MMLoP: Multi-Modal Low-Rank Prompting for Efficient Vision-Language Adaptation

Sajjad Ghiasvand[1], Haniyeh Ehsani Oskouie[2], Mahnoosh Alizadeh[1], and Ramtin Pedarsani[1]

[1] University of California, Santa Barbara
{sajjad,alizadeh,ramtin}@ucsb.edu
[2] University of California, Los Angeles
haniyeh@cs.ucla.edu

Abstract. Prompt learning has become a dominant paradigm for adapting vision-language models (VLMs) such as CLIP to downstream tasks without modifying pretrained weights. While extending prompts to both vision and text encoders across multiple transformer layers significantly boosts performance, it dramatically increases the number of trainable parameters, with state-of-the-art methods requiring millions of parameters and abandoning the parameter efficiency that makes prompt tuning attractive. In this work, we propose **MMLoP** (**M**ulti-**M**odal **Lo**w-Rank **P**rompting), a framework that achieves deep multi-modal prompting with only **11.5K trainable parameters**, comparable to early text-only methods like CoOp. MMLoP parameterizes vision and text prompts at each transformer layer through a low-rank factorization that constrains prompts to a compact subspace, providing parameter efficiency while motivating the need for our complementary regularization components. To further close the accuracy gap with state-of-the-art methods, we introduce three complementary components: a self-regulating consistency loss that anchors prompted representations to frozen zero-shot CLIP features at both the feature and logit levels, a uniform drift correction that removes the global embedding shift induced by prompt tuning to preserve class-discriminative structure, and a shared up-projection that couples vision and text prompts through a common low-rank factor to enforce cross-modal alignment. Extensive experiments across three benchmarks and 11 diverse datasets demonstrate that MMLoP achieves a highly favorable accuracy-efficiency tradeoff, outperforming the majority of existing methods including those with orders of magnitude more parameters, while achieving a harmonic mean of 79.70% on base-to-novel generalization. Code is available at https://github.com/sajjad-ucsb/MMLoP.

Keywords: Vision-Language Models · Prompt Learning · Parameter Efficiency · Low-Rank Adaptation · Few-Shot Learning

1 Introduction

Large-scale vision-language models (VLMs) such as CLIP [43] and ALIGN [21] have emerged as powerful foundations for multi-modal understanding, enabling

strong zero-shot transfer across a wide range of downstream tasks including image classification [74], object detection [30, 37, 38], and semantic segmentation [32, 44]. Trained on hundreds of millions of image-text pairs via contrastive learning, these models acquire rich, generalizable representations that can be readily applied to new tasks without any additional training. However, fully fine-tuning such models on downstream tasks often degrades their original generalization ability, while simple linear probing yields suboptimal adaptation performance [24]. This has motivated the development of parameter-efficient adaptation strategies that preserve pretrained representations while enabling task-specific learning.

Prompt learning has emerged as a dominant paradigm for adapting VLMs without modifying pretrained weights [23,34,49]. Early methods such as CoOp [73] and CoCoOp [72] optimize continuous context vectors in the text branch of CLIP, achieving strong few-shot adaptation with as few as 2K–8K trainable parameters. Subsequent works extended this idea to both modalities through multi-modal deep prompting [13, 24, 47], where separate prompt tokens are learned at every transformer layer of both vision and text encoders. While this consistently boosts performance, it comes at a steep cost: MaPLe [24] requires over 3.5M trainable parameters, and more recent methods such as CoPrompt [47] push this even further. In pursuing higher accuracy, these methods abandon one of the core promises of prompt tuning: *parameter efficiency*.

This tension between accuracy and efficiency motivates our work. As illustrated in Fig. 1, existing methods that achieve competitive base-to-novel generalization and few-shot performance do so with orders of magnitude more trainable parameters than early text-only methods, while methods that remain parameter-efficient fall significantly short in accuracy. This raises a natural question: *can deep multi-modal prompting be made as parameter-efficient as early text-only methods, without sacrificing competitive accuracy?* A natural candidate is low-rank factorization [19, 65], which has proven highly effective for parameter-efficient adaptation of large language models. However, naively applying low-rank factorization to multi-modal prompts reduces expressive capacity without any mechanism for cross-modal alignment, leaving a significant accuracy gap that must be addressed through careful regularization design.

We present **MMLoP** (**M**ulti-**M**odal **Lo**w-Rank **P**rompting), a parameter-efficient framework for vision-language adaptation that achieves competitive accuracy with only **11.5K trainable parameters** — comparable to CoOp, yet benefiting from deep multi-modal prompting across both encoders. MMLoP parameterizes deep prompts via low-rank factorization, reducing the parameter count by over 300× relative to MaPLe. To compensate for the reduced expressiveness of the low-rank subspace, we introduce three complementary components: (i) a **Self-Regulating Consistency Loss** ($\mathcal{L}_{\text{SCL}}$) that anchors prompted features to the frozen zero-shot CLIP representations at both the feature and logit levels, preventing overfitting to base classes; (ii) a **Uniform Drift Correction** (UDC) that identifies and removes the global embedding shift induced by prompt tuning, preserving class-discriminative structure and improving generalization to novel classes; and (iii) a **Shared Up-Projection** that couples vision and text prompts

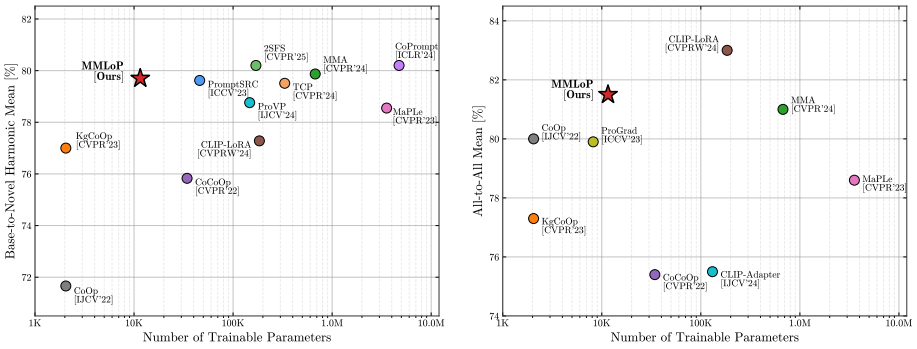

Fig. 1: Accuracy vs. number of trainable parameters for prompt learning methods on base-to-novel generalization (left) and all-to-all few-shot classification (right).

through a common low-rank factor at each layer, enforcing cross-modal alignment at virtually no additional parameter cost.

Our main contributions are as follows:

- We propose MMLoP, a multi-modal prompt learning framework that achieves deep vision-language prompting at CoOp-level parameter cost through low-rank factorization of prompt matrices across transformer layers.
- We introduce three regularization components, a self-regulating consistency loss, uniform drift correction, and shared up-projection, that together recover the accuracy gap introduced by low-rank constraints while improving generalization to novel classes.
- We conduct extensive experiments across three benchmarks (base-to-novel generalization, domain generalization, and all-to-all few-shot classification) on 11 diverse datasets. MMLoP outperforms 16 of 19 compared methods, including those requiring up to $\sim$300$\times$ more trainable parameters, achieving a harmonic mean of 79.70% on base-to-novel generalization and a mean accuracy of 81.5% on all-to-all few-shot classification, all with only 11.5K trainable parameters.

2 Related Work

Vision-Language Models. Large-scale vision-language models (VLMs) [21, 31, 43, 63] have emerged as powerful foundations for multi-modal understanding by learning aligned visual and textual representations from massive image-text datasets. Models such as CLIP [43] and ALIGN [21] are trained on hundreds of millions of image-text pairs using contrastive learning, enabling strong zero-shot transfer to a wide range of downstream tasks including image classification [74], object detection [30, 37], and semantic segmentation [32]. However, fully fine-tuning these large-scale models on downstream tasks often degrades their original

generalization ability, while simple linear probing yields suboptimal adaptation performance [24]. This has motivated the development of parameter-efficient adaptation strategies that preserve the pretrained representations while enabling task-specific learning.

Prompt Learning. Prompt learning, originating in NLP [23, 34, 49], has become a dominant paradigm for adapting VLMs to downstream tasks without modifying the pretrained weights. These methods can be broadly categorized into three forms: textual prompt learning [3, 7, 35, 36, 41, 51, 61, 62, 72, 73, 75] that optimizes continuous prompt vectors in the language branch of CLIP; visual prompt learning [1, 22, 29, 54, 57] that introduces learnable tokens into the visual input space while keeping pretrained backbones frozen; and multi-modal prompt learning [4, 11, 14, 24, 25, 28, 33, 47, 58, 60, 67, 70] that applies prompts to both vision and language branches for improved cross-modal alignment. Text-only methods such as CoOp [73], CoCoOp [72], and KgCoOp [61] are highly parameter-efficient but adapt only a single modality. Multi-modal approaches such as MaPLe [24], CoPrompt [47], and MMRL [13] achieve stronger performance by learning deep prompts across both encoders, but at the cost of significantly increased parameter counts. *In this work, we aim to bridge this gap by reducing the trainable parameter count of deep multi-modal prompting to be comparable with early text-only methods like CoOp, while maintaining competitive accuracy and generalization with state-of-the-art approaches.*

Low-Rank Factorization for VLMs. Inspired by LoRA [19], recent work has explored low-rank decompositions for parameter-efficient adaptation of VLMs, falling into two camps depending on whether low-rank constraints are applied to backbone *weight matrices* or to learnable *prompt tokens*. Weight-space methods modify the pretrained weights during training: CLIP-LoRA [65] adapts the attention projections of both CLIP encoders, Comp-LoRA [53] mitigates the resulting forgetting via orthogonal gradient constraints, Block-LoRA [71] shares sub-matrices across layers to reduce redundancy, and AdvCLIP-LoRA [12] extends CLIP-LoRA to the adversarial setting. Prompt-space methods instead keep the backbone frozen: DIP [15] applies low-rank constraints to CLIP prompt matrices with specialized initialization, reaching competitive accuracy with fewer than 0.5K parameters. MMLoP also belongs to the prompt-space category, but uniquely applies low-rank factorization to deep prompts in *both* encoders and couples them via a shared up-projection that enforces cross-modal alignment at no additional parameter cost. *Unlike CLIP-LoRA and its variants, MMLoP never modifies the frozen CLIP backbone, retaining full zero-shot compatibility while reducing trainable parameters by* $\sim$16$\times$ *relative to CLIP-LoRA and outperforming it on base-to-novel generalization.*

3 Preliminaries

CLIP. We denote the CLIP image and text encoders as f and g, respectively, with pretrained parameters $\theta_{\text{CLIP}} = \{\theta_f, \theta_g\}$. Given an input image $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$, it is divided into M patches and projected to produce patch tokens. A learnable

class token $\boldsymbol{e}_{\text{cls}}$ is appended, forming $\tilde{\boldsymbol{X}} = \{\boldsymbol{e}_{\text{cls}}, \boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_M\}$. The image encoder produces a visual feature $\tilde{\boldsymbol{f}} = f(\tilde{\boldsymbol{X}}, \theta_f) \in \mathbb{R}^d$. On the text side, a class label c_k is wrapped in a template such as "a photo of a {class}", forming $\tilde{\boldsymbol{Y}}_k = \{\boldsymbol{t}_{SOS}, \boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_N, c_k, \boldsymbol{t}_{EOS}\}$, where $\{t_n\}_{n=1}^N$ are word embeddings of the template, and t_{SOS}, t_{EOS} are the start and end tokens. The text encoder produces a textual feature $\tilde{\boldsymbol{g}}_k = g(\tilde{\boldsymbol{Y}}_k, \theta_g) \in \mathbb{R}^d$. For zero-shot inference, predictions are made by matching image features with textual features of all C classes:

$$p(y = k \mid \boldsymbol{X}) = \frac{\exp(\text{sim}(\tilde{\boldsymbol{g}}_k, \tilde{\boldsymbol{f}})/\tau)}{\sum_{i=1}^{C} \exp(\text{sim}(\tilde{\boldsymbol{g}}_i, \tilde{\boldsymbol{f}})/\tau)}, \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is the temperature.

Prompt Learning for CLIP. Instead of hand-crafted templates, prompt learning [72, 73] replaces the fixed text tokens with T learnable context vectors $\boldsymbol{P}_t = \{\boldsymbol{p}_t^1, \boldsymbol{p}_t^2, \cdots, \boldsymbol{p}_t^T\}$, so that the textual input becomes $\tilde{\boldsymbol{Y}}_p = \{\boldsymbol{t}_{SOS}, \boldsymbol{P}_t, \boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_N, c_k, \boldsymbol{t}_{EOS}\}$. The prompted textual feature is then $\tilde{\boldsymbol{g}}_p = g(\tilde{\boldsymbol{Y}}_p, \theta_g)$. Only the prompt vectors $\boldsymbol{P}_t$ are optimized via the cross-entropy loss while θ_{CLIP} remains frozen.

Independent Vision-Language Prompting (IVLP). IVLP [45] extends prompt learning to both modalities by appending V visual prompts $\boldsymbol{P}_v = \{\boldsymbol{p}_v^1, \boldsymbol{p}_v^2, \cdots, \boldsymbol{p}_v^V\}$ and T textual prompts $\boldsymbol{P}_t = \{\boldsymbol{p}_t^1, \boldsymbol{p}_t^2, \cdots, \boldsymbol{p}_t^T\}$. The image encoder processes $\tilde{\boldsymbol{X}}_p = \{\boldsymbol{P}_v, \boldsymbol{e}_{\text{cls}}, \boldsymbol{e}_1, \cdots, \boldsymbol{e}_M\}$ to produce $\tilde{\boldsymbol{f}}_p = f(\tilde{\boldsymbol{X}}_p, \theta_f)$, while the text encoder processes $\tilde{\boldsymbol{Y}}_p = \{\boldsymbol{t}_{SOS}, \boldsymbol{P}_t, \boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_N, c_k, \boldsymbol{t}_{EOS}\}$ to produce $\tilde{\boldsymbol{g}}_p = g(\tilde{\boldsymbol{Y}}_p, \theta_g)$. In its deep prompting variant, separate prompt sets $\boldsymbol{P}_v^{(l)}$ and $\boldsymbol{P}_t^{(l)}$ are learned at each transformer layer $l \in \{1, \cdots, L\}$. The full set of learnable parameters is $\boldsymbol{P} = \{\boldsymbol{P}_v^{(l)}, \boldsymbol{P}_t^{(l)}\}_{l=1}^L$, optimized with:

$$\mathcal{L}_{\text{CE}} = \arg\min_{\boldsymbol{P}} \; \mathbb{E}_{(\boldsymbol{X}, y) \sim \mathcal{D}} \; \mathcal{L}\big(\text{sim}(\tilde{\boldsymbol{f}}_p, \tilde{\boldsymbol{g}}_p), \; y\big). \tag{2}$$

While effective, IVLP learns vision and text prompts *independently*—the parameters $\boldsymbol{P}_v^{(l)}$ and $\boldsymbol{P}_t^{(l)}$ share no structure, providing no mechanism for cross-modal interaction during prompt optimization. This independence limits the model's ability to learn coordinated vision-language representations and can lead to overfitting on base classes at the expense of generalization to novel classes.

4 Proposed Algorithm

4.1 Motivation

Deep multi-modal prompting [13, 24, 59] significantly boosts few-shot recognition performance over text-only methods [61, 73], but dramatically increases the number of trainable parameters — MaPLe [24] requires over 3.5M parameters, while CoPrompt [47] pushes this even further. In pursuing higher accuracy, these approaches abandon one of the core promises of prompt tuning: *parameter efficiency*. MMLoP addresses this by parameterizing deep prompts through a

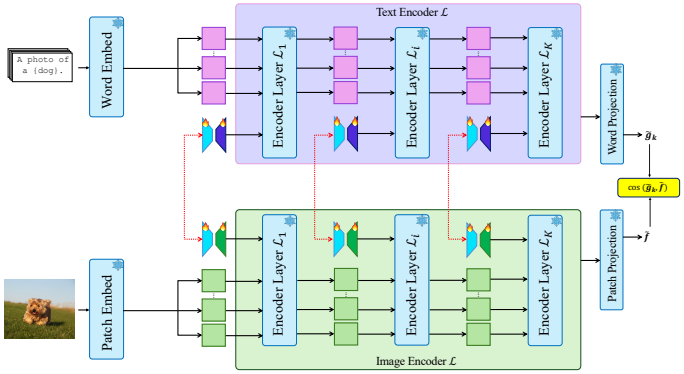

Fig. 2: Overview of MMLoP. Both text and image encoders are equipped with deep low-rank prompts at each transformer layer, with vision and text prompts sharing a common up-projection matrix $\boldsymbol{U}^{(l)}$ for cross-modal alignment. Snowflake icons indicate frozen CLIP parameters. The self-regulating consistency loss is omitted for clarity.

low-rank factorization inspired by [12, 19], reducing the parameter count to CoOp-level while maintaining competitive performance. To further close the gap with state-of-the-art methods, we introduce three key components: (i) a *self-regulating consistency loss* that prevents the prompted model from drifting away from CLIP's pretrained representations, (ii) a *uniform drift correction* that removes the global embedding shift induced by prompt tuning, and (iii) a *shared up-projection* that couples vision and text prompts through a common low-rank factor, enforcing cross-modal alignment.

4.2 Low-Rank Prompt Parameterization

Instead of learning full-rank prompt matrices $\boldsymbol{P}_v^{(l)} \in \mathbb{R}^{V \times d_v}$ and $\boldsymbol{P}_t^{(l)} \in \mathbb{R}^{T \times d_t}$ at each layer l, we parameterize the prompts through a low-rank factorization. Drawing inspiration from [19], we decompose each prompt matrix into the product of two low-rank factors:

$$\boldsymbol{P}_v^{(l)} = \boldsymbol{U}_v^{(l)} \boldsymbol{V}_v^{(l)}, \quad \boldsymbol{P}_t^{(l)} = \boldsymbol{U}_t^{(l)} \boldsymbol{V}_t^{(l)}, \tag{3}$$

where $\boldsymbol{U}_v^{(l)} \in \mathbb{R}^{V \times r}$ and $\boldsymbol{U}_t^{(l)} \in \mathbb{R}^{T \times r}$ are the up-projection matrices, $\boldsymbol{V}_v^{(l)} \in \mathbb{R}^{r \times d_v}$ and $\boldsymbol{V}_t^{(l)} \in \mathbb{R}^{r \times d_t}$ are the down-projection matrices, and $r \ll \min(d_v, d_t)$ is the rank.

This factorization constrains each prompt to lie in a rank-r subspace, reducing expressive capacity while providing a foundation for parameter efficiency. As shown in Table 4, this constraint alone is insufficient for competitive performance, which motivates our three complementary regularization components. The prompted image and text inputs at layer l follow the same structure as IVLP:

Algorithm 1 MMLoP: Multi-Modal Low-Rank Prompting

Require: Training data $\mathcal{D}$, frozen CLIP encoders f, g, frozen zero-shot text features $\{\tilde{g}_k\}_{k=1}^{C}$, number of classes C, rank r, prompt depth L, loss weights λ_1, λ_2
Ensure: Trained low-rank factors $\{U^{(l)}, V_v^{(l)}, V_t^{(l)}\}_{l=1}^{L}$

Initialize $U^{(l)} \in \mathbb{R}^{T\times r}$, $V_v^{(l)} \in \mathbb{R}^{r\times d_v}$, $V_t^{(l)} \in \mathbb{R}^{r\times d_t}$ with $\mathcal{N}(0, 0.05)$ for $l = 1, \ldots, L$
for each training iteration **do**
  Sample mini-batch $\{(x_i, y_i)\}$ from $\mathcal{D}$
  // Construct low-rank prompts via shared up-projection
  for $l = 1, \ldots, L$ **do**
    $P_v^{(l)} \leftarrow U^{(l)} V_v^{(l)}$ ▷ Vision prompt at layer l
    $P_t^{(l)} \leftarrow U^{(l)} V_t^{(l)}$ ▷ Text prompt at layer l
  end for
  // Extract prompted features
  $\tilde{f}_p \leftarrow f\big(\{P_v^{(l)}\}, x_i\big), \quad \tilde{g}_k^p \leftarrow g\big(\{P_t^{(l)}\}, c_k\big)$ for all k
  // Uniform Drift Correction (UDC)
  $r_k \leftarrow \tilde{g}_k^p - \tilde{g}_k$ for all k
  $\bar{r} \leftarrow \frac{1}{C}\sum_{k=1}^{C} r_k$
  $\hat{g}_k \leftarrow \tilde{g}_k + (r_k - \bar{r}), \quad \hat{g}_k \leftarrow \hat{g}_k / \|\hat{g}_k\|$ for all k
  // Compute losses
  $\mathcal{L}_{\text{CE}} \leftarrow \text{CrossEntropy}\big(\{\text{sim}(\tilde{f}_p, \hat{g}_k)\}, y_i\big)$
  $\mathcal{L}_{\text{SCL-text}} \leftarrow \lambda_1 \|\hat{g}_k - \tilde{g}_k\|_1$
  $\mathcal{L}_{\text{SCL-image}} \leftarrow \lambda_2 \|\tilde{f}_p - \tilde{f}\|_1$
  $\mathcal{L}_{\text{SCL-logits}} \leftarrow \frac{1}{2}\mathcal{D}_{\mathcal{KL}}\big(\text{sim}(\tilde{f}_p, \hat{g}_k) \| \text{sim}(\tilde{f}, \tilde{g}_k)\big) + \frac{1}{2}\mathcal{D}_{\mathcal{KL}}\big(\text{sim}(\tilde{f}, \tilde{g}_k) \| \text{sim}(\tilde{f}_p, \hat{g}_k)\big)$
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{SCL-text}} + \mathcal{L}_{\text{SCL-image}} + \mathcal{L}_{\text{SCL-logits}}$
  Update $\{U^{(l)}, V_v^{(l)}, V_t^{(l)}\}_{l=1}^{L}$ via SGD on $\mathcal{L}$
end for
Inference: compute $\hat{g}_k$ via UDC, predict $\hat{y} = \arg\max_k \text{sim}(\tilde{f}_p, \hat{g}_k)$

$$\tilde{\boldsymbol{X}}_p^{(l)} = \{\boldsymbol{P}_v^{(l)}, \boldsymbol{e}_{\text{cls}}^{(l)}, \boldsymbol{e}_1^{(l)}, \ldots, \boldsymbol{e}_M^{(l)}\},$$
$$\tilde{\boldsymbol{Y}}_p^{(l)} = \{t_{\text{SOS}}^{(l)}, \boldsymbol{P}_t^{(l)}, t_1^{(l)}, \ldots, t_N^{(l)}, c_k^{(l)}, t_{\text{EOS}}^{(l)}\}, \tag{4}$$

where the superscript (l) denotes representations at layer l of the transformer. During training, the entire CLIP backbone θ_{CLIP} remains frozen and only the low-rank factors $\{\boldsymbol{U}_v^{(l)}, \boldsymbol{V}_v^{(l)}, \boldsymbol{U}_t^{(l)}, \boldsymbol{V}_t^{(l)}\}_{l=1}^{L}$ are optimized.[3]

4.3 Self-Regulating Consistency Loss (SCL)

While the low-rank parameterization reduces the risk of overfitting, prompt-tuned models can still drift away from the pretrained CLIP representations, harming

[3] This is the general (independent up-projection) parameterization of Eq. (3). In our default model, the vision and text prompts share a single up-projection (Sec. 4.5, Eq. (9)), so the optimized set reduces to $\{\boldsymbol{U}^{(l)}, \boldsymbol{V}_v^{(l)}, \boldsymbol{V}_t^{(l)}\}_{l=1}^{L}$, consistent with Algorithm 1.

generalization to unseen classes. To mitigate this, we incorporate a consistency regularization inspired by [25] that anchors the learned features to the frozen zero-shot CLIP features. Let $\tilde{\boldsymbol{f}}_p$ and $\tilde{\boldsymbol{g}}_p$ denote the image and text features produced by the prompted model, and let $\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{g}}$ denote the features from the frozen zero-shot CLIP model. We define three consistency terms.

Feature-level consistency. We penalize the deviation between the prompted and zero-shot features in both modalities using the L1 norm:

$$\mathcal{L}_{\text{SCL-image}} = \|\tilde{\boldsymbol{f}}_p - \tilde{\boldsymbol{f}}\|_1, \quad \mathcal{L}_{\text{SCL-text}} = \|\tilde{\boldsymbol{g}}_p - \tilde{\boldsymbol{g}}\|_1. \tag{5}$$

Logit-level consistency. We also regularize the output logit distributions to remain close to those of the zero-shot model. Unlike [25] which employs a standard (asymmetric) KL divergence for this purpose, we adopt a symmetric KL divergence, as we find it more effective in practice:

$$\mathcal{L}_{\text{SCL-logits}} = \frac{1}{2}\mathcal{D}_{\mathcal{KL}}\Big(\text{sim}(\tilde{\boldsymbol{f}}_p, \tilde{\boldsymbol{g}}_p) \parallel \text{sim}(\tilde{\boldsymbol{f}}, \tilde{\boldsymbol{g}})\Big) + \frac{1}{2}\mathcal{D}_{\mathcal{KL}}\Big(\text{sim}(\tilde{\boldsymbol{f}}, \tilde{\boldsymbol{g}}) \parallel \text{sim}(\tilde{\boldsymbol{f}}_p, \tilde{\boldsymbol{g}}_p)\Big). \tag{6}$$

This symmetric formulation penalizes divergence equally in both directions, treating the prompted and zero-shot distributions more uniformly and avoiding the asymmetric gradient behavior of the standard KL. An empirical comparison of both variants is provided in Table A3 in the Appendix. The total consistency loss is:

$$\mathcal{L}_{\text{SCL}} = \lambda_1 \, \mathcal{L}_{\text{SCL-text}} + \lambda_2 \, \mathcal{L}_{\text{SCL-image}} + \mathcal{L}_{\text{SCL-logits}}, \tag{7}$$

where λ_1 and λ_2 are weighting hyperparameters. The overall training objective is then: $\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{SCL}}$.

4.4 Uniform Drift Correction (UDC)

While the self-regulating consistency loss encourages the prompted model to stay close to CLIP's pretrained representations, prompt tuning can still induce a systematic shift that affects all class embeddings uniformly. Such a shift does not improve discrimination between classes—it is shared across the entire embedding space and therefore carries no class-specific signal. Rather, it reflects a form of base-class bias absorbed into the prompt during few-shot training, which disproportionately harms generalization to novel classes.

To address this, we propose a correction that removes the uniform component of the learned text feature shift throughout both training and evaluation. Let $\tilde{g}_k^p$ denote the prompted text feature for class k, and let $\tilde{g}_k$ denote the corresponding frozen zero-shot feature. We decompose the prompted feature into a zero-shot anchor and a residual: $r_k = \tilde{g}_k^p - \tilde{g}_k$. The mean residual $\bar{r} = \frac{1}{C}\sum_{k=1}^{C} r_k$ captures the uniform drift shared across all C training classes. We subtract this common component and reconstruct the corrected feature:

$$\hat{g}_k = \tilde{g}_k + (r_k - \bar{r}), \qquad \hat{g}_k \leftarrow \frac{\hat{g}_k}{\|\hat{g}_k\|}. \tag{8}$$

The corrected features $\hat{g}_k$ retain the class-specific adaptations learned by the prompt while eliminating the shared bias. Furthermore, since UDC is applied during training, $\mathcal{L}_{\text{SCL-text}}$ is computed on drift-corrected features $\hat{g}_k$ rather than raw prompted features. This means the consistency regularization acts on the class-discriminative residuals alone, making UDC and $\mathcal{L}_{\text{SCL}}$ mutually reinforcing rather than redundant: the consistency loss encourages meaningful class-specific adaptation while UDC ensures that any global offset is continuously removed. The correction requires no additional parameters and introduces no asymptotic overhead: it reuses the frozen zero-shot features already computed for $\mathcal{L}_{\text{SCL}}$ and the per-iteration text-encoder forward over all C classes that every CLIP-style few-shot method already performs to compute the cross-entropy loss, adding only an $\mathcal{O}(Cd)$ mean subtraction on top.

4.5 Cross-Modal Coupling via Shared Up-Projection

The low-rank parameterization in Eq. (3) with independent up-projections $\boldsymbol{U}_v^{(l)}$ and $\boldsymbol{U}_t^{(l)}$ already provides parameter efficiency. However, the vision and text prompts remain structurally decoupled—each modality's low-rank subspace is learned independently. We observe that by tying the up-projection matrices across modalities, we can introduce cross-modal interaction at virtually no additional cost. Specifically, we set $T = V$ and constrain the two modalities to share a single up-projection:

$$\boldsymbol{P}_v^{(l)} = \boldsymbol{U}^{(l)}\boldsymbol{V}_v^{(l)}, \quad \boldsymbol{P}_t^{(l)} = \boldsymbol{U}^{(l)}\boldsymbol{V}_t^{(l)}, \tag{9}$$

where $\boldsymbol{U}^{(l)} \in \mathbb{R}^{T\times r}$ is now shared between modalities. This design means that the vision and text prompts at each layer are constrained to share the same row space. The shared factor $\boldsymbol{U}^{(l)}$ determines the token-wise activation pattern common to both modalities, while the modality-specific factors $\boldsymbol{V}_v^{(l)}$ and $\boldsymbol{V}_t^{(l)}$ project this shared structure into the respective embedding spaces.

This coupling acts as an additional regularizer: gradient updates to $\boldsymbol{U}^{(l)}$ must simultaneously benefit both modalities, which discourages overfitting to modality-specific noise in the few-shot training data. As shown in our ablation study (Table 4), introducing the shared up-projection on top of the consistency loss and UDC further improves novel class accuracy by +0.86% and the harmonic mean by +0.50%, demonstrating that cross-modal structural alignment provides complementary benefits to feature-level regularization.

5 Empirical Results

5.1 Experiments Setup

We evaluate MMLoP on three benchmark settings: base-to-novel generalization, domain generalization, and all-to-all few-shot classification.

Base-to-Novel Generalization. Following the protocol established by CoOp [73] and CoCoOp [72], each dataset is equally split into base and novel

classes. The model is trained exclusively on base classes with 16 shots per class and evaluated on both base and novel classes. The harmonic mean (HM) of base and novel accuracy is reported as the primary metric. This setting evaluates whether the model can learn task-specific representations without sacrificing CLIP's inherent zero-shot generalization ability. We conduct experiments on 11 diverse image recognition datasets: ImageNet [6] and Caltech101 [9] for generic object recognition; OxfordPets [42], StanfordCars [26], Flowers102 [40], Food101 [2], and FGVCAircraft [39] for fine-grained classification; SUN397 [55] for scene recognition; UCF101 [50] for action recognition; DTD [5] for texture classification; and EuroSAT [16] for satellite image classification.

Domain Generalization. To assess robustness to distribution shifts, we train the model on ImageNet [6] and directly evaluate on four out-of-distribution variants: ImageNetV2 [46], ImageNet-Sketch [52], ImageNet-A [18], and ImageNet-R [17], each introducing a different type of domain shift.

All-to-All Few-Shot Classification. In this setting, train and test categories coincide, and we evaluate performance across different numbers of shots ($K = 1, 2, 4, 8, 16$) and different CLIP backbones (ViT-B/16 and ViT-B/32) on the same 11 datasets.

Implementation Details. We use a ViT-B/16-based CLIP model as the default backbone and report results averaged over 3 seeds. We adopt deep prompting with $T = V = 4$ prompt tokens and parameterize prompts through a rank-1 factorization ($r = 1$) with the shared up-projection. For domain generalization, we train the ImageNet source model on all classes with $K = 16$ shots in the first 3 transformer layers. For few-shot and base-to-novel settings, prompts are learned in the first 9 transformer layers. Low-rank prompts are randomly initialized with a normal distribution with standard deviation 0.05. The consistency loss weights are set to $\lambda_1 = 25$ and $\lambda_2 = 10$ for $\mathcal{L}_{\text{SCL-text}}$ and $\mathcal{L}_{\text{SCL-image}}$, respectively. We train for 30 epochs for base-to-novel and 50 epochs for the few-shot and domain generalization settings, using SGD with a learning rate of 0.0025. All hyperparameters are fixed across all datasets and benchmarks. For the zero-shot text features used in $\mathcal{L}_{\text{SCL}}$, we use an ensemble of $N = 60$ standard text templates provided in [43]. All experiments are conducted on four NVIDIA A6000 GPUs.

5.2 Main Results

Base-to-Novel Generalization. Table 1 reports results across 11 datasets under the base-to-novel generalization protocol, where models are trained on base classes with 16 shots and evaluated on both base and novel classes. MMLoP achieves an average harmonic mean of **79.70%**, outperforming a wide range of recent and more sophisticated methods, including PromptSRC (79.62%, 46K parameters), TCP (79.51%, 332K parameters), MetaPrompt (79.09%, 31K parameters), LoGoPrompt (79.03%), LASP (78.61%), MaPLe (78.55%, 3.55M parameters), ProVP (78.76%, 147K parameters), RPO (77.78%), and CLIP-LoRA (77.28%, 184K parameters). MMLoP also remains competitive with MMA (79.87%, 581K parameters), 2SFS (80.20%, 170K parameters), and CoPrompt (80.48%, 3.82M parameters) — methods that require **51**×, **15**×, and **332**× more

Table 1: Base-to-novel generalization results across multiple datasets. HM refers to harmonic mean.

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [43]	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp [73]	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp [72]	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
ProDA [36]	81.56	72.30	76.65	75.40	70.23	72.72	98.27	93.23	95.68	95.43	97.83	96.62
ProGrad [75]	82.48	70.75	76.16	77.02	66.66	71.46	98.02	93.89	95.91	95.07	97.63	96.33
KgCoOp [61]	80.73	73.60	77.00	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18
MaPLe [24]	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
IVLP [45]	84.21	71.79	77.51	77.00	66.50	71.37	98.30	93.20	95.68	94.90	97.20	96.04
PromptSRC [25]	84.23	75.48	79.62	77.60	70.73	74.01	97.93	93.83	95.84	95.47	97.40	96.43
LASP [3]	82.70	74.90	78.61	76.20	70.95	73.48	98.10	94.24	96.16	95.90	97.93	96.90
RPO [27]	81.13	75.00	77.78	76.60	71.57	74.00	97.97	94.37	96.03	94.63	97.50	96.05
ProVP [56]	85.20	73.22	78.76	75.82	69.21	72.36	98.92	94.21	96.51	95.87	97.65	96.75
LoGoPrompt [48]	84.47	74.24	79.03	76.74	70.83	73.66	98.19	93.78	95.93	96.07	96.31	96.18
MetaPrompt [69]	83.65	75.48	79.09	77.52	70.83	74.02	98.13	94.58	96.32	95.53	97.00	96.26
CLIP-LoRA [65]	85.32	70.63	77.28	77.58	68.76	72.91	98.19	93.05	95.55	94.36	95.71	95.03
TCP [62]	84.13	75.36	79.51	77.27	69.87	73.38	98.23	94.67	96.42	94.67	97.20	95.92
MMA [59]	83.20	76.80	79.87	77.31	71.00	74.02	98.40	94.00	96.15	95.40	98.07	96.72
CoPrompt [47]	84.00	77.23	80.48	77.67	71.27	74.33	98.27	94.90	96.55	95.67	98.10	96.87
2SFS [8]	85.55	75.48	80.20	77.71	70.99	74.20	98.71	94.43	96.52	95.32	97.82	96.55
MMLoP (Ours)	83.79	75.98	79.70	77.00	70.50	73.61	98.27	93.93	96.05	95.47	96.97	96.21

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [43]	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp [73]	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp [72]	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
ProDA [36]	74.70	71.20	72.91	97.70	68.68	80.66	90.30	88.57	89.43	36.90	34.13	35.46
ProGrad [75]	77.68	68.63	72.88	95.54	71.87	82.03	90.37	89.59	89.98	40.54	27.57	32.82
KgCoOp [61]	71.76	75.04	73.36	95.00	74.73	83.65	90.50	91.70	91.09	36.21	33.55	34.83
MaPLe [24]	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	91.38	37.44	35.61	36.50
IVLP [45]	79.53	71.47	75.28	97.97	72.10	83.07	89.37	90.30	89.83	42.60	25.23	31.69
PromptSRC [25]	78.50	75.40	76.92	97.90	76.37	85.80	90.87	91.53	91.20	42.70	37.07	39.69
LASP [3]	75.17	71.60	73.34	97.00	74.00	83.95	91.20	91.70	91.44	34.53	30.57	32.43
RPO [27]	73.87	75.53	74.69	94.13	76.67	84.50	90.33	90.83	90.58	37.33	34.20	35.70
ProVP [56]	80.43	67.96	73.67	98.42	72.06	83.20	90.32	90.91	90.61	47.08	29.87	36.55
LoGoPrompt [48]	78.36	72.39	75.26	99.05	76.52	86.34	90.82	91.41	91.11	45.98	34.67	39.53
MetaPrompt [69]	76.34	75.01	75.48	97.66	74.49	84.52	90.74	91.85	91.29	40.14	36.51	38.24
CLIP-LoRA [65]	83.93	65.54	73.60	97.91	68.61	80.68	86.84	86.67	86.76	50.10	26.03	34.26
TCP [62]	80.80	74.13	77.32	97.73	75.57	85.23	90.57	91.37	90.97	41.97	34.43	37.83
MMA [59]	78.50	73.10	75.70	97.77	75.93	85.48	90.13	91.30	90.71	40.57	36.33	38.33
CoPrompt [47]	76.97	74.40	75.66	97.27	76.60	85.71	90.73	92.07	91.4	40.20	39.33	39.76
2SFS [8]	82.50	74.80	78.46	98.29	76.17	85.83	89.11	91.34	90.21	47.48	35.51	40.63
MMLoP (Ours)	77.77	74.83	76.27	97.63	76.73	85.93	90.70	91.70	91.19	42.17	34.60	38.01

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [43]	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp [73]	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp [72]	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
ProDA [36]	78.67	76.93	77.79	80.67	56.48	66.44	83.90	66.00	73.88	85.23	71.97	78.04
ProGrad [75]	81.26	74.17	77.55	77.35	52.35	62.45	90.11	60.89	72.67	84.33	74.94	79.35
KgCoOp [61]	80.29	76.53	78.36	77.55	54.99	64.35	85.64	64.34	73.48	82.89	76.67	79.65
MaPLe [24]	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
IVLP [45]	81.60	75.50	78.43	82.40	56.20	66.82	96.73	67.83	79.74	85.93	74.17	79.62
PromptSRC [25]	82.67	78.47	80.52	83.37	62.47	71.42	93.17	68.33	78.84	86.37	78.70	82.36
LASP [3]	80.70	78.60	79.63	81.40	58.60	68.14	94.60	77.78	85.36	84.77	78.03	81.26
RPO [27]	80.60	77.80	79.18	76.70	62.13	68.61	86.63	68.97	76.79	83.67	75.43	79.34
ProVP [56]	80.67	76.11	78.32	83.95	59.06	69.34	97.12	72.91	83.29	88.56	75.55	81.54
LoGoPrompt [48]	81.20	78.12	79.63	82.87	60.14	69.70	93.67	69.44	79.75	86.19	73.07	79.09
MetaPrompt [69]	82.26	79.04	80.62	83.10	58.05	68.35	93.53	75.21	83.38	85.33	77.72	81.35
CLIP-LoRA [65]	81.11	74.53	77.68	83.95	62.84	71.39	97.04	62.50	76.03	87.52	72.74	79.45
TCP [62]	82.63	78.20	80.35	82.77	58.07	68.25	91.63	74.73	82.32	87.13	80.77	83.83
MMA [59]	82.27	78.57	80.38	83.20	65.63	73.38	85.46	82.34	83.87	86.23	80.03	82.20
CoPrompt [47]	82.63	80.03	81.31	83.13	64.73	72.79	94.60	78.57	85.84	86.90	79.57	83.07
2SFS [8]	82.59	78.91	80.70	84.60	65.01	73.52	96.91	67.09	79.29	87.85	78.19	82.74
MMLoP (Ours)	82.40	78.13	80.21	82.67	60.87	70.11	92.37	79.10	85.22	85.23	78.37	81.66

Table 2: Domain generalization performance on ImageNet variants.

Dataset	CLIP [43]	UPT [66]	CoOp [73]	CoCoOp [72]	ProGrad [75]	KgCoOp [61]	TaskRes [64]	MaPLe [24]	RPO [27]	MMA [59]	CoPrompt [47]	**MMLoP**
ImageNet	66.73	72.63	71.51	71.02	72.24	71.20	73.90	70.72	71.67	71.00	70.80	71.00
ImageNet-V2	60.83	64.35	64.20	64.07	64.73	64.10	65.85	64.07	65.13	64.33	64.25	64.30
ImageNet-S	46.15	48.66	47.99	48.75	47.61	48.97	47.70	49.15	49.27	49.13	49.43	49.07
ImageNet-A	47.77	50.66	49.71	50.63	49.39	50.69	49.17	50.90	50.13	51.12	50.50	50.83
ImageNet-R	73.96	76.24	75.21	76.18	74.58	76.70	75.23	76.98	76.57	77.32	77.51	77.63
Average	57.18	59.98	59.28	59.91	59.07	60.11	59.49	60.28	60.27	60.48	60.42	60.46

trainable parameters, respectively. With only **11.5K trainable parameters**, MMLoP operates at a scale comparable to early text-only methods like CoOp (8.2K), while benefiting from deep multi-modal prompting across both encoders. Notably, MMLoP demonstrates strong novel class accuracy of **75.98%** on average, reflecting a +4.19% gain over the IVLP baseline and confirming that our regularization components effectively prevent overfitting to base classes. On Food101 it achieves a novel accuracy of 91.70%, and on EuroSAT a harmonic mean of 85.22%, substantially surpassing MaPLe (82.35%) and CoCoOp (71.21%). The performance gap relative to the top-performing methods is most apparent on fine-grained datasets such as FGVCAircraft, where discriminative visual features are harder to capture within a low-rank subspace. Overall, these results show that MMLoP lies on the *accuracy–efficiency Pareto frontier*: every compared method with higher HM uses substantially more parameters, and every method with fewer parameters achieves lower HM. MMLoP thus outperforms the majority of existing methods while using a fraction of their parameter budgets. We further compare against dedicated low-rank adaptation methods in Appendix C.

Domain Generalization. Table 2 evaluates robustness to distribution shift by training on ImageNet and directly evaluating on four out-of-distribution variants: ImageNet-V2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. MMLoP achieves a competitive average target accuracy of **60.46%**, outperforming the majority of compared methods including MaPLe (60.28%), CoPrompt (60.42%), and RPO (60.27%), while using only 11.5K trainable parameters. Notably, MMLoP attains the highest accuracy on ImageNet-R (**77.63%**) among all methods, suggesting that our consistency regularization and low-rank parameterization effectively preserve CLIP's pretrained representations and prevent source-domain overfitting. While MMA (60.48%) achieves a marginally higher average target accuracy by **0.02%**, it requires **51**× more trainable parameters. These results demonstrate that MMLoP generalizes robustly across domain shifts without sacrificing parameter efficiency.

All-to-All Few-Shot Classification. Table 3 reports all-to-all few-shot results on 11 datasets using the ViT-B/16 backbone across $K \in \{4, 8, 16\}$ shots, where train and test categories coincide. MMLoP demonstrates consistently strong performance across all shot settings, achieving mean accuracies of **81.5%**, **79.6%**, and **77.5%** for 16, 8, and 4 shots, respectively. At 16 shots, MMLoP ranks second overall, surpassing MMA (81.0%), MaPLe (78.6%), and all non-adapter baselines, while remaining competitive with CLIP-LoRA (83.0%) which uses dedicated LoRA adapters on the full backbone rather than prompt tokens. At 4 shots,

Table 3: *All-to-all* experiments, where train/test categories coincide, with the ViT-B/16 backbone, using $k = 4, 8, 16$ shots per class.

Shots	Method	ImageNet	SUN	AIR	ESAT	CARS	FOOD	PETS	FLWR	CAL	DTD	UCF	Mean
	Zero-Shot	66.7	62.6	24.7	47.5	65.3	86.1	89.1	71.4	92.9	43.6	66.7	65.1
	CoOp [73] (ctx=16)	71.9	74.9	43.2	85.0	82.9	84.2	92.0	96.8	95.8	69.7	83.1	80.0
	CoCoOp [72]	71.1	72.6	33.3	73.6	72.3	87.4	93.4	89.1	95.1	63.7	77.2	75.4
	TIP-Adapter-F [68]	73.4	76.0	44.6	85.9	82.3	86.8	92.6	96.2	95.7	70.8	83.9	80.7
	CLIP-Adapter [10]	69.8	74.2	34.2	71.4	74.0	87.1	92.3	92.9	94.9	59.4	80.2	75.5
	KgCoOp [61]	70.4	73.3	36.5	76.2	74.8	87.2	93.2	93.4	95.2	68.7	81.7	77.3
16	TaskRes [64]	73.0	76.1	44.9	82.7	83.5	86.9	92.4	97.5	95.8	71.5	84.0	80.8
	MaPLe [24]	71.9	74.5	36.8	87.5	74.3	87.4	93.2	94.2	95.4	68.4	81.4	78.6
	ProGrad [75]	72.1	75.1	43.0	83.6	82.9	85.8	92.8	96.6	95.9	68.8	82.7	79.9
	LP++ [20]	73.0	76.0	42.1	85.5	80.8	87.2	92.6	96.3	95.8	71.9	83.9	80.5
	CLIP-LoRA [65]	73.6	76.1	54.7	92.1	86.3	84.2	92.4	98.0	96.4	72.0	86.7	83.0
	MMA [59]	73.2	76.6	44.7	85.0	80.2	87.0	93.9	96.8	95.8	72.7	85.0	81.0
	MMLoP	71.9	75.8	45.5	92.2	80.4	87.4	93.9	97.2	95.7	72.6	84.3	81.5
	Zero-Shot	66.7	62.6	24.7	47.5	65.3	86.1	89.1	71.4	92.9	43.6	66.7	65.1
	CoOp [73] (ctx=16)	70.6	71.9	38.5	77.1	79.0	82.7	91.3	94.9	94.5	64.8	80.0	76.8
	CoCoOp [72]	70.8	71.5	32.4	69.1	70.4	87.0	93.3	86.3	94.9	60.1	75.9	73.8
	TIP-Adapter-F [68]	71.7	73.5	39.5	81.3	78.3	86.9	91.8	94.3	95.2	66.7	82.0	78.3
	CLIP-Adapter [10]	69.1	71.7	30.5	61.6	70.7	86.9	91.9	83.3	94.5	50.5	76.2	71.5
	KgCoOp [61]	70.2	72.6	34.8	73.9	72.8	87.0	93.0	91.5	95.1	65.6	80.0	76.0
8	TaskRes [64]	72.3	74.6	40.3	77.5	79.6	86.4	92.0	96.0	95.3	66.7	81.6	78.4
	MaPLe [24]	71.3	73.2	33.8	82.8	71.3	87.2	93.1	90.5	95.1	63.0	79.5	76.4
	ProGrad [75]	71.3	73.0	37.7	77.8	78.7	86.1	92.2	95.0	94.8	63.9	80.5	77.4
	LP++ [20]	72.1	75.1	39.0	78.2	76.4	86.8	91.8	95.2	95.5	67.7	81.9	78.2
	CLIP-LoRA [65]	72.3	74.7	45.7	89.7	82.1	83.1	91.7	96.3	95.6	67.5	84.1	80.3
	MMA [59]	71.9	74.7	38.9	69.7	76.8	86.4	92.9	94.6	95.6	66.9	82.9	77.4
	MMLoP	71.5	74.7	40.7	88.2	77.8	87.0	93.3	95.8	95.3	68.9	82.7	79.6
	Zero-Shot	66.7	62.6	24.7	47.5	65.3	86.1	89.1	71.4	92.9	43.6	66.7	65.1
	CoOp [73] (ctx=16)	68.8	69.7	30.9	69.7	74.4	84.5	92.5	92.2	94.5	59.5	77.6	74.0
	CoCoOp [72]	70.6	70.4	30.6	61.7	69.5	86.3	92.7	81.5	94.8	55.7	75.3	71.7
	TIP-Adapter-F [68]	70.7	70.8	35.7	76.8	74.1	86.5	91.9	92.1	94.8	59.8	78.1	75.6
	CLIP-Adapter [10]	68.6	68.0	27.9	51.2	67.5	86.5	90.8	73.1	94.0	46.1	70.6	67.7
	KgCoOp [61]	69.9	71.5	32.2	71.8	69.5	86.9	92.6	87.0	95.0	58.7	77.6	73.9
4	TaskRes [64]	71.0	72.7	33.4	74.2	76.0	86.0	91.9	85.0	95.0	60.1	76.2	74.7
	MaPLe [24]	70.6	71.4	30.1	69.9	70.1	86.7	93.3	84.9	95.0	59.0	77.1	73.5
	ProGrad [75]	70.2	71.7	34.1	69.6	75.0	85.4	92.1	91.1	94.4	59.7	77.9	74.7
	LP++ [20]	70.8	73.2	34.0	73.6	74.0	85.9	90.9	93.0	95.1	62.4	79.2	75.6
	CLIP-LoRA [65]	71.4	72.8	37.9	84.9	77.4	82.7	91.0	93.7	95.2	63.8	81.1	77.4
	MMA [59]	70.5	72.9	35.0	42.4	73.3	86.0	92.9	91.3	94.5	60.1	79.0	72.5
	MMLoP	70.8	73.1	36.1	84.5	74.9	86.1	93.2	93.3	95.2	64.7	80.2	77.5

MMLoP achieves the **highest mean accuracy of 77.5%** among all compared methods, outperforming CLIP-LoRA (77.4%) and LP++ (75.6%), highlighting the strength of our low-rank parameterization and consistency regularization in the extremely low-data regime. Notably, MMLoP achieves particularly strong results on EuroSAT across all shot settings (92.2%, and 84.5% for 16 and 4 shots respectively), demonstrating robust adaptation to specialized domains. These results validate that MMLoP achieves strong few-shot adaptation across diverse datasets and data regimes, with only 11.5K trainable parameters. Additional results using the ViT-B/32 backbone are reported in Table A1. Few-shot curves for both ViT-B/16 and ViT-B/32 are visualized in Figs. A2 and A3 (Appendix).

5.3 Ablation Study

Table 4 validates the contribution of each proposed component through an incremental ablation study averaged over 11 datasets. Starting from the IVLP baseline (77.51% HM), adding the low-rank factorization alone reduces both base and novel accuracy (77.06% HM), as the rank-1 constraint limits the prompt's

Table 4: Effect of our proposed regularization techniques. Results are averaged over 11 datasets.

Method	Base Acc.	Novel Acc.	HM
1: Independent V-L prompting	**84.21**	71.79	77.51
2: + low-rank prompting	83.70	71.39	77.06
3: + $\mathcal{L}_{\text{SCL}}$	83.60	74.48	78.78
4: + UDC	83.78	75.12	79.20
5: + Shared Up-Projection	83.79	**75.98**	**79.70**

Table 5: Hyperparameter sensitivity analysis. Results are averaged over 9 datasets.

Depth	Base	Novel	HM
7	83.63	74.19	78.63
9	84.44	**75.85**	**79.91**
11	**84.87**	74.58	79.39

Length	Base	Novel	HM
2	84.16	74.99	79.31
4	84.44	**75.85**	**79.91**
8	**84.83**	74.38	79.26

Rank	Base	Novel	HM
1	84.44	**75.85**	**79.91**
2	84.76	75.01	79.59
4	**84.80**	75.20	79.71

expressive capacity without any compensating regularization. Incorporating the Self-Regulating Consistency Loss ($\mathcal{L}_{\text{SCL}}$) yields a substantial gain in novel class accuracy from 71.39% to 74.48%, boosting the harmonic mean to 78.78% — confirming that explicit feature-level anchoring to the frozen CLIP representations is critical for generalization to unseen classes. Adding Uniform Drift Correction (UDC) further improves novel accuracy to 75.12% (HM: 79.20%) by removing the global embedding shift shared across all class embeddings. Finally, the Shared Up-Projection pushes novel accuracy to **75.98%** and the harmonic mean to **79.70%** by coupling vision and text prompts through a common low-rank factor, enforcing cross-modal alignment at virtually no additional parameter cost. We note that these gains in novel class generalization (+4.19% over IVLP) come at the cost of a modest reduction in base class accuracy (−0.42%), which is expected given that our regularization components explicitly discourage the model from over-specializing to the base training classes in favor of preserving CLIP's broader representational capacity. Per-dataset results are in Table A4 in Appendix.

5.4 Hyperparameter Sensitivity

Table 5 analyzes the sensitivity of MMLoP to three key hyperparameters: prompt depth, prompt length, and prompt factorization rank. For prompt depth, performance peaks at depth 9 (HM: 79.91%), with shallower prompting (depth 7) yielding a notable drop to 78.63%, while deeper prompting (depth 11) improves base accuracy (84.87%) but hurts novel accuracy (74.58%), suggesting overfitting to base classes when prompts are applied too deeply. For prompt length, depth 4 achieves the best harmonic mean (79.91%), with both shorter (length 2: 79.31%) and longer (length 8: 79.26%) configurations performing slightly worse, indicating that the model is not highly sensitive to this parameter around the chosen value. For prompt factorization rank, rank 1 achieves the best harmonic mean (79.91%) and the highest novel accuracy (75.85%), with higher ranks (2 and 4) improving

Table 6: Efficiency comparison averaged over 11 datasets (ViT-B/16, single A6000).

Method	Params (M)	VRAM (GB)	Train (min)	Train (ms/img)	Infer (ms/img)	FPS	GFLOP	HM
CoOp [73]	0.008	1.96	58.1	19.39	1.94	561.6	324.6	71.66
PromptSRC [25]	0.046	2.99	44.7	30.16	2.29	530.0	324.9	79.62
MaPLe [24]	3.555	2.27	7.8	19.75	2.03	545.4	324.7	78.55
MMA [59]	0.581	2.57	7.9	25.84	2.39	489.6	326.5	79.87
CoPrompt [47]	3.818	2.62	9.6	20.49	1.39	791.7	342.3	80.48
CLIP-LoRA [65]	0.184	4.39	41.4	34.32	2.27	487.8	324.7	77.28
Comp-LoRA [53]	0.184	4.62	41.6	33.20	1.83	552.0	346.2	78.77
Block-LoRA [71]	0.230	4.39	56.4	33.53	1.82	555.5	324.7	78.82
MMLoP (Ours)	0.012	2.33	24.7	20.28	1.79	638.0	324.9	79.70

base accuracy marginally but consistently hurting novel generalization — confirming that the stronger implicit regularization of a rank-1 factorization is beneficial for generalization to unseen classes. Overall, MMLoP is reasonably robust to hyperparameter choices, and all hyperparameters are fixed across all datasets and benchmarks without any dataset-specific tuning. Detailed per-dataset results are provided in Tables A5, A6, and A7 in the appendix.

5.5 Efficiency Analysis

Tab. 6 reports trainable parameters, peak VRAM, training and inference cost, and HM averaged over 11 datasets (ViT-B/16, single A6000). MMLoP achieves the lowest parameter count (0.012M, $\sim$50–330$\times$ fewer than multi-modal baselines) while remaining competitive across all system-level metrics: VRAM (2.33 GB) is lower than 6 of 8 baselines; per-iteration training cost (20.28 ms/img) matches CoOp/MaPLe and is $\sim$40% faster than the LoRA family. The lower total training time (Train min) of MaPLe, MMA, and CoPrompt reflects their use of very few epochs (5–8) to avoid overfitting, not per-iteration efficiency. Inference latency (1.79 ms/img) is faster than CoOp, MaPLe, PromptSRC, MMA, and CLIP-LoRA, demonstrating that UDC and the consistency loss add no inference-time overhead.

6 Conclusion

We presented MMLoP, a parameter-efficient multi-modal prompt learning framework that achieves deep vision-language prompting with only 11.5K trainable parameters. By parameterizing prompts through a low-rank factorization with a shared up-projection, and combining this with a self-regulating consistency loss and uniform drift correction, MMLoP enforces cross-modal alignment while preventing overfitting to base classes and preserving CLIP's pretrained representations. Extensive experiments across base-to-novel generalization, domain generalization, and few-shot classification show that MMLoP outperforms most existing methods at a fraction of their parameter budgets.

Acknowledgments

This work was supported by the National Science Foundation under Grant 2419982, Grant 2342253, and Grant 2236483.

References

1. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 (2022)
2. Bossard, L., Guillaumin, M., Gool, L.V.: Food-101–mining discriminative components with random forests. In: ECCV (2014)
3. Bulat, A., Tzimiropoulos, G.: Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In: CVPR (2023)
4. Cheng, S., Han, K.: Vamp: Variational multi-modal prompt learning for vision-language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
5. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR (2014)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
7. Du, Y., Sun, W., Snoek, C.: Ipo: Interpretable prompt optimization for vision-language models. In: NeurIPS (2024)
8. Farina, M., Mancini, M., Iacca, G., Ricci, E.: Rethinking few-shot adaptation of vision-language models in two stages. In: CVPR (2025)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: CVPR workshop (2004)
10. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. IJCV (2024)
11. Ghiasvand, S., Alizadeh, M., Pedarsani, R.: pfedmma: Personalized federated fine-tuning with multi-modal adapter for vision-language models. In: The Fourteenth International Conference on Learning Representations
12. Ghiasvand, S., Oskouie, H.E., Alizadeh, M., Pedarsani, R.: Few-shot adversarial low-rank fine-tuning of vision-language models. arXiv preprint arXiv:2505.15130 (2025)
13. Guo, Y., Gu, X.: Mmrl: Multi-modal representation learning for vision-language models. In: CVPR (2025)
14. Hao, F., He, F., Wu, F., Wang, T., Song, C., Cheng, J.: Task-aware clustering for prompting vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14745–14755 (2025)
15. Hao, T., Lyu, M., Chen, H., Zhao, S., Ding, X., Han, J., Ding, G.: Towards efficient vision-language tuning: More information density, more generalizability. arXiv preprint arXiv:2312.10813 (2023)
16. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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 (2019)
17. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021)
18. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15262–15271 (2021)
19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

20. Huang, Y., Shakeri, F., Dolz, J., Boudiaf, M., Bahig, H., Ben Ayed, I.: Lp++: A surprisingly strong linear probe for few-shot clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23773–23782 (2024)
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
22. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: ECCV (2022)
23. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? TACL (2020)
24. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: CVPR (2023)
25. Khattak, M.U., Wasim, S.T., Naseer, M., Khan, S., Yang, M.H., Khan, F.S.: Self-regulating prompts: Foundational model adaptation without forgetting. In: ICCV (2023)
26. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: ICCV (2013)
27. Lee, D., Song, S., Suh, J., Choi, J., Lee, S., Kim, H.J.: Read-only prompt optimization for vision-language few-shot learning. In: 2023 IEEE/CVF International Conference on Computer Vision, ICCV 2023. pp. 1401–1411. Institute of Electrical and Electronics Engineers Inc. (2023)
28. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: Multimodal prompting with missing modalities for visual recognition. In: CVPR (2023)
29. Li, F., Jiang, Q., Zhang, H., Ren, T., Liu, S., Zou, X., Xu, H., Li, H., Yang, J., Li, C., et al.: Visual in-context prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12861–12871 (2024)
30. Li, J., Zhang, J., Li, J., Li, G., Liu, S., Lin, L., Li, G.: Learning background prompts to discover implicit knowledge for open vocabulary object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16678–16687 (2024)
31. Li, J., He, X., Wei, L., Qian, L., Zhu, L., Xie, L., Zhuang, Y., Tian, Q., Tang, S.: Fine-grained semantically aligned vision-language pre-training. Advances in neural information processing systems **35**, 7290–7303 (2022)
32. Li, X., Yuan, H., Li, W., Ding, H., Wu, S., Zhang, W., Li, Y., Chen, K., Loy, C.C.: Omg-seg: Is one model good enough for all segmentation? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 27948–27959 (2024)
33. Li, Y., Quan, R., Zhu, L., Yang, Y.: Efficient multimodal fusion via interactive prompting. In: CVPR (2023)
34. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM computing surveys **55**(9), 1–35 (2023)
35. Liu, X., Wang, D., Fang, B., Li, M., Xu, Y., Duan, Z., Chen, B., Zhou, M.: Patch-prompt aligned bayesian prompt tuning for vision-language models. In: The 40th Conference on Uncertainty in Artificial Intelligence
36. Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: CVPR (2022)
37. Maaz, M., Rasheed, H., Khan, S., Khan, F.S., Anwer, R.M., Yang, M.H.: Class-agnostic object detection with multi-modal transformer. In: European conference on computer vision. pp. 512–531. Springer (2022)

38. Mahjourian, N., Nguyen, V.: Multimodal object detection using depth and image data for manufacturing parts. In: International Manufacturing Science and Engineering Conference. vol. 89022, p. V002T16A001. American Society of Mechanical Engineers (2025)
39. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
40. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP (2008)
41. Park, J., Ko, J., Kim, H.J.: Prompt learning via meta-regularization. In: CVPR (2024)
42. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: CVPR (2012)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
44. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: CVPR (2022)
45. Rasheed, H., Khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Fine-tuned clip models are efficient video learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6545–6554 (2023)
46. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
47. Roy, S., Etemad, A.: Consistency-guided prompt learning for vision-language models. In: The Twelfth International Conference on Learning Representations
48. Shi, C., Yang, S.: Logoprompt: Synthetic text images can be good visual prompts for vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2932–2941 (2023)
49. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: EMNLP (2020)
50. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
51. Tian, X., Zou, S., Yang, Z., Zhang, J.: Argue: Attribute-guided prompt tuning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28578–28587 (2024)
52. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in neural information processing systems **32** (2019)
53. Wang, Z., Dai, J., Li, K., Li, X., Guo, Y., Xiang, M.: Complementary subspace low-rank adaptation of vision-language models for few-shot classification. arXiv preprint arXiv:2501.15040 (2025)
54. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: CVPR (2022)
55. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)
56. Xu, C., Zhu, Y., Shen, H., Chen, B., Liao, Y., Chen, X., Wang, L.: Progressive visual prompt learning with contrastive feature re-formation. International Journal of Computer Vision **133**(2), 511–526 (2025)

57. Yang, L., Wang, Y., Li, X., Wang, X., Yang, J.: Fine-grained visual prompting. Advances in Neural Information Processing Systems **36**, 24993–25006 (2023)
58. Yang, L., Zhang, R.Y., Chen, Q., Xie, X.: Learning with enriched inductive biases for vision-language models. IJCV (2025)
59. Yang, L., Zhang, R.Y., Wang, Y., Xie, X.: Mma: Multi-modal adapter for vision-language models. In: CVPR (2024)
60. Yao, H., Zhang, R., Lyu, H., Zhang, Y., Xu, C.: Bi-modality individual-aware prompt tuning for visual-language model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
61. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: CVPR (2023)
62. Yao, H., Zhang, R., Xu, C.: Tcp: Textual-based class-aware prompt tuning for visual-language model. In: CVPR (2024)
63. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. In: International Conference on Learning Representations
64. Yu, T., Lu, Z., Jin, X., Chen, Z., Wang, X.: Task residual for tuning vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10899–10909 (2023)
65. Zanella, M., Ben Ayed, I.: Low-rank few-shot adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1593–1603 (2024)
66. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225 (2022)
67. Zhang, F., Zhou, T., Yao, J., Zhang, Y., Tsang, I.W., Wang, Y.: Decouple before align: Visual disentanglement enhances prompt tuning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
68. Zhang, R., Fang, R., Gao, P., Zhang, W., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. In: ECCV (2022)
69. Zhao, C., Wang, Y., Jiang, X., Shen, Y., Song, K., Li, D., Miao, D.: Learning domain invariant prompt for vision-language models. IEEE Transactions on Image Processing **33**, 1348–1360 (2024)
70. Zheng, H., Yang, S., He, Z., Yang, J., Huang, Z.: Hierarchical cross-modal prompt learning for vision-language models. In: ICCV (2025)
71. Zhou, C., Shen, Q., Yu, Z., Bu, J., Wang, H.: One head eight arms: Block matrix based low rank adaptation for clip-based few-shot learning. arXiv preprint arXiv:2501.16720 (2025)
72. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR (2022)
73. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV (2022)
74. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV (2022)
75. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: ICCV (2023)