an interesting pattern: In label skew scenarios, **weighted aggregation** leads to better performance, as larger clients contribute more meaningfully to the global prompt optimization. In domain shift scenarios, **equal aggregation** performs better, as it prevents overfitting to the data-rich clients and helps maintain diversity across domains. However, it is important to note that the overall difference in performance between these aggregation methods is **relatively small**. More importantly, both methods still outperform *CLIP*'s zero-shot inference, reinforcing the robustness of federated prompt learning to different aggregation strategies. ggests that while aggregation choice may slightly influence results, FPL remains stable and effective across different settings.

Scale of Clients. To evaluate the impact of client scale in federated learning settings, we conduct experiments on Cifar-100 under label skew conditions and DomainNet under domain shift conditions. The results are presented in Fig. 6.

Our findings indicate that in label skew scenarios, model accuracy gradually decreases as the number of clients increases. This is likely due to the increasing divergence in class distributions across clients, making it more challenging to learn a globally consistent prompt representation. In contrast, under domain shift scenarios, there is no clear pattern in accuracy fluctuations as the client count changes. Nevertheless, across both settings, the variation in client scale does not significantly degrade model performance, and in all cases, FPL consistently outperforms *CLIP*'s zero-shot inference. This demonstrates the robustness of FPL to different client scales, making it a reliable approach even in privacy-friendly environments with varying numbers of participants.

Prompt Lengths. We further examine the effect of prompt length as a critical parameter in prompt-based learning. We conduct experiments on Cifar-100 and Office-Home, using visual prompt tuning (VPT), and present the results in Tab. 4. Key observations include: (a) Across multiple label skew settings, increasing the prompt length (e.g., doubling its size) generally leads to a slight improvement in accuracy. (b) Even with minimal prompt tuning, FPL consistently surpasses *CLIP*'s zero-shot inference, demonstrating its robustness to variations in prompt length. While longer prompts can provide slight accuracy gains, they also increase the computational and memory overhead required for training. In real-world deployment, an optimal trade-off must be considered, balancing improved performance with resource constraints.

The above experiments collectively demonstrate that changes in communication epochs, aggregation strategies, scale of clients, and prompt length lead to minor variations in performance. In all cases, FPL outperforms CLIP's zero-

Methods	Office-Home					Office31				Domainnet						
	A	*C*	*PR*	*RW*	*AVG*	*AM*	*D*	*W*	*AVG*	*C*	*I*	*P*	*Q*	*R*	*S*	*AVG*
ZS-CLIP	84.30	66.17	89.18	89.66	82.32	80.96	72.45	74.05	75.82	87.69	69.66	79.77	28.11	91.91	84.82	73.66
Lable skew: $\beta = 0.5$																
VPT	85.08	74.89	93.21	92.25	86.62	85.66	90.82	89.62	88.70	88.86	74.59	83.41	49.70	93.25	86.74	79.43
LPT	86.45	75.99	93.94	93.17	87.39	88.15	93.47	94.43	92.02	88.89	73.75	84.89	43.74	93.64	87.00	78.65
VLPT	85.95	78.19	95.22	93.01	**88.09**$_{\uparrow 5.77}$	87.94	94.90	95.44	**92.76**$_{\uparrow 16.94}$	90.41	74.52	84.78	53.98	93.65	87.82	**80.86**$_{\uparrow 7.20}$
Lable skew: $\beta = 1.0$																
VPT	84.92	73.78	93.78	91.93	86.10	85.98	90.20	89.49	88.56	88.95	75.27	82.51	49.80	93.35	86.61	79.41
LPT	87.44	74.01	94.66	92.53	87.16	87.97	91.63	95.44	91.68	89.74	75.32	83.41	45.94	93.76	87.03	79.20
VLPT	87.07	77.34	94.99	92.53	**87.98**$_{\uparrow 5.66}$	87.19	94.29	95.06	**92.18**$_{\uparrow 16.36}$	91.40	73.94	84.70	55.12	93.46	87.10	**80.95**$_{\uparrow 7.29}$

Table 3: Performance of different prompt learning in complex scenarios containing both **Label Skew and Domain Shift**. Please refer to Sec. 5.3

Length	Cifar100			
	$\beta = 0.3$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 5.0$
ZS-CLIP	64.88			
16	78.74$_{\uparrow 13.86}$	79.44$_{\uparrow 14.56}$	79.38$_{\uparrow 14.50}$	79.76$_{\uparrow 14.88}$
32	79.95$_{\uparrow 15.07}$	79.87$_{\uparrow 14.99}$	80.49$_{\uparrow 15.61}$	80.32$_{\uparrow 15.44}$
64	80.01$_{\uparrow 15.13}$	80.10$_{\uparrow 15.22}$	80.58$_{\uparrow 15.70}$	80.92$_{\uparrow 16.04}$
128	**80.32**$_{\uparrow 15.44}$	**80.37**$_{\uparrow 15.49}$	**81.21**$_{\uparrow 16.33}$	**81.56**$_{\uparrow 16.68}$
Length	Office-Home			
	$\beta = 0.3$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 5.0$
ZS-CLIP	82.32			
16	86.11$_{\uparrow 3.79}$	86.36$_{\uparrow 4.04}$	86.10$_{\uparrow 3.78}$	86.71$_{\uparrow 4.39}$
32	86.09$_{\uparrow 3.77}$	86.38$_{\uparrow 4.06}$	**86.47**$_{\uparrow 4.15}$	86.48$_{\uparrow 4.16}$
64	86.47$_{\uparrow 4.15}$	86.52$_{\uparrow 4.20}$	86.36$_{\uparrow 4.04}$	86.54$_{\uparrow 4.22}$
128	**86.62**$_{\uparrow 4.30}$	**86.74**$_{\uparrow 4.42}$	86.22$_{\uparrow 3.90}$	**86.85**$_{\uparrow 4.53}$

Table 4: Effects of different **Length of Prompt** on VPT in Cifar-100 and Office-Home. See details in Sec. 5.2.

shot inference, reinforcing its adaptability and effectiveness in heterogeneous federated learning environments. We can answer **Q2** in Sec. 1 that: **A2**: *Federated Prompt Learning* ***remains robust*** *under different settings.*

5.3 Collaboration Effect in Complex Scenarios

In real-world environments, label skew and domain shift often coexist, creating highly complex learning conditions for federated prompt learning. To address this, we explore strategies to enhance Federated Prompt Learning (FPL) under such challenging conditions. Specifically, we investigate whether combining both vision prompt learning (VPT) and language prompt learning (LPT) can improve performance when computational resources allow. This combined approach, referred to as VLPT (Vision-Language Prompt Tuning), jointly leverages both textual and visual prompts to better align multimodal representations.

We evaluate VLPT across multiple datasets, including Office-Home, Office31, and DomainNet, where we introduce label skew on top of the existing domain shift settings. The experimental results are presented in Tab. 3 and several key findings are recapped below:

- In all tested complex scenarios, VLPT consistently achieves the best performance, outperforming both standalone VPT and LPT across datasets.
- The results in Sec. 5.1 suggest that LPT is more effective under domain shift, whereas VPT performs better in label skew scenarios. By combining both, VLPT benefits from their complementary strengths, improving the overall generalization capability.
- As seen in Tab. 3, VLPT provides noticeable improvements across diverse domains, particularly in DomainNet, where domain shifts are more severe.

Practical Implications. As for **Q3** in Sec. 1, we have the conclusion that: **A3**: *In real-world federated learning applications, where both label skew and domain shift are frequently common, VLPT provides an* ***effective solution for boosting*** *model performance.* If computational resources permit, utilizing both vision and language prompts simultaneously can significantly enhance the adaptability and overall robustness in federated learning settings.

6 Conclusion and Future Work

In this paper, we present a comprehensive empirical study on prompt learning for Vision Language Models (VLMs) in federated settings. Based on the experiments above, we can draw the key conclusions corresponding to the main research objectives as below:

- **A1:** *Under label skew and domain shift, LPT and VPT show significant* **Behavior Discrepancies**. *LPT excels in domain shift scenarios, while VPT works better in label skew since more similar updating directions are not always the "antidote" to data heterogeneity.*
- **A2:** *Federated prompt learning exhibits* **Robustness Patterns** *when federated and prompt settings vary, such as communication epochs, aggregation ways, client scales and prompt lengths.*
- **A3:** *Integrating LPT and VPT helps to adapt in both branches of language and vision, thus showing* **Collaboration Effect** *in complex scenarios containing label skew and domain shift.*

While our findings highlight the feasibility and advantages of prompt learning in federated environments, there remain several avenues for future exploration. Current limitations in training time and computational resources restrict the scalability of our experiments, and further research is needed to enhance the efficiency, personalization, and adaptability of federated prompt learning.

Contribution Statement

* Equal contribution. † Corresponding author.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2024YFC3308400), National Natural Science Foundation of China under Grant (62032016, 62361166629, 62176188, 623B2080), Zhongguancun Laboratory, and the Wuhan University Undergraduate Innovation Research Fund Project.

References

[Balakrishnan *et al.*, 2019] Narayanaswamy Balakrishnan, Samuel Kotz, and Norman L. Johnson. Continuous multivariate distributions, volume 1: Models and applications. 2019.

[Bang *et al.*, 2024] Jihwan Bang, Sumyeong Ahn, and Jae-Gil Lee. Active prompt learning in vision language models. In *CVPR*, 2024.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

[Chen *et al.*, 2023] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt learning with optimal transport for vision-language models. In *ICLR*, 2023.

[Dai *et al.*, 2023] Yutong Dai, Zeyuan Chen, Junnan Li, Shelby Heinecke, Lichao Sun, and Ran Xu. Tackling data heterogeneity in federated learning with class prototypes. In *AAAI*, 2023.

[Feng *et al.*, 2023] Chun-Mei Feng, Bangjun Li, Xinxing Xu, Yong Liu, Huazhu Fu, and Wangmeng Zuo. Learning federated visual prompt in null space for mri reconstruction. In *CVPR*, 2023.

[Gu *et al.*, 2023] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

[Guo *et al.*, 2023a] Tao Guo, Song Guo, and Junxiao Wang. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *WWW*, pages 1364–1374, 2023.

[Guo *et al.*, 2023b] Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE TMC*, 2023.

[Hu *et al.*, 2024] Ming Hu, Yue Cao, Anran Li, Zhiming Li, Chengwei Liu, Tianlin Li, Mingsong Chen, and Yang Liu. Fedmut: Generalized federated learning via stochastic mutation. In *AAAI*, 2024.

[Huang *et al.*, 2022] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*, 2022.

[Huang *et al.*, 2023a] Wenke Huang, Mang Ye, Zekun Shi, and Bo Du. Generalizable heterogeneous federated cross-correlation and instance similarity learning. *TPAMI*, 2023.

[Huang *et al.*, 2023b] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, pages 16312–16322, 2023.

[Huang *et al.*, 2024] Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. Federated learning for generalization, robustness, fairness: A survey and benchmark. *IEEE PAMI*, 2024.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022.

[Kairouz *et al.*, 2019] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[khattak *et al.*, 2023a] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023.

[Khattak *et al.*, 2023b] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, 2023.

[Konečný *et al.*, 2016] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016.

[Krizhevsky and Hinton, 2009] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[Lester *et al.*, 2021] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.

[Li *et al.*, 2020a] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process Mag*, pages 50–60, 2020.

[Li *et al.*, 2020b] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process Mag*, 2020.

[Li *et al.*, 2020c] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, 2020.

[Li *et al.*, 2021a] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, 2021.

[Li *et al.*, 2021b] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fed{bn}: Federated learning on non-{iid} features via local batch normalization. In *ICLR*, 2021.

[Li *et al.*, 2022] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *ICDE*, 2022.

[Li *et al.*, 2023] Yizhe Li, Yu-Lin Tsai, Chia-Mu Yu, Pin-Yu Chen, and Xuebin Ren. Exploring the benefits of visual prompting in differential privacy. In *ICCV*, pages 5158–5167, 2023.

[Li *et al.*, 2024a] Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *CVPR*, 2024.

[Li *et al.*, 2024b] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *CVPR*, 2024.

[Liu *et al.*, 2020] Wei Liu, Li Chen, Yunfei Chen, and Wenyi Zhang. Accelerating federated learning via momentum gradient descent. *IEEE TPDS*, 2020.

[Lu *et al.*, 2023] Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning. *IEEE DEB*, 2023.

[Luo *et al.*, 2021] Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. In *NeurIPS*, 2021.

[Ma *et al.*, 2022] Xiaodong Ma, Jia Zhu, Zhihao Lin, Shanxuan Chen, and Yangjie Qin. A state-of-the-art survey on solving non-iid data in federated learning. *Future Gener Comput Syst*, 2022.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282, 2017.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Peng *et al.*, 2019] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019.

[Qiu *et al.*, 2024] Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Federated text-driven prompt generation for vision-language models. In *ICLR*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[Sattler *et al.*, 2021] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE TNNLS*, 2021.

[Shin *et al.*, 2020] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020.

[Su *et al.*, 2024] Shangchao Su, Mingzhao Yang, Bin Li, and Xiangyang Xue. Federated adaptive prompt tuning for multi-domain collaborative learning. In *AAAI*, pages 15117–15125, 2024.

[Tan *et al.*, 2023] Yue Tan, Yixin Liu, Guodong Long, Jing Jiang, Qinghua Lu, and Chengqi Zhang. Federated learning on non-iid graphs via structural knowledge sharing. In *AAAI*, 2023.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017.

[Wang *et al.*, 2025] Zhihao Wang, He Bai, Wenke Huang, Duantengchuan Li, Jian Wang, and Bing Li. Federated recommendation with explicitly encoding item bias. In *AAAI*, 2025.

[Wei *et al.*, 2023] Guoyizhe Wei, Feng Wang, Anshul Shah, and Rama Chellappa. Dual prompt tuning for domain-aware federated learning. *arXiv preprint arXiv:2310.03103*, 2023.

[Yang *et al.*, 2023] Fu-En Yang, Chien-Yi Wang, and Yu-Chiang Frank Wang. Efficient model personalization in federated learning via client-specific prompt generation. In *ICCV*, pages 19159–19168, 2023.

[Yao *et al.*, 2023] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, 2023.

[Ye *et al.*, 2023] Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Comput. Surv.*, 2023.

[Zhao *et al.*, 2018] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022.