ProLoG: Hybrid Prompt and LoRA Based Adaptation of Vision-Language Models for OOD Generalization

Jungwuk Park[1], **Dong-Jun Han**[2], **Jaekyun Moon**[1]

[1]Korea Advanced Institute of Science and Technology (KAIST), School of Electrical Engineering
[2]Yonsei University, Department of Computer Science and Engineering
savertm9@gmail.com, djh@yonsei.ac.kr, jmoon@kaist.edu

Abstract

While vision-language models (VLMs) achieve remarkable performance when fine-tuned on downstream in-distribution (ID) data, this process compromises their generalization ability on out-of-distribution (OOD) data that deviate from the downstream tasks due to overfitting. To address this, we propose ProLoG, a new adaptation method that effectively fine-tunes VLMs on downstream tasks while achieving high OOD performance. Specifically, we design a unique integration of prompt tuning and LoRA, offering a robust hybrid platform to improve performance. During training, we propose an augmentation-based regularization loss that enhances the generalization of our hybrid network by using augmented image features aligned with LLM-generated texts containing key attributes of each class. By leveraging our hybrid design, we also introduce an adaptive inference strategy that flexibly applies trained prompts and LoRA based on a task similarity score to effectively handle both ID and OOD data. Experimental results demonstrate that our proposed method outperforms existing works on various datasets, confirming its advantages.

Code & Appendix — github.com/savertm/ProLoG_VLM

1 Introduction

Vision-language models (VLMs) (Radford et al. 2021; Jia et al. 2021; Zhai et al. 2022) have demonstrated exceptional performance in various vision and multimodal applications such as image perception (Vidit, Engilberge, and Salzmann 2023; Liang et al. 2023; Xu et al. 2023; Zhou et al. 2022a) and text-to-image generation (Gal et al. 2022; Wang et al. 2022; Ganz and Elad 2024). These large-scale foundation models are trained on extensive datasets of image-text pairs, aligning image and text samples within a shared feature space. This comprehensive training on diverse data enables VLMs not only to achieve superior generalization performance on zero-shot tasks but also to serve as high-quality *foundation models* for various downstream tasks.

To efficiently adapt VLMs to downstream tasks, recent studies have focused on parameter-efficient tuning methods that introduce learnable parameters while preserving the weights of the pre-trained models. Adapter-based methods (Gao et al. 2024; Chen et al. 2022b) insert small learnable modules into the pre-trained model; low-rank adaptation (LoRA) (Hu et al. 2021; Zanella and Ben Ayed 2024) injects trainable low-rank matrices into existing weights; and prompt-based methods (Zhou et al. 2022b,a) introduce learnable prompts into the input or feature space of models.

Among them, prompt tuning (Zhou et al. 2022b; Zang et al. 2022) has been actively explored due to its effectiveness and simplicity of integration. CoOp (Zhou et al. 2022b) pioneers this direction by introducing learnable context vectors into the text input space. While these methods improve performance on downstream tasks, which are considered as in-distribution (ID), they compromise the model's original generalization to out-of-distribution (OOD) data that deviate from the ID data. This is because prompts tend to overfit to the downstream tasks during fine-tuning, which disrupts CLIP's original image–text alignment (Zhou et al. 2022a; Yao, Zhang, and Xu 2023). To address this, subsequent methods built on CoOp to improve OOD robustness via constraint-based optimization. Approaches such as (Yao, Zhang, and Xu 2023; Zhu et al. 2023; Roy and Etemad 2024; Khattak et al. 2023b; Cho, Kim, and Kim 2023) regularize prompts to preserve the original alignment, using text augmentations, gradient information, or distribution matching to enhance OOD generalization.

However, in most prompt-based methods, the same prompt is jointly optimized across all ID classes (i.e., classes in the downstream task) during fine-tuning, as shown in Fig. 1 (top). This shared design limits the model's ability to capture class-specific characteristics, as the same prompt is used to encode text features regardless of class. Moreover, as shown in Fig.1 (bottom), applying such prompts (optimized only for ID classes) to OOD samples that differ significantly at inference may result in inaccurate predictions, as the prompts may not semantically align with OOD classes. Although a few studies (Zhou et al. 2022a; Kan et al. 2023; Yao, Zhang, and Xu 2024) (e.g., CoCoOp) explore class-specific prompt tuning, they are limited to unimodal (text-only) settings and still struggle to generalize to unseen OOD classes.

Contributions. To address these key challenges in prior prompt-based methods, we propose **ProLoG**, a novel hybrid **Pro**mpt- and **Lo**RA-based adaptation method for OOD **G**eneralization in VLMs. ProLoG leverages the synergistic strengths of prompt tuning and LoRA to achieve strong performance on both ID and OOD data. In our hybrid design, LoRA captures class-specific features while prompts encode shared contextual information within a multimodal frame-

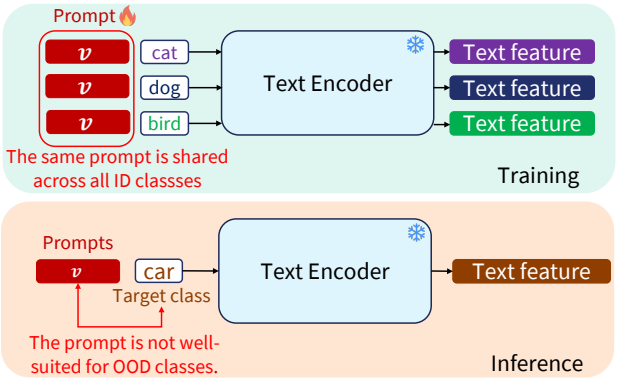


Figure 1: Limitations of existing prompt-based methods. Training (top) and inference (bottom) of prompt-based methods. During training, the same prompt is jointly optimized across all ID classes (e.g., cat, dog, bird), which limits the model's ability to capture class-specific features. At inference time, applying the same prompts to distant OOD samples (e.g., car) can lead to inaccurate predictions.

work, through a unique integration not explored in prior work. At inference, trained prompts and LoRA are flexibly applied in a task-adaptive manner to robustly handle both ID and OOD classes, as detailed in the following three components:

1) Hybrid structure: We introduce a new hybrid structure that strategically integrates prompt tuning with LoRA to provide an effective platform for handling ID/OOD data. Simply combining them tends to bias LoRA toward shared prompts, which can limit generalization. To address this, we propose a masking strategy that selectively applies LoRA only to class-related patches (not shared prompts) in image and text encoders, mitigating overfitting of LoRA. As a result, LoRA is trained to capture information specific to each ID class. In contrast, common prompts are inserted into the text input space for all classes to encode text features, playing a key role in capturing contextual knowledge across ID classes during training. As described in the next component, this hybrid structure also enables our adaptive inference strategy.

2) Regularized training via feature-aligned augmentation: The introduced prompts and LoRA are optimized on downstream tasks during training. However, as shown in (Khattak et al. 2023b; Shu et al. 2023), fine-tuning solely with a cross-entropy loss could distort the pre-trained CLIP's image-text alignment (due to overfitting on downstream tasks), hampering generalization to OOD data. To address this, we propose an augmentation-based regularization loss that minimizes the distortion of the original image-text alignment while encouraging the model to learn more generalized features. Our key idea is to leverage algebraic operations on the text feature space to augment image features using the pre-trained CLIP, aligning them with LLM-generated text augmentations that contain diverse class attributes. These augmentations are used to compute the ℓ_1 distance from the hybrid network's non-augmented features to regularize training. Unlike prior work (Roy and Etemad 2024) that applies random image augmentation, our method leverages semantic relationships between text features to guide meaningful image augmentation.

3) Adaptive inference strategy: Leveraging the design of our hybrid network, we propose a new inference strategy to effectively handle ID/OOD data, which flexibly applies trained prompts and LoRA in a task-adaptive manner. The core idea is to compute a task similarity score between training and test tasks to evaluate the extent of context shared between the training and test classes. Based on this score, the model makes predictions by adaptively applying the prompts and LoRA for each test sample. Building on the improvements from our training strategy, this inference scheme further enhances generalization to OOD data that deviate significantly from ID classes, addressing a key challenge of prior work.

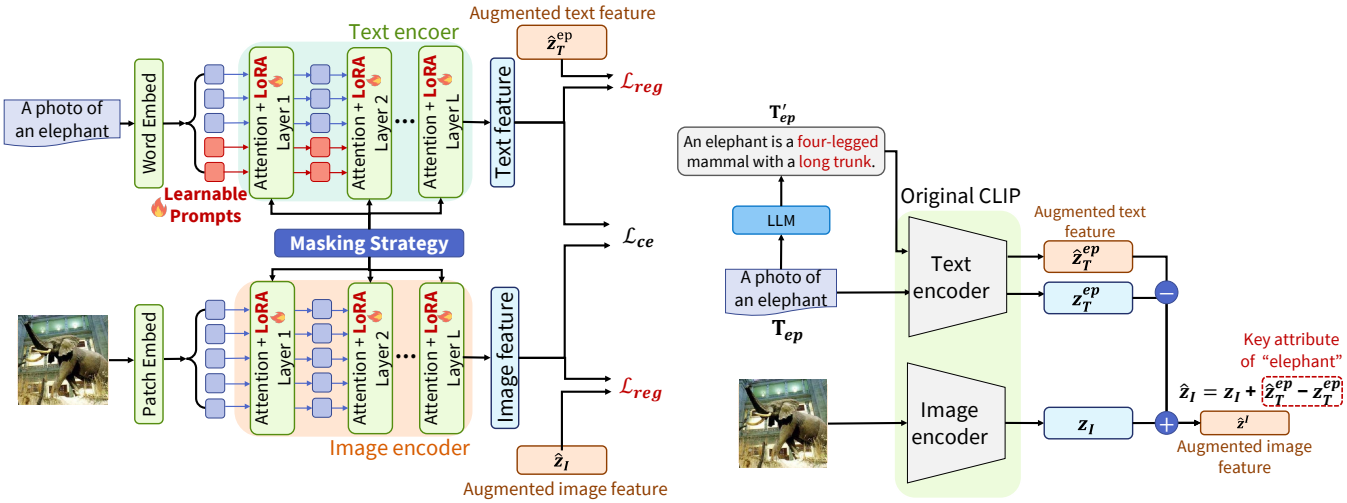
We demonstrate the effectiveness of ProLoG on common benchmark settings, including base-to-new generalization, cross-dataset generalization, and domain generalization. ProLoG outperforms existing baselines on various datasets, achieving strong performance on both ID and OOD samples.

2 Related Works

Vision-language models. Various vision-language models, including ViLT (Kim, Son, and Kim 2021), PaLI (Chen et al. 2022a), CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), and OpenCLIP (Cherti et al. 2023), have been proposed to bridge the modality gap between images and texts, showing remarkable performance in diverse tasks (Vidit, Engilberge, and Salzmann 2023; Liang et al. 2023; Ganz and Elad 2024). These networks are trained on large-scale image-text datasets, providing them with superior generalization ability. However, their performance on user-specific downstream tasks remains limited, requiring further task adaptation.

Paremeter-efficient fine-tuning. Instead of updating all the parameters of pre-trained models, a rich line of recent work has explored parameter-efficient tuning for VLMs by introducing learnable parameters, such as adapters, low-rank matrices, and prompts while preserving the original weights of VLMs. (Gao et al. 2024) proposed an adapter-based scheme that adds learnable weights within CLIP to perform residual-style feature blending, enabling the model to learn diverse features. (Zanella and Ben Ayed 2024) proposed a LoRA-based fine-tuning method that applies LoRA to the image and text encoders of CLIP for efficient fine-tuning.

Prompt tuning. Another approach for parameter-efficient fine-tuning is prompt-based methods, which add learnable prompts (or prefixes) in the input space or feature space of the text encoder to guide model outputs. CoOp (Zhou et al. 2022b) pioneered this approach by introducing context vectors into the text input space and optimizing them. Subsequent works improved CoOp in various ways: CoCoOp (Zhou et al. 2022a) used image features as conditional information for context vectors; KgCoOp (Yao, Zhang, and Xu 2023) integrated hand-crafted prompts to guide context vectors; TCP (Yao, Zhang, and Xu 2024) introduced textual class-aware prompts based on a textual knowledge embedding module. Recently, the authors of (Huang et al. 2024) proposed generating image-dependent text prompts by aggregating LLM-generated descriptions, improving generalization. However, due to its prompt predictions conditioned on the input image, this method exhibits low inference speed.



(a) Proposed hybrid structure and training process

(b) Proposed augmentation method

Figure 2: Proposed hybrid network and augmentation method: Fig.2a shows the hybrid network. Learnable prompts inserted into the text input space capture common context for the base task. LoRA is applied to each attention layer of image and text encoders with a new masking strategy, enabling effective training for each class. Fig. 2b shows how augmented image/text features for $\mathcal{L}_{reg}$ are generated. In this example, an augmented image feature is generated from an elephant image with two legs occluded, capturing the key characteristic of four legs.

Multimodal Prompting. Beyond text-only prompt tuning, recent works have explored multimodal prompt tuning. MaPLe (Khattak et al. 2023a) added learnable parameters to input and feature spaces of both image and text branches. PromptSRC (Khattak et al. 2023b) proposed a self-regulating method that aligns prompt features with original CLIP features at both the feature and logit levels, enhancing generalization. Recently, CoPrompt (Roy and Etemad 2024) used LLM-generated text augmentations with class-specific traits, applying consistency losses to preserve the original image-text alignment of CLIP. Another approach proposed in (Zang et al. 2024) introduced feature synthesis for unknown classes along with adaptive self-distillation, which can be combined with prior methods.

LoRA for prompt adaptation. Jain et al. (2024) employs LoRA to facilitate prompt adaptation by adjusting prompt vectors, with a focus on language tasks. In contrast, we address the distinct challenges of OOD generalization in vision by designing a unique integration of prompt tuning and LoRA through a masking strategy in a multimodal framework.

Despite prior efforts, it remains challenging to achieve strong performance on both ID and OOD data, due to the overfitting of adapters or prompts, as shown in (Khattak et al. 2023b; Shu et al. 2023), especially with OOD data that deviate significantly from downstream tasks. To address this, we propose a new hybrid network that leverages the strengths of both prompt tuning and LoRA, along with an augmentation-based regularization loss and an adaptive inference strategy tailored to our hybrid network. This approach suggests a new direction, and the experimental results indicate its effectiveness.

3 Proposed Algorithm

Problem Setup. In this work, we use the term *base task* to refer to the downstream task on which foundation models are fine-tuned. The base dataset is defined as $D_{base} = \{x_i^b, y_i^b\}_{i=1}^{N_b}$, consisting of N_b image-label pairs. Our goal is to effectively fine-tune VLMs on D_{base} to achieve high performance on the base task, while handling any *unseen target task* with OOD classes during inference.

3.1 Preliminaries

We adopt the pre-trained CLIP in (Radford et al. 2021) as the base model in this work. CLIP consists of an image encoder f_I and a text encoder f_T, which map image and text samples into a shared embedding space. For the image encoder, the input image $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ is divided into N patches, with a class token added. After the patch embedding layer, the resulting input is represented as $\mathbf{I} = [i_{cls}, i_1, \ldots, i_N]$. For the text encoder, the k-th class label is combined with a text template such as 'a photo of a {class k}', after which each word is tokenized and passes through the word embedding layer to form $\mathbf{T}_k = [t_{SOS}, t_1, \ldots, t_M, c_k, t_{EOS}]$, where $\{t_e|_{e=1}^{M}\}$ denote the template embeddings, c_k represents the k-th class label embedding, and t_{SOS} and t_{EOS} are the start and end embeddings in CLIP. The image and text inputs are then encoded via their respective encoders to obtain the image feature $\mathbf{z}_I = f_I(\mathbf{I}) \in \mathbb{R}^d$ and the text feature $\mathbf{z}_T^k = f_T(\mathbf{T}_k) \in \mathbb{R}^d$, both of which are normalized.

During inference, given an image feature $\mathbf{z}_I$ and a set of K text features $\{\mathbf{z}_T^k\}_{k=1}^K$ corresponding to all classes, the logit of the k-th class is computed as:

$$\text{Logit}_k(\mathbf{z}_I, \{\mathbf{z}_T^k\}_{k=1}^K) = \frac{\exp(\text{sim}(\mathbf{z}_I, \mathbf{z}_T^k)/\tau)}{\sum_{c=1}^K \exp(\text{sim}(\mathbf{z}_I, \mathbf{z}_T^c)/\tau)}, \quad (1)$$

where K is the number of classes, $\text{sim}(\cdot)$ denotes the cosine similarity, and τ is the temperature parameter. Here, we define the logit computed by the original CLIP as $\text{Logit}_k^{ori} :=$

$\text{Logit}_k(\mathbf{z}_I, \{\mathbf{z}_T^k\}_{k=1}^K)$. The corresponding logit vector for all classes is denoted as $\text{Logit}^{ori} = [\text{Logit}_k^{ori}]_{k=1}^K$.

Low-rank adaptation. Low-Rank Adaptation (LoRA) (Hu et al. 2021) enables efficient fine-tuning of foundation models by introducing two learnable low-rank matrices, $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$ ($r \ll d$). Given an input feature x, LoRA updates the pre-trained weight (W) to W' as:

$$W'x = Wx + \gamma \Delta W x, \quad \text{where} \quad \Delta W = BA \qquad (2)$$

and γ is a scaling factor. LoRA is applied to every patch in the attention layers of the image and text encoders of CLIP. For example, given a text input $\mathbf{T}_k$, let $\mathbf{h}_T^{k,l} = [h_{SOS}^l, h_1^l, \ldots, h_M^l, h_{c_k}^l, h_{EOS}^l]$ be the intermediate feature before the $(l+1)$-th attention layer. In the attention layer, LoRA is applied to the features of all patches in $\mathbf{h}_{k,l}^T$ for each key, query, and value matrix in the attention layers.

3.2 Hybrid Prompt and LoRA Architecture

To offer an effective platform for handling both ID and OOD data, we propose a new hybrid network that strategically integrates prompt tuning with LoRA, leveraging the synergistic strengths of both approaches. In our structure, prompts capture shared contextual knowledge across all classes in the base task, while LoRA modules learn class-specific information. This hybrid design also enables our inference strategy to handle various target scenarios, which will be discussed later. The overall architecture is depicted in Fig. 2a.

Prompt insertion. We insert prompts into the text input space to learn common contextual knowledge across all classes in the base task during training. Specifically, v learnable prompts, denoted as $\mathbf{P} = \{p_1, p_2, \ldots, p_v\}$, are added to the text input, resulting in an augmented text input: $\mathbf{T}_{k,p} = [t_{SOS}, t_1, \ldots, t_M, c_k, \mathbf{P}, t_{EOS}]$. The inserted prompts $\mathbf{P}$ are shared across all class inputs $\{\mathbf{T}_{k,p}\}_{k=1}^K$, ensuring that common information is injected into the text features of all classes by the text encoder. As a result, the prompts are optimized on the base task, capturing shared contextual knowledge.

LoRA with a masking strategy. However, since these prompts are commonly applied across all classes, their ability to learn fine-grained information specific to each class is limited. To address this, we additionally apply LoRA modules to the projection matrices for the query, key, and value in both the image and text encoders of CLIP. Nevertheless, naively applying LoRA to all patches, including the prompts and start/end tokens, may lead to overfitting, limiting the model's generalization ability, as these patches are also shared across text features of all classes in both the base and target tasks (see Appendix B for more details). This is because, after being trained on these shared patches for the base task, LoRA may inject biased information into predictions for the target task that contains OOD classes. To mitigate this, we propose a masking strategy that applies LoRA only to class-specific patches. In the text encoder, given an intermediate feature $\mathbf{h}_T^{k,l}$, we define a binary mask $\mathbf{S}_T$ corresponding to each token: $\mathbf{S}_T^k = [s_{SOS}^T, s_1^T, \ldots, s_M^T, s_{c_k}^T, s_{\mathbf{P}}^T, s_{EOS}^T]$ where

$$s_i^T = \begin{cases} 1, & \text{if } i \text{ corresponds to the class token } c_k, \\ 0, & \text{otherwise.} \end{cases}$$

These masks are applied to the intermediate feature ($\mathbf{h}_T^{k,l}$) before each LoRA layer is applied. Then, the final LoRA output in Eq. (2) is modified as follows: $W'\mathbf{h}_T^{k,l} = W\mathbf{h}_T^{k,l} + \gamma BA(\mathbf{h}_T^{k,l} \odot \mathbf{S}_T^k)$, where $\odot$ denotes patch-wise multiplication. Details on the binary mask in the image encoder are provided in Appendix B. By incorporating these masks into LoRA modules for both the text and image encoders, we effectively mitigate the overfitting caused by shared patches.

3.3 Regularized Training via Feature-Aligned Augmentation

Our hybrid network with prompts and LoRA is fine-tuned on the base task during training. However, fine-tuning solely with a cross-entropy loss can distort the image-text alignment of the original CLIP, hampering generalization to OOD samples (Khattak et al. 2023b; Shu et al. 2023). To address this, we propose an augmentation-based regularization loss. Our key idea is to augment image features in the original CLIP's embedding space, aligning them with LLM-generated text augmentations, which contain diverse key attributes of each class, by leveraging algebraic operations on the text feature space. These augmentations are used to regularize the model during training, ensuring that the hybrid network's features remain aligned with the original space while learning more generalized features, as illustrated in Fig. 2.

Text augmentation. Inspired by (Roy and Etemad 2024; Yao, Zhang, and Xu 2023), we first use LLMs to generate text augmentations containing diverse key attributes of each class to enhance generalization (see Appendix A for more details). These augmented texts, along with the original text inputs, are processed through the text encoder of the pre-trained CLIP to extract their corresponding features on the original CLIP embedding space. For example, given a text input $\mathbf{T}_{ep}$: 'A photo of an elephant', we generate multiple text augmentations per class, such as $\mathbf{T}'_{ep}$: 'An elephant is a four-legged mammal with a long trunk', describing its key characteristics. Each text is encoded and normalized as $\mathbf{z}_T^{ep}$ and $\hat{\mathbf{z}}_T^{ep}$ for $\mathbf{T}_{ep}$ and $\mathbf{T}'_{ep}$, respectively.

Image augmentation via feature alignment. Given the text augmentations, we generate image feature augmentations with key class attributes aligned to the augmented text features, to enable the model to learn more generalized features during training. Unlike prior methods that do not consider image augmentation (Khattak et al. 2023a,b) or simply apply random image augmentation (Roy and Etemad 2024), we propose a simple yet effective method that considers the relationships between augmented image and text features to facilitate image-text alignment. Our key idea is to leverage the property that algebraic operations on two textual representations can capture the semantic relationship between them. As observed in (Ramesh et al. 2022; Vidit, Engilberge, and Salzmann 2023), the difference between the two features, $\hat{\mathbf{z}}_T^{ep} - \mathbf{z}_T^{ep}$, captures key attributes (such as four-legged and long trunk) present in $\mathbf{T}'_{ep}$ but not in $\mathbf{T}_{ep}$. This semantic difference in the text space can be transferred to the image space, as image and text features are aligned in the embedding space. Using this property, we transform an elephant image feature $\mathbf{z}_I$, extracted from the original CLIP, into $\mathbf{z}_I + (\hat{\mathbf{z}}_T^{ep} - \mathbf{z}_T^{ep})$,

followed by normalization, to obtain a new image feature $\hat{\mathbf{z}}_I$ that reflects the key attributes of the augmented text feature $\hat{\mathbf{z}}_T^{ep}$, as illustrated in Fig. 2b and detailed in Appendix H.

Regularization loss. Given an image sample $\mathbf{I}$ from the k-th class, we compute the regularization losses using the augmented text and image features as follows:

$$\mathcal{L}_{reg} = \left\| \hat{\mathbf{z}}_I - \tilde{f}_I(\mathbf{I}) \right\| + \left\| \hat{\mathbf{z}}_T^k - \tilde{f}_T(\mathbf{T}_k, \mathbf{P}) \right\|, \quad (3)$$

where $\tilde{f}_I$ and $\tilde{f}_T$ are the image and text encoders with LoRA applied in our hybrid network, $\mathbf{P}$ is the prompt added to the text input, and $\|\cdot\|$ denotes the L_1 norm. This loss aligns the image and text features from our hybrid network with their corresponding augmented image and text features, emphasizing the direction of diverse key attributes for each class in the original feature space. This guides the model to produce improved image and text features for better generalization. Furthermore, since the augmented features are extracted from the original CLIP, the loss also prevents significant deviation from the original embeddings during fine-tuning, preserving CLIP's original knowledge.

Additionally, we apply a cross-entropy loss to directly align image samples with their labels using the features extracted by our proposed network. Given a sample with the k-th label, the cross-entropy loss is computed as $\mathcal{L}_{ce} = -\log \text{Logit}_k^{pl}$, where Logit_k^{pl} is computed by Logit_k in Eq. (1) as:

$$\text{Logit}_k^{pl} \coloneqq \text{Logit}_k \left(\tilde{f}_I(\mathbf{I}), \left\{ \tilde{f}_T(\mathbf{T}_k, \mathbf{P}) \right\}_{k=1}^{K} \right). \quad (4)$$

With both losses and a hyperparameter λ, the final loss is defined as $\mathcal{L}_{final} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{reg}$.

3.4 Adaptive Inference Based on Task Similarity

By strategically leveraging the design of our hybrid network, we propose a new inference strategy that adaptively applies the trained prompts and LoRA based on a task similarity score to effectively handle diverse target task scenarios during inference. This strategy utilizes class information from the target task, provided prior to the inference phase, as required for CLIP's prediction. Given a target task, the key idea is to compute a task similarity score between base and target tasks, categorize the target task into predefined scenarios, and adaptively apply the prompts and LoRA accordingly.

Task similarity score (S_{TS}): Let $\mathbf{Z}_T^b \in \mathbb{R}^{K_b \times d}$ and $\mathbf{Z}_T^t \in \mathbb{R}^{K_t \times d}$ be the text features of the base and target classes, respectively, obtained using the text encoder of the pre-trained CLIP, where K_b and K_t denote the numbers of classes in each task. Using these features, we compute a similarity score ($S \in \mathbb{R}$) for the i-th target class with respect to the base task as: $S(\mathbf{z}_{(t,i)}, \mathbf{Z}_T^b) \coloneqq \text{sim}(\text{Proj}_{\mathbf{Z}_T^b}(\mathbf{z}_{(t,i)}), \mathbf{z}_{(t,i)})$ where $\mathbf{z}_{(t,i)} \in \mathbb{R}^d$ is the i-th class feature in the target task and $\text{Proj}_{\mathbf{Z}_T^b}(\mathbf{z}_{(t,i)})$ is its projection onto the base class feature space (see Eq. (1) in Appendix). The task similarity score $S_{\text{TS}} \in \mathbb{R}$ is then defined as the average over all target classes: $S_{\text{TS}}(\mathbf{Z}_T^t, \mathbf{Z}_T^b) \coloneqq \frac{1}{K_t} \sum_{i=1}^{K_t} S(\mathbf{z}_{(t,i)}, \mathbf{Z}_T^b)$. A higher score indicates that the target classes are semantically closer to the base task and share a similar context with it. Based on this score, we categorize target tasks into three scenarios:

1) Scenario I (Sufficient context sharing): If S_{TS} exceeds a predefined threshold δ_1, indicating a high similarity between the base and target tasks, we hypothesize that the target classes share sufficient contextual knowledge with the base classes. Since the target task closely aligns with the knowledge captured by the prompts and LoRA from the base task, we make predictions using both of them as follows: $\text{Logit}^{S_{\text{I}}} = \text{Logit}^{pl}$ where Logit^{pl} is the logit vector consisting of Logit_k^{pl} in Eq. (4).

2) Scenario II (Partial context sharing): If S_{TS} lies within the predefined thresholds δ_1 and δ_2 ($\delta_2 < S_{\text{TS}} < \delta_1$), we hypothesize that the target classes partially share contextual knowledge with the base classes. This case represents a practical scenario where the target task includes both ID and OOD classes. In our hybrid structure, the prompts are designed to capture the contextual knowledge of the base task, while LoRA modules focus on class-specific information. Although both are trained to improve generalization on OOD samples via the regularization loss, the base task's context knowledge captured by the prompts may offer limited benefits for OOD samples that deviate significantly from the base context, as it may not align with considerably unrelated classes. For example, when the base task is to classify types of flowers, the contextual knowledge about flowers learned by the prompts tends to be effective for ID or (potentially) OOD flower classes but less effective for unrelated classes, such as elephants or cats. Further insights are in Appendix C.

Process in Scenario II. To further enhance performance in this scenario where ID and diverse OOD samples are mixed, the core idea is to estimate whether each test sample belongs to distant OOD classes or not and adaptively apply the prompts and LoRA accordingly. For a given target image, we first compute the logit Logit_k^{pl} using the prompts and LoRA as in Scenario I. The predicted label $\hat{y}^{pl}$ and its confidence C_{pl} are obtained from $\arg\max$ and $\max$ of the logits, respectively, to estimate if the sample belongs to distant OOD classes. If the similarity score $S(\mathbf{z}_{(t,\hat{y}^{pl})}, \mathbf{Z}_T^b)$ for the predicted label is below δ_1, and the confidence C_{pl} is below a predefined threshold C_0, the sample is likely an OOD sample substantially different from the base task. For such samples, we perform ensemble predictions by combining the logits from predictions with and without the learned prompts. This approach integrates the contextual knowledge from the base task with a context-agnostic perspective, enabling the model to consider diverse viewpoints and thereby improve performance on uncertain OOD data. Specifically, we compute logits without the prompts as $\text{Logit}_k^l \coloneqq \text{Logit}_k(\tilde{f}_I(\mathbf{I}_j), \{\tilde{f}_T(\mathbf{T}_k)\}_{k=1}^K)$. The final prediction is obtained by ensembling Logit^l, which excludes context knowledge, with Logit^{pl}, which incorporates it, improving performance on challenging OOD samples. Note that the extra computational cost for the ensemble prediction is negligible, as the text features can be precomputed and remain fixed during inference (see Appendix F).

Conversely, if a sample has a high similarity score or high confidence, it is considered as an ID sample or an OOD sample related to the base task. This is because, even with a low similarity score, high confidence suggests that the model

	Average on 11 datasets			ImageNet			Flowers102			OxfordPets		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
CLIP (Zhou et al. 2022b)	66.95	71.54	69.17	67.46	64.04	65.71	72.27	**74.18**	73.21	90.75	96.70	93.63
CoOp (Zhou et al. 2022b)	79.33	63.48	70.52	71.05	62.10	66.27	94.83	56.56	70.85	92.53	95.47	93.97
CoCoOp (Zhou et al. 2022a)	77.25	69.35	73.09	71.21	**66.80**	68.93	91.07	68.65	78.29	93.09	96.70	94.86
KgCoOp (Yao, Zhang, and Xu 2023)	77.31	70.74	73.88	70.64	65.62	68.04	91.07	70.35	79.38	93.67	96.42	95.03
TCP (Yao, Zhang, and Xu 2024)	80.69	71.84	76.01	71.79	65.79	68.66	95.92	71.74	82.09	92.88	96.67	94.74
CLIPood (Shu et al. 2023)	80.34	71.78	75.82	71.91	65.59	68.60	90.79	71.92	80.26	94.47	95.53	95.00
MaPLe (Khattak et al. 2023a)	78.99	70.69	74.61	71.50	65.63	68.44	93.21	72.16	81.34	93.44	93.54	93.46
PromptSRC (Khattak et al. 2023b)	**81.28**	71.29	75.95	72.58	65.73	68.99	95.73	71.85	82.07	93.78	96.20	94.97
CoPrompt (Roy and Etemad 2024)	79.15	70.14	74.38	71.21	66.45	68.75	93.21	69.12	79.37	93.54	94.66	94.08
ProLoG (ours)	81.05	**73.05**	**76.84**	**73.02**	66.48	**69.60**	**96.02**	71.88	**82.21**	**94.77**	**97.40**	**96.06**

Table 1: Results in the base-to-new generalization setting. Due to limited space, the full results on 11 datasets are reported in Appendix I. The best and second-best results are bolded and underlined. Based on the task similarity score, tasks on base and new class across all datasets are categorized as Scenario I and II, respectively, with ProLoG applying the corresponding inference strategy.

has encountered similar patterns before, making it more likely that the sample shares attributes with the base task. For such samples, the model uses Logit_k^{pl} to make predictions as in Scenario I. Overall, the prediction for Scenario II is given by:

$$\text{Logit}^{S_{\text{II}}} = \begin{cases} \frac{\left(\text{Logit}^{pl}+\text{Logit}^{l}\right)}{2}, & \text{if } C_{pl}<C_0 \ \&\ S(\mathbf{z}_{(t,\hat{y}^{pl})},\mathbf{Z}_b)<\delta_1, \\ \text{Logit}^{pl}, & \text{otherwise.} \end{cases}$$

3) Scenario III (No shared context): If S_{TS} is below the threshold δ_2 (i.e., $S_{\text{TS}} < \delta_2$), indicating a significantly low similarity, the target classes are considered to share no context with the base task. In this case, although the hybrid network trained on the base task preserves generalization ability, it does not offer a significant advantage over the original CLIP, as the target task consists of OOD classes that deviate substantially from the base classes. Thus, for efficiency, predictions are made using the original CLIP: $\text{Logit}^{S_{\text{III}}} = \text{Logit}^{ori}$.

Remark. During inference, the model adaptively applies the prompts and LoRA based on the task similarity score, effectively handling diverse target task scenarios. Even when scenarios are not ideally categorized, performance remains stable due to the strong generalization of the prompts and LoRA (trained with our regularization loss), as well as the inherent robustness of the pre-trained CLIP, making our method insensitive to hyperparameters. For all experiments, we fix δ_1, δ_2 for scenario categorization, and C_0 for model confidence to 0.98, 0.85, and 0.5, respectively, demonstrating robustness under consistent settings. See Fig. 3 and Appendix D for ablation results. Our method introduces only a small number of additional parameters and incurs low computational overhead compared to the baselines, as detailed in Appendix F.

4 Experimental Results

4.1 Experimental Setup

To validate ProLoG, which integrates both the proposed training and inference strategies within the hybrid network, we adopt the standard evaluation framework of CoOp (Zhou et al. 2022b,a). Specifically, we consider three conventional benchmark settings: 1) base-to-new generalization, 2) cross-dataset generalization, and 3) domain generalization (in Appendix D). We use CLIP with ViT-B/16 and ViT-B/32 backbones, and provide results with a different backbone in Appendix D. Further implementation details are in Appendix A. All results in the tables represent average values over three random seeds. We consider 11 image classification datasets encompassing a diverse range of object classes: ImageNet (Deng et al. 2009), Caltech101 (Fei-Fei, Fergus, and Perona 2004), OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Van Gool 2014), FGVCAircraft (Maji et al. 2013), SUN397 (Xiao et al. 2010), UCF101 (Soomro 2012), EuroSAT (Helber et al. 2019) and DTD (Cimpoi et al. 2014). Additionally, we consider the Waterbirds dataset (Sagawa et al. 2019) to evaluate robustness to spurious correlations.

Baselines. We compare ProLoG with recent baselines: a full fine-tuning method (CLIPood (Shu et al. 2023)); LoRA-based (CLIP-LoRA (Zanella and Ben Ayed 2024)); text prompt methods (CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), KgCoOp (Yao, Zhang, and Xu 2023), TCP (Yao, Zhang, and Xu 2024)); and multimodal prompt methods (MaPLe (Khattak et al. 2023a), PromptSRC (Khattak et al. 2023b), CoPrompt (Roy and Etemad 2024)). For fair comparison, we reproduce their results using the official code.

4.2 Base-to-New Generalization

In this setup, classes in each dataset are divided into two disjoint categories: base and new classes. The model is trained on the base classes, and evaluated on two separate test datasets: one for base classes and another for new classes. Table 1 shows the results across 11 datasets using ViT-B/32, reporting accuracies on base and new classes, and their harmonic mean (HM). Due to limited space, the full results on 11 datasets using ViT-B/32 and ViT-B/16 are in Appendix I.

Inference scenario in ProLoG. In Table 1, target tasks are categorized into Scenario I or II based on the task similarity score. For all datasets, according to δ_1 and δ_2, when the target classes are base classes, they are categorized as Scenario I. When the target classes are new classes, they are categorized as Scenario II, as they are split from the same dataset as the base classes, ensuring a certain level of similarity. The corresponding inference strategy is applied accordingly.

Results. As shown in Table 1, ProLoG consistently outperforms all baselines in average HM across 11 datasets, regard-

	Base	Target										
Scenario	UCF I	Caltech II	Pets III	Cars III	Flowers III	Food II	Aircraft III	SUN II	DTD II	EuSAT II	ImNet III	Avg.
MaPLe	**78.18**	87.97	82.19	58.04	54.65	82.73	15.71	57.10	37.74	47.82	61.54	60.33
PromptsSRC	76.20	91.22	82.49	63.88	66.41	**85.30**	20.70	**64.18**	42.65	**50.75**	66.63	63.42
CoPrompt	77.80	90.39	84.57	60.19	54.53	82.04	21.27	62.19	38.95	43.43	65.04	60.26
ProLoG (Ours)	77.20	**92.56**	**88.03**	**65.39**	**67.32**	85.24	**23.82**	63.56	**45.07**	44.19	**66.73**	**64.19**

(a) Results in a cross-dataset generalization setting using the UCF101 dataset as the base task.

	Base	Target										
Scenario	DTD I	Caltech III	Pets III	Cars III	Flowers III	Food III	Aircraft III	SUN III	UCF III	EuSAT II	ImNet III	Avg.
MaPLe	**63.91**	89.86	84.16	58.65	65.06	82.78	13.39	58.88	59.57	44.51	62.72	61.96
PromptsSRC	61.94	92.06	86.43	64.31	**67.56**	85.28	18.73	62.71	63.78	**49.71**	66.27	65.69
CoPrompt	62.00	92.17	82.47	61.24	63.30	82.55	19.83	**63.05**	62.20	49.15	64.95	64.09
ProLoG (Ours)	63.67	**92.90**	**88.03**	**65.39**	67.32	**85.41**	**23.82**	62.56	**65.00**	43.82	**66.73**	**66.10**

(b) Results in a cross-dataset generalization setting using the DTD dataset as the base task.

Table 2: Results in the cross-dataset generalization setting. (I), (II), and (III) indicate that the target task is categorized as Scenario I, II, and III respectively, based on the proposed task similarity score. The corresponding inference strategy is then applied in ProLoG.

less of the backbone. Specifically, ProLoG outperforms most baselines and achieves performance comparable to PromptSRC in the average base-class accuracy. In terms of average new-class accuracy, it significantly surpasses all baselines with a gain of over 1.2%, resulting in the highest average HM score with a margin of over 0.83%.

4.3 Cross-Dataset Generalization

In the cross-dataset generalization setup, the model is trained on one dataset (base) and evaluated on different datasets (target). This setup is more challenging than the base-to-new generalization, as the base dataset is entirely different from target datasets, and the task-specific knowledge learned from the base task is generally not shared with target tasks. We adopt UCF101 and DTD as base datasets and use the ViT-B/16 backbone for this setting. Table 2 compares our method with competitive multimodal prompt learning baselines (MaPLe, PromptSRC, and CoPrompt) in this setup.

Inference scenario in ProLoG. Each target task is categorized as Scenario II or III based on task similarity, and the corresponding inference strategy is applied. Due to the specificity of UCF101 and DTD, which focus on human action and texture recognition, respectively, target datasets (e.g., OxfordPets) that differ substantially from each base dataset are categorized as III (see Tables 2a–2b).

Results. In Tables 2a and 2b, ProLoG outperforms MaPLe and CoPrompt on average target accuarcy, achieving over a 2% gain. While PromptSRC shows improved performance over other baselines, it still lags behind ProLoG and shows over a 1% drop in base accuracy, highlighting the benefits of ProLoG. Additional results on ImageNet are in Appendix D.

4.4 Robustness to Spurious Correlations

The Waterbirds dataset is used to assess spurious correlations, consisting of bird images where each foreground class (i.e., waterbird or landbird) appears with two background types

Method	Average	Worst†
CLIP (Radford et al. 2021)	64.53	40.34
CoOP (Zhou et al. 2022b)	77.10	43.93
ERM (Vapnik and Vapnik 1998)	78.18	47.98
PromptSRC (Khattak et al. 2023b)	76.40	50.15
CoPrompt (Roy and Etemad 2024)	**83.49**	52.83
ProLoG (ours)	83.46	**54.75**

Table 3: Results on Waterbirds.

(i.e., water or land). During training, waterbirds largely appear on water and landbirds on land, which can lead models to rely on backgrounds. The key metric is *worst-group accuracy*, which captures spurious correlations by evaluating accuracy on rare class-background pairs (e.g., waterbirds on land), while average accuracy reflects overall performance. In Table 3, ProLoG outperforms baselines in worst-group accuracy, validating its robustness to spurious correlations.

4.5 Ablation Studies

Effect of each component. To see the effect of each component in ProLoG, we apply them one by one in the base-to-new setting. As shown in Table 4, using only the hybrid network (without masking) achieves the best base performance, but shows low performance on new classes, as the model only focuses on the base task during training. The proposed regularization enhances generalization to new classes by minimizing the distortion of the image-text alignment of the original CLIP, while slightly compromising performance on base classes. The masking strategy mitigates overfitting in LoRA modules, improving performance on both base and new classes. Finally, the inference strategy further boosts performance on new classes, resulting in the best HM score.

Effect of hyperparameters. We conduct an ablation study on key hyperparameters (δ_1, C_0, δ_2, and λ), to confirm the robustness to varying values. Fig. 3 shows performance variations with different δ_1 and C_0 on the average HM score on

Method	Average on 11 datasets		
	Base	New	HM
CoCoOp (Zhou et al. 2022a)	80.47	71.69	75.83
PromptSRC (Khattak et al. 2023b)	84.26	76.10	79.97
CLIP-LoRA (Zanella and Ben Ayed 2024)	84.98	72.58	78.29
CLIP	69.34	74.22	71.70
+ (1) Hybrid network (w/o masking)	**85.21**	73.57	78.96
+ (1) + (2) Regularization loss	84.52	76.18	80.13
+ (1) + (2) + (3) Masking strategy	84.77	76.32	80.33
+ (1) + (2) + (3) + (4) Inference strategy	84.77	**76.89**	**80.64**

Table 4: Effect of each component using ViT-B/16.

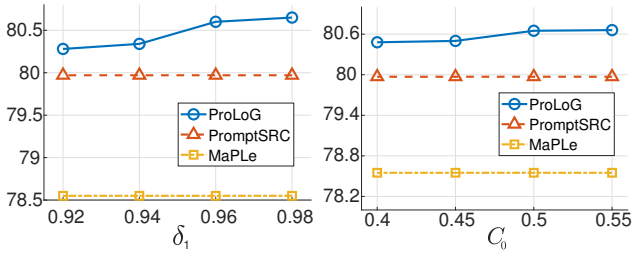


Figure 3: Ablation study on δ_1 (left) and C_0 (right).

ViT-B/16. ProLoG consistently surpasses baselines across a wide range of values. Additional results for other hyperparameters, such as δ_2 and λ, are provided in Appendix D.

5 Conclusion and Future Work

In this work, we propose ProLoG, a hybrid adaptation method combining prompt tuning and LoRA for OOD generalization. It integrates augmentation-based regularization that considers semantic relationships between images and texts, along with a task similarity-based inference scheme to improve generalization. Extensive experiments validate the effectiveness of ProLoG, offering a promising solution for downstream tasks requiring both adaptability and generalization. While we focus on image classification in this work, we consider extending our approach to tasks beyond classification to be a promising direction for future research.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. NRF RS-2024-00340966 and NRF RS-2024-00408003), and by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government(MSIT) (No. IITP RS-2024-00444862). This research was additionally supported by the Yonsei University Research Fund of 2025-22-0438. Dong-Jun Han is the corresponding author.

References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, 446–461. Springer.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. 2022a. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; and Qiao, Y. 2022b. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2818–2829.

Cho, E.; Kim, J.; and Kim, H. J. 2023. Distribution-aware prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22004–22013.

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

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, 178–178. IEEE.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Ganz, R.; and Elad, M. 2024. Clipag: Towards generator-free text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3843–3853.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, C.; Seto, S.; Abnar, S.; Grangier, D.; Jaitly, N.; and Susskind, J. 2024. Aggregate-and-Adapt Natural Language Prompts for Downstream Generalization of CLIP. *Advances in Neural Information Processing Systems*, 37: 81077–81104.

Jain, A.; Chaudhuri, S.; Reps, T.; and Jermaine, C. 2024. Prompt tuning strikes back: Customizing foundation models with low-rank prompt adaptation. *Advances in Neural Information Processing Systems*, 37: 47297–47316.

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

Kan, B.; Wang, T.; Lu, W.; Zhen, X.; Guan, W.; and Zheng, F. 2023. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15670–15680.

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

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

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, 5583–5594. PMLR.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7061–7070.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

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

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

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

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Roy, S.; and Etemad, A. 2024. Consistency-guided Prompt Learning for Vision-Language Models. In *The Twelfth International Conference on Learning Representations*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Shu, Y.; Guo, X.; Wu, J.; Wang, X.; Wang, J.; and Long, M. 2023. Clipood: Generalizing clip to out-of-distributions. In *International Conference on Machine Learning*, 31716–31731. PMLR.

Soomro, K. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Vapnik, V.; and Vapnik, V. 1998. Statistical learning theory Wiley. *New York*, 1(624): 2.

Vidit, V.; Engilberge, M.; and Salzmann, M. 2023. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3219–3229.

Wang, Z.; Liu, W.; He, Q.; Wu, X.; and Yi, Z. 2022. Clip-gen: Language-free training of a text-to-image generator with clip. *arXiv preprint arXiv:2203.00386*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2945–2954.

Yao, H.; Zhang, R.; and Xu, C. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6757–6767.

Yao, H.; Zhang, R.; and Xu, C. 2024. TCP: Textual-based Class-aware Prompt tuning for Visual-Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23438–23448.

Zanella, M.; and Ben Ayed, I. 2024. Low-Rank Few-Shot Adaptation of Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1593–1603.

Zang, Y.; Goh, H.; Susskind, J.; and Huang, C. 2024. Overcoming the pitfalls of vision-language model finetuning for OOD generalization. *arXiv preprint arXiv:2401.15914*.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18123–18133.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

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

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