DiPrompT: Disentangled Prompt Tuning for Multiple Latent Domain Generalization in Federated Learning

Sikai Bai[1*], Jie Zhang[2*], Song Guo[1†],
Shuaicheng Li[3 †‡], Jingcai Guo[2], Jun Hou[3], Tao Han[1], Xiaocheng Lu[1]
[1]The Hong Kong University of Science and Technology, [2]The Hong Kong Polytechnic University,
[3]SenseTime Research

Abstract

Federated learning (FL) has emerged as a powerful paradigm for learning from decentralized data, and federated domain generalization further considers the test dataset (target domain) is absent from the decentralized training data (source domains). However, most existing FL methods assume that domain labels are provided during training, and their evaluation imposes explicit constraints on the number of domains, which must strictly match the number of clients. Because of the underutilization of numerous edge devices and additional cross-client domain annotations in the real world, such restrictions may be impractical and involve potential privacy leaks. In this paper, we propose an efficient and novel approach, called ***Di****sentangled* ***Promp****t* ***T****uning (****DiPrompT****), a method that tackles the above restrictions by learning adaptive prompts for domain generalization in a distributed manner. Specifically, we first design two types of prompts, i.e., global prompt to capture general knowledge across all clients and domain prompts to capture domain-specific knowledge. They eliminate the restriction on the one-to-one mapping between source domains and local clients. Furthermore, a dynamic query metric is introduced to automatically search the suitable domain label for each sample, which includes two-substep text-image alignments based on prompt tuning without labor-intensive annotation. Extensive experiments on multiple datasets demonstrate that our DiPrompT achieves superior domain generalization performance over state-of-the-art FL methods when domain labels are not provided, and even outperforms many centralized learning methods using domain labels.*

1. Introduction

Federated learning (FL) is an emerging privacy-preserving machine-learning technique [26], which enables multiple clients (e.g., mobile devices or organizations) to collaboratively learn a global model without exchanging their private data. However, a practical concern with conventional FL methods is that they usually ignore the possible domain shift between training data (source domains) and test data (target domain) [1], which can incur poor performance on unseen target domains due to domain discrepancies.

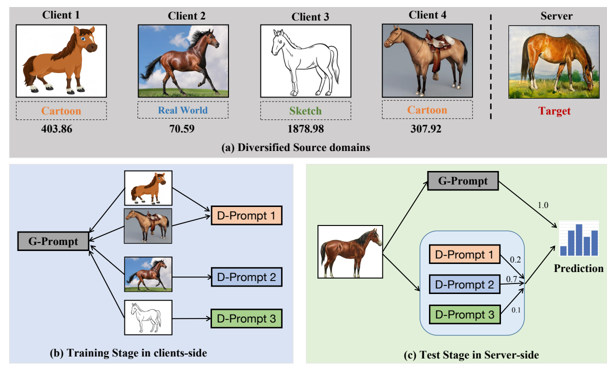


Figure 1. The motivation example and our main idea. (a) When clients outnumber source domains, learning domain-invariant features may become challenging due to imbalanced contributions across domains/clients. Note that the contribution imbalance of local data is measured through its feature distances with the target domain. (b) DiPrompT separates domain-specific features and general knowledge during local training. (c) DiPrompT adaptive ensembles for generic and valuable specific knowledge for better target domain prediction during inference.

Recently, some research efforts have attempted to incorporate domain generalization into the FL framework. For example, FedDG [21] shares the amplitude spectrum of images among local clients for medical image segmentation. FedSR [29] builds a simple algorithm that utilizes two local regularizers for domain generalization. These methods extract domain-invariant features across all source domains. Nevertheless, most of these methods only focus on domain-invariant features across clients, and they rely on the assumption of one-to-one mapping of client and domain.

∗ Equal contribution, † Corresponding authors, ‡ Project leader

When the quantity of local clients increases regardless of source domains, a more severe and real-world challenge emerges. In this scenario, data from a source domain can be dispersed among multiple clients, resulting in initially general knowledge that may be limited to certain clients and become specific. Thus learning sufficient invariant knowledge across all clients becomes infeasible. In Figure 1(a), for the horse class, the only general thing across all clients is shape. However, some specific knowledge from certain clients possess a smaller feature distance with the target domain than others (e.g., the horse features distance between client 2 and the target domain is only 70.59). Intuitively, we can additionally harness these features throughout the training phase to facilitate performance in the target domain. Unfortunately, due to interference from spurious specific information and the entanglement between generic and specific features, it is non-trivial to separately extract generic and specific knowledge for each domain with a single model. Moreover, domain information is indispensable for each local sample in these methods, but it is prohibitively expensive and may risk privacy leakage to annotate to which domain each sample belongs in a decentralized setting.

To tackle the above issues, we propose a novel and efficient method termed Disentangled Prompt Tuning (DiPrompT) for domain generalization in FL settings. The motivation is to simultaneously spin off domain-specific and general features with different lightweight components in local training, minimizing their interference and removing irrelevant specifics. Furthermore, we can gain crucial complementary information by adaptive ensembling of specific and generic knowledge for the target domain in the test stage. Specifically, DiPrompT first introduces two types of prompts: 1) **Global prompt (G-Prompt)** maintains the domain-invariant representation across all clients. It would be invariant to the domain shift brought by all clients in decentralized training. 2)**Domain prompts (D-Prompts)**: Inspired by prototype learning [36], we construct a prototypical prompt for each predefined source domain, encapsulating discriminative specific knowledge from each source domain into the respective prompts. It resolves the count consistency restriction between clients and domains by domain-wise optimization and aggregation, while enhancing cross-client representation for each domain. Furthermore, when domain labels are unknown during both training and test periods (i.e., latent domains), we design an adaptive query mechanism to explore the potential domain for each sample. An additional prompt (**Q-prompt**) is introduced, which automatically queries the domain label for each sample from all possible options by image-text alignments after excluding the interference of semantic categories. Finally, at inference time, we leverage a collaborative ensemble metric to provide valuable complementary information from G-Prompt and D-Prompts for better target prediction.

- To the best of our knowledge, this is the first lightweight work that handles federated domain generalization through prompt tuning. We aspire that our research and findings can offer a fresh perspective toward solving cutting-edge challenges in domain generalization and federated learning.
- We propose DiPrompT, a novel federated domain generalization framework based on prompt tuning. It removes two impractical restrictions and provides complementary knowledge for generalization on unseen target domains with small-scale operations.
- Extensive experiments on multiple domain generalization tasks verify the superiority of DiPrompT over state-of-the-art methods, which even outperforms some domain generalization methods with domain labels.

2. Related Work

Federated learning (FL) is a decentralized training technique that leaves training data distributed on multiple decentralized clients and learns a global model by aggregating the locally uploaded parameters in a central server [40]. In FL, clients protect data privacy as raw data are always kept locally[4]. FedAvg [26] is the first and most common FL work, which aggregates model updates by weighted averaging. One of the important challenges in FL is statistical heterogeneity among clients, in which each client contains different local data distribution with each other. A plethora of research has been done to tackle this problem, such as FedProx[17], MOON[14], and FCCL[11]. While these FL studies have dealt with distribution heterogeneity among local clients (source domains), they ignored generalization to unseen target domains, which is the problem we mainly focus on in this paper.

Domain generalization (DG) generalizes a learned model from multiple source domains to unseen target domains and motivates extensive studies in a centralized setting [45]. Representative methods either learn domain-invariant knowledge across multiple source domains [2, 18, 28, 39] or employ the idea of meta-learning [7, 13]. However, these DG methods require access to data from all source domains in a centralized server and ground-truth domain labels for each sample, which is usually impractical in FL scenarios due to privacy protection [29]. Moreover, some methods perform domain generalization without domain labels [25, 44], assuming the latent domain of images is reflected in their style or divided into multiple latent domains using clustering algorithms, then assigning pseudo domain labels to each sample. However, they cannot utilize text information about underlying domains. Recently, several works [29, 37, 43] try to solve the DG task in the FL context. In this paper, we consider a more realistic and challenging scenario, where the number of clients is more flexible regardless of source domains, and the domain labels are unknown.

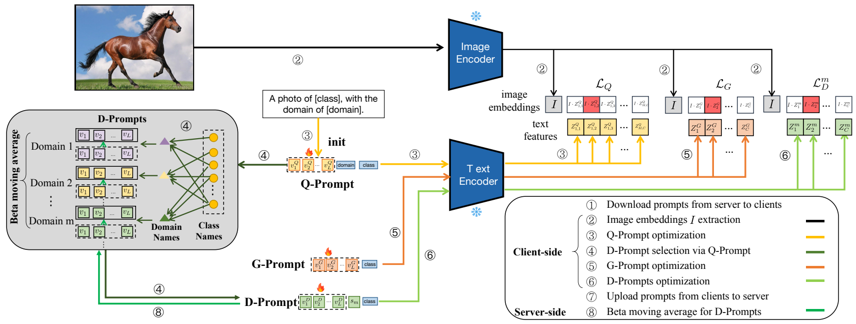


Figure 2. Illustration of **Disentangled Prompt Tuning (DiPrompT) during traing**. We devise an alternative optimization strategy to update two key modules (disentangled prompt learning and dynamic query scheme), which mainly contain six steps except communication between clients and server. We first generate image embeddings and update Q-Prompt via steps 2 and 3, respectively. Then the suitable D-Prompt is selected using Q-prompt in step 4. G-Prompt and D-Prompt in disentangled prompt learning are simultaneously optimized using steps 5 and 6. Finally, we perform the beta moving average update for D-Prompts to avoid client drift in the central server in step 8.

Prompt Tuning is a super efficient transfer learning paradigm [20, 34], whose core idea is to add little learnable embeddings at the input tokens and fast adaptation for the large language model to various downstream tasks without re-training model parameters. Early works [30, 31] aim to manually design prompt templates based on human prior knowledge. Furthermore, CoOp [47] and its extended versions [6, 41, 46] utilize a set of continuous vectors in the language branch of a large vision-language pre-trained model (CLIP), and achieve great performance improvement on multiple few-shot visual recognition tasks[3, 15, 16, 23, 24]. Compared with these methods, DiPrompT further extends prompt tuning into federated domain generalization by extracting multiple valuable knowledge based on prompt tuning and achieves superior generalization ability.

3. Methodology

In this section, we present preliminary knowledge, followed by the introduction of two key modules in training time: disentangled prompt learning and a dynamic query metric for domain queries. The overall pipeline is illustrated in Figure 2, with accompanying pseudo-code in the supplementary materials. Finally, we introduce a collaborative ensemble scheme during inference.

3.1. Preliminary and Notations

Federated Learning (FL) aims to utilize K local models to train a global model f_g with parameter θ_g through R global rounds, in which each randomly selected local model is trained with T local iterations per global round. Domain generalization (DG) trains a model using M source domains $D^S = \{D^S_m\}_{m=1}^{M}$ and is targeted to achieve decent performance on the unseen target domain D^t. Each source domain has a unique data distribution but shares the same task (e.g., image classification) with other domains. m-th domain D^S_m can be represented as $\{x_j, y_j, d_m\}_{j=1}^{M_m}$ if the domain label of each sample is known and denoted as d_m.

For conventional federated DG settings, the client/domain number is limited and there exists a strict one-to-one mapping between clients and domains (i.e., $K = M$). In contrast, we relax the above limitation into cross-device scenarios, allowing for greater flexibility in choosing the number of clients. K can significantly exceed M (i.e., $K \gg M$) and multiple clients may possess datasets originating from a common domain. Meanwhile, we consider a more practical requirement called latent domain generalization, where $D^S_m = \{x_j, y_j\}_{j=1}^{M_m}$ since domain label d_m is not given.

3.2. Disentangled Prompt Learning

Although learning a single shared prompt across all source domains/clients enables extracting invariant knowledge, it may be too casual and filter out valuable information that only appeared on a subset of clients, especially when the number of clients dramatically exceeds the volume of source domains. Thus, we present disentangled prompt learning, which includes two key subcomponents: global prompt tuning and domain prompt tuning. The former aims to achieve global optimization across all clients, and the latter is utilized to extract valuable specific knowledge from different domains. We will make a detailed introduction for them as follows.

3.2.1 Global Prompt Tuning (G-Prompt)

To capture generic characteristics shared among all clients, G-Prompt performs global optimization using a single prompt. Concretely, each client is equipped with a pre-trained vision-language model (CLIP), which includes a frozen text encoder g and visual encoder f. Given an image x along with its label y, the visual embeddings $I = f(x)$ can be extracted by the visual encoder. For text embeddings, the context prompt is a set of L learnable vectors $V^G = \left\{v_1^G, v_2^G, ..., v_L^G\right\}$, where the embedding dimension of each element is d. For the j-th class, the whole input for the text encoder is $t_j = \{V_g, c_j\}$, where c_j indicates the word embedding corresponding to j-th class name with the same embedding dimension d. Thus, we can obtain corresponding text embeddings as $Z_j^G = g(t_j)$.

Furthermore, the local model computes the prediction probability $P(y = j|x)$ using the extracted image-text embeddings pair, which maximizes the cosine similarity score $\text{sim}(\cdot, \cdot)$ for the correct pairs while minimizing those incorrect pairs. It can be formularized as:

$$P_g(y = j|x) = \frac{\exp(\text{sim}(I, Z_j)/\tau)}{\sum_{i=1}^{C} \exp(\text{sim}(I, Z_i)/\tau)}, \tag{1}$$

where C is the number of categories. We optimize the global prompt with the cross-entropy loss between the prediction probability and its label during local iterations as:

$$\mathcal{L}_G(x, y) = \mathbb{E}_{x,j}\mathcal{L}_{ce}(y, P_g(y = j|x)). \tag{2}$$

3.2.2 Domain Prompts Tuning (D-Prompts)

While G-Prompt acquires generic knowledge across all clients, it may overlook other valuable information from diverse clients. This is particularly evident when the number of clients significantly surpasses the domains, and diverse clients may hold data originating from a shared domain. To this end, we devise domain-specific prototypical prompts, to separately extract the distinct knowledge from the corresponding domains. Specifically, we build a pool of domain-wise prompt pool $V^D = \left\{V_1^D, ..., V_M^D\right\}$, where $V_m^D = \left\{v_1^D, v_2^D, ..., v_L^D, s_m\right\}$ is the prompt for m-th source domains and s_m denotes the text embeddings corresponding to m-th source domain name. For j-th class from m source domain, the text encoder produce embeddings $Z_j^m = g(t_j^m)$, where $t_j^m = \left\{V_m^D, c_j\right\}$.

To train D-Prompts in local clients, we first optimize each element using cross-entropy loss between the prediction probability of image-text pairs and the ground-truth label. Moreover, to prevent them from progressively converging towards the same point, a contrastive loss is employed, where positive pairs involve hand-crafted prompts with the same domain names and negative pairs incorporate D-Prompts from other domains. Formally, the optimization of m-th domain prompt can be expressed as follows:

$$\begin{aligned} \mathcal{L}_D^m(x, y) &= \mathcal{L}_{ce}^m + \mathcal{L}_{cont}^m \\ &= \mathcal{L}_{ce}^m - \log \frac{\exp(\text{sim}(V_m^D \cdot \tilde{V}_m)))}{\sum_{i=1}^{M} \exp(\text{sim}(V_m^D \cdot V_i^D))}, \end{aligned} \tag{3}$$

where $\mathcal{L}_{ce}^m = \mathbb{E}_{x,j}\mathcal{L}_{ce}(y, P_m(y = j|x))$. and $P_m(y = j|x) = \frac{\exp(\text{sim}(I, Z_j^m))/\tau)}{\sum_{i=1}^{C} \exp(\text{sim}(I, Z_i^m)/\tau)}$ denotes the prediction probability with m-th domain prompt for j-th class. Besides, $\tilde{V}_m$ indicates the hand-crafted prompt for m-th domain, i.e., "a photo of a [class] with the domain of <domain >", where "<domain >" is the text of m-th domain and "[class]" denotes all potential category options.

Furthermore, since multiple clients may hold data originating from a shared domain. When $K > M$, we devise a domain-wise aggregation strategy to aggregate knowledge from the same domains. It can be represented as a weighted combination of those prompts from the same domain:

$$V_m^{D,r+1} = V_m^{D,r} + \frac{\sum_{i=1}^{K}(|\mathcal{D}_i| * \mathcal{I}_{m,i}) \cdot \triangle V_{m,i}^{D,r+1}}{\sum_{i=1}^{K}(|\mathcal{D}_i| * \mathcal{I}_{m,i})}, \tag{4}$$

where $\triangle V_{m,i}^{D,r+1} = V_{m,i}^{D,r+1} - V_m^{D,r}$. For m-th domain, $V_{m,i}^{D,r+1}$ is uploaded prompt from i-th client in $r + 1$-th global round, and $V_m^{D,r}$ denotes updated prompt after aggregation in r-th global round. $\mathcal{I}_{m,i}$ is the output (0 or 1) of indicator function $\mathcal{I}(|\triangle V_{m,i}^{D,r+1}|)$ and $\mathcal{D}_i$ indicate the local data in i-th client. The operation only aggregates those updated prompts and filters those unchanged ones to ensure valid learning.

Despite diversified knowledge extracted from different domains, separately optimizing the domain prompt for each domain risks client drift. Inspired by the pre-trained vision-language update in CLIPood [35], We employ a beta momentum averaging mechanism to update domain prompts. Unlike exponential moving averages that underweight initial parameters, beta momentum averaging can avoid forgetting domain information appending at the initial period. Specifically, when m-th domain prompt is updated from $V_m^{D,r}$ to $V_m^{D,r+1}$ between two adjacent rounds r and $r + 1$, a momentum average prompt $\hat{V}_m^{D,r+1}$ can be computed by:

$$\hat{V}_m^{D,r+1} = \frac{\sum_{i=0}^{r} \alpha_i}{\sum_{i=0}^{r+1} \alpha_i}\hat{V}_m^{D,r} + \frac{\alpha_r}{\sum_{i=0}^{r}}V_m^{D,r+1}, \tag{5}$$

where $\alpha_i = \mathbf{B}(\beta, \beta)(\frac{i+0.5}{R+1})$ and β is a hyper-parameter for beta distribution $\mathbf{B}(\cdot, \cdot)$. We set $\beta = 0.2$ to preserve current optimization and initial domain knowledge attached at the beginning of training. R is the number of global rounds.

After T local iterations, only G-Prompt and D-Prompts are uploaded to the central server. Like FedAvg, the server updates the global prompt using a weighted aggregation of prompts from received clients in the current round. The process continues for next round by sending updated G-Prompt and momentum-averaged D-Prompts to newly selected clients.

3.3. Dynamic Query Scheme (Q-Prompt)

To efficiently learn prompts with unknown domain labels, we design a dynamic query scheme based on prompt tuning, which automatically selects appropriate domain prompts for different source inputs. Considering data is naturally separated according to semantic categories, most prior methods cannot avoid interference from the semantic category labels [45]. Our query scheme adopts a two-substep strategy as shown in Figure 2 middle left (gray area), where each input image is first classified into a category and then grouped to an underlying domain. Practically, we construct a learnable prompt called Q-Prompt $V^Q = \left\{v_1^Q, v_2^Q, ..., v_L^Q\right\}$ that is shared among all classes and domains. For j-th class and m-th domain, the output text embedding is $Z_{j,m}^Q = g(t_{j,m})$, where $t_{j,m} = \{V^Q, c_j, s_m\}$. We perform class and domain similarity matching on the input text-image pairs by:

$$P(y=j, d=m|x) = \frac{\exp(\text{sim}(I, Z_{j,m}^Q)/\tau)}{\sum_{p=1}^{C}\sum_{q=1}^{M}\exp(\text{sim}(I, Z_{p,q}^Q)/\tau)}, \quad (6)$$

where we select the domain with the highest probability under the ground-truth class during training and across all categories when testing on the target domain.

To ensure the effectiveness of the Q-prompt during the early period training, we initialize it with word embeddings derived from the hand-crafted prompt, i.e., "a photo of a [class] with the domain of [domain]". Furthermore, we utilize two self-consistent regularization terms to perform Q-prompt tuning for better adaptation to downstream tasks. Specifically, MSE loss regulates feature-level alignments between the current prompt and beta momentum averaging counterpart, while we impose logits-level constraints by minimizing the Kullback-Leibler divergence between input images with current and momentum prompts, respectively. Formally, the optimization can be formalized as:

$$\begin{aligned}\mathcal{L}_Q = \mathcal{L}_{mse} + \mathcal{L}_{KL} = \mathbb{E}_{x,y}\sum_{m=1}^{M}[(Z_{y,m}^Q) - (\hat{Z}_{y,m}^Q)]^2 \\ + \mathcal{D}_{KL}(\text{sim}(I, Z_{y,m}^Q), \text{sim}(I, \hat{Z}_{y,m}^Q))],\end{aligned} \quad (7)$$

where $Z_{y,m}^Q = g(\{V^Q, c_y, s_m\})$ and $\hat{Z}_{y,m}^Q = g(\left\{\hat{V}^Q, c_y, s_m\right\})$. $\hat{V}^Q$ is the beta momentum average for Q-Prompt V^Q. Given input data (x, y), the overall optimization for disentangled prompt tuning is $\mathcal{L}(x,y) = \mathcal{L}_G(x,y) + \lambda\mathcal{L}_D^m(x,y)$, where λ is a weighted coefficient to balance G-Prompt and D-Prompts, and m is the predicted domain via Q-Prompt.

3.4. Collaborative Ensemble Process

During training, we can learn general and specific knowledge from different source domains via G-Prompt and D-Prompts. In inference time, it is essential to extend this valuable knowledge into the target domain. One straightforward way is to average the information from all optimized prompts. However, this method ignores the relationship (feature distance) differences between target samples and different source domains. To this end, we build a dynamic ensemble metric in our DiPrompT, which considers the above vary while effectively exploiting valuable knowledge in global prompt and various domain prompts for better target inference. Concretely, we construct ensembled text features $Z = \{Z_1, ..., Z_C\}$ for each target domain sample, which is a dynamic weighted combination of knowledge from different prompts. The ensembled text embeddings Z_j for j-th class is defined as:

$$Z_j = \sum_{m=1}^{M} w_m Z_j^m + w_g Z_j^G, \quad (8)$$

where we set $w_g = 1$. $w_m = \frac{\max_j \text{sim}(I \cdot Z_j^m)}{\sum_{i=1}^{M}\max_j I \cdot Z_j^i}$ represents the highest prediction for m-th domain aross all categories.

4. Experiments

In this section, we conduct extensive experiments in three benchmark datasets with domain distribution shifts to demonstrate the effectiveness and robustness of our method. More details and additional experiments can be found in the supplementary material.

4.1. Experimental Setup

4.1.1 Datasets.

We perform comprehensive experiments on three widely used datasets in domain generalization tasks, including PACS (4 domains: photo, art-painting, cartoon, and sketch) [12], Officehome (4 domains: Art, Clipart, Product, and Real World) [38], and VLCS (4 domains: Pascal, Labelme, Caltech and Sun) [8]. To conduct our analysis on each dataset, we adopt the "leave-one-domain-out" strategy, wherein we select one domain as the target domain and utilize the remaining domains as source domains. We provide detailed dataset descriptions in supplementary materials.

4.1.2 Baselines.

We compare our DiPrompT with the following state-of-the-art methods. SWAD [5], I2ADR[27], PCL [42], Fishr [33], ITL-Net [9] and VAUE [19] are sota centralized learning methods that learned a generalized model by using all sources domains in a data pool regardless of data privacy. Meanwhile, we consider recent FL algorithms as our main competitors, including FedAvg[26], FedProx [17], FedADG[43], FedSR[29], FedGMA[37], PromptFL [10] and FedCLIP[22].

Methods		PACS					Officehome				
		Art	Cartoon	Photo	Sketch	Avg	Art	Clipart	Product	Real	Avg
Centralized Learning	SWAD	89.30	83.40	98.10	82.60	88.79	66.10	59.90	78.70	80.70	70.60
	I2ADR	82.90	80.80	95.00	83.50	85.60	70.30	55.10	80.70	79.20	71.40
	PCL	90.20	83.90	98.10	82.60	88.70	67.30	**59.90**	78.70	80.70	71.60
FL	FedAvg	80.41	77.55	92.33	63.31	78.39	61.84	51.15	76.41	77.39	66.69
	FedProx	79.19	78.96	94.92	64.28	79.33	62.34	52.00	76.62	78.56	67.37
FL+DG	FedADG	78.02	79.24	88.50	64.25	77.50	62.75	51.43	74.07	77.98	66.55
	FedSR	82.00	82.95	93.53	66.29	81.19	62.75	49.85	72.24	74.10	64.73
	FedGMA	83.88	80.04	95.78	69.17	82.21	65.34	52.11	75.98	79.59	68.25
FL+ Adapter/Prompt	PromptFL	92.77	94.24	99.40	82.13	92.13	72.18	56.88	82.13	84.78	73.99
	FedCLIP	92.93	94.80	99.52	82.26	92.37	71.03	56.13	83.76	84.14	73.76
	Ours	**94.97**	**96.25**	**99.56**	**84.72**	**93.88**	**74.21**	58.90	**85.51**	**86.12**	**76.18**

Table 1. Performance comparison of our proposed DiPrompT with state-of-the-art methods on PACS and Officehome datasets. FedDure outperforms all other methods

Methods	VLCS				
	Caltech	Labelme	Pascal	Sun	Avg
Fishr	98.90	64.0	71.50	76.80	77.80
ITL-NET	98.30	65.40	75.10	76.80	78.90
VAUE	99.00	64.70	75.10	79.40	79.40
FedAvg	93.54	60.69	72.22	74.66	75.27
FedProx	94.55	61.55	73.75	75.52	76.34
FedADG	95.96	61.43	66.30	72.44	74.03
FedSR	96.37	60.15	69.74	73.40	74.91
FedGMA	97.88	61.47	73.76	76.19	77.32
PromptFL	99.49	65.36	78.54	78.75	80.53
FedCLIP	99.39	66.70	82.24	78.62	81.73
Ours	**99.70**	**69.23**	**84.16**	**81.72**	**83.70**

Table 2. Performance comparison of our proposed DiPrompT with state-of-the-art methods on VLCS dataset.

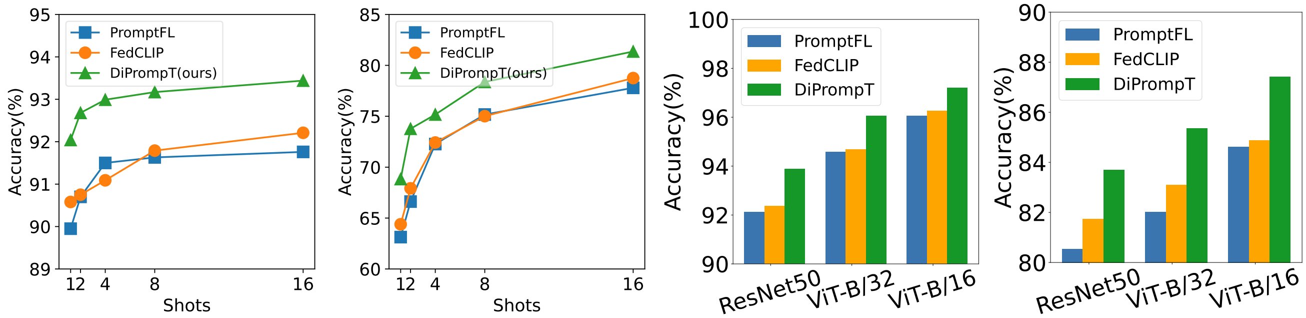
4.1.3 Implementation Details.

In our experiment, we conducted model training using PyTorch on GeForce RTX 3090 GPU. The training process for all models involved the utilization of the Adam optimizer with a learning rate set at 5e-4. It's worth noting that this uniform configuration applied to all methods under consideration, with the exception of FedGMA. For FedGMA, we tailored its optimizer parameters to align with its best-reported hyperparameters in [37], ensuring a comprehensive and consistent approach across our experimental design. Furthermore, our default settings also include the weighted coefficient $\lambda = 1.0$, batch size $b = 16$, local iterations $T = 1$, the selected clients in each round $H = 5$, the number of clients $K = 20$, and the number of global rounds $R = 100$. Following CLIP [32], we adopt ResNet50 as the backbone architecture of an image encoder and use a masked self-attention Transformer as a text encoder. Note that if there is no specified description, we use the same hyper-parameters and backbone architecture for DiPrompT and other methods in all experiments to implement fair comparison.

4.2. Performance Comparison

We report the experimental results of DiPrompT and other state-of-the-art methods in Table 1 and Table 2, which involves 12 domain generalization tasks on the PACS, OfficeHome, and VLCS datasets. All results are the average of three runs, and bold text represents the best results. Specifically, as mentioned earlier, we considered a more flexible scenario when $K > M$, thus there is no one-to-one mapping between local clients and domains. It can be observed that the previous federated domain generalization (FL+DG) methods perform poorly, even worse than the FL methods designed for client heterogeneity. FedGMA's performance is comparable to FedAvg and FedProx, while FedADG and FedSR exhibit lower performance in many DG tasks compared to FedAvg and FedProx. The phenomenon might stem from FL+DG methods struggling with valuable knowledge as client volume significantly surpasses the number of source domains.

Furthermore, we compare our DiPrompT with state-of-the-art methods from centralized learning, FL, and FL+DG. It is noted that due to the dependencies on domain labels and interrelationships between source domains, centralized learning methods cannot be applied to this domain-separated setting. They are not direct competitors to our model due to the different setups (FL vs. centralized learning), and we adopt the results taken from their original papers. With prompt tuning from the textual branch, DiPrompT outperforms centralized methods as well as FL and FL+DG techniques by a big margin, where we adopt the same image backbone (i.e. ResNet50) even when domain labels are unknown for DiPrompT. Moreover, due to



(a) PACS (b) VLCS (c) PACS (d) VLCS

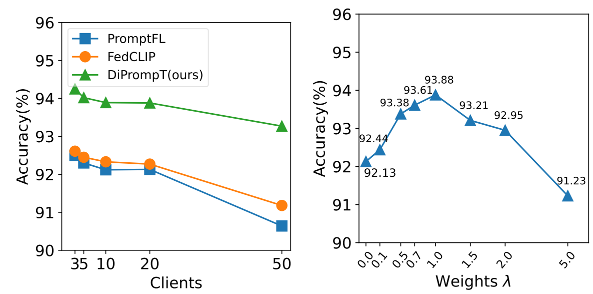
Figure 3. Analysis in terms of few shots settings and different backbone architectures on PACS and VLCS datasets.

complementary knowledge from G-Prompt and D-Prompts, our DiPrompT shows significant superiority over PromptFL and FedCLIP, which utilize the same pre-trained vision-language model (CLIP) as our methods. Overall, DiPrompT achieves the best average accuracy across four benchmark datasets. When examining the results for each target domain setting within each dataset, DiPrompT outperforms previous methods in 11 out of 12 settings. These quantitative results demonstrate the effectiveness of our approach.

4.3. Ablation Study

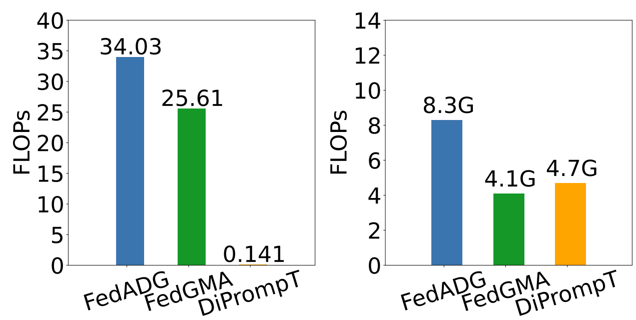
4.3.1 Effectiveness of Components.

To measure the importance of each component in our DiPrompT, we conduct an ablation analysis for the following variants using the PACS dataset in Table 3. Row 1 denotes that DiPrompT removes G-Prompts in each local model. Row 2 indicates the variant that eliminates D-Prompts in our DiPrompT, while only updating G-Prompt, which is essentially equivalent to PromptFL. Compared with DiPrompT, the dramatic performance drop shows that the two components are both effective and can provide value complementary knowledge by ensembling them. In Row 3, we remove the built contrastive loss during D-Prompts optimization and find a slight drop. Instead of learnable Q-Prompt, we use static prompts in Row 4 (i.e., “a photo of a [CLASS] with the domain of [Domain].”) to look up domain prompts for each sample. As results show, learnable Q-Prompt plays an important role in exploring domain information in latent domain learning. In Row 5, we only choose the domain prompt with the highest weight rather than ensembling knowledge from different source domains during inference. The performance decrease suggests that the collaborative ensemble metric enables the sufficient exploitation of more valuable knowledge. Moreover, the results in Row 6 and 7 implicitly showcase the benefit of L_{KL} and L_{mse} for Q-Prompt optimization. When the domain labels are given for all local training data, we can observe that our DiPrompT is close to its upper bound. These



(a) The number of Client K (b) Weight coefficient λ

Figure 4. Hyperparameters analysis in terms of the number of clients K and weight coefficient λ on PACS.



(a) Communication cost (b) Computation cost

Figure 5. Comparison of computation and communication cost of DiPrompT and other federated domain generalization methods.

evaluations verify the effectiveness of each component, and DiPrompT can obtain complementary knowledge and eliminate the dependency on the domain label for better generalization on the unseen target domain.

4.3.2 Results on Few-shot Recognition.

To explore the effectiveness of our framework with extremely limited data, we compare the generalization performance of DiPrompT with some state-of-the-art methods across various few-shot settings (i.e., 1, 2, 4, 8, 16 shots). As depicted in Figure 3a and shots 3b, we observe that as the number of training samples per class on each client increases, the performance of all methods is enhanced. Meanwhile, DiPrompT achieves significant performance improvements across all shot settings on PACS and VLCS datasets, and shows higher gains in minimal data cases (e.g. 1, 2 shots). Thereby, these results demonstrate the robustness of DiPrompT in acquiring complementary knowledge in data-scarce scenarios.

4.3.3 Impacts of Backbone Architecture.

To investigate how the choice of backbone architecture impacts generalization performance, we summarize the generalization effect of multiple methods based on different image encoders, including Reset50, ViT-B/32, and ViT-B/16. As shown in Figure 3c and 3d, the more advanced the backbone architecture is, the better the performance of all methods can achieve. More importantly, our DiPrompT is significantly better than other state-of-the-art methods across all backbone on PACS and VLCS datasets, which show the robustness of our DiPrompT in terms of backbones.

4.3.4 Effect of Hyperparameters.

There are two key hyperparameters, including the number of clients K and the value of the weight λ between G-Prompt and D-Prompts. We first investigate the performance impact of the number of total clients, which varies in $\{3, 5, 10, 20, 50\}$. Note that we assign a source domain to a single client when $K = 3$ and randomly select 5 clients per round when $K >= 5$. As illustrated in Figure 4a, although there is inevitable performance degradation for all methods by increasing total clients, our DiPrompT only has a light performance drop (about 1%) and outperforms other methods across all numerical settings. This phenomenon demonstrates the challenges posed by a large number of clients in federated domain generalization as well as the effectiveness of DiPrompT. Finally, we examine the impacts of the different values of the weight hyperparameter λ between G-Prompt and D-Prompts on performance. As shown in Figure 4b, we observe that the optimal performance is achieved when setting $\lambda = 1$.

4.3.5 Cost Analysis.

Finally, we analyze the efficiency of DiPrompT in terms of the computation and communication cost in Figure 5. We measure the communication cost by the size of uploaded data per round. It can be observed that DiPrompT can save at most 830 times communication cost per round compared to FedADG and 625 times for FedGM, indicating that DiPrompT can significantly reduce communication burden. For the computation cost, we report the comparison of GPU time as in the same mini-batch, where DiPrompT outperforms FedADG around 2 times and is comparable computation cost with FedGMA. It is noted that we don't compare with other FL and FL+DG methods (i.e FedAvg, FedProx, and FedSR) since they have almost the same communication and computation costs as FedGMA, while PrompFL and FedCLIP have similar costs with our DiPrompT. These results demonstrate the efficiency of our DiPrompT, which can be applied to many real-world scenarios.

Ablated components	PACS				
	Art	Cartoon	Photo	Sketch	Avg
w/o G-Prompt	88.87	91.42	98.22	75.47	88.50
w/o D-Prompts	92.77	94.24	99.40	82.13	92.13
w/o L_{cont}	94.57	95.61	99.52	84.27	93.49
w/o Q-Prompt	93.35	93.69	99.44	83.16	92.40
w/o ensemble	94.19	94.88	99.52	83.4	93.00
w/o L_{KL}	94.04	95.56	99.40	81.19	92.54
w/o L_{mse}	93.75	95.01	99.34	80.80	92.23
w domain	95.82	96.81	99.64	84.80	94.26
Ours	94.47	96.25	99.56	84.72	93.88

Table 3. Quantitative analysis of components of DiPrompT.

5. Conclusion

In this work, we propose a novel framework named DiPrompT, the first attempt to introduce prompt tuning to federated domain generalization. Specifically, DiPrompT aims to learn general knowledge as well as valuable specific information across all clients, especially when the number of clients and source domains is inconsistent during training and different clients may store data originating from a shared domain. It provides more complementary knowledge for unseen target prediction during inference. Moreover, we build an adaptive query mechanism based on prompt tuning, which automatically searches the suitable domain for each sample when domain labels are not given for all local data. Extensive experiments show that our DiPrompT outperforms state-of-the-art methods, and is even better than many centralized learning strategies using domain labels.

Acknowledgement

This research was supported by fundings from the Key-Area Research and Development Program of Guangdong Province (No. 2021B0101400003), Hong Kong RGC Research Impact Fund (No. R5060-19, No. R5034-18), Areas of Excellence Scheme (AoE/E-601/22-R), General Research Fund (No. 152203/20E, 152244/21E, 152169/22E, 152228/23E), Shenzhen Science and Technology Innovation Commission (JCYJ20200109142008673).

References

[1] Ruqi Bai, Saurabh Bagchi, and David I Inouye. Benchmarking algorithms for federated domain generalization. *arXiv preprint arXiv:2307.04942*, 2023. 1

[2] Sikai Bai, Junyu Gao, Qi Wang, and Xuelong Li. Multi-domain synchronous refinement network for unsupervised cross-domain person re-identification. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 2

[3] Sikai Bai, Qi Wang, and Xuelong Li. Mfi: Multi-range feature interchange for video action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6664–6671. IEEE, 2021. 3

[4] Sikai Bai, Shuaicheng Li, Weiming Zhuang, Kunlin Yang, Jun Hou, Shuai Yi, Shuai Zhang, Junyu Gao, Jie Zhang, and Song Guo. Combating data imbalances in federated semi-supervised learning with dual regulators. 2024. 2

[5] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *European Conference on Computer Vision*, pages 440–457. Springer, 2022. 5

[6] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. 2023. 3

[7] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 200–216, 2020. 2

[8] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013. 5

[9] Boyan Gao, Henry Gouk, Yongxin Yang, and Timothy Hospedales. Loss function learning for domain generalization by implicit gradient. In *International Conference on Machine Learning*, pages 7002–7016, 2022. 5

[10] Tao Guo, Song Guo, Junxiao Wang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. *arXiv preprint arXiv:2208.11625*, 2022. 5

[11] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10143–10153, 2022. 2

[12] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 5

[13] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[14] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021. 2

[15] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13668–13677, 2021. 3

[16] Shuaicheng Li, Feng Zhang, Kunlin Yang, Lingbo Liu, Shinan Liu, Jun Hou, and Shuai Yi. Probing visual-audio representation for video highlight detection via hard-pairs guided contrastive learning. *arXiv preprint arXiv:2206.10157*, 2022. 3

[17] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 2, 5

[18] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[19] Jianxin Lin, Yongqiang Tang, Junping Wang, and Wensheng Zhang. Mitigating both covariate and conditional shift for domain generalization. In *2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pages 437–443, 2022. 5

[20] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 3

[21] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. 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*, pages 1013–1023, 2021. 1

[22] Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning. *arXiv preprint arXiv:2302.13485*, 2023. 5

[23] Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23560–23569, 2023. 3

[24] Xiaocheng Lu, Ziming Liu, Song Guo, Jingcai Guo, Fushuo Huo, Sikai Bai, and Tao Han. Drpt: Disentangled and recurrent prompt tuning for compositional zero-shot learning. *arXiv preprint arXiv:2305.01239*, 2023. 3

[25] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11749–11756, 2020. 2

[26] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 5

[27] Rang Meng, Xianfeng Li, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Mingli Song, Di Xie, and Shiliang Pu. Attention diversification for domain generalization. In *European conference on computer vision*, pages 322–340. Springer, 2022. 5

[28] A Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. *Advances in Neural Information Processing Systems*, 34:5264–5275, 2021. 2

[29] A Tuan Nguyen, Philip Torr, and Ser Nam Lim. Fedsr: A simple and effective domain generalization method for federated learning. *Advances in Neural Information Processing Systems*, 35:38831–38843, 2022. 1, 2, 5

[30] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019. 3

[31] Nina Poerner, Ulli Waltinger, and Hinrich Schütze. E-bert: Efficient-yet-effective entity embeddings for bert. *arXiv preprint arXiv:1911.03681*, 2019. 3

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[33] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377, 2022. 5

[34] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. pages 4222–4235, 2020. 3

[35] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. *arXiv preprint arXiv:2302.00864*, 2023. 4

[36] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017. 2

[37] Irene Tenison, Sai Aravind Sreeramadas, Vaikkunth Mugunthan, Edouard Oyallon, Eugene Belilovsky, and Irina Rish. Gradient masked averaging for federated learning. *arXiv preprint arXiv:2201.11986*, 2022. 2, 5, 6

[38] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 5

[39] Qi Wang, Sikai Bai, Junyu Gao, Yuan Yuan, and Xuelong Li. Unsupervised domain adaptive learning via synthetic data for person re-identification. *arXiv preprint arXiv:2109.05542*, 2021. 2

[40] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. 2

[41] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 3

[42] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7097–7107, 2022. 5

[43] Liling Zhang, Xinyu Lei, Yichun Shi, Hongyu Huang, and Chao Chen. Federated learning with domain generalization. *arXiv preprint arXiv:2111.10487*, 2021. 2, 5

[44] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 2

[45] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 5

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3