# DiSa: Directional Saliency-Aware Prompt Learning for Generalizable Vision-Language Models

Niloufar Alipour Talemi
nalipou@clemson.edu
Clemson University
Clemson, SC, United States

Hossein Kashiani
hkashia@clemson.edu
Clemson University
Clemson, SC, United States

Hossein R. Nowdeh
hrajoli@clemson.edu
Clemson University
Clemson, SC, United States

Fatemeh Afghah
fafghah@clemson.edu
Clemson University
Clemson, SC, United States

## Abstract

Prompt learning has emerged as a powerful paradigm for adapting vision-language models such as CLIP to downstream tasks. However, existing methods often overfit to seen data, leading to significant performance degradation when generalizing to novel classes or unseen domains. To address this limitation, we propose DiSa, a Directional Saliency-Aware Prompt Learning framework that integrates two complementary regularization strategies to enhance generalization. First, our Cross-Interactive Regularization (CIR) fosters cross-modal alignment by enabling cooperative learning between prompted and frozen encoders. Within CIR, a saliency-aware masking strategy guides the image encoder to prioritize semantically critical image regions, reducing reliance on less informative patches. Second, we introduce a directional regularization strategy that aligns visual embeddings with class-wise prototype features in a directional manner to prioritize consistency in feature orientation over strict proximity. This approach ensures robust generalization by leveraging stable prototype directions derived from class-mean statistics. Extensive evaluations on 11 diverse image classification benchmarks demonstrate that DiSa consistently outperforms state-of-the-art prompt learning methods across various settings, including base-to-novel generalization, cross-dataset transfer, domain generalization, and few-shot learning.

## CCS Concepts

• **Computing methodologies** → **Computer vision**.

## Keywords

Vision-Language Models, Few-shot Learning, Prompt Learning, Domain Generalization

**ACM Reference Format:**
Niloufar Alipour Talemi, Hossein Kashiani, Hossein R. Nowdeh, and Fatemeh Afghah. 2025. DiSa: Directional Saliency-Aware Prompt Learning for Generalizable Vision-Language Models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3711896.3736911

This work is licensed under a Creative Commons Attribution 4.0 International License.
*KDD '25, Toronto, ON, Canada*
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3736911

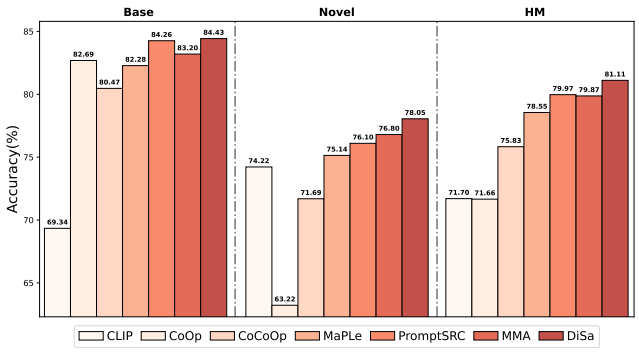


**Figure 1: Performance comparison of base-to-novel generalization: DiSa outperforms state-of-the-art methods across 11 diverse image recognition datasets for both base and novel classes.**

## 1 Introduction

Large-scale pre-trained Vision-Language (VL) models have become foundational tools across a broad range of downstream tasks including applications such as few-shot image recognition [12, 21], object detection [11], and image segmentation [9, 13]. These models, trained on enormous web-scale datasets, show exceptional generalization ability across diverse downstream tasks without task-specific tuning. Among these models, Contrastive Language-Image Pre-training (CLIP) [29] stands out as an innovative model, leveraging contrastive learning on massive image-text pairs from the internet. While CLIP excels in zero-shot recognition, fine-tuning it for downstream tasks remains challenging, particularly in the few-shot setting, where limited data can cause overfitting and compromise its generalization ability [46].

The adaptation of CLIP to downstream recognition tasks has been explored through various methods, with prompt engineering [29] being a straightforward yet labor-intensive approach. It involves crafting input queries, such as "a photo of a" or "a close-up

photo of a", to condition the model's text encoder for category-specific embeddings. However, this process demands extensive manual tuning and does not guarantee optimal prompts for target tasks [46]. As a more efficient alternative, prompt learning has been introduced to adapt CLIP to downstream tasks without modifying its pre-trained weights [3, 19, 30, 36, 37, 45, 46]. This method incorporates a small number of learnable prompt vectors, allowing the model to better align with downstream objectives. However, since these prompts are optimized based on task-specific objectives, the adapted model can overfit to the training distribution [45]. Consequently, while the model may achieve strong performance on seen classes, its performance on unseen classes or domains can degrade significantly.

Recently, some studies [20, 33] have introduced regularization approaches to jointly optimize for both task-specific objectives and task-agnostic general representations to mitigate prompt overfitting. These approaches primarily focus on maximizing score similarity between the prompted and frozen models. However, they often overlook the significant contributions of individual modalities when leveraging their cross-modal alignment. These approaches treat vision and text modalities in isolation, neglecting their inherent cross-modal relationships. This limitation reduces the effectiveness of cross-modal learning, as the model cannot fully leverage the complementary strengths of modalities. We argue that effective regularization should also preserve cross-modal alignment. In this work, we propose a Cross-Interactive Regularization (CIR) framework that fosters a deeper interaction between the modality-specific branches of each model. This CIR framework enables cooperative learning by facilitating the interaction of image embeddings from the prompted model with text embeddings from the frozen model, and vice versa. This structure encourages the prompted encoders to independently engage with the cross-modal representations from the pre-trained encoders, leveraging the generalization capabilities of the pre-trained VL model while adapting prompts for downstream tasks. As such, the CIR framework prevents modality-specific overfitting by anchoring the prompted encoders to the frozen encoder's cross-modal relationships.

Furthermore, training with few samples often results in the prompted image encoder developing a dependency on specific patches, which undermines its ability to generalize to unseen classes and data. To address this issue, we equip our CIR framework with a saliency-aware masking approach that guides the prompted image encoder to focus on critical patches under the guidance of the frozen model. More precisely, the masking process involves identifying less informative patches based on the guidance of the frozen model and then applying random masking within this subset of patches. This selective randomness ensures that the masked images retain the most salient regions. Within the CIR framework, this approach effectively distills representations derived from the full image into those predicted from the masked image, encouraging the prompted encoder to focus on salient regions rather than depending solely on global alignment. By leveraging this saliency-aware masking approach, the model's ability to generalize to novel classes and unseen data is significantly enhanced.

In addition to our proposed saliency-aware CIR framework, which introduces a novel score-based regularization, we incorporate a directional feature alignment constraint to ensure the prompted features remain aligned with the pre-trained encoder's embeddings. To achieve this, we employ two complementary strategies. First, unlike prior works [20, 44] that rely on sample-based alignment, we align the prompted feature with its corresponding class-wise prototype extracted from the frozen model. By using these prototypes, which represent broader class-level features, the prompted model can more effectively utilize the available data and consider diverse samples with various characteristics of the same class under the alignment constraint. Second, to maintain a balance between generalization and task adaptation, we enforce directional feature alignment rather than strict proximity. Experimental results confirm that aligning only the directional component of the prompted embeddings with the prototypes, instead of minimizing the absolute distance between embeddings that constrain both magnitude and direction, allows the model greater flexibility to adapt while preserving its generalization capabilities. This directional alignment, combined with robust prototypes, enhances the prompted model's ability to generalize to unseen classes and datasets. In summary, the main contributions of this work include:

- We introduce CIR, a novel regularization-based prompt learning framework that promotes interaction between the modality-specific branches of prompted and frozen models, enabling cooperative learning and enhancing cross-modal alignment. Additionally, CIR employs a saliency-aware masking strategy to preserve the pre-trained model's generalization capabilities.
- We introduce a novel directional regularization approach that aligns the prompted features with class-wise prototypes, represented as mean embeddings from the frozen model. By utilizing class-wise means as reliable prototypes and emphasizing feature direction alignment over strict proximity, our proposed method improves generalization by leveraging pre-trained model capabilities while avoiding limitations inherent in distance-based metrics.
- We conduct extensive evaluations on 11 popular image classification benchmarks. The results demonstrate the effectiveness of DiSa in all the base-to-novel generalization, cross-dataset transfer, domain generalization, and few-shot learning settings.

## 2 Related Work

### 2.1 Prompt Learning for Vision-Language Models

VL models [18, 29, 43] integrate both visual and textual information to produce rich multi-modal representations. These models are generally pre-trained on vast datasets; for example, CLIP [29] and ALIGN [18] are trained on approximately 400 million and 1 billion image-text pairs, respectively. By leveraging self-supervised learning, VL models are able to construct joint vision and language representations that greatly improve their ability to learn and generalize across different modalities. This capability enables them to excel in various tasks, particularly in few-shot [4, 24, 26] and zero-shot visual recognition [29]. However, one of the key challenges lies in adapting these large-scale pre-trained models to specialized downstream tasks without compromising their inherent generalization ability. Prompt learning has emerged as an effective technique, leveraging learnable embeddings, known as prompt tokens, that

are incorporated into model inputs [8, 24, 45]. This approach offers notable advantages, including parameter efficiency and rapid convergence, making it particularly appealing for fine-tuning foundational models like CLIP [29] across both vision and VL tasks. For instance, CoOp [46] introduced prompt learning for CLIP by optimizing continuous prompt vectors within the language branch for few-shot image recognition. Subsequently, MaPLe [19] extended this concept by proposing a multi-modal prompt-tuning framework that enhances transferability through the hierarchical joint learning of prompts across both vision and language branches. Although prompt learning has seen remarkable advancements, overfitting remains a significant challenge, limiting its potential for new classes and domains. This work addresses this issue by introducing a novel regularization-based prompt learning approach that preserves the generalization capabilities of the pre-trained model while enabling effective adaptation to downstream tasks.

## 2.2 Regularization for Prompt Learning

Regularization techniques play a critical role in mitigating overfitting and enhancing the ability of models to generalize effectively to unseen data. These methods can be categorized into two primary groups. The first group, known as constraint-based techniques, incorporates additional restrictions into the training process, such as weight decay [23] and adversarial training [1]. The second group, termed input and parameter modification strategies, includes methods like dropout [35], model ensembling [17], label smoothing [17], and data augmentation [5]. In the specific context of prompt learning for vision-language models, overfitting is one of the most challenging problems. Some recent works have explored self-regularization strategies that consider constraints to preserve the generalization potential of pre-trained models while fine-tuning them for downstream tasks. For instance, CoPrompt [33] introduces a consistency constraint that aligns predictions between the prompted and pre-trained models, thus mitigating overfitting in downstream tasks. Similarly, PromptSRC [20] integrates feature-based regularization alongside the prediction-based constraint to enhance model robustness. However, these approaches often treat vision and text encoders separately, limiting cross-modal interactions and overlooking the individual contributions of each modality. To address this, we propose a novel prompt learning framework that enhances generalization by integrating cross-interactive regularization for improved cross-modal learning and directional regularization for embedding alignment with class-wise prototypes.

# 3 Method

## 3.1 Preliminaries

**Contrastive Language-Image Pre-training (CLIP).** The CLIP model, as described in [29], is a pre-trained system designed to align visual and textual representations by leveraging large-scale image-text datasets from the web. It is composed of two primary components: an image encoder $f(.)$ and a text encoder $g(.)$. The pre-trained model's parameters are expressed as $\theta_{CLIP} = \theta_f, \theta_g$, where $\theta_f$ represents the parameters of the image encoder, and $\theta_g$ represents those of the text encoder. For any input image $x$, the feature extraction process starts by partitioning the image into $V$ fixed-size patches, which are subsequently projected into the patch feature space. A learnable class token, represented as $CLS$, is then appended to these features, forming the sequence $X = \{CLS, e_1, e_2, \ldots, e_V\}$. This sequence is passed through $L$ layers of a transformer network, producing a visual feature representation $\mathbf{f} \in \mathbb{R}^d$. For the corresponding class label $y$, it is embedded within a text template, such as "a photo of a class name", represented as $Y = \{t_{SOS}, t_1, t_2, ..., t_T, c_k, t_{EOS}\}$, where $c_k$ is the word embedding for the class label, and $t_{SOS}$ and $t_{EOS}$ are learnable start and end tokens, respectively. Similarly to the image encoder, these text embeddings are processed through multiple transformer layers to generate the latent text feature $\mathbf{g} \in \mathbb{R}^d$. For zero-shot inference, text features of the text template with class labels $\{1, 2, ..., N_c\}$ are matched with the image feature as $\frac{\exp(\text{sim}(\mathbf{g}^i, \mathbf{f})/\tau)}{\sum_{i=1}^{N_c} \exp(\text{sim}(\mathbf{g}^i, \mathbf{f})/\tau)}$, where sim() represents cosine similarity, and $\tau$ is the temperature parameter of the softmax function.

**Prompt Learning for CLIP.** Inspired by prompt learning in natural language processing, numerous studies have explored the VL models by incorporating learnable prompt tokens during end-to-end training. In this study, we utilize hierarchical learnable prompt tokens independently for the text and image encoders, following the simple baseline method called Independent Vision-Language Prompting (IVLP) [31]. We concatenate learnable language prompts, denoted as $P_t = \{p_t^1, p_t^2, \cdots, p_t^T\}$, and visual prompts, denoted as $P_v = \{p_v^1, p_v^2, \cdots, p_v^V\}$, into the respective sets of textual and visual input tokens. Thus, the image encoder generates the prompted visual feature $\mathbf{f}_p = f(X_p, \theta_f)$ from input tokens $X_p = \{P_v, e_{CLS}, e_1, e_2, \cdots, e_V\}$, and the textual feature $\mathbf{g}_p = g(Y_p, \theta_g)$ is obtained from $Y_p = \{t_{SOS}, P_t, t_1, t_2, \cdots, t_T, c_k, t_{EOS}\}$. It should be noted that in this work, we employ deep prompting, which involves learning distinct sets of prompts for each transformer layer. For image classification on a downstream dataset $D_s$, consisting of an $K$-shot $N_c$-class training set, prompts interact with frozen $\theta_f$ and $\theta_g$ and are optimized with the cross-entropy loss as:

$$\mathcal{L}_{CE} = \frac{1}{KN_c} \sum_{(x,y)\in\mathcal{D}_s} \frac{\exp\Big(\text{sim}(\mathbf{f}_p, \mathbf{g}_p^y)/\tau\Big)}{\sum_{i=1}^{N_c} \exp\Big(\text{sim}(\mathbf{f}_p, \mathbf{g}_p^i)/\tau\Big)}, \tag{1}$$

where $\mathbf{f}_p$ and $\mathbf{g}_p \in \mathbb{R}^d$ represent the vision and text embeddings derived from the prompted model, respectively.

## 3.2 Directional Saliency-Aware Prompt Learning

As illustrated in Fig. 2, our proposed framework incorporates two complementary regularization approaches: saliency-aware cross-interactive regularization and directional regularization. The subsequent subsections provide a detailed explanation of each.

*3.2.1 Saliency-aware Cross-Interactive Regularization.* Training prompts solely with a supervised loss specific to the task tends to undermine the general features encapsulated within the frozen CLIP model. As a result, although the weights of the CLIP image and text encoders remain unchanged, their effectiveness on unseen domains diminishes. To address this challenge, we introduce a cross-interactive regularization framework that forces prompted

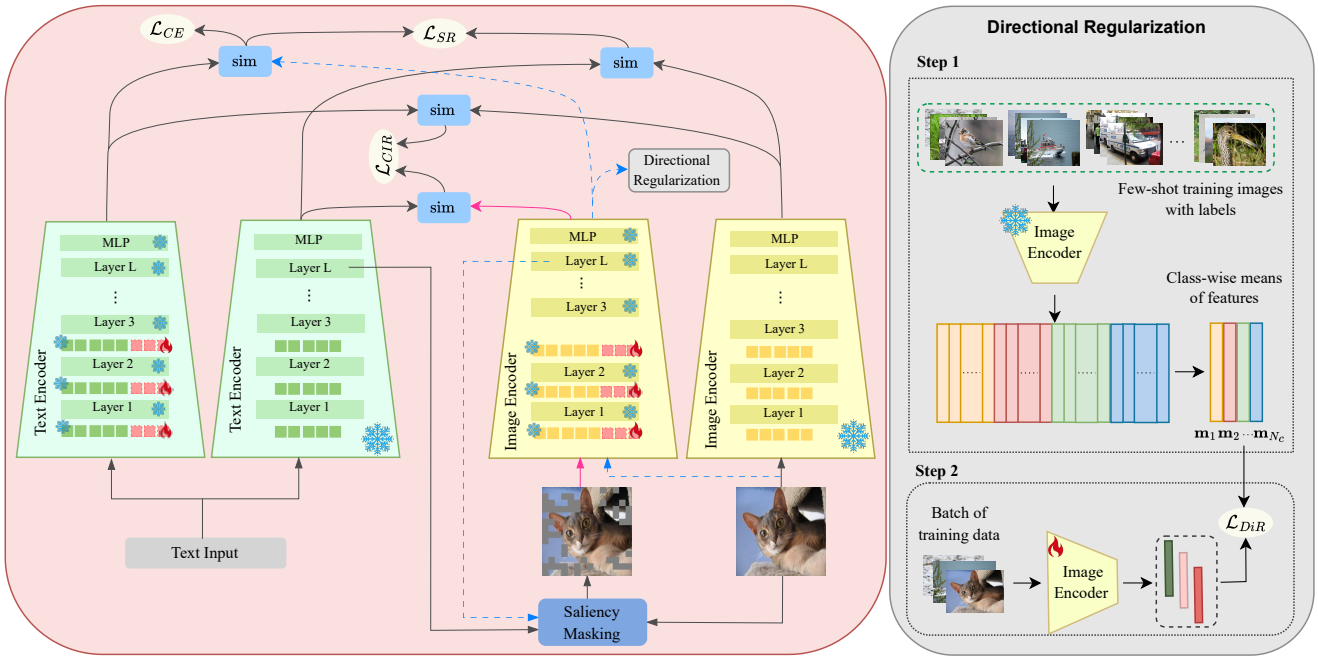


**Figure 2: Overview of the proposed DiSa. The DiSa employs two complementary regularization approaches: saliency-aware cross-interactive regularization and directional regularization. The saliency-aware cross-interactive framework ensures that the prompted encoders establish independent, saliency-based interactions with the cross-modality outputs of the frozen encoders. Meanwhile, the directional regularization aligns prompted features with classification goals using class-wise feature means from the frozen model as robust prototypes. Note that the saliency masking component evaluates the importance of image tokens by computing attention scores between image patch tokens and the *CLS* token from the frozen model's text encoder. For visual clarity, it should also be noted that we employ distinct arrow styles to illustrate different data flows: dashed blue arrows indicate the path of the full (unmasked) image through the prompted vision encoder; solid pink arrows represent the saliency-masked image path; and solid black arrows denote flows without multiple input dependencies.**

encoders to independently interact with the cross-modality from the pre-trained encoders. This guidance from the cross-modality helps the model to preserve its inherent generalization while adapting prompts to downstream tasks. Additionally, benefiting from its large-scale training data, the frozen image encoder can identify and prioritize key patches for decision-making. In contrast, the prompted encoder, constrained by limited training data, risk over-relying on specific patches, resulting in significant performance degradation when encountering unseen classes or domains. To mitigate this issue, we enforce the prompted image encoder to focus on critical patches under the guidance of the frozen model. This saliency-aware regularization enhances the model's ability to generalize effectively across novel classes and diverse domains.

**Cross-Interactive Regularization (CIR).** Our proposed CIR introduces an innovative method to regularize the prompted model with the frozen CLIP model by leveraging interactive learning across both visual and textual encoders. Unlike recent regularization-based methods that focus primarily on maximizing score-based similarity between the prompted and frozen models [33], CIR fosters a deeper interaction between the modality-specific branches of each model. This cross-interaction mechanism in CIR enables cooperative learning between the two models' modality branches, where the image embeddings from the prompted model interact with the text embeddings from the frozen model, and the text embeddings from the prompted model interact with the image embeddings of the frozen model. This structure is designed to optimize the mutual information shared across models, allowing the prompted model to capture cross-modal dependencies more effectively. For this purpose, CIR employs Kullback-Leibler (KL) divergence to align two critical probability scores. The first score, $q^{\mathbf{f}_p\mathbf{g}_o}$, is computed as the cosine similarity between the image embeddings from the prompted model and text embeddings from the frozen model. The second score, $q^{\mathbf{f}_o\mathbf{g}_p}$, is derived from the cosine similarity between the text embeddings from the prompted model and image embeddings from the frozen model. We define the CIR loss as:

$$\mathcal{L}_{CIR} = \mathcal{D}_{\mathrm{KL}}(q^{\mathbf{f}_p\mathbf{g}_o}, q^{\mathbf{f}_o\mathbf{g}_p}), \tag{2}$$

$$q^{\mathbf{f}_p\mathbf{g}_o} = \mathrm{sim}(\mathbf{f}_p, \mathbf{g}_o),\ q^{\mathbf{f}_o\mathbf{g}_p} = \mathrm{sim}(\mathbf{f}_o, \mathbf{g}_p), \tag{3}$$

where the function sim represents cosine similarity.

Furthermore, to ensure consistency between the prompted text and image encoders, we align the score predictions of the prompted model ($q^{\mathbf{f}_p\mathbf{g}_p}$) with those of the frozen CLIP model ($q^{\mathbf{f}_o\mathbf{g}_o}$) using the original image as follows:

$$\mathcal{L}_{SR} = \mathcal{D}_{\mathrm{KL}}(q^{\mathbf{f}_p\mathbf{g}_p}, q^{\mathbf{f}_o\mathbf{g}_o}), \tag{4}$$

$$q^{\mathbf{f}_p\mathbf{g}_p} = \mathrm{sim}(\mathbf{f}_p, \mathbf{g}_p),\ q^{\mathbf{f}_o\mathbf{g}_o} = \mathrm{sim}(\mathbf{f}_o, \mathbf{g}_o). \tag{5}$$

This integrated alignment approach enables a more sophisticated cross-modal interaction, ultimately enhancing the generalizability of the prompted model across both visual and textual modalities.

**Saliency Masking.** To ensure the prompted image encoder focuses on critical patches rather than over-relying on specific regions, we leverage the frozen model's guidance to randomly mask less important patches, providing the masked image as input to the prompted model. The *CLS* token from the text encoder captures the semantic content of the text, and its attention weights on image tokens serve as an effective indicator of each token's relevance to the linguistic semantics. Therefore, to identify the importance of different image tokens, we compute the attention value between the image patch tokens and the *CLS* token from the frozen model's text encoder. By masking random, less important patches in the input image, the prompted model is encouraged to produce embeddings aligned with the frozen model, which has access to the full image. This approach encourages the model to focus on essential image regions, enhancing its ability to generalize by reducing reliance on less informative features. As a result, the model becomes more adaptable to varying data distributions and better equipped to handle challenging scenarios involving domain shifts or novel classes.

Considering the final layer of the prompted model, let $z(n) \in \mathbb{R}^d$ denote the embedding of the $n$-th image patch from the image encoder, and let $z(CLS) \in \mathbb{R}^d$ represent the embedding of the *CLS* token from the text encoder of the frozen model. The importance of each token is computed as:

$$\alpha_n = \frac{1}{H} \sum_{h=1}^{H} \text{Softmax}\left( \frac{z_h^Q(\text{CLS}) \cdot z_h^K(n)}{\sqrt{C}} \right), \tag{6}$$

where $h$ indicates the attention head index; $z_h^Q(\text{CLS})$ denotes the query embedding of the *CLS* token at head $h$; $z_h^K(n)$ is the key embedding of the $n$-th image token at head $h$, and $C$ is the dimensionality of the query and key embeddings. Image tokens to mask are selected based on the scores $\alpha_n$. We randomly mask $\gamma$ percent of the image tokens with the lowest scores.

*3.2.2 Directional Regularization.* To complement our proposed saliency-aware CIR framework, we incorporate a directional feature alignment constraint to ensure the prompted features remain aligned with the pre-trained model's embeddings. This is achieved through two synergistic approaches. In contrast to previous methods [20, 44] that rely on sample-based alignment, we align the prompted feature with the corresponding prototype derived from the frozen model. These prototypes capture broader class-level features, enabling the prompted model to leverage diverse samples more effectively under the alignment constraint. To maintain generalization while enhancing adaptability, our approach aligns feature directions instead of enforcing strict proximity. Our findings indicate that aligning only the directional component of prompted embeddings with prototypes, rather than constraining both magnitude and direction, enables greater flexibility for adaptation. Thus, we employ a cosine similarity-based loss to align the direction of the prompted feature $\mathbf{f}_p$ with its corresponding class-wise prototype $\mathbf{m}_i$. Accordingly, the directional regularization loss is defined as:

$$\mathcal{L}_{DiR} = \left| 1 - \cos(\mathbf{f}_p, \mathbf{m}_i) \right|, \tag{7}$$

where $\mathbf{m}_i$ represents the class-mean embedding for the $i^{\text{th}}$ class, calculated as:

$$\mathbf{m}_i = \frac{1}{|I_j|} \sum_{j \in I_j} \mathbf{f}_{o_j}, \tag{8}$$

with $I_j$ denotes the set of indices corresponding to training examples from class $i$. Finally, we employ a weighted combination of all specified loss terms for end-to-end training. Consequently, our final loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{SR} + \mathcal{L}_{CIR} + \lambda \mathcal{L}_{DiR}, \tag{9}$$

where $\lambda$ controls the contribution of the directional regularization loss to the total loss function.

# 4 Experiments

## 4.1 Datasets and Implementation Details

Consistent with prior research on prompt tuning [19, 31, 45], we assess our proposed method across four evaluation scenarios: generalization from base-to-novel classes, cross-dataset evaluation, domain generalization, and few-shot image recognition. Our experiments incorporate a variety of datasets, including two generic-object datasets (ImageNet [7] and Caltech101 [10]), fine-grained datasets (OxfordPets [28], StanfordCars [22], Flowers102 [27], Food101 [2], and FGVCAircraft[25]), a remote sensing classification dataset (EuroSAT[14]), a scene recognition dataset (SUN397 [39]), an action recognition dataset (UCF101[34]), and a texture dataset (DTD[6]). For domain generalization, we employ ImageNet as the source dataset, and its four variants as target datasets, including ImageNetV2 [32], ImageNetSketch [38], ImageNet-A[16], and ImageNet-R [15].

In all our experiments, to align the comparisons with previous approaches [19, 20], we utilize a ViT-B/16-based CLIP model and employ deep prompting with $V = T = 4$ VL prompts. It should be noted that the prompts are randomly initialized using a normal distribution, except for the text prompts in the first layer, which are initialized with the word embeddings of "a photo of a". For domain generalization and cross-dataset evaluation, learnable prompts are injected into the first three transformer layers. For the few-shot and base-to-novel settings, prompts are injected into the first nine transformer layers. Each model is trained with a batch size of 4 and a learning rate of 0.0025 using the SGD optimizer on a single NVIDIA RTX 4090 GPU. Training is conducted for 50 epochs in the few-shot setting, while for all other settings, the model is trained for 20 epochs under a 16-shot configuration, where only 16 training samples per category are provided. The results are averaged over three runs. The parameter $\lambda$ is set to 12. For masking less informative patches, we identify the 50% of patches ($\gamma = 50\%$) with the lowest scores (calculated as per Eq. 6) and randomly mask half of these less important patches.

## 4.2 Base-to-Novel Generalization

We evaluate the generalization ability of our proposed approach across 11 different image classification datasets. Following the experimental setup outlined in prior studies [19, 24, 45], each dataset

**Table 1: Comparison of different methods in 16-shot base-to-novel generalization. We report the accuracy (%) on both base and novel classes, as well as their harmonic mean. The best results are presented in bold.**

**(a) Average over 11 datasets**

| | Base | New | HM |
|---|---|---|---|
| CLIP [29] | 69.34 | 74.22 | 71.70 |
| CoOp [46] | 82.69 | 63.22 | 71.66 |
| CoCoOp [45] | 80.47 | 71.69 | 75.83 |
| Maple [19] | 82.28 | 75.14 | 78.55 |
| PromptSRC [20] | 84.26 | 76.10 | 79.97 |
| CoPrompt [33] | 84.00 | 77.23 | 80.48 |
| MMA [40] | 83.20 | 76.80 | 79.87 |
| APEX [41] | 83.99 | 76.76 | 80.04 |
| TCP [42] | 84.13 | 75.36 | 79.51 |
| DiSa (Ours) | **84.43** | **78.05** | **81.11** |

**(b) ImageNet**

| | Base | New | HM |
|---|---|---|---|
| CLIP[29] | 72.43 | 68.14 | 70.22 |
| CoOp[46] | 76.47 | 67.88 | 71.92 |
| CoCoOp[45] | 75.98 | 70.43 | 73.10 |
| Maple [19] | 76.66 | 70.54 | 73.47 |
| PromptSRC [20] | 77.60 | 70.73 | 74.01 |
| CoPrompt [33] | **77.67** | 71.27 | 74.33 |
| MMA [40] | 77.31 | 71.00 | 74.02 |
| APEX [41] | 77.12 | 71.10 | 73.99 |
| TCP [42] | 77.27 | 69.87 | 73.38 |
| DiSa (Ours) | 77.56 | **71.65** | **74.49** |

**(c) Caltech101**

| | Base | New | HM |
|---|---|---|---|
| CLIP[29] | 96.84 | 94.00 | 95.40 |
| CoOp[46] | 98.00 | 89.81 | 93.73 |
| CoCoOp [45] | 97.96 | 93.81 | 95.84 |
| Maple [19] | 97.74 | 94.36 | 96.02 |
| PromptSRC [20] | 98.10 | 94.03 | 96.02 |
| CoPrompt [33] | 98.27 | 94.90 | 96.55 |
| MMA [40] | **98.40** | 94.00 | 96.15 |
| APEX [41] | 98.18 | 95.06 | 96.59 |
| TCP [42] | 98.23 | 94.67 | 96.42 |
| DiSa (Ours) | 98.29 | **95.41** | **96.83** |

**(d) OxfordPets**

| | Base | New | HM |
|---|---|---|---|
| CLIP[29] | 91.17 | 97.26 | 94.12 |
| CoOp[46] | 93.67 | 95.29 | 94.47 |
| CoCoOp[45] | 95.20 | 97.69 | 96.43 |
| Maple [19] | 95.43 | 97.76 | 96.58 |
| PromptSRC [20] | 95.33 | 97.30 | 96.30 |
| CoPrompt [33] | **95.67** | 98.10 | 96.87 |
| MMA [40] | 95.40 | 98.07 | 96.72 |
| APEX [41] | 95.11 | 97.27 | 96.18 |
| TCP [42] | 94.67 | 97.20 | 95.92 |
| DiSa (Ours) | 95.48 | **98.67** | **97.05** |

**(e) StanfordCars**

| | Base | New | HM |
|---|---|---|---|
| CLIP[29] | 63.37 | 74.89 | 68.65 |
| CoOp[46] | 78.12 | 60.40 | 68.13 |
| CoCoOp[45] | 70.49 | 73.59 | 72.01 |
| Maple [19] | 72.94 | 74.00 | 73.47 |
| PromptSRC [20] | 78.27 | 74.97 | 76.58 |
| CoPrompt [33] | 76.97 | 74.40 | 75.66 |
| MMA [40] | 78.50 | 73.10 | 75.70 |
| APEX [41] | 80.53 | **75.08** | **77.71** |
| TCP [42] | **80.80** | 74.13 | 77.32 |
| DiSa (Ours) | 78.54 | 75.07 | 76.77 |

**(f) Flowers102**

| | Base | New | HM |
|---|---|---|---|
| CLIP[29] | 72.08 | 77.80 | 74.83 |
| CoOp[46] | 97.60 | 59.67 | 74.06 |
| CoCoOp[45] | 94.87 | 71.75 | 81.71 |
| Maple [19] | 95.92 | 72.46 | 82.56 |
| PromptSRC [20] | 98.07 | 76.50 | 85.95 |
| CoPrompt [33] | 97.27 | 76.60 | 85.71 |
| MMA [40] | 97.77 | 75.93 | 85.48 |
| APEX [41] | 97.47 | **77.58** | **86.40** |
| TCP [42] | 97.73 | 75.57 | 85.23 |
| DiSa (Ours) | **98.14** | 76.77 | 86.15 |

**(g) Food101**

| | Base | New | HM |
|---|---|---|---|
| CLIP[29] | 92.43 | 91.22 | 90.66 |
| CoOp[46] | 88.33 | 82.26 | 85.19 |
| CoCoOp[45] | 90.70 | 91.29 | 90.99 |
| Maple [19] | 90.71 | 92.05 | 91.38 |
| PromptSRC [20] | 90.67 | 91.53 | 91.10 |
| CoPrompt [33] | 90.73 | 92.07 | 91.40 |
| MMA [40] | 90.13 | 91.30 | 90.71 |
| APEX [41] | 89.60 | 92.06 | 90.81 |
| TCP [42] | 90.57 | 91.37 | 90.97 |
| DiSa (Ours) | **90.81** | **92.32** | **91.56** |

**(h) FGVCAircraft**

| | Base | New | HM |
|---|---|---|---|
| CLIP[29] | 27.19 | 36.29 | 31.09 |
| CoOp[46] | 40.44 | 22.30 | 28.75 |
| CoCoOp[45] | 33.41 | 23.71 | 27.74 |
| Maple [19] | 37.44 | 35.61 | 36.50 |
| PromptSRC [20] | **42.73** | 37.87 | 40.15 |
| CoPrompt [33] | 40.20 | 39.33 | 39.76 |
| MMA [40] | 40.57 | 36.33 | 38.33 |
| APEX [41] | 42.69 | 35.21 | 38.59 |
| TCP [42] | 41.97 | 34.43 | 37.83 |
| DiSa (Ours) | 42.65 | **39.38** | **40.95** |

**(i) SUN397**

| | Base | New | HM |
|---|---|---|---|
| CLIP[29] | 69.36 | 75.35 | 72.23 |
| CoOp [46] | 80.60 | 65.89 | 72.51 |
| CoCoOp[45] | 79.74 | 76.86 | 78.27 |
| Maple [19] | 80.82 | 78.70 | 79.75 |
| PromptSRC [20] | 82.67 | 78.47 | 80.52 |
| CoPrompt [33] | 82.63 | 80.03 | 81.31 |
| MMA [40] | 82.27 | 78.57 | 80.38 |
| APEX [41] | 81.17 | 78.98 | 80.06 |
| TCP [42] | 82.63 | 78.20 | 80.35 |
| DiSa (Ours) | **82.69** | **80.53** | **81.60** |

**(j) DTD**

| | Base | New | HM |
|---|---|---|---|
| CLIP[29] | 53.24 | 59.90 | 56.37 |
| CoOp[46] | 79.44 | 41.18 | 54.24 |
| CoCoOp[45] | 77.01 | 56.00 | 64.85 |
| Maple [19] | 80.36 | 59.18 | 68.16 |
| PromptSRC [20] | **83.37** | 62.97 | 71.75 |
| CoPrompt [33] | 83.13 | 64.73 | 72.79 |
| MMA [40] | 83.20 | 65.63 | 73.38 |
| APEX [41] | 82.45 | 63.80 | 71.94 |
| TCP [42] | 82.77 | 58.07 | 68.25 |
| DiSa (Ours) | 83.33 | **65.71** | **73.49** |

**(k) EuroSAT**

| | Base | New | HM |
|---|---|---|---|
| CLIP[29] | 56.48 | 64.05 | 60.03 |
| CoOp[46] | 92.19 | 54.74 | 68.69 |
| CoCoOp [45] | 87.49 | 60.04 | 71.21 |
| Maple [19] | 94.07 | 73.23 | 82.35 |
| PromptSRC [20] | 92.90 | 73.90 | 82.32 |
| CoPrompt [33] | **94.60** | 78.57 | 85.84 |
| MMA [40] | 85.46 | 82.34 | 83.87 |
| APEX [41] | 92.83 | 79.89 | 85.85 |
| TCP [42] | 91.63 | 74.73 | 82.32 |
| DiSa (Ours) | 94.10 | **82.69** | **88.03** |

**(l) UCF101**

| | Base | New | HM |
|---|---|---|---|
| CLIP [29] | 70.53 | 77.50 | 73.85 |
| CoOp [46] | 84.69 | 56.05 | 67.46 |
| CoCoOp[45] | 82.33 | 73.45 | 77.64 |
| Maple [19] | 83.00 | 78.66 | 80.77 |
| PromptSRC [20] | 87.10 | 78.80 | 82.74 |
| CoPrompt [33] | 86.90 | 79.57 | 83.07 |
| MMA [40] | 86.23 | 80.03 | 82.20 |
| APEX [41] | 86.74 | 78.37 | 82.34 |
| TCP [42] | 87.13 | **80.77** | **83.83** |
| DiSa (Ours) | **87.16** | 80.45 | 83.67 |

is divided into base and novel classes. The model is trained on the base classes using a few-shot learning strategy with 16 shots and is later tested on both the base and novel classes. Performance results, detailed in Table 1, compare our proposed method with state-of-the-art methods, including zero-shot CLIP [29], CoOp [46], CoCoOp [45], MaPLe [19], CoPrompt [33], PromptSRC [20], and MMA [40], across the 11 datasets. The proposed DiSa framework delivers remarkable advancements over existing methods, with an average improvement of 0.82% in the demanding novel-class setting. Beyond its core goal of advancing generalization to novel scenarios, DiSa demonstrates a 0.17% average performance gain over PromptSRC for base classes, illustrating its superior adaptability while maintaining strong generalization capabilities.

## 4.3 Cross-Dataset Evaluation

We further consider a more challenging setting to assess the generalization capability of our method across different datasets. Similar to the state-of-the-art methods [24, 33, 45], we train the model on

**Table 2: Comparison of our proposed method with state-of-the-art approaches in cross-dataset evaluation. Our method achieves superior average performance across 10 datasets, highlighting its strong zero-shot adaptability.**

| | Source | Target | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ImageNet | Caltech101 | OxfordPets | StanfordCars | Flowers102 | Food101 | Aircraft | SUN397 | DTD | EuroSAT | UCF101 | Average |
| **CoOp** [46] | 71.51 | 93.70 | 89.14 | 64.51 | 68.71 | 85.30 | 18.47 | 64.15 | 41.92 | 46.39 | 66.55 | 63.88 |
| **CoCoOp** [45] | 71.02 | 94.43 | 90.14 | 65.32 | 71.88 | 86.06 | 22.94 | 67.36 | 45.73 | 45.37 | 68.21 | 65.74 |
| **MaPLe** [19] | 70.72 | 93.53 | 90.49 | 65.57 | 72.23 | 86.20 | 24.74 | 67.01 | 46.49 | 48.06 | 68.69 | 66.30 |
| **PromtSCR** [20] | 71.27 | 93.60 | 90.25 | 65.70 | 70.25 | 86.15 | 23.90 | 67.10 | 46.87 | 45.50 | 68.75 | 65.81 |
| **CoPrompt** [33] | 70.80 | 94.50 | 90.73 | 65.67 | 72.30 | 86.43 | 24.00 | 67.57 | 47.07 | **51.90** | 69.73 | 67.00 |
| **MMA** [40] | 71.00 | 93.80 | 90.30 | 66.13 | 72.07 | 86.12 | **25.33** | 68.17 | 46.57 | 49.24 | 68.32 | 66.61 |
| **APEX** [41] | **72.00** | 94.46 | 90.06 | 65.46 | 71.58 | 86.44 | 24.44 | 67.20 | 45.70 | 47.58 | 68.80 | 66.16 |
| **TCP** [42] | 71.40 | 93.97 | **91.25** | 64.69 | 71.21 | **86.69** | 23.45 | 67.15 | 44.35 | 51.45 | 68.73 | 66.29 |
| **DiSa** | 71.21 | **94.62** | 90.94 | **66.22** | **72.51** | 86.64 | 25.26 | **68.32** | **47.23** | 50.84 | **69.93** | **67.25** |

**Table 3: Comparison of our proposed method with state-of-the-art studies in the domain generalization setting.**

| | Source | Target | | | | |
|---|---|---|---|---|---|---|
| | ImageNet | -V2 | -S | -A | -R | Avg. |
| **CLIP** [29] | 66.73 | 60.83 | 46.15 | 47.77 | 73.96 | 57.18 |
| **CoOp** [46] | 71.51 | 64.20 | 47.99 | 49.71 | 75.21 | 59.28 |
| **CoCoOp** [45] | 71.02 | 64.07 | 48.75 | 50.63 | 76.18 | 59.91 |
| **MaPLe** [19] | 70.72 | 64.07 | 49.15 | 50.90 | 76.98 | 60.27 |
| **PromptSRC** [20] | 71.27 | 64.35 | 49.55 | 50.90 | 77.80 | 60.65 |
| **CoPrompt** [33] | 70.80 | 64.81 | 49.54 | 51.51 | 77.34 | 60.80 |
| **MMA** [40] | 71.00 | 64.33 | 49.13 | 51.12 | 77.32 | 60.77 |
| **APEX** [41] | **72.00** | 64.70 | 48.48 | 50.68 | 76.76 | 60.16 |
| **TCP** [42] | 71.20 | 64.60 | 49.50 | 51.20 | 76.73 | 60.51 |
| **DiSa** | 71.21 | **65.60** | **50.29** | **51.90** | **77.94** | **61.43** |

ImageNet [7] and directly evaluate it on other datasets with all learnable parameters frozen. Table 2 reports the results of DiSa and other methods on out-of-distribution datasets. DiSa achieves the highest average performance among all methods, with an average improvement of 0.25%, showcasing the effectiveness of our CIR framework in enabling adaptability across diverse datasets.

## 4.4 Domain Generalization

In this evaluation setting, consistent with earlier studies [24, 33, 33, 40, 45], we assess the transferability of the ImageNet-trained model to several ImageNet variant datasets. As illustrated in Table 3, our proposed method outperforms existing approaches, achieving a top average accuracy of 61.43%. These findings highlight the impact of cross-interactive regularization and saliency guidance in enabling DiSa to generalize effectively across out-of-distribution domains.

## 4.5 Few-shot Experiments

We further extend our experiments to the few-shot learning benchmark to assess the impact of our regularization framework on the ability of prompts to acquire task-specific knowledge. Figure 3 presents a comparative analysis of our proposed method against existing approaches across various K-shot settings (K = 1, 2, 4, 8, 16). As observed, our method consistently outperforms existing methods in terms of average performance, achieving gains of 1.18%, 0.95%, 0.80%, 0.45%, and 0.36% for 1, 2, 4, 8, and 16 shots across 11 datasets. Furthermore, our approach exhibits relatively larger performance gains in low-shot scenarios, particularly for K=1, 2, and 4 across almost all datasets. This indicates that DiSa effectively mitigates overfitting while allowing the prompts to capture task-specific knowledge.

**Table 4: Analysis of the effectiveness of each component in DiSa. $\mathcal{L}_{DiR}$ with *Sample* is included for comparison only.**

| Approach | | | | | Accuracy | | |
|---|---|---|---|---|---|---|---|
| $\mathcal{L}_{CIR}$ | *Masking* | $\mathcal{L}_{SR}$ | $\mathcal{L}_{DiR}$ *Sample* | $\mathcal{L}_{DiR}$ *Prototype* | Base | Novel | HM |
| | | | | | 84.21 | 71.79 | 77.51 |
| ✓ | | | | | 84.35 | 74.23 | 78.97 |
| ✓ | ✓ | | | | 84.31 | 74.86 | 79.30 |
| ✓ | ✓ | ✓ | | | 84.27 | 76.53 | 80.21 |
| ✓ | ✓ | ✓ | ✓ | | 84.25 | 77.09 | 80.51 |
| ✓ | ✓ | ✓ | | ✓ | **84.43** | **78.05** | **81.11** |

## 4.6 Ablation and Analysis

**Effectiveness of regularization constraints.** To clarify the impact of our various regularization constraints, we perform a series of experiments. The first row in Table 4 presents the performance of the baseline, which corresponds to the IVLP prompt learning approach. In the second through fourth rows, we apply the cross-interactive regularization, cross-interactive regularization with masking strategies, and score-based regularizations, respectively. The final two rows assess the contribution of the proposed directional regularization. Specifically, in the penultimate row, we align the feature direction of the prompted model with that of the frozen model derived from a single sample. In contrast, the final row aligns the prompted features with class-wise prototypes. Comparing these two settings highlights the considerable contribution of employing class-wise prototypes, which effectively adapt the model for both base and novel classes. The results clearly show that each regularization constraint plays a vital role in enhancing the model's generalization, leading to substantial improvements for novel classes. Furthermore, while the primary goal is to boost generalization capabilities, the integration of CIR and robust prototypes with directional alignment also delivers a measurable 0.22% improvement for base classes over the baseline, further demonstrating the overall robustness of the proposed approach.

**Analysis of the saliency-masking approach.** The results in Table 4 demonstrate that integrating the proposed saliency-masking approach with the CIR framework significantly enhances model

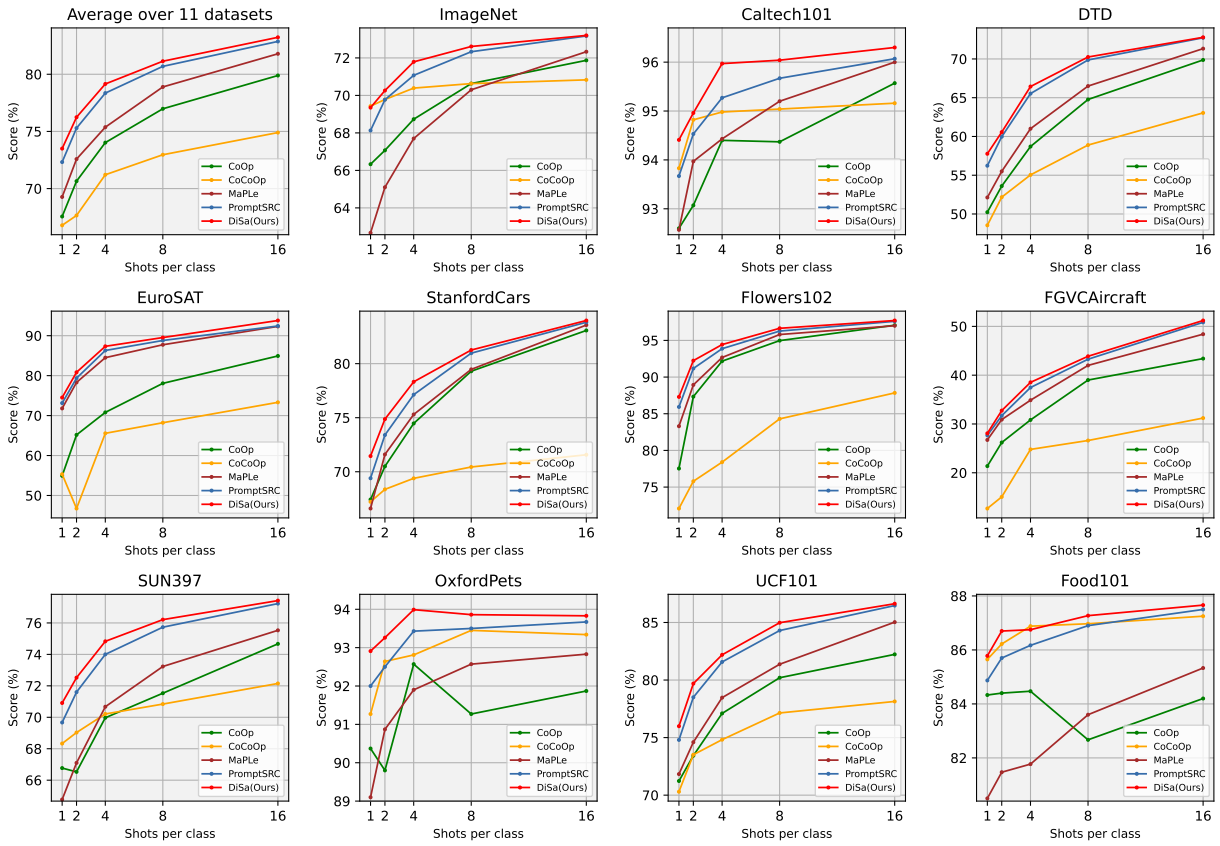


**Figure 3: Performance comparison across K-shot settings (K = 1, 2, 4, 8, 16). Our approach consistently achieves superior average performance, with notable gains in low-shot scenarios, particularly for K=1, 2, 4.**

performance. This integration achieves over a 0.63% improvement for novel classes, which pose the greatest challenge. In this section, we provide a detailed analysis of the proposed saliency-masking approach and explore the impact of its hyper-parameters. To this end, we conduct two sets of experiments to evaluate the impact of masking on model performance. In the first set (Fig. 4 (a)), we investigate the effect of masking different percentages of the input image by selectively masking the least informative patches. The results indicate a slight improvement for novel classes when 25% to 35% of the patches are masked. However, masking less than 25% yields negligible changes in performance, while masking more than 35% leads to performance degradation for both base and novel classes. Therefore, moderate masking levels (25–35%) effectively guide the prompted image encoder to attend to the most relevant regions. However, increasing the masking ratio beyond this range starts to eliminate patches containing important contextual cues, which in turn disrupts the model's ability to capture meaningful patterns. In the second set of experiments (Fig. 4 (b)), we extend the masking strategy by first identifying the least informative patches (defined as the bottom 50% in importance, as described in Subsection 3.2) and then randomly masking a specific percentage of these patches. The results indicate that masking 25% of the least informative patches (equivalent to 50% masking within this subset) significantly enhances performance for novel classes while maintaining stable performance for base classes. These findings validate the effectiveness of randomness in patch selection.

**Analysis of directional regularization.** To clarify the motivation for aligning the prompted feature direction with class-mean features, we conduct additional experiments. Each feature vector, $\mathbf{f} \in \mathbb{R}^d$, can be considered as consisting of two main components: the norm ($||\mathbf{f}||_2$) and the direction ($\mathbf{f}/||\mathbf{f}||_2$) components. Based on this, we define three types of alignment. The first involves aligning the norm components by minimizing the distance $|\|\mathbf{f}\|_2 - \|\mathbf{m}_i\|_2|$. The second addresses the alignment of the entire feature by minimizing the MSE loss $(1/d)\|\mathbf{f} - \mathbf{m}_i\|_2^2$ which includes both norm and direction information. The third focuses exclusively on aligning the directional components, as formalized in Eq. 7. As shown in Fig. 4 (c), norm alignment degrades performance on novel classes, while directional alignment yields the most improvement. This underscores the importance of directional alignment in preventing overfitting and enhancing generalization.

**Sensitivity study.** This section presents a sensitivity analysis of the proposed method with respect to key hyperparameters. We first examine the influence of the weighting factor, $\lambda$, on model performance. Given that the magnitude of the $\mathcal{L}_{\text{DiR}}$ loss is considerably smaller than that of other components in the total loss

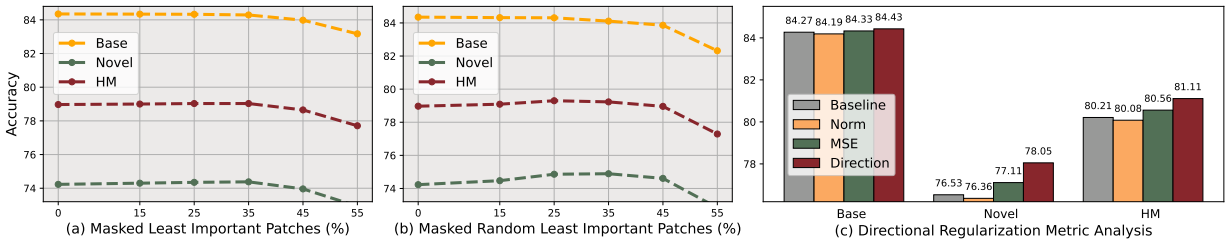


**Figure 4: Analysis of saliency masking and the directional regularization. (a) Accuracy vs. percentage of least informative patches masked, showing optimal performance for novel classes at 25–35% masking. (b) Random masking of 50% least informative patches improves novel-class accuracy, with stability for base classes at 25% masking. (c) Comparison of feature alignment strategies, highlighting directional alignment as most effective for improving generalization.**

**Table 5: Impact of $\lambda$ on DiSa's average performance across 11 datasets.**

| $\lambda$ | 0 | 1 | 4 | 8 | 12 | 16 |
|---|---|---|---|---|---|---|
| HM | 80.21 | 80.41 | 80.62 | 80.90 | **81.11** | 80.98 |

**Table 6: Ablation on prompt depth for different benchmarks.**

| Layer Depth | Avg. Domain Generalization | HM Base-to-Novel |
|---|---|---|
| layer 1 | 60.82 | 77.93 |
| layers 1-3 | **61.43** | 79.08 |
| layers 1-6 | 61.30 | 80.24 |
| layers 1-9 | 60.57 | **81.11** |
| layers 1-12 | 59.14 | 81.02 |

function, we assign it a higher weight to ensure its effective contribution to the overall optimization objective. Table 5 summarizes the impact of varying $\lambda$ values on the average performance across 11 datasets. The results clearly indicate that increasing $\lambda$ improves performance, with the best average accuracy across all datasets achieved at $\lambda = 12$.

Additionally, we analyze the impact of the number of prompted layers on DiSa's performance. As summarized in Table 6, injecting learnable prompts into the first three transformer layers yields the best results for domain generalization setting. In the case of base-to-novel setting, the best performance is achieved by injecting learnable prompts into the first nine layers. These findings demonstrate that our model maintains stable and robust performance across a range of hyperparameter values, highlighting its practicality for real-world applications without the need for extensive tuning.

**Computational complexity analysis.** Table 7 shows the complexity of DiSa in comparison with existing prompt learning approaches, including CoOp, CoCoOp, IVLP, and PromptSRC, for both training and inference phases. The comparison includes GFLOPs, training time, and throughput (FPS). Notably, DiSa incurs only a 0.11x increase in training GFLOPs compared to the baseline IVLP, while maintaining identical GFLOPs and throughput during inference. In addition, DiSa consistently outperforms these methods in terms of accuracy, highlighting its ability to balance computational efficiency with strong performance.

**Table 7: Computational cost comparison on the SUN397 dataset. Training times are reported over 10 epochs. DiSa achieves competitive testing efficiency and higher accuracy than all prior methods.**

| Method | GFLOP (train) | GFLOP (test) | Train time (min) | FPS | HM |
|---|---|---|---|---|---|
| CoOp [46] | 162.5 | 162.5 | 10.08 | 1344 | 71.66 |
| CoCoOp [45] | 162.5 | 162.5 | 39.53 | 15.08 | 75.83 |
| IVLP [31] | 162.8 | 162.8 | 12.01 | 1380 | 77.51 |
| PromptSRC [20] | 179.6 | 162.8 | 13.13 | 1380 | 79.97 |
| DiSa | 179.9 | 162.8 | 13.22 | 1380 | **80.95** |

## 5 Conclusion

In this work, we propose a novel directional saliency-aware prompt learning framework designed to enhance the adaptability and generalization of VL models. Our proposed method, DiSa, employs the novel CIR framework to facilitate cooperative learning between prompted and frozen models. The proposed CIR framework leverages a saliency-aware masking approach to ensure the prompted image encoder focuses on semantically critical regions. We also complement our CIR framework with the directional regularization that aligns prompted features with class-wise prototypes in a manner that prioritizes feature orientation over strict proximity. Extensive evaluations across 11 diverse image classification benchmarks demonstrate the superiority of DiSa over state-of-the-art methods, with considerable improvements in generalization to new classes and domains.

## Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant Numbers CNS-2232048, and CNS-2204445.

## References


[1] Maximilian Augustin, Alexander Meinke, and Matthias Hein. 2020. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision*. 228–245.

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*. 446–461.

[3] Adrian Bulat and Georgios Tzimiropoulos. 2023. LASP: Text-to-text optimization for language-aware soft prompting of vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23232–23241.

[4] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. 2022. PLOT: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253* (2022).

[5] Hyeong Kyu Choi, Joonmyung Choi, and Hyunwoo J Kim. 2022. TokenMixup: Efficient attention-guided token-level data augmentation for transformers. *Advances in Neural Information Processing Systems* 35 (2022), 14224–14235.

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3606–3613.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Ieee, 248–255.

[8] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor G Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. 2023. Bayesian prompt learning for image-language model generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15237–15246.

[9] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. 2022. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11583–11592.

[10] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 178–178.

[11] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. 2022. PromptDet: Towards open-vocabulary detection using uncurated images. In *European Conference on Computer Vision*. 701–717.

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* 132, 2 (2024), 581–595.

[13] Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. 2023. Clip-S4: Language-guided self-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11207–11216.

[14] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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 (2019), 2217–2226.

[15] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8340–8349.

[16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15262–15271.

[17] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. 2022. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems* 35 (2022), 29262–29277.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 4904–4916.

[19] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19113–19122.

[20] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15190–15200.

[21] Konwoo Kim, Michael Laskin, Igor Mordatch, and Deepak Pathak. 2021. How to adapt your large-scale vision-and-language model. (2021).

[22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 554–561.

[23] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[24] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. 2022. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5206–5215.

[25] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151* (2013).

[26] Muhammad Ferjad Naeem, Yongqin Xian, Luc V Gool, and Federico Tombari. 2022. I2DFormer: Learning image to document attention for zero-shot image classification. *Advances in Neural Information Processing Systems* 35 (2022), 12283–12294.

[27] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. 722–729.

[28] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3498–3505.

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

[30] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. 2022. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18082–18091.

[31] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Fine-tuned CLIP models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6545–6554.

[32] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet?. In *International Conference on Machine Learning*. PMLR, 5389–5400.

[33] Shuvendu Roy and Ali Etemad. 2023. Consistency-guided prompt learning for vision-language models. *arXiv preprint arXiv:2306.01195* (2023).

[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[35] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[36] Niloufar Alipour Talemi, Hossein Kashiani, and Fatemeh Afghah. 2025. StylePro: Style-Guided Prompt Learning for Generalizable Vision-Language Models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, 6207–6216.

[37] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.

[38] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems* 32 (2019).

[39] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3485–3492.

[40] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. 2024. MMA: Multi-Modal Adapter for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23826–23837.

[41] Yongjin Yang, Jongwoo Ko, and Se-Young Yun. 2023. Towards Difficulty-Agnostic Efficient Transfer Learning for Vision-Language Models. *arXiv preprint arXiv:2311.15569* (2023).

[42] Hantao Yao, Rui Zhang, and Changsheng Xu. 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.

[43] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783* (2021).

[44] Zhaohui Zheng, Rongguang Ye, Qibin Hou, Dongwei Ren, Ping Wang, Wangmeng Zuo, and Ming-Ming Cheng. 2023. Localization distillation for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 10070–10083.

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

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