Self-regulating Prompts: Foundational Model Adaptation without Forgetting

Muhammad Uzair Khattak[1*✉] Syed Talal Wasim[1*] Muzammal Naseer[1]
Salman Khan[1,2] Ming-Hsuan Yang[4,5] Fahad Shahbaz Khan[1,3]

[1]Mohamed bin Zayed University of AI [2]Australian National University
[3]Linköping University [4]University of California, Merced [5]Google Research

Abstract

Prompt learning has emerged as an efficient alternative for fine-tuning foundational models, such as CLIP, for various downstream tasks. Conventionally trained using the task-specific objective, i.e., cross-entropy loss, prompts tend to overfit downstream data distributions and find it challenging to capture task-agnostic general features from the frozen CLIP. This leads to the loss of the model's original generalization capability. To address this issue, our work introduces a self-regularization framework for prompting called PromptSRC (Prompting with Self-regulating Constraints). PromptSRC guides the prompts to optimize for both task-specific and task-agnostic general representations using a three-pronged approach by: (a) regulating prompted representations via mutual agreement maximization with the frozen model, (b) regulating with self-ensemble of prompts over the training trajectory to encode their complementary strengths, and (c) regulating with textual diversity to mitigate sample diversity imbalance with the visual branch. To the best of our knowledge, this is the first regularization framework for prompt learning that avoids overfitting by jointly attending to pre-trained model features, the training trajectory during prompting, and the textual diversity. PromptSRC explicitly steers the prompts to learn a representation space that maximizes performance on downstream tasks without compromising CLIP generalization. We perform extensive experiments on 4 benchmarks where PromptSRC overall performs favorably well compared to the existing methods. Our code and pre-trained models are publicly available at: https://github.com/muzairkhattak/PromptSRC.

1. Introduction

Vision-Language (VL) models, such as CLIP [35] and ALIGN [20], have demonstrated remarkable generalization capabilities for downstream tasks. These VL models are trained on large-scale web data with a contrastive loss, which allows them to encode open-vocabulary concepts by aligning pairs of images and texts in a shared embedding space. The resulting model is suited for downstream tasks such as open-vocabulary image recognition [23], object detection [11], and image segmentation [29].

Prompt learning has emerged as a more efficient alternative to fine-tuning large-scale models, as shown in recent studies [58, 59, 3, 17, 40, 28]. This approach introduces a few learnable prompt vectors to adapt models like CLIP for downstream tasks while keeping the pre-trained model weights fixed. However, since the prompts are optimized with respect to the task-specific objective [59], such as the cross-entropy loss for ImageNet [6] classification, the prompted model tends to overfit to the task-specific data distribution as the training progresses. This can result in the prompted model losing the original generalization capability of the frozen CLIP model towards new tasks. Therefore, learning prompts that can model both task-specific and task-agnostic representations remain a major challenge for adapting foundational VL models.

This work seeks to self-regulate prompts to address the issue of prompt overfitting. To this end, we propose a self-regularizing framework that guides the prompts to jointly optimize for both task-specific and task-agnostic general representations using a three-pronged approach. **a)** *Regulating via Mutual Agreement Maximization:* We observe that generalizable zero-shot knowledge is preserved within frozen pre-trained VL model features but they lack task-specific knowledge. In contrast, prompts achieve better adaptation to a given task but with reduced generalizability to new tasks. Therefore, we propose to regulate learned prompts by maximizing the agreement between prompted and frozen VL model features while adapting them to the downstream task. **b)** *Regulating with the Self-ensemble:* In the early epochs, prompts act are not mature to capture contextual information. As the training progresses, prompts tend to become more task-specific. Therefore we deploy a weighted prompt aggregation technique to prompts during training to regulate them using their self-ensemble over the

[*] Joint first authors.
[✉] uzair.khattak@mbzuai.ac.ae

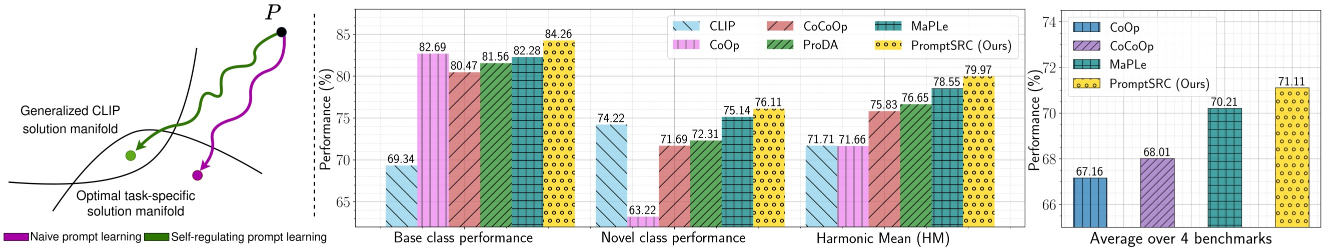

Figure 1: **(Left)**: Existing prompt learning approaches rely on task-specific objectives that restrict prompt learning to learn a feature space suitable only for downstream tasks and consequently lose the generalized knowledge of CLIP (shown in purple). Our self-regulating framework explicitly guides the training trajectory of prompts towards the closest point between two optimal solution manifolds (solid line) to learn task-specific representations while also retaining generalized CLIP knowledge (shown in green). **(Middle)**: Averaged across 11 image recognition datasets, PromptSRC surpasses existing methods on the base-to-novel generalization setting. **(Right)**: We evaluate our approach on four diverse image recognition benchmarks and it overall shows competitive results compared to the previous state-of-the-art.

training phase. The weights are sampled from a Gaussian distribution which suitably aggregates the useful knowledge learned by prompts at different training epochs. **c)** *Regulating with Textual Diversity:* We note that unlike having multiple image samples per category for the vision encoder, there is only a single textual label available for each class. Therefore, imposing the mutual agreement constraints on multi-modal features results in sub-optimal performance due to the lack of diversity in text-side labels for the text encoder. We overcome this disparity and regulate the prompts through diverse text label templates for each class.

Overall, our approach explicitly steers prompts to learn a representation space that maximizes its performance on downstream tasks without compromising pre-trained CLIP generalization (Fig. 1: Left). We demonstrate the effectiveness of PromptSRC on four representative tasks. On the base-to-novel generalization benchmark across 11 datasets (Fig. 1: Middle), our method achieves average gains of +1.42% in harmonic-mean over the state-of-the-art MaPLe [22] and +8.26% over CLIP. Further, PromptSRC achieves competitive results in cross-dataset transfer, domain generalization, and few-shot image recognition (Fig. 1:Right).

In summary, our self-regulating prompt learning framework has the following main contributions:

- We address the inherent problem of prompt overfitting for adapting foundational models through self-regularization. Our framework explicitly guides the prompts to jointly acquire both *task-specific knowledge* and *task-agnostic generalized knowledge* by maximizing the mutual agreement between prompted and frozen VL model features. (§3.2.1)
- We suggest a weighted self-ensembling strategy for prompts that captures their complementary features learned at different epochs during training and enhances their generalization performance. (§3.2.2)
- To overcome the significant diversity mismatch between the text and visual domains, we propose text-side diversity which complements limited textual labels via multiple text augmentations and regularizes prompts to learn more generalized contexts. (§3.2.3)

2. Related Work

Vision Language models: Foundational vision-language (VL) models [35, 20, 54, 49, 51] leverage both visual and textual modalities to encode rich multi-modal representations. These models are pre-trained on a large corpus of image-text pairs available on the internet in a self-supervised manner. For instance, CLIP [35] and ALIGN [20] utilize around 400M and 1B image-text pairs, respectively, to train their multi-modal networks. During pre-training, contrastive loss is commonly used as a self-supervision loss. This loss pulls together the features of paired images and texts while pushing away the unpaired image-text features. VL models possess a strong understanding of open-vocabulary concepts, making them suitable for various downstream vision and vision-language applications [12, 56, 38, 30, 60, 13, 32, 53, 26, 36, 8]. However, transferring these foundational models for downstream tasks without compromising on their original generalization ability still remains a major challenge. Our work aims to address this problem by proposing a novel regularization framework to adapt VL models via prompt learning.
Prompt learning: Prompt learning is an alternative fine-tuning method for transferring a model towards downstream tasks without re-learning the trained model parameters. This approach adapts a pre-trained model by adding a small number of new learnable embeddings at the input known as prompt tokens. Due to its efficiency in terms of parameters and convergence rate, prompt learning is found to be of great interest for adapting foundational models like CLIP for vision [21, 57, 45, 46] and vision-language tasks [59, 58, 61, 7]. CoOp [59] fine-tunes CLIP by optimizing a continuous set of prompt vectors in its language branch for few-shot image recognition. Bahng *et al.* [1] perform visual prompt tuning on CLIP by learning prompts

on the vision branch. [3] and [28] propose to learn multiple sets of prompts for learning different contextual representations. CoCoOp [58] highlights the overfitting problem of CoOp and proposes to condition prompts based on visual features for improved performance on generalization tasks. MaPLe [22] proposes a multi-modal prompt learning approach by learning hierarchical prompts jointly at the vision and language branches of CLIP for better transfer. Our approach builds on a variant [37] where prompts are learned at both the vision and language encoder of CLIP.

Network regularization: Incorporating regularization techniques in neural networks has been proven to enhance their generalization capabilities [25]. Regularization strategies can be broadly classified into two streams. The first stream consists of constraint-based regularization methods, such as weight decay [27] and adversarial training [50]. These techniques introduce additional constraints to the learning process, which helps to prevent overfitting. The second stream of regularization techniques involves modifying the inputs, model parameters, or annotations. This category includes methods such as data augmentations [52, 55, 5], dropout [42], model ensembling [18, 47], label smoothing [43] and batch normalization [19]. Our method aims to enhance the generalization performance of learned prompts via a multi-stage regularization framework, which takes inspiration from both streams of regularization techniques mentioned above. However, to the best of our knowledge, this is the first effort to regularize prompts during adaptation by jointly attending to the original VL model feature space, the training trajectory of prompts as well as the diversity of textual inputs for the multi-modal models.

3. Proposed Method

Prompt learning aims to adapt the general knowledge of VL foundational models like CLIP without full fine-tuning [59, 58, 3]. Since prompts are the only learnable vectors, this strategy aims to retain the pretrained generalized feature representations of CLIP while re-purposing them for downstream task-specific data via prompts. Although effective, they are susceptible to overfitting on the supervised downstream task (see Fig. 2) and their generalization towards new classes and datasets reduces as compared to the original zero-shot pre-trained CLIP.

Our work seeks to address the overfitting behavior of prompts. Unlike prior prompting approaches that improve generalization mainly from the model architecture perspective [58, 22], we motivate our work from the regularization perspective. As evidenced by the strong zero-shot performance, pre-trained CLIP features possess robust generalization characteristics. However, naively training prompts with the supervised task-specific loss struggles to retain these general attributes from the frozen CLIP. To this end, we propose a self-regularizing framework to explicitly guide the training trajectory of prompts to maximize its interaction with the pre-trained knowledge stored in the frozen CLIP.

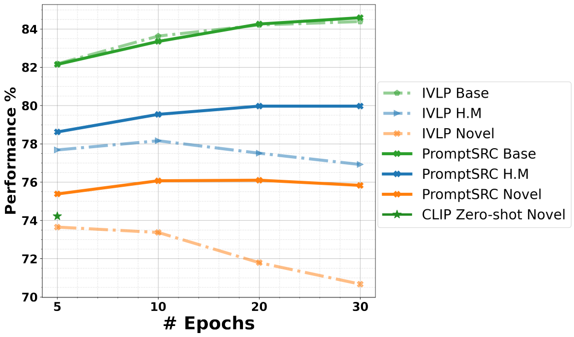

Figure 2: Naively training prompts with standard supervised objectives improves supervised class performance but leads to poor generalization as training schedule increases. Our PromptSRC method with explicit prompts consistency constraints improves on base classes as well as shows improvements on novel classes.

Fig. 3 shows our overall methodology which optimizes the prompts as follows. **a)** *Regularization through mutual agreement maximization:* We impose an explicit consistency constraint between prompted features and the pre-trained CLIP features within the CLIP embedding space. **b)** *Regularization through prompt self-ensembling:* To further reduce overfitting, we propose a Gaussian weighted average of the prompt vectors learned at different training epochs. This ensemble-level regularization aggregates information from learned prompts across different epochs for improved generalization. **c)** *Regularization through textual diversity:* Unlike having multiple images for each class, the text labels during fine-tuning are limited and bounded by the number of class categories. We incorporate textual augmentations by defining multiple text label templates for a given class. The ensemble of textual labels regularizes the prompts for better generalization during optimization.

We now continue by explaining our methodology in detail. We first revisit CLIP and CLIP-based prompt learning in Sec. 3.1. This is followed by the explanation of our self-regulating prompt learning approach in Sec. 3.2.

3.1. Preliminaries

We denote the CLIP image and text encoders as f and g, respectively and their pretrained parameters as $\theta_{\texttt{CLIP}} = \{\theta_f, \theta_g\}$ where θ_f and θ_g refer to the image and text encoder parameters, respectively. The input image $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$ is divided into M patches followed by a projection to produce patch tokens. Further, a learnable class token $\boldsymbol{e}_{cls}$ is appended with the input patches as $\tilde{\boldsymbol{X}} = \{\boldsymbol{e}_{cls}, \boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_M\}$. The image encoder f encodes the input patches via multiple transformer blocks to produce a latent visual feature representation $\tilde{\boldsymbol{f}} = f(\tilde{\boldsymbol{X}}, \theta_f)$, where $\tilde{\boldsymbol{f}} \in \mathbb{R}^d$. Next, the corresponding class label

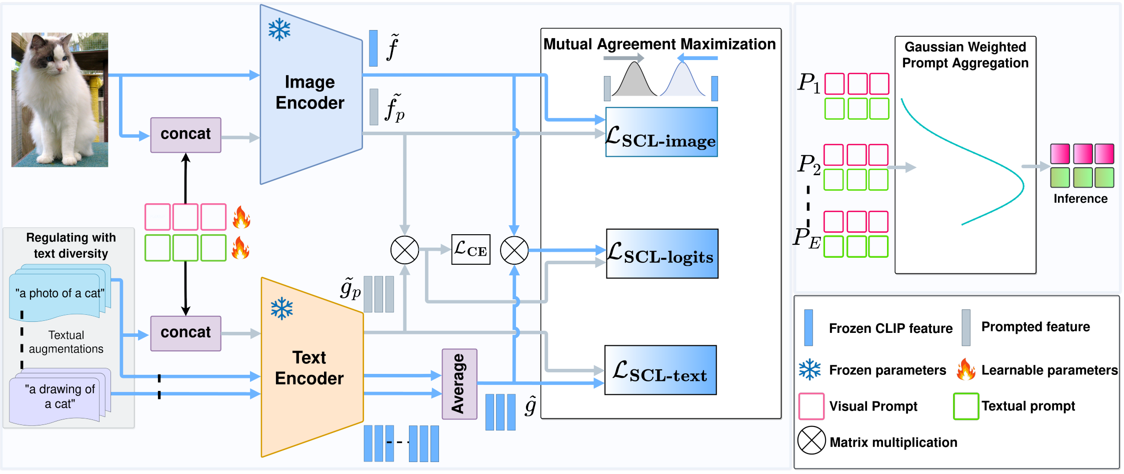

Figure 3: Our proposed PromptSRC framework for self-regulating prompt learning. CLIP encoders are used to generate **prompted** ($\tilde{\boldsymbol{f}}_{\boldsymbol{p}}, \tilde{\boldsymbol{g}}_{\boldsymbol{p}}$) and **pre-trained** ($\tilde{\boldsymbol{f}}, \tilde{\boldsymbol{g}}$) features at the image and text sides. First, we introduce textual diversity (§3.2.3) and define textual augmentations to produce a diverse set of frozen VL textual features, which are averaged to represent the pre-trained VL text features ($\tilde{\boldsymbol{g}}$). Next, we employ Mutual Agreement Maximization constraints ($\mathcal{L}_{\text{SCL}}$) to regulate the prompts, which ensure that the prompted features align well with the pre-trained VL representations at both the feature and logit levels (§3.2.1). As CLIP is frozen, we use the same VL encoders to obtain both types of features. Further, our prompt self-ensembling combines the strengths of prompts learned at different epochs ($P_1, P_2 \cdots P_E$) during training via Gaussian weighted sampling (§3.2.2). The ensembled **visual** and **textual** prompts are then used for the final inference.

y is wrapped within a text template such as ‘a photo of a {class label}’ which can be formulated as $\tilde{\boldsymbol{Y}} = \{\boldsymbol{t}_{SOS}, \boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_L, \boldsymbol{c}_k, \boldsymbol{t}_{EOS}\}$. Here $\{\boldsymbol{t}_l|_{l=1}^{L}\}$ and $\boldsymbol{c}_k$ are the word embeddings corresponding to the text template and the class label, respectively while $\boldsymbol{t}_{SOS}$ and $\boldsymbol{t}_{EOS}$ are the learnable start and end token embeddings. The text encoder g encodes $\tilde{\boldsymbol{Y}}$ via multiple transformer blocks to produce the latent textual feature as $\tilde{\boldsymbol{g}} = g(\tilde{\boldsymbol{Y}}, \theta_g)$, where $\tilde{\boldsymbol{g}} \in \mathbb{R}^d$. For zero-shot inference, textual features of text template with class labels $\{1, 2, \cdots, C\}$ are matched with image feature $\tilde{\boldsymbol{f}}$ as $\frac{\texttt{exp}(\texttt{sim}(\tilde{\boldsymbol{g}}\cdot\tilde{\boldsymbol{f}})\tau)}{\sum_{i=1}^{C}\texttt{exp}(\texttt{sim}(\tilde{\boldsymbol{g}}_{\boldsymbol{i}}\cdot\tilde{\boldsymbol{f}})\tau)}$, where $\texttt{sim}()$ denotes the cosine similarity and τ is the temperature.

Prompt Learning for CLIP: Prompt learning approaches append learnable prompt tokens at either the text [59, 58] encoder or image [1] encoder. We use a simple baseline method [37] that learns hierarchical prompt tokens on both the text and image encoders separately, named as Independent Vision-Language Prompting (IVLP).

Specifically, we append learnable T language and V visual prompts given as $\boldsymbol{P_t} = \{\boldsymbol{p}_{\boldsymbol{t}}^1, \boldsymbol{p}_{\boldsymbol{t}}^2, \cdots, \boldsymbol{p}_{\boldsymbol{t}}^T\}$ and $\boldsymbol{P_v} = \{\boldsymbol{p}_{\boldsymbol{v}}^1, \boldsymbol{p}_{\boldsymbol{v}}^2, \cdots, \boldsymbol{p}_{\boldsymbol{v}}^V\}$ with the textual and visual input tokens, respectively. Therefore, the image encoder processes the following input tokens $\tilde{\boldsymbol{X}}_{\boldsymbol{p}} = \{\boldsymbol{P_v}, \boldsymbol{e}_{cls}, \boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_M\}$ to generate prompted visual feature represented as $\tilde{\boldsymbol{f}}_{\boldsymbol{p}} = f(\tilde{\boldsymbol{X}}_{\boldsymbol{p}}, \theta_f)$. Similarly, textual feature is obtained as $\tilde{\boldsymbol{g}}_{\boldsymbol{p}} = g(\tilde{\boldsymbol{Y}}_{\boldsymbol{p}}, \theta_g)$, where $\tilde{\boldsymbol{Y}}_{\boldsymbol{p}} = \{\boldsymbol{t}_{SOS}, \boldsymbol{P_t}, \boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_L, c_k, \boldsymbol{t}_{EOS}\}$. In contrast to shallow prompting where learnable prompts are introduced only at the first transformer block of the image and text encoders, our approach uses deep prompting which learns separate sets of prompts at every transformer block. The vision and language prompts are jointly represented as $\boldsymbol{P} = \{\boldsymbol{P_v}, \boldsymbol{P_t}\}$. The feature representations obtained using these learnable prompts are referred to as *prompted features*.

For image classification on a downstream dataset $\mathcal{D}$, prompts $\boldsymbol{P}$ interact with pre-trained and frozen θ_f and θ_g and are optimized with the cross-entropy loss, $\mathcal{L}_{\text{CE}}$, as:

$$\mathcal{L}_{\text{CE}} = \arg\min_{\boldsymbol{P}} \mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}}\, \mathcal{L}(\text{sim}(\tilde{\boldsymbol{f}}_{\boldsymbol{p}}, \tilde{\boldsymbol{g}}_{\boldsymbol{p}}), y). \tag{1}$$

3.2. Self-Regularization for Prompt Learning

The $\mathcal{L}_{\text{CE}}$ objective employs ground truth labels to optimize the prompts for the downstream task. As a result, the prompts adapt and learn *task-specific knowledge*. During training, prompts interact with pre-trained and frozen CLIP tokens through self-attention layers in the transformer blocks. This interaction of prompts tokens with pre-trained CLIP weights θ_{CLIP} provides implicit regularization and encourages retaining the *task-agnostic generalized knowledge* within learned prompts. However, as shown in Fig. 2, prompts tend to overfit on the supervised task and drift away from the generalized CLIP space as the training schedule increases. Consequently, new task performance is degraded, despite the fact that CLIP image and text encoder weights θ_f and θ_g are kept frozen. As prompts undergo further training, the implicit generalization constraint becomes weaker against the task-specific $\mathcal{L}_{\text{CE}}$ objective.

One naive approach to address this issue is to reduce the training schedule to balance the performance between the base and new tasks. However, training the prompts for fewer iterations to prevent losing generalization comes at the cost of relatively lower performance on the supervised task. Here, we present a prompt learning approach that maximizes supervised task performance without sacrificing performance on novel tasks and classes. We propose to anchor prompt training with self-regularization which constitutes three main components as discussed below.

3.2.1 Mutual agreement maximization

As discussed above, the strong downstream dataset transfer constraint imposed by $\mathcal{L}_{\text{CE}}$ causes the prompts to overfit on task-specific data and it struggles to effectively utilize the general information from the frozen CLIP. We propose to explicitly guide the training trajectory by imposing a constraint to maximize its mutual agreement between the prompted and the frozen CLIP features. We achieve this by explicitly conditioning the prompted features to be consistent with the CLIP features obtained without learnable prompts. As we do not require any second model for such conditioning, we call this regularizing constraint as a self-consistency loss (SCL). For a given input sample and its corresponding textual label, we obtain visual features using learnable prompts and pre-trained visual features, $\tilde{\boldsymbol{f}}_{\boldsymbol{p}}$ and $\tilde{\boldsymbol{f}}$ within the frozen CLIP latent space. Similarly, we obtain textual features $\tilde{\boldsymbol{g}}_{\boldsymbol{p}}$ and $\tilde{\boldsymbol{g}}$.

We then impose a constraint on the prompted visual and text features to ensure their consistency with the CLIP pre-trained features as follows,

$$\mathcal{L}_{\text{SCL-image}} = \sum_{i=1}^{d} |\tilde{\boldsymbol{f}}_{\boldsymbol{p}} - \tilde{\boldsymbol{f}}|, \; \mathcal{L}_{\text{SCL-text}} = \sum_{i=1}^{d} |\tilde{\boldsymbol{g}}_{\boldsymbol{p}} - \tilde{\boldsymbol{g}}|. \quad (2)$$

As shown in Eq. 2, we utilize $L1$ loss to impose the feature level consistency. Note that our self-consistency constraint is also compatible with other variants of matching losses such as cosine similarity or MSE loss which we study in our ablations (Sec. 4.7).

To further complement the regularization constraint and maximize the alignment between the general features and the prompted features, we impose logit level self-consistency regularization and condition the prompted logits distribution on pre-trained CLIP logits distribution by minimizing the Kullback-Leibler divergence as follows,

$$\mathcal{L}_{\text{SCL-logits}} = \mathcal{D}_{\mathcal{KL}}(\text{sim}(\tilde{\boldsymbol{f}}_{\boldsymbol{p}}, \tilde{\boldsymbol{g}}_{\boldsymbol{p}}), \text{sim}(\tilde{\boldsymbol{f}}, \tilde{\boldsymbol{g}})). \quad (3)$$

Overall, the self-consistency training objectives guide the prompts to gain complementary knowledge from pre-trained CLIP features, therefore providing strongly generalized prompts,

$$\mathcal{L}_{\text{SCL}} = \lambda_1 \mathcal{L}_{\text{SCL-image}} + \lambda_2 \mathcal{L}_{\text{SCL-text}} + \mathcal{L}_{\text{SCL-logits}}, \quad (4)$$

where λ_1 and λ_2 are loss balancing hyper-parameters. Our overall training objective thus becomes,

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{SCL}}. \quad (5)$$

Discussion on $\mathcal{L}_{\text{final}}$: $\mathcal{L}_{\text{SCL}}$ loss guides the prompts to converge at solutions that are generalized. On the other hand, $\mathcal{L}_{\text{CE}}$ guides the prompts to maximize performance on the downstream supervised tasks. The combination of these losses conditions the prompts to maximize their performance on supervised tasks and at the same time guides the prompts learning trajectory toward a weight space that is consistent with the CLIP zero-shot features. As shown in Fig. 2, our proposed methodology maximizes the supervised tasks' performance while also improving the generalization. This shows that the proposed training objectives for prompt learning setup are complementary to each other.

3.2.2 Regularization with prompt self-ensembling

The second component in our self-regularizing framework enforces regularization using prompt self-ensembling. Model ensembling in the weight space has been shown to improve both the performance and generalization of a model [47, 18]. However, it has not been actively studied in the context of prompt learning, where prompts are only learnable parameters with frozen model parameters.

To effectively utilize the prompts knowledge from the previous training iterations, we propose prompts aggregation for a generalizable solution. For a training schedule with E total epochs, prompts at every epoch are given by $\{\boldsymbol{P}\}_{t=1}^{E}$. Aggregated prompts (AP) are then calculated as,

$$\{\boldsymbol{P}\}^{\text{AP}} = \sum_{t=1}^{E} \frac{w_t.\boldsymbol{P}}{\sum_{i=1}^{E} w_i}, \quad (6)$$

where w_i is the weight assigned to prompts at each epoch t.

In the early epochs, prompts are not mature to capture contextual information due to their random initialization. During aggregation, they should be given less weight as they act as noise which is carried along with the input tokens. On the other hand, the prompts learned in the last few epochs are task specific and highly favours the supervised downstream task distribution. We propose to perform Gaussian weighted prompt aggregation (GPA), where small aggregation weights are given to prompts at initial epochs, higher weights to prompts at middle epochs, and relatively lower weights to prompts at final epochs, resulting in optimal prompt representations that improve generalization to downstream tasks. GPA provides optimal weight values w_i by sampling from a Gaussian distribution $w_i \sim \mathcal{N}(\mu, \sigma^2)$, where σ^2 and μ are hyper-parameters and $\sum_{i=1}^{E} w_i = 1$. Gaussian distribution is defined over the epochs and its mean is dictated by the epoch number. We formulate this weighting as a moving average to avoid saving multiple copies of prompts by keeping one additional copy which is updated via aggregation at every epoch i,

$$\boldsymbol{P}^{\text{GPA}} = \sum_{i=1}^{E} w_i.\boldsymbol{P_i}. \tag{7}$$

3.2.3 Regulating prompts with textual diversity

Through the $\mathcal{L}_{\text{SCL}}$ loss, the visual prompted features to instill *diverse generalized contexts* from pre-trained CLIP visual features as multiple image samples are present for each label category. This provides a natural source of augmentations at the image side and promotes additional regularization. However, as opposed to having multiple images per category, we note that the text space during fine-tuning is limited, and prompted features are learned based on pre-trained CLIP text features, with only one feature representation per category. This mismatch between the available diversity at the image and text side leads to sub-optimal learning of prompted textual features. To address the diversity mismatch, we incorporate textual diversity in the text encoder. Specifically, we use a pool of textual prompt templates $\{PT|_{l=1}^{N}\}$, containing N augmentations to form multiple text features per category. The pre-trained CLIP textual features are now obtained as an ensemble of multiple prompts templates $\tilde{\boldsymbol{g}} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{g}}^{\boldsymbol{i}}$. As pre-trained CLIP textual features are now represented by the ensemble of multiple augmentations for each label, the prompted textual features learn more *diverse generalized contexts* from the frozen CLIP. We note that the proposed textual diversity is different from the standard prompt ensembling technique explored by CLIP authors. CLIP uses ensemble of text prompts during inference for classification. In contrast, we utilize them during training for self-regularization by enforcing mutual agreement of ensembled features with prompted features, and prompted features are used at inference. Next, we show the efficacy of our proposed components via comprehensive experiments provided below.

4. Experiments

4.1. Evaluation settings

We extensively evaluate our approach and present a comparison with other methods on four benchmark settings.

Base-to-novel class generalization: In this setting, we equally split the datasets into base and novel classes. The model is trained on base classes and evaluated on both base classes and novel classes. This benchmark evaluates the generalization ability of a method within a dataset.

Few-shot learning: We incorporate this setting to compare the learning capacity of the model under extremely limited supervision and verify if our approach learns complementary task-specific and task-agnostic knowledge. For each dataset, we test the model's generalization for different K-shots per category, where $K = 1, 2, 4, 8, 16$.

Dataset		CLIP [35]	CoOp [59]	CoCoOp [58]	ProDA [28]	MaPLe [22]	PromptSRC (Ours)	Δ
Average on 11 datasets	Base	69.34	82.69	80.47	81.56	82.28	**84.26**	+2.0
	Novel	74.22	63.22	71.69	72.30	75.14	**76.10**	+1.0
	HM	71.70	71.66	75.83	76.65	78.55	**79.97**	+1.4
ImageNet	Base	72.43	76.47	75.98	75.40	76.66	**77.60**	+0.9
	Novel	68.14	67.88	70.43	70.23	70.54	**70.73**	+0.2
	HM	70.22	71.92	73.10	72.72	73.47	**74.01**	+0.5
Caltech101	Base	96.84	98.00	97.96	**98.27**	97.74	98.10	+0.4
	Novel	94.00	89.81	93.81	93.23	**94.36**	94.03	-0.3
	HM	95.40	93.73	95.84	95.68	**96.02**	**96.02**	+0.0
OxfordPets	Base	91.17	93.67	95.20	**95.43**	**95.43**	95.33	-0.1
	Novel	97.26	95.29	97.69	**97.83**	97.76	97.30	-0.5
	HM	94.12	94.47	96.43	**96.62**	96.58	96.30	-0.3
Stanford Cars	Base	63.37	78.12	70.49	74.70	72.94	**78.27**	+5.3
	Novel	74.89	60.40	73.59	71.20	74.00	**74.97**	+1.0
	HM	68.65	68.13	72.01	72.91	73.47	**76.58**	+3.1
Flowers102	Base	72.08	97.60	94.87	97.70	95.92	**98.07**	+2.1
	Novel	**77.80**	59.67	71.75	68.68	72.46	76.50	+4.1
	HM	74.83	74.06	81.71	80.66	82.56	**85.95**	+3.4
Food101	Base	90.10	88.33	90.70	90.30	**90.71**	90.67	-0.1
	Novel	91.22	82.26	91.29	88.57	**92.05**	91.53	-0.5
	HM	90.66	85.19	90.99	89.43	**91.38**	91.10	-0.3
FGVC Aircraft	Base	27.19	40.44	33.41	36.90	37.44	**42.73**	+5.3
	Novel	36.29	22.30	23.71	34.13	35.61	**37.87**	+2.3
	HM	31.09	28.75	27.74	35.46	36.50	**40.15**	+3.7
SUN397	Base	69.36	80.60	79.74	78.67	80.82	**82.67**	+1.9
	Novel	75.35	65.89	76.86	76.93	**78.70**	78.47	-0.2
	HM	72.23	72.51	78.27	77.79	79.75	**80.52**	+0.8
DTD	Base	53.24	79.44	77.01	80.67	80.36	**83.37**	+3.0
	Novel	59.90	41.18	56.00	56.48	59.18	**62.97**	+3.8
	HM	56.37	54.24	64.85	66.44	68.16	**71.75**	+3.6
EuroSAT	Base	56.48	92.19	87.49	83.90	**94.07**	92.90	-1.2
	Novel	64.05	54.74	60.04	66.00	73.23	**73.90**	+0.7
	HM	60.03	68.69	71.21	73.88	**82.35**	82.32	-0.1
UCF101	Base	70.53	84.69	82.33	85.23	83.00	**87.10**	+4.1
	Novel	77.50	56.05	73.45	71.97	78.66	**78.80**	+0.1
	HM	73.85	67.46	77.64	78.04	80.77	**82.74**	+2.0

Table 1: Accuracy comparison on Base-to-novel generalization of PromptSRC with previous methods. The prompts learned with our self-regularizing approach show overall consistent improvements on base classes, without losing generalization. Absolute gains over MaPLe [22] are shown in blue.

Domain generalization setting: We train a source model on ImageNet [6] and evaluate on out-of-distribution datasets to test performance under domain shifts.

Cross-dataset evaluation: In cross-dataset transfer, we train the models on ImageNet [6] and directly evaluate it on other datasets without any data-specific fine-tuning.

Datasets: For base to novel class generalization, few-shot setting and cross-dataset evaluation, we follow CoOp [59] and CoCoOp [58], and use 11 image recognition datasets. The datasets cover multiple recognition tasks including ImageNet [6] and Caltech101 [10] which consists of generic objects; OxfordPets [34], StanfordCars [24], Flowers102 [33], Food101 [2], and FGVCAircraft [31] for fine-grained classification, SUN397 [48] for scene recognition, UCF101 [41] for action recognition, DTD [4] for texture classification, and EuroSAT [14] which consists of satellite images. For domain generalization benchmark, we use ImageNet [6] as a source dataset and use ImageNet-A [16], ImageNet-R [15], ImageNet-Sketch [44] and ImageNetV2 [39] as out of distribution datasets.

Implementation details: We use a ViT-B/16 based CLIP model in our experiments and report results averaged over 3 runs. We use deep prompting with $V = T = 4$ VL prompts and train for 50 epochs for few-shot setting and 20 epochs the rest of the 3 benchmarks respectively. For domain generalization and cross-dataset evaluation, we train the ImageNet source model on all classes with $K = 16$ shots using $V = T = 4$ VL prompts in the first 3 transformer layers. For few-shot and base-to-novel setting, prompts are learned in the first 9 transformer layers. Prompts are randomly initialized with a normal distribution except the text prompts of the first layer which are initialized with the word embeddings of "a photo of a". We fix the learning rate to 0.0025. We set $\lambda_1 = 10$ and $\lambda_2 = 25$ to weight $\mathcal{L}_{\text{SCL-image}}$ and $\mathcal{L}_{\text{SCL-text}}$ respectively. The corresponding hyperparameters are fixed across all datasets and benchmarks. For textual diversity, we use a total of $N = 60$ standard prompt templates provided in [35]. For comparison with ProDA [28], we report their results produced by [7]. Refer to Appendix A for additional implementation details.

4.2. Effectiveness of Self-regulating Prompts

We first disentangle the regularization components in our self-regulating prompting framework and show the individual contributions in Table 2. Baseline IVLP provides high base class performance but suffers from poor generalization (row-1). By enforcing mutual agreement through $\mathcal{L}_{\text{SCL}}$ (row-2), novel class performance significantly increases by 3.95% while maintaining base class gains. This suggests that $\mathcal{L}_{\text{SCL}}$ explicitly enforces the prompts to capture the generalizable features from frozen CLIP. Integrating GPA (row-3) which suitably aggregates prompts across the training cycle further reduces overfitting and improves the novel class performance. Finally, combined with textual diversity to overcome the diversity mismatch between the text and visual domains (row-4), PromptSRC achieves improvements on both base and novel classes, leading to the average novel class and harmonic mean gains of +4.31% and +2.46% respectively. The averaged results on 11 datasets are summarized in Table 2. Note that even small improvements in these metrics correspond to significant gains. We refer the readers to Appendix B for results on individual datasets.

Method	Base Acc.	Novel Acc.	HM
1: Independent V-L prompting	84.21	71.79	77.51
2: + $\mathcal{L}_{\text{SCL}}$	84.21	75.38	79.55
3: + GPA	84.16	75.69	79.70
4: + Textual diversity	**84.26**	**76.10**	**79.97**

Table 2: Effect of our proposed regularization techniques. Results are averaged over 11 datasets. HM refers to harmonic mean.

4.3. Base-to-Novel Generalization

We compare the performance of our approach with zero-shot CLIP [35], CoOp [59], CoCoOp [58], ProDA [28] and MaPLe [22], in Table 1. Overall, all existing approaches outperform zero-shot CLIP on base classes but show inferior performance on novel classes except MaPLe. This suggests that they overall tend to lose the generalizable features stored in the frozen CLIP model. In contrast, PromptSRC significantly improves base class performance while improving the zero-shot CLIP novel class accuracy by 1.88%. This shows the importance of explicit guidance provided by PromptSRC in learning complementary task-specific and task-agnostic representations which aid base and novel classes respectively.

CoOp is heavily trained on base classes and consequently compromises on its generalization. For instance, on EuroSAT [14], CoOp provides a substantial 92.19% base class accuracy and inferior novel class accuracy of 54.74%. On the other hand, PromptSRC which learns self-regulating prompts provides the highest base and novel class accuracies of 92.90% and 73.90% on EuroSAT respectively.

In comparison to CoCoOp and ProDA, PromptSRC shows gains on the 10/11 datasets respectively. Against the recent MaPLe approach, PromptSRC improves performance on 8/11 datasets while using 77x less tunable parameters (3.55M of MaPLe vs 46K of PromptSRC). With respect to the averaged results, PromptSRC provides the best results of 84.26%, 76.10%, and 79.97% on the base class, novel class, and harmonic mean respectively.

4.4. Few-shot Experiments

To explicitly verify if our regularization framework restricts the prompts to learn task-specific knowledge or not, we compare our few-shot results with existing methods in Fig. 4. In general, all prompt learning approaches perform better than the linear probe, especially in scenarios with lesser shots *i.e.*, $K = 1, 2, 4$. PromptSRC overall provides consistent improvements on all shots in comparison with all existing methods. When compared with the existing best method MaPLe, PromptSRC consistently provides absolute gains of 3.05%, 2.72%, 2.59%, 1.80%, and, 1.07% on 1, 2, 4, 8, and 16 shots respectively which are averaged over 11 datasets. Furthermore, we note that our approach achieves relatively larger gains in minimal data cases such

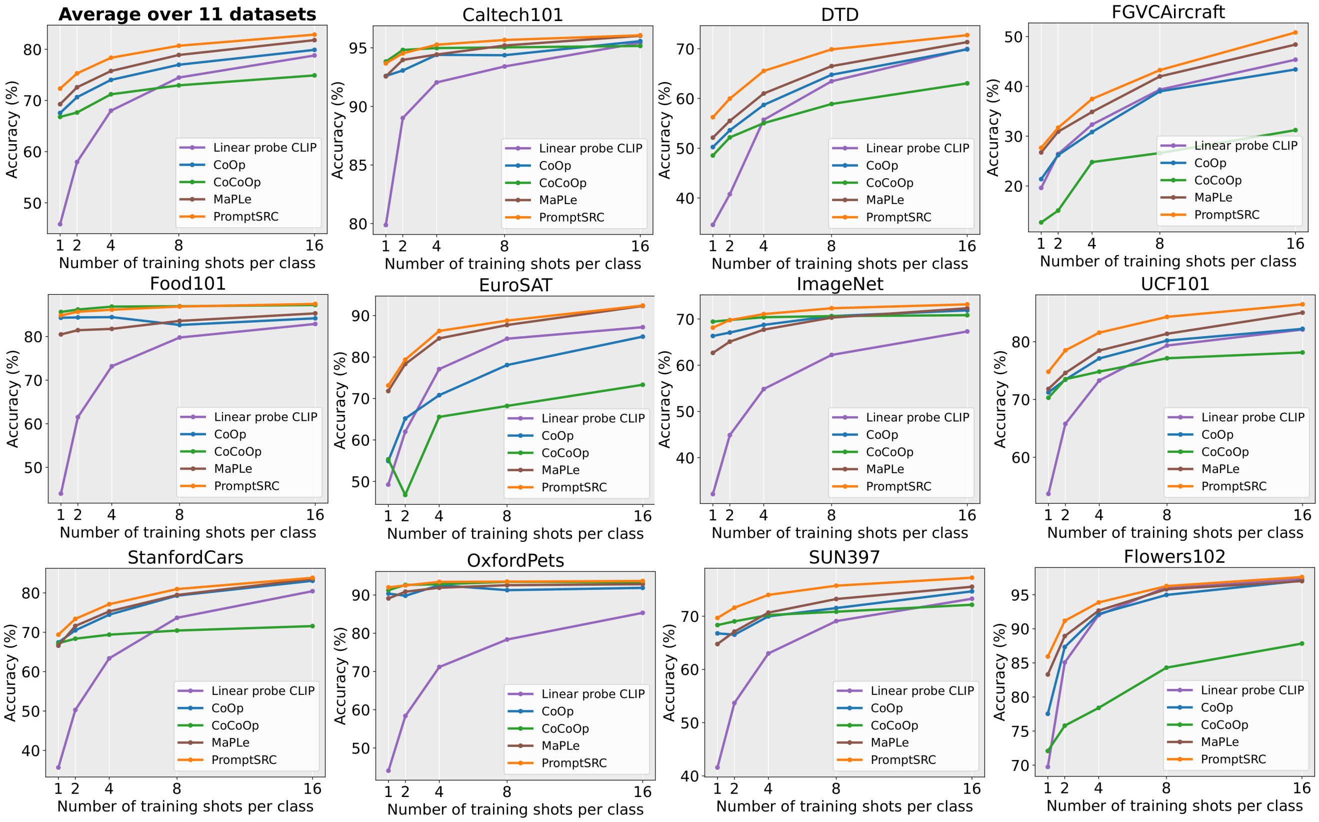

Figure 4: PromptSRC performance comparison in few-shot image recognition setting. All methods are trained on ViT-B/16 CLIP backbone using their best settings. PromptSRC demonstrates consistent improvements over existing methods specifically for lesser shots *i.e.* $K = 1, 2, 4$. On average, PromptSRC provides the highest performance gains for all shots. These results demonstrate that PromptSRC learns complementary task-agnostic general features from frozen CLIP without being restricted from learning downstream task representations.

	Source	**Target**										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	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	70.72	93.53	**90.49**	65.57	**72.23**	**86.20**	**24.74**	67.01	46.49	**48.06**	68.69	**66.30**
PromptSRC	71.27	**93.60**	90.25	**65.70**	70.25	86.15	23.90	67.10	**46.87**	45.50	**68.75**	65.81

Table 3: Cross-dataset benchmark evaluation. PromptSRC achieves overall favourable performance.

as for $K = 1, 2$ for almost all datasets. This demonstrates that PromptSRC regulates prompts against overfitting without restricting the prompts to learn task-specific knowledge.

4.5. Cross Dataset Evaluation

We compare our cross-dataset performance with previous methods in Table 3. On the source dataset, PromptSRC performs comparably to other methods. In comparison with CoOp and CoCoOp, PromptSRC shows competitive performance and achieves better generalization in 8/10 and 7/10 datasets respectively. Compared with MaPLe, PromptSRC shows improved performance in 5/10 datasets while utilizing significantly less tunable parameters (46K vs 3.55M).

	Source	**Target**				
	ImageNet	-V2	-S	-A	-R	Avg.
CLIP	66.73	60.83	46.15	47.77	73.96	57.18
CoOp	**71.51**	64.20	47.99	49.71	75.21	59.28
Co-CoOp	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe	70.72	64.07	49.15	50.90	76.98	60.27
PromptSRC	71.27	**64.35**	**49.55**	**50.90**	**77.80**	**60.65**

Table 4: Domain generalization. Prompt learning methods are trained on imageNet and evaluated on datasets with domain shifts.

4.6. Domain Generalization Experiments

Table 4 summarizes the results of PromptSRC and previous methods on out-of-distribution datasets. We directly evaluate our model trained on ImageNet. On target datasets, PromptSRC consistently outperforms all existing methods, with an overall highest average accuracy of 60.65%. This suggests that our self-regulating framework favors better generalization for datasets with domain shifts.

Method	Base Acc.	Novel Acc.	HM
1: Independent V-L prompting (IVLP)	84.21	71.79	77.51
2: IVLP + Cosine similarity	84.47	74.51	79.17
3: IVLP + Mean square error (MSE)	**84.59**	74.68	79.33
4: IVLP + $L1$	84.42	**74.99**	**79.43**

Table 5: Effect of matching losses for $\mathcal{L}_{\text{SCL-image}}$ and $\mathcal{L}_{\text{SCL-image}}$ consistency objectives. $L1$ matching loss provides highest HM.

Method	Base Acc.	Novel Acc.	HM
1: Exponential moving average	83.09	76.15	79.47
2: Equal weighting (averaging)	83.50	**76.47**	79.83
3: GPA (Ours)	**84.26**	76.10	**79.97**

Table 6: Ablation on prompt ensembling techniques. Gaussian weighted prompt aggregation (GPA) provides better performance.

Method	GFLOP (train)	GFLOP (test)	Train time (min)	FPS	HM
CoOp	162.5	162.5	10.08	1344	71.66
CoCoOp	162.5	162.5	39.53	15.08	75.83
IVLP	162.8	162.8	12.01	1380	77.51
PromptSRC	179.6	162.8	13.13	1380	**79.97**

Table 7: PromptSRC compute cost comparison using SUN397 dataset. Training time for all methods is calculated for 10 epochs on a single A100 GPU on SUN397 dataset.

4.7. Ablative Analysis

Embedding consistency loss ablation: In Table 5, we ablate on the choice of matching loss metric used in our proposed feature level $\mathcal{L}_{\text{SCL}}$ loss constraints. For simplicity, we only incorporate $\mathcal{L}_{\text{SCL-image}}$ and $\mathcal{L}_{\text{SCL-text}}$ on top of the IVLP baseline. Generally, distance-based matching metrics outperform the cosine similarity metric in terms of generalization as they impose a much harder constraint. Overall, the $L1$ matching metric provides the highest HM.

Prompt ensembling: Table 6 shows ablation on various prompt ensembling techniques. Using equal weights for prompts reduces base class results as initial epoch prompts are not mature enough. In contrast, our proposed Gaussian weighted prompt aggregation results in the highest performance. Detailed ablation experiments for other hyperparameters are provided in Appendix C.

Training and inference compute cost analysis: In Table 7, we show the compute cost analysis of our approach and compare it with other prompting methods. PromptSRC's overall training GFLOPs are only 0.13x higher than baseline IVLP, while it maintains the same GFLOPs and throughput during inference. Pre-trained CLIP textual features are pre-computed and a single additional forward pass is required through image encoder to compute pre-trained CLIP visual features for our mutual agreement maximization technique. Training time of PromptSRC is 9.3% longer than IVLP which is significantly lower than CoCoOp. We use 4 vision and text prompts similar to the IVLP.

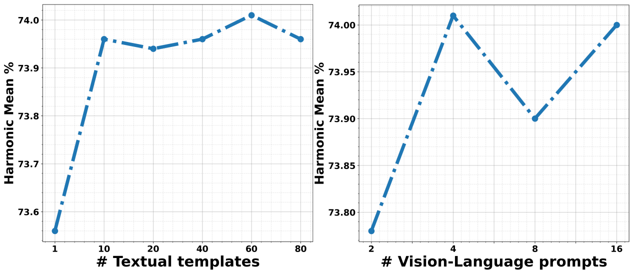

Figure 5: Ablation study on the number of textual prompts for textual diversity (left) and prompt token length (right) on ImageNet.

Prompt Length: Fig. 5 (right) shows the effect of prompt token length on the harmonic mean. Overall, the performance increases as prompt length increases. Using 4 vision-language prompts provides the highest harmonic mean.

No. of templates in textual diversity: In Fig. 5 (left), we ablate on the number of text prompt templates for textual diversity. We note that increasing the number of textual templates for textual diversity generally increases the performance. This suggests that adding textual diversity using multiple templates for pre-trained features provides more rich supervision for the learned prompted features.

5. Conclusion

Prompt learning has emerged as an effective paradigm for adapting foundational VL models like CLIP. However, the prompts learned by the majority of existing methods inherently tend to overfit task-specific objectives and consequently compromise the inherent generalization ability of CLIP. Our work proposes a self-regulating prompt learning framework that addresses the prompt overfitting problem for better generalization. We show it is critical to guide the training trajectory of prompts by explicitly encouraging its mutual agreement with the frozen model through self-consistency constraints supplemented by incorporating textual diversity. We also propose a self-ensembling strategy for prompts that appropriately aggregates them via a Gaussian-weighted approach over the course of training. Extensive evaluations on multiple benchmarks show the benefit of our self-regulating approach for prompt learning.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022. 2, 4

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 7

[3] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with op-

timal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 1, 3

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 7

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshop*, pages 702–703, 2020. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1, 6, 7

[7] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Variational prompt tuning improves generalization of vision-language models. *arXiv preprint arXiv:2210.02390*, 2022. 2, 7

[8] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, pages 11583–11592, 2022. 2

[9] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pages 19358–19369, 2023. 13, 14

[10] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshop*, pages 178–178. IEEE, 2004. 7

[11] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *ECCV*, 2022. 1

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 2

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[14] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *J-STARS*, 12(7):2217–2226, 2019. 7

[15] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021. 7

[16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 7

[17] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 1

[18] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *arXiv preprint arXiv:2208.05592*, 2022. 3, 5

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456. pmlr, 2015. 3

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 1, 2

[21] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2

[22] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 2, 3, 6, 7

[23] Konwoo Kim, Michael Laskin, Igor Mordatch, and Deepak Pathak. How to adapt your large-scale vision-and-language model, 2022. 1

[24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV*, pages 554–561, 2013. 7

[25] Kyungmoon Lee, Sungyeon Kim, and Suha Kwak. Cross-domain ensemble distillation for domain generalization. In *ECCV*, pages 1–20. Springer, 2022. 3

[26] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 2

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[28] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, pages 5206–5215, 2022. 1, 3, 6, 7

[29] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, pages 7086–7096, 2022. 1

[30] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *ECCV*. Springer, 2022. 2

[31] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 7

[32] Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications. *arXiv preprint arXiv:2302.00389*, 2023. 2

[33] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE, 2008. 7

[34] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 7

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 6, 7, 12

[36] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18082–18091, 2022. 2

[37] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, pages 6545–6554, 2023. 3, 4, 13

[38] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *NeurIPS*, 2022. 2

[39] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 7

[40] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2209.07511*, 2022. 1

[41] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7

[42] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014. 3

[43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 3

[44] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, volume 32, 2019. 7

[45] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, 2022. 2

[46] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pages 139–149, 2022. 2

[47] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, pages 7959–7971, 2022. 3, 5

[48] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 7

[49] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[50] Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhiming Ma. Improved ood generalization via adversarial training and pretraing. In *ICML*, pages 11987–11997. PMLR, 2021. 3

[51] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[52] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *CVPR*, pages 6023–6032, 2019. 3

[53] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *ECCV*, 2022. 2

[54] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18123–18133, 2022. 2

[55] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3

[56] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *ECCV*, 2022. 2

[57] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022. 2

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1, 2, 3, 4, 6, 7

[59] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1, 2, 3, 4, 6, 7, 13

[60] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2

[61] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022. 2