Soft Prompt Generation for Domain Generalization

Shuanghao Bai[1⋆]   Yuedi Zhang[1⋆]   Wanqi Zhou[1]
Zhirong Luan[2]   Badong Chen[1†]

[1]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
[2]School of Electrical Engineering, Xi'an University of Technology, Xi'an, China
{baishuanghao, zyd993, zwq785915792}@stu.xjtu.edu.cn,
luanzhirong@xaut.edu.cn, chenbd@mail.xjtu.edu.cn

Abstract. Large pre-trained vision language models (VLMs) have shown impressive zero-shot ability on downstream tasks with manually designed prompt. To further adapt VLMs to downstream tasks, soft prompt is proposed to replace manually designed prompt, which undergoes fine-tuning based on specific domain data. Prior prompt learning methods primarily learn a fixed prompt or residuled prompt from training samples. However, the learned prompts lack diversity and ignore information about unseen domains. In this paper, we reframe the prompt learning framework from a generative perspective and propose a simple yet efficient method for the Domain Generalization (DG) task, namely **S**oft **P**rompt **G**eneration (SPG). Specifically, SPG consists of a two-stage training phase and an inference phase. During the training phase, we introduce soft prompt label for each domain, aiming to incorporate the generative model domain knowledge. During the inference phase, the generator of the generative model is employed to obtain instance-specific soft prompts for the unseen target domain. Extensive experiments on five domain generalization benchmarks of three DG tasks demonstrate that SPG achieves state-of-the-art performance. The code is available at https://github.com/renytek13/Soft-Prompt-Generation-with-CGAN.

Keywords: Domain Generalization · Visual Language Models · Prompt Learning · Generative Models

1 Introduction

Large vision language models, such as CLIP [32] and ALIGN [16], have attracted significant attention owing to their effective adaptation to downstream tasks with manually designed prompts. However, manually designed prompts are not always optimal for domain-specific tasks. Instead of manually designed prompts, the soft prompt [19, 31, 40] can be optimized in a data-driven manner through back-propagation. The soft prompt, serving as a learning vector, is refined through

---

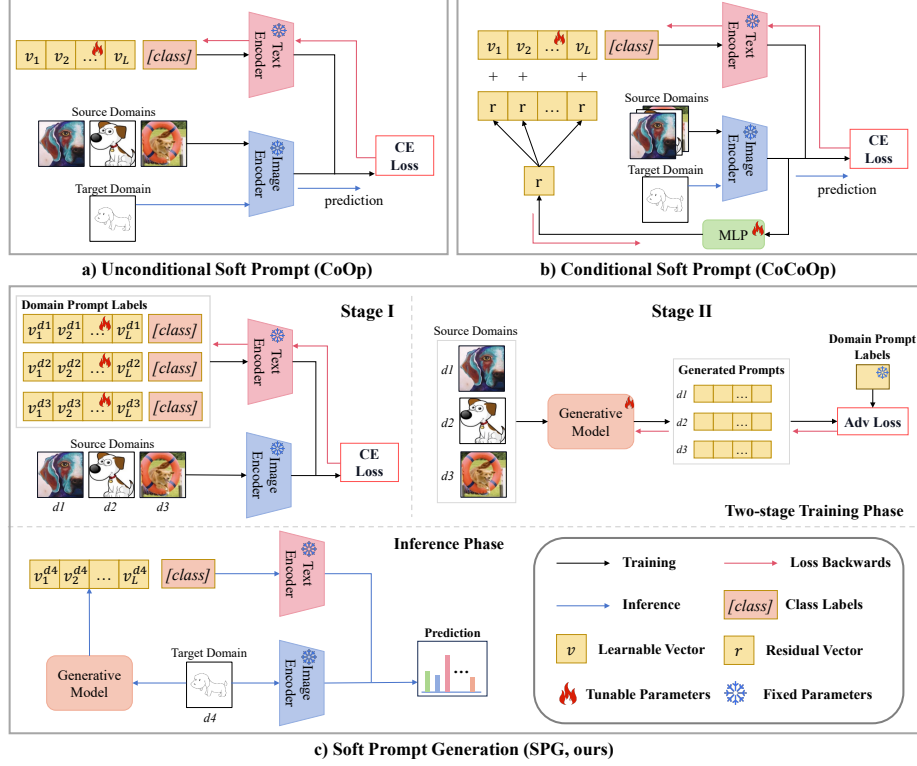⋆ Equal contribution.
† Corresponding author.

Fig. 1: The difference between previous work and our work. We reframe the prompt learning framework from a generative perspective. We exclusively rely on a generative model to directly produce soft prompts, ensuring their diversity. Essentially, we transfer the prompt's adaptability to the generative model by incorporating domain knowledge.

fine-tuning with domain-specific data to better adapt to downstream tasks, including base-to-novel generalization [18, 40] and domain adaptation [2, 13, 35]. Despite the progress made in prompt learning, generalization performance tends to decline significantly when facing distribution shifts.

Many efforts have been made to address the distribution shift problem by domain generalization (DG) [25, 38]. The objective of DG is to train a model using data from one or multiple related but distinct source domains in a manner that enables the model to generalize effectively to any out-of-distribution (OOD) unseen target domain. Due to the extensive learning in various domains, vision language models like CLIP exhibit a generalization performance that can even rival state-of-the-art traditional supervised learning algorithms. Taking a step, many works adopt prompt learning techniques [6, 7, 29, 36, 39] to enhance the generalization performance of CLIP. Niu et al. [29] introduce domain bank to incorporate textual domain knowledge into soft prompts. Both Zhou et al. [39] and Zhang et al. [36] establish a lightweight neural network (*i.e.*, multilayer

perceptron) to generate soft prompts conditioned on images with a residual or concatenation manner. However, these methods either lack diversity and visibility into the information from the target domain or rely on simple models to approximate the relationship between images and soft prompts. Consequently, the learned prompts may still fall short of being transferable.

As shown in Figure 1, unlike the previous prompt learning methods, such as unconditional prompt learning [18,40] and conditional prompt learning [8,36,39] methods, which typically rely on fixed prompts obtained from the training set for downstream tasks, or employ a straightforward multi-layer perceptron to learn a residual vector to enhance the richness of the fixed prompts, our approach takes a different direction. We reframe the prompt learning framework by adopting a generative perspective, marking the first integration of generative models into VLMs for prompt learning. We propose a new prompt learning paradigm **S**oft **P**rompt **G**eneration (SPG), which offers a straightforward yet effective solution for Domain Generalization. SPG is designed to exclusively harness a generative model for prompt generation, leveraging the model's inherent capability to encode domain and content knowledge directly into the generated prompts.

Our Proposed SPG: SPG method consists of a two-stage training phase and an inference phase. Specifically, in the initial training phase, we introduce domain prompt labels, representing optimal prompts for each domain. Images and corresponding domain prompt labels are then input into a simple yet effective generative model, Conditional Generative Adversarial Net (CGAN) [27], to incorporate domain and content knowledge into the prompt generator model in the second training phase. As a result, domain knowledge is stored not in soft prompts but in the generative model. During inference, the generator of the generative model is employed to obtain domain-specific soft prompts for the target domain data, enabling enhanced diversity and transferability for prompts.

Our main contributions are as follows:

- To the best of our knowledge, we are the first to introduce the generative model into prompt learning in VLMs. Then, we propose a new paradigm of prompt tuning, namely **S**oft **P**rompt **G**eneration (SPG).
- We design a two-stage training phase to align the generative model with domain prompt labels. It incorporates domain knowledge into the generated prompts, enhancing the transferability across unseen domains.
- Extensive experiments on five datasets for three DG tasks demonstrate that the proposed SPG achieves state-of-the-art performance.

2   Related Work

2.1   Domain Generalization

Domain Generalization (DG) aims to train a model using data from one or multiple related yet different source domains, enabling the model to generalize effectively to any out-of-distribution (OOD) target domain. Existing works mainly focus on learning domain-invariant features across one or multiple source

domains. One line of research focuses on domain alignment methods, primarily aiming to minimize moments [28], KL Divergence [23], and Maximum Mean Discrepancy [22] to learn domain-invariant features. Another line of work involves leveraging data augmentation to enrich images or their features, thereby contributing to the learning of invariant predictors [24,34]. Additional strategies encompass adversarial learning [12], ensemble learning [26], meta-learning [20], gradient operations [4], and more. Recently, large vision language models (VLMs) such as CLIP [32] have been applied to various downstream tasks, demonstrating their potential for remarkable generalization performance. One of the most efficient fine-tuning paradigms for VLMs is prompt learning. Building upon the success of prompt learning, our work delves deeper into exploring this highly efficient paradigm. We propose a novel prompt learning framework that departs from the previous prompt learning methods of utilizing fixed prompts [40] or residualed prompts [39] directly during inference. Instead, our framework leverages a generative model to dynamically obtain soft prompts for the inference process, introducing a new paradigm of prompt tuning.

2.2   Prompt Learning for Vision Language Models

Prompt learning for VLMs, also referred to as prompt tuning, aims to tune the model on domain-specific downstream tasks by only training a small set of parameters which might be a set of newly added parameters with the input. CoOp [40] firstly introduces soft prompt in VLMs and demonstrates a suitable prompt that can improve performance for the image recognition task. CoCoOp [39] solves the overfitting problem of CoOp with conditioned prompt learning, i.e., residualed prompt, ensuring the diversity of prompt. DAPL [13] and PDA [2] introduce domain labels and domain alignment into prompt learning for domain adaptation, respectively. For DG problem, CAE [29] aims to obtain domain-unified text representation with domain bank. DPL [36] generates prompts through MLP to add domain-specific features to the prompt template. Different from previous prompt learning methods in VLMs, we reframe the prompt learning framework from a generative perspective, i.e., exclusively relying on a generative model to dynamically produce soft prompts. For the DG setting, we further introduce a two-stage training paradigm and domain prompt labels to effectively embed domain knowledge within the generative model.

3   Preliminaries

3.1   Problem Setup of Domain Generalization

In DG setting, there are M source domains $D_s = \{D_s^1, D_s^2, ..., D_s^M\}$ and N target domains $D_t = \{D_t^1, D_t^2, ..., D_t^N\}$, all of which follow different distributions. For each source domain, $D_s^i = \{(x_j^i, y_j^i)\}_{j=1}^{n_i}$ where n_i denotes the size of samples in D_s^i and (x_j^i, y_j^i) denotes the pair of input and target label for the jth sample in the ith domain. Then we denote the input space as X and denote the label set as Y.

Assuming that all domains share the same label space, previous methods mainly focus on learning a domain-invariant model to map $F : X \rightarrow Y$ from images to labels, with the aspiration that this mapping can generalize to unseen target domains. With the advent of VLMs, prompt learning methods are proposed to incorporate soft prompts V into the input, and the mapping is rephrased as $F : \{X, V\} \rightarrow Y$. In contrast to previous methods that involved fine-tuning the model, prompt learning methods center on the fine-tuning of the prompt V. Unlike these methods above, we introduce a new prompt learning paradigm that solely relies on a generative model G to produce soft prompts directly. Thus, the mapping is rephrased as $F : \{X, G(X)\} \rightarrow Y$. Our goal is to learn a generative model that can capture both domain-invariant and domain-specific features to produce generalized prompts for unseen target domains dynamically.

3.2 Prompt Learning Methods in Generalization

Contrastive Language-Image Pre-Training (CLIP) [32] model is pre-trained on 400 million image-text pairs collected from the internet with contrastive learning. It consists of an image encoder f and a text encoder g, which encodes images and corresponding natural language descriptions, respectively.

Zero-shot CLIP directly incorporates a template text description as a prompt into the text encoder, such as "a photo of a [CLASS]" where [CLASS] denotes the class token. The image features and text features w of manually designed prompts are extracted from the image encoder and text encoder, respectively. The prediction of the class of the image is $\hat{y} = \underset{k}{\arg\max} \langle f(\mathbf{x}), w_k \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity.

CoOp [40] introduces a set of continuous learnable context vectors v concatenated with the template prompt c, namely soft prompt, then the ith class of text prompt t^i can be defined as $t^i = [v, c^i]$. Therefore the whole framework can be updated through the frozen CLIP model via tuning these prompts with cross-entropy loss $\mathcal{L}_{ce} = -\mathbb{E}_y \left[\sum_i y_i \log (\hat{y}_i | \mathbf{x}_i, t_i) \right]$, where $\hat{y}_i$ denotes ith element of the model's predicted probability distribution. However, CoOp learns a fixed prompt from training samples, which may lead to overfitting to the training distribution and degraded performance on test distribution.

CoCoOp [39] **and DPL** [36] attempt to overcome distribution shifts by learning an instance-specific continuous prompt that is conditioned on the input image with an MLP layer ϕ. Both of them learn an image-conditional vector $r = \phi(f(\mathbf{x}))$, which is then either added for CoCoOp with the learnable context vectors v or concatenated for DPL. Then the ith class of text prompt t^i can be defined as $t^i = [v + r, c^i]$ or $t^i = [r, c^i]$. By tuning the prompts and MLP layer, the loss is formulated as $\mathcal{L}_{ce} = -\mathbb{E}_y \left[\sum_i y_i \log (\hat{y}_i | \mathbf{x}_i, t_i, \phi) \right]$.

4 Method

Different from previous prompt learning methods, we abandon the training paradigm of the fixed prompt and residualed prompt. Instead, we introduce a
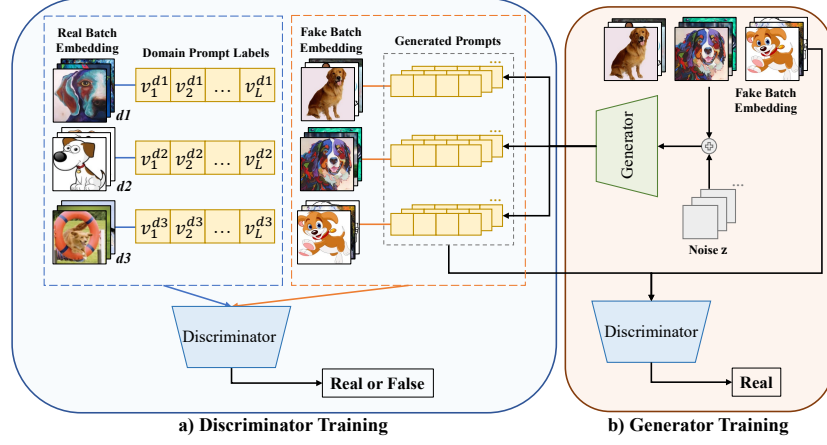
Fig. 2: The design of the second stage of the training phase. The condition generative adversarial net is the backbone of the generative model. The generator is guided by images to produce prompts. Meanwhile, the discriminator evaluates the authenticity of the prompt labels and the generated prompts with image data.

novel generative perspective to prompt learning, namely **S**oft **P**rompt **G**eneration (SPG). Specifically, our method includes a two-stage training phase to train a generative model incorporated with domain information and an inference phase to generate prompts directly. In our method, the generation of soft prompts is exclusively handled by a generative model. Notably, the domain knowledge is stored within the generative model, making it possible for each image to generate an instance-specific prompt. Our method ensures the diversity of prompts and allows for the incorporation of domain-specific information. We introduce our SPG method as follows.

Training Stage I: Domain Prompt Labels Learning. To better adapt our SPG method to the DG problem, we introduce the concept of domain prompt labels. Each domain corresponds to a domain prompt label $\mathbf{v}^{d_i}$ where d_i denotes ith domain, derived from training on the data of each domain with cross-entropy:

$$\mathbf{v}^{d_i*} = \arg\min_{\mathbf{v}} \mathbb{E}_{\mathbf{x}_j^{d_i}, y_j^{d_i}} \left[-\log p\left(y_j^{d_i} \mid \mathbf{x}_j^{d_i}, \mathbf{v}^{d_i}\right) \right], \tag{1}$$

where $(\mathbf{x}_j^{d_i}, y_j^{d_i})$ denotes images and labels of training samples in ith domain. The prompt design follows Zhou et al. [40] with text prompt tuning. These domain prompt labels represent optimal prompts for each domain, which encapsulate rich domain information.

Training Stage II: Generative Model Pre-training. We adopt a simple yet efficient generative model, conditional generative adversarial net (CGAN) [27],

to demonstrate the effectiveness of our method. CGAN, an extension of the conventional generative adversarial network (GAN) framework, operates by conditioning the generation process on additional information, usually presented in the form of auxiliary input data or labels. The CGAN architecture includes a generator G and a discriminator D. The generator G is guided by additional information, ensuring that the generated outputs align with the specified conditions. Meanwhile, the discriminator D evaluates the authenticity of the real data and fake output, fostering a dynamic adversarial interplay that refines the overall generative capabilities of the model. In this work, the fake batch of images is randomly sampled from the dataset.

As shown in Figure 2, we adapt the CGAN model to serve the DG task of prompt generation, aligning the generated prompts with their corresponding images. Our objective is to learn the generator's distribution p_g over soft prompt $\mathbf{v}$, transferring the transferability from prompts to the generator. In the generator, we define a prior on input noise variables $\mathbf{z}$, and the joint hidden representation $[\mathbf{z}, f(\mathbf{x})]$ combines these noise variables with image embeddings $f(\mathbf{x})$ with concatenation operation. In the discriminator, domain prompt labels and the generated prompt with image embeddings $f(\mathbf{x})$ serve as inputs to a discriminative function. Therefore, the objective function $V(G, D)$ of a two-player minimax game can be formulated as:

$$
\begin{aligned}
\min_G \max_D V(G, D) = {} & \mathbb{E}_{\mathbf{v} \sim p_\mathbf{v}(\mathbf{v})}[\log D(\mathbf{v} \mid f(\mathbf{x}))] \\
& + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z}|f(\mathbf{x}))))].
\end{aligned}
\tag{2}
$$

The pre-trained CGAN aims to capture domain-invariant and domain-specific features, ensuring consistency across diverse domains. Additionally, its generator maintains prompt diversity, enhancing the model's adaptability and generalization capabilities to handle varied inputs.

Inference. The generator of CGAN is employed to produce a domain-specific soft prompt for each image of the target domain. The probability of the image belonging to the ith class can be formulated as:

$$
p(y = i \mid \mathbf{x}) = \frac{\exp(\langle w_i, f(\mathbf{x}) \rangle / \tau)}{\sum_{j=1}^{K} \exp(\langle w_j, f(\mathbf{x}) \rangle / \tau)},
\tag{3}
$$

$$
\text{s.t.} \qquad w_i = g([G(\mathbf{z} \mid f(\mathbf{x})), c_i]),
\tag{4}
$$

where τ denotes temperature parameter, K denotes the number of classes, g denotes the text encoder, $G(\mathbf{z} \mid f(\mathbf{x}))$ denotes the domain-specific soft prompt, c_i denotes ith tokenized class token, and $[\cdot, \cdot]$ denotes the concatenation operation. In this way, exclusively relying on a generative model for soft prompt production provides benefits in terms of adaptability and dynamic prompt generation. Our method fosters diversity in prompt generation and enhances generalization across various tasks and domains.

5   Experiments

5.1   Experimental Setting

Datasets. Experiments are conducted on five popular benchmark datasets of DG, namely PACS [21], VLCS [11], OfficeHome [33], TerraIncognita (TerraInc.) [3] and DomainNet [30]. The dataset details can be seen in the appendix.

Baselines. We consider two types of methods for comparisons. For the traditional DG methods, we report the performance of ERM [14], SWAD [4], MIRO [5]. For CLIP-based methods, we compare SPG method with zero-shot CLIP (ZS-CLIP) [32], linear probing of CLIP (Lin. Prob.) [32], CoOp [40], Co-CoOp [39], VPT [17], VP [1], MaPLe [18], DPL [36].

Experimental Setup. ResNet50 (RN50) [15] and ViT-B/16 [10] are adopted as our backbones. For our SPG method, in the first stage, we train the domain prompt labels using the SGD optimizer with a batch size of 32 and the context length of the prompt label is 4. In the second stage, we train the CGAN model using the AdamW optimizer with weight decay 1e-4 and betas (0.9, 0.999), starting with a learning rate of 2e-3 for PACS, VLCS and TerraIncognita dataset, and a learning rate of 2e-4 for OfficeHome and DomainNet dataset. To enhance the stability of the CGAN, we incorporate a gradient clipping strategy to impose conditional constraints on the learnable parameter. For other CLIP-based methods, we set the learning rate initially to around 2e-3 and decay it using a cosine annealing rule with 20 epochs. We also set the batch size to 32. Moreover, the context tokens length is set to 2 for MaPLe, 10 for VPT and VP, and 16 for CoOp and CoCoOp. We employ the training-domain validation set method for model selection for all methods, selecting the model that achieves the highest accuracy on the overall validation set.

5.2   Comparison with SOTA Methods

Multi-source Domain Generalization. We follow the leave-one-domain-out evaluation protocol [14] for multi-source domain generalization. In this protocol, the model excludes one domain from the training set in each evaluation round and is then tested on the excluded domain. This iterative process continues until each domain has been excluded once.

In Table 1, we report the mean leave-one-domain-out performance of PACS, VLCS, Office-Home, TerraIncognita, and DomainNet for both ResNet50 and ViT-B/16 backbones. We can observe that SPG outperforms all other CLIP-based methods on five datasets with two backbones. For instance, with ResNet50 as the backbone, SPG can surpass zero-shot CLIP and the state-of-the-art (SOTA) CLIP-based method with a large margin of 13.7% and 3.4% for TerraIncognita. SPG also achieves a large improvement of around 4.0% and 2.2% compared with zero-shot CLIP and the SOTA method CoCoOp for VLCS. SPG

Table 1: Comparisons with SOTA methods on five domain generalization benchmark datasets for multi-source DG in terms of mean leave-one-domain-out performance with ResNet50 and ViT-B/16 as the backbone (B.). The results marked by † are the reported numbers from the original papers. Average accuracies and standard errors are reported from three trials. Bold denotes the best score.

Method	B.	PACS	VLCS	OfficeHome	TerraInc.	DomainNet	Avg
ERM† [14]	RN50	85.7±0.5	77.4±0.3	67.5±0.5	47.2±0.4	41.2±0.2	63.8
MIRO† [5]		85.4±0.4	79.0±0.0	70.5±0.4	50.4±1.1	44.3±0.2	65.9
SWAD† [4]		**88.1**±0.1	**79.1**±0.1	**70.6**±0.2	**50.0**±0.3	**46.5**±0.1	**66.9**
ZS-CLIP [32]	CLIP RN50	90.7±0.0	80.0±0.0	70.8±0.0	23.8±0.0	46.4±0.0	62.3
Lin. Prob. [32]		90.6±0.3	79.8±0.4	65.5±0.2	33.0±1.2	27.1±0.2	59.2
CoOp [40]		91.3±0.3	81.4±0.2	73.5±0.2	33.2±3.4	49.7±0.2	65.9
CoCoOp [39]		91.9±0.6	81.8±0.3	73.4±0.4	34.1±3.0	49.7±0.1	66.3
DPL [36]		91.8±0.7	80.8±0.8	73.6±0.4	34.4±1.0	49.6±0.2	66.0
VP [1]		90.2±0.1	80.5±0.3	70.2±0.2	25.6±1.0	45.8±0.1	62.4
SPG (ours)		**92.8**±0.2	**84.0**±1.1	**73.8**±0.5	**37.5**±1.8	**50.1**±0.2	**67.5**
ZS-CLIP [32]	CLIP ViT-B/16	96.1±0.0	82.3±0.0	81.8±0.0	33.8±0.0	56.6±0.0	70.2
Lin. Prob. [32]		94.9±1.4	77.5±0.7	79.3±0.2	44.6±2.1	48.2±0.2	68.9
CoOp [40]		96.4±0.3	80.8±0.3	83.0±0.1	46.8±0.7	59.5±0.2	73.6
CoCoOp [39]		96.7±0.2	80.3±0.3	83.4±0.2	45.3±2.4	59.4±0.2	73.2
DPL [36]		96.4±0.3	80.9±0.5	83.0±0.3	46.6±0.8	59.5±0.3	73.6
VP [1]		95.8±0.1	82.2±0.0	81.2±0.2	34.9±0.2	56.5±0.0	70.1
VPT [17]		96.9±0.2	82.0±0.2	83.2±0.1	46.7±0.6	58.5±0.2	73.6
MaPLe [18]		96.5±0.2	82.2±0.2	83.4±0.0	**50.2**±0.9	59.5±0.3	74.4
SPG (ours)		**97.0**±0.5	**82.4**±0.4	**83.6**±0.4	**50.2**±1.2	**60.1**±0.5	**74.7**

achieves a convincing improvement of approximately 1.5% on the averaged results of five benchmarks, establishing a new SOTA for the multi-source DG task. The results underscore the potential of generative models for prompt learning.

Meanwhile, SPG with ResNet50 as the backbone outperforms the SOTA traditional DG methods by a large margin of 4.7%, 4.9%, 3.2%, and 3.6% for PACS, VLCS, OfficeHome, and DomainNet datasets, respectively. We also observe that traditional DG methods achieve the SOTA performance on the TerraIncognita dataset, surpassing both the zero-shot CLIP model and our method. This discrepancy may arise from the fact that CLIP was not pre-trained on data similar to TerraIncognita, highlighting the need for further exploration of VLMs on unseen domain data and other downstream tasks.

Single-source Domain Generalization. The leave-all-but-one-domain-out evaluation protocol is adopted for single-source domain generalization. Under this protocol, all domains except one are included in the training set, and the model is then tested on the remaining domain.

Table 2 shows the experimental results of the leave-all-but-one-domain-out performance of PACS, VLCS, Office-Home, TerraIncognita and DomainNet for ResNet50 backbone. SPG outperforms other CLIP-based fine-tuning methods on five datasets. We observe averaged improvements of 19.0% and 1.4% compared with linear probing of CLIP and the SOTA method CoOp, respectively. This

Table 2: Comparisons with CLIP-based fine-tuning methods on five domain generalization benchmark datasets for single-source DG in terms of leave-all-but-one-domain-out performance with ResNet50 as the backbone. Bold denotes the best scores.

Method	PACS	VLCS	OfficeHome	TerraInc.	DomainNet	Avg
Lin. Prob. [32]	77.3	65.5	46.4	23.3	6.3	43.8
CoOp [40]	86.0	75.3	70.7	30.7	44.9	61.4
CoCoOp [39]	88.1	68.2	70.6	25.6	45.5	59.6
DPL [36]	86.8	75.2	70.8	28.4	43.8	61.0
SPG (ours)	**88.8**	**76.5**	**70.9**	**32.3**	**45.6**	**62.8**

Table 3: Comparisons with CLIP-based methods on four domain generalization benchmark datasets for cross-dataset generalization performance with ViT-B/16 as the backbone. Bold denotes the best scores.

Method	PACS	VLCS	OfficeHome	TerraInc.	Avg
ZS-CLIP [32]	94.5	**80.5**	82.1	30.9	72.0
CoOp [40]	96.2	72.8	83.1	32.1	71.1
CoCoOp [39]	95.8	72.6	83.3	34.2	71.5
DPL [36]	96.0	72.7	83.6	30.2	70.6
VPT [17]	95.7	77.7	82.5	26.2	70.5
MaPLe [18]	95.1	74.4	**83.9**	27.3	70.2
SPG (ours)	**96.7**	76.7	**83.9**	**38.0**	**73.8**

demonstrates that by leveraging the generative model for prompt generation, SPG is able to produce more domain-relevant and adaptive prompts, leading to improved performance across different domains. The domain-wise results for both multi-source DG and single-source DG are provided in the appendix.

Cross-Dataset Generalization. We split the DomainNet subset into training and validation datasets. With training-domain validation set model selection, we train on training data of DomainNet and select the best model on validation data of DomainNet. Then we test the best model on four downstream datasets, i.e., PACS, VLCS, OfficeHome and TerraIncognita.

As shown in Table 3, SPG outperforms all other CLIP-based methods, indicating the effectiveness of our generative prompt learning method in tackling distribution shifts. SPG achieves an average accuracy improvement of 1.8% compared with the zero-shot CLIP. It is also noteworthy that there is a stable improvement in the performance of baseline methods for the PACS and Office-Home datasets, such as CoOp, CoCoOp, etc. We attribute this to the similarity in the distribution between DomainNet and these downstream datasets. However, when encountering a significant distribution shift, such as observed with the VLCS and TerraIncognita, the performance of these methods may decline due to the learned prompts not being able to generalize to the new distribution. For TerraIncognita dataset, our SPG method can mitigate this challenge to some extent, the generative model is capable of dynamically generating prompts for
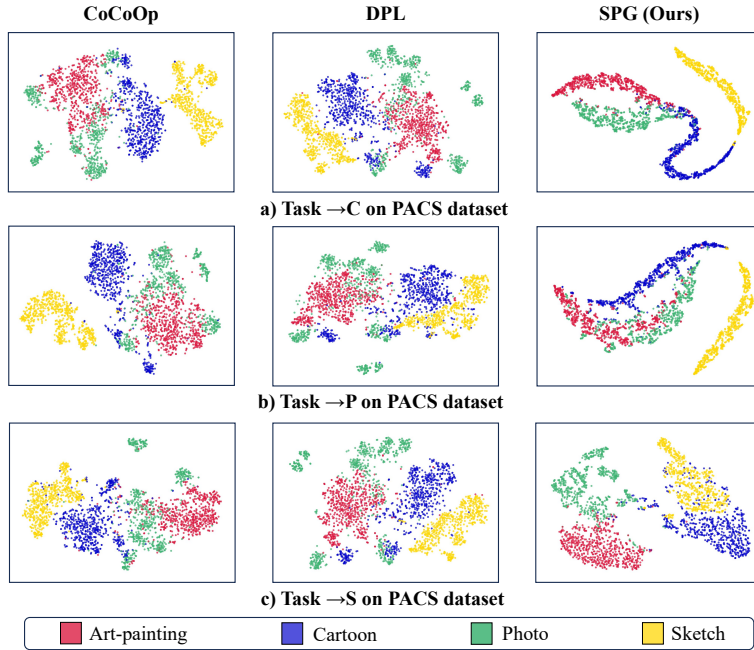
Fig. 3: The t-sne visualization of the prompt embeddings for CoCoOp, DPL, and our SPG method. Multi-source domain generalization models on 3 tasks of PACS dataset are employed to obtain prompt embeddings. Different colors denote different classes. All the domains including the target domain are highly clustered in SPG.

each image of unseen distribution, thus enabling more adaptable prompt generation. Nevertheless, it remains particularly challenging for the VLCS dataset, which indicates a need for further exploration.

5.3 Visualization

The t-SNE visualization. In Figure 3, we qualitatively evaluate prompt embeddings synthesized by baseline methods and our SPG method for three multi-source domain generalization tasks of PACS using t-SNE visualization [9]. We aim to generate a domain-specific prompt for each image, which ensures that prompt diversity and domain knowledge benefit the DG tasks. As illustrated in Figure 3, the prompts generated by our SPG method demonstrate a higher degree of clustering compared to prompts generated by alternative methods such as CoCoOp and DPL. This improved clustering indicates a higher degree of discriminative ability of the prompts, suggesting that our SPG method is more effective at capturing domain knowledge during prompt learning. The clustered domain prompts may also facilitate a better understanding of the semantic relationship between images and text, allowing the model to focus more on comprehending class concepts when aligning with the image encoder.
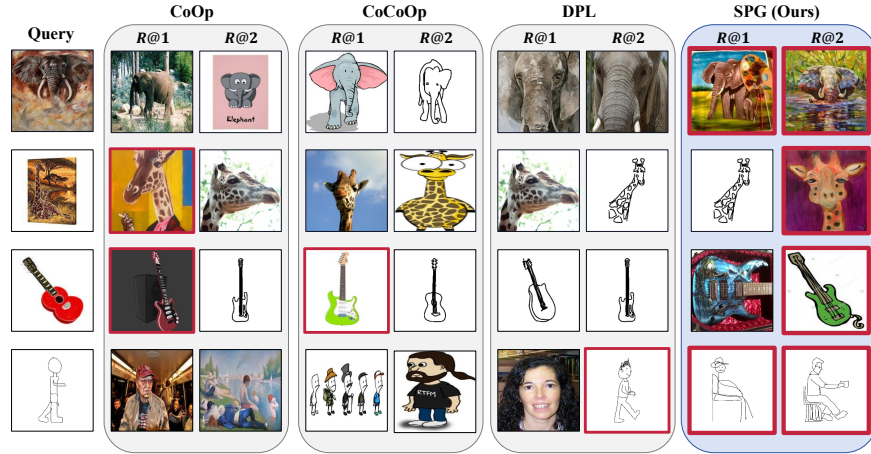
Fig. 4: The image-prompt-image retrieval experiment is designed to demonstrate the correlation between the prompts and the images. We present the top-2 results of image retrieval conducted using CoOp, CoCoOp, DPL, and our SPG method. Images encased in red rectangles indicate instances where the query image and the retrieval image belong to the same domain.

Image retrieval. We designed an image-prompt-image retrieval experiment to demonstrate the correlation between the generated prompt and the images in each domain. Specifically, we first sampled N-way K-shot images from each domain in the dataset and combined them into an image library, where N is 7 and K is 10 for PACS. Then, we randomly select one image as a query from the remaining images. The generated prompt of each query image is concatenated with the class token, which is passed through a text encoder to calculate the probability distribution with the features of each image in the image library. Subsequently, we select the probability value corresponding to the class of the query image as the confidence score. These scores are then sorted in sequence, and the top-2 images are obtained as the results of the image retrieval.

In Figure 4, we can observe that the prompts generated by our SPG method are capable of retrieving images that are more closely aligned with the query image, implying that the prompts generated by our model for the query image are more closely related to domain-specific information. It shows the effectiveness of our SPG method in encoding domain-relevant information within the prompts with a generative model. While previous methods have not emphasized the importance of domain knowledge, our SPG method stands out by explicitly integrating domain-specific information into the prompt generation process.

5.4 Ablation Study

Ablation on the different design of prompt. We have summarized four characteristics of prompts: fixed, learnable, conditional, and generative. Based

Table 4: Comparisons with different design of prompt on five domain generalization benchmark datasets for multi-source DG performance with ResNet50 as the backbone. O.H. denotes OfficeHome, and Do.Net denotes DomainNet. Bold denotes the best scores.

Method	Prompt Type				PACS	VLCS	O.H.	TerraInc.	Do.Net
	Fixed	Learnable	Conditional	Generative					
Manual Prompt	✓				90.7	80.0	70.8	23.8	46.4
Mix-domain-prompt	✓	✓			92.0	76.2	72.5	29.6	47.9
All-domain-prompt	✓	✓			91.3	81.4	73.2	33.2	49.7
CoCoOp [39]		✓	✓		91.9	81.8	73.4	34.1	49.7
DPL [36]		✓	✓		91.8	80.8	73.6	34.4	49.6
SPG (Ours)		✓		✓	**92.8**	**84.0**	**73.8**	**37.5**	**50.1**

Table 5: Ablation on the context length of prompt on PACS for multi-DG performance with ResNet50 as the backbone.

Context Length	Art	Cartoon	Photo	Sketch	Avg
2	91.5	**94.4**	99.1	74.6	89.9
4	92.8	93.8	**99.5**	**85.1**	**92.8**
4 (random)	**93.0**	93.4	**99.5**	82.9	92.2
8	83.1	91.7	99.2	79.2	88.3
16	84.7	93.6	98.3	82.0	89.7

on these characteristics, we selected five methods for comparison: (1) Manual prompt: manually designed prompt as "a photo of a [CLS]", where [CLS] is the class token. (2) Mix-domain-prompt: weighted sum of domain prompts obtained from each source domain. We set equal weights for each domain. (3) All-domain-prompt: prompt obtained from all source domain data. (4) CoCoOp. (5) DPL.

As shown in Table 4, in the multi-source DG setting, the manual prompt performs the worst, highlighting the necessity of fine-tuning for downstream tasks. For the learnable fixed prompt, the all-domain-prompt shows better performance than mix-domain-prompt. However, if the weights for mix-domain-prompt are more reasonable, there might be some improvement of performance, as some works have already attempted to explore this aspect [37]. Additionally, the conditional prompt methods outperform the fixed prompt methods, likely due to the integration of image information and a more diverse prompt. Finally, our generative prompt method performs the best, demonstrating its potential in prompt learning in VLMs.

Sensitivity of context length of prompt. As shown in Table 5, we evaluate the effects of different context lengths of prompt for multi-source DG on PACS using the ResNet50 backbone. We mainly initialize the soft prompt in two ways. One is to initialize it with the embedding of the text "a photo of a", namely word embeddings-based initialization. The other one is to sample from a zero-mean Gaussian distribution with a standard deviation of 0.02 (marked as random). In Table 5, we find that this initialization leads to slight improvement. Overall, we

14   Bai et al.

Table 6: Ablation on the training samples of the generative model (CGAN) on PACS for multi-DG performance with ResNet50 as the backbone.

Training Samples	Art	Cartoon	Photo	Sketch	Avg
20	30.1	34.9	23.4	34.5	30.7
40	51.2	55.3	34.3	65.5	51.6
60	73.8	76.7	56.8	81.1	72.1
80	89.2	93.5	84.8	**86.4**	88.5
full	**92.8**	**93.8**	**99.5**	85.1	**92.8**

observe that the prompt context length of 4 provides the optimal performance for our SPG method.

Sensitivity to the number of training samples. We sample different proportions of the training data for the generative model, including 20%, 40%, 60%, 80%, and all samples. Table 6 presents the results of our multi-source domain generalization experiments on the PACS dataset, and we observe a consistent increase in accuracy for all four tasks on the PACS as the amount of data for training the generative model increases. The performance reached its optimum when using all the data.

6   Conclusion

In this paper, we reframe the prompt learning framework from a generative perspective, and are the first to introduce the generative model into prompt learning in Vision Language Models (VLMs). We propose a simple yet efficient Soft Prompt Generation (SPG) method for the Domain Generalization (DG). SPG is a new paradigm of prompt tuning, which consists of a two-stage training phase and an inference phase. In the training phase, we introduce the concept of domain prompt labels, which are adopted to incorporate the generative model with domain knowledge. During the inference phase, the generative model is employed to obtain domain-specific soft prompts for target domain data. SPG relies exclusively on a generative model to directly produce soft prompts. This preserves the diversity of generated soft prompts, aiming to learn domain information with the generative model, including the target domain. Extensive experiments on three DG tasks of five DG benchmark datasets confirm the effectiveness of our proposed SPG method. Compared with traditional DG methods and CLIP-based approaches, SPG achieves new state-of-the-art performance for domain generalization. We hope this research inspires future exploration into tapping the potential of generative models in prompt generation and learning.

Acknowledgments

References

1. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 (2022)
2. Bai, S., Zhang, M., Zhou, W., Huang, S., Luan, Z., Wang, D., Chen, B.: Prompt-based distribution alignment for unsupervised domain adaptation. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI 2024). AAAI Press (2024)
3. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018)
4. Cha, J., Chun, S., Lee, K., Cho, H.C., Park, S., Lee, Y., Park, S.: Swad: Domain generalization by seeking flat minima. Advances in Neural Information Processing Systems **34**, 22405–22418 (2021)
5. Cha, J., Lee, K., Park, S., Chun, S.: Domain generalization by mutual-information regularization with pre-trained models. In: European Conference on Computer Vision. pp. 440–457. Springer (2022)
6. Cheng, D., Xu, Z., Jiang, X., Wang, N., Li, D., Gao, X.: Disentangled prompt representation for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23595–23604 (2024)
7. Cho, J., Nam, G., Kim, S., Yang, H., Kwak, S.: Promptstyler: Prompt-driven style generation for source-free domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15702–15712 (2023)
8. Derakhshani, M.M., Sanchez, E., Bulat, A., da Costa, V.G.T., Snoek, C.G., Tzimiropoulos, G., Martinez, B.: Bayesian prompt learning for image-language model generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15237–15246 (2023)
9. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655. PMLR (2014)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (2021)
11. Fang, C., Xu, Y., Rockmore, D.N.: 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. pp. 1657–1664 (2013)
12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016)
13. Ge, C., Huang, R., Xie, M., Lai, Z., Song, S., Li, S., Huang, G.: Domain adaptation via prompt learning. IEEE Transactions on Neural Networks and Learning Systems (2023)
14. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: International Conference on Learning Representations (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)

17. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)

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

19. Li, C., Liu, X., Wang, Y., Li, D., Lan, Y., Shen, C.: Dialogue for prompting: A policy-gradient-based discrete prompt generation for few-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18481–18489 (2024)

20. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Learning to generalize: Meta-learning for domain generalization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)

21. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE international conference on computer vision. pp. 5542–5550 (2017)

22. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5400–5409 (2018)

23. Li, H., Wang, Y., Wan, R., Wang, S., Li, T.Q., Kot, A.: Domain generalization for medical imaging classification with linear-dependency regularization. Advances in neural information processing systems **33**, 3118–3129 (2020)

24. Li, P., Li, D., Li, W., Gong, S., Fu, Y., Hospedales, T.M.: A simple feature augmentation for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8886–8895 (2021)

25. Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 624–639 (2018)

26. Li, Z., Ren, K., Jiang, X., Shen, Y., Zhang, H., Li, D.: Simple: Specialized model-sample matching for domain generalization. In: The Eleventh International Conference on Learning Representations (2022)

27. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

28. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: International conference on machine learning. pp. 10–18. PMLR (2013)

29. Niu, H., Li, H., Zhao, F., Li, B.: Domain-unified prompt representations for source-free domain generalization. arXiv preprint arXiv:2209.14926 (2022)

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

31. Qin, G., Eisner, J.: Learning How to Ask: Querying LMs with Mixtures of Soft Prompts. In: Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2021), https://aclanthology.org/2021.naacl-main.410/

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

33. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5018–5027 (2017)
34. Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. Advances in neural information processing systems **31** (2018)
35. Wang, Q., Lin, Y., Chen, Y., Schmidt, L., Han, B., Zhang, T.: Do clips always generalize better than imagenet models? arXiv preprint arXiv:2403.11497 (2024)
36. Zhang, X., Gu, S.S., Matsuo, Y., Iwasawa, Y.: Domain prompt learning for efficiently adapting clip to unseen domains. Transactions of the Japanese Society for Artificial Intelligence **38**(6), B–MC2_1 (2023)
37. Zheng, Z., Yue, X., Wang, K., You, Y.: Prompt vision transformer for domain generalization. arXiv preprint arXiv:2208.08914 (2022)
38. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
39. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
40. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)