MMRL: Multi-Modal Representation Learning for Vision-Language Models

Yuncheng Guo[1], Xiaodong Gu[2*]
Department of Electronic Engineering, Fudan University, Shanghai 200438, China
[1]23210720033@m.fudan.edu.cn, [2]xdgu@fudan.edu.cn

Abstract

*Large-scale pre-trained **V**ision-**L**anguage **M**odels (VLMs) have become essential for transfer learning across diverse tasks. However, adapting these models with limited few-shot data often leads to overfitting, diminishing their performance on new tasks. To tackle this issue, we propose a novel **M**ulti-**M**odal **R**epresentation **L**earning (MMRL) framework that introduces a shared, learnable, and modality-agnostic representation space. MMRL projects the space tokens to text and image representation tokens, facilitating more effective multi-modal interactions. Unlike previous approaches that solely optimize class token features, MMRL integrates representation tokens at higher layers of the encoders—where dataset-specific features are more prominent—while preserving generalized knowledge in the lower layers. During training, both representation and class features are optimized, with trainable projection layer applied to the representation tokens, whereas the class token projection layer remains frozen to retain pre-trained knowledge. Furthermore, a regularization term is introduced to align the class features and text features with the zero-shot features from the frozen VLM, thereby safeguarding the model's generalization capacity. For inference, a decoupling strategy is employed, wherein both representation and class features are utilized for base classes, while only the class features, which retain more generalized knowledge, are used for new tasks. Extensive experiments across 15 datasets demonstrate that MMRL outperforms state-of-the-art methods, achieving a balanced trade-off between task-specific adaptation and generalization. Code is available at https://github.com/yunncheng/MMRL.*

1. Introduction

Vision-Language Models (VLMs) [1, 15, 16, 24, 33, 34, 50], such as CLIP [34], have gained significant attention for their ability to leverage the rich, complementary information inherent in both textual and visual modalities.

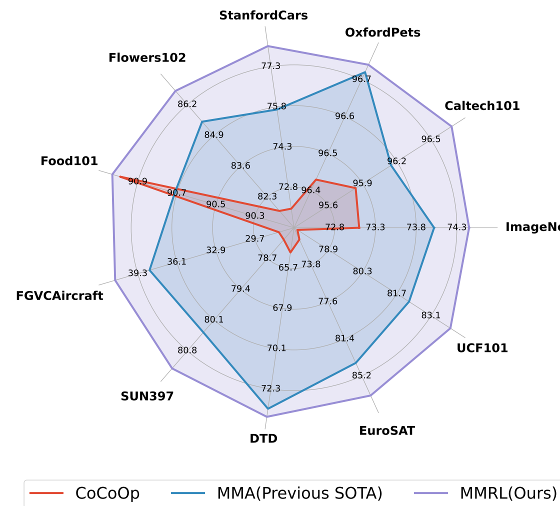

Figure 1. Comprehensive comparison of the harmonic mean performance between the previous sota method MMA and our proposed MMRL across 11 diverse datasets for base-to-novel generalization. Our method achieves the best on all datasets.

By constructing distinct encoders for images and text, and employing contrastive learning [39] on over 400 million image-text pairs, CLIP effectively captures complex visual-text relationships, demonstrating strong performance across various downstream tasks, such as medical image analysis [14, 44, 54], image and video captioning [2, 28, 37], and visual question answering [31, 41, 51]. Despite their versatility, VLMs encounter limitations in adapting to new tasks, as fine-tuning their large-scale architectures demands considerable computational resources.

To facilitate efficient adaptation of VLMs, strategies such as prompt engineering and ensembling [34] have shown potential. Specifically, prompt engineering involves crafting dataset-specific prompts, such as "A photo of a [CLASS], a type of pet." for the OxfordPets [32] dataset. Alternatively, ensembling can integrate multiple zero-shot classifiers by varying context prompts, *e.g.*, "A photo of a big [CLASS]." and "A photo of a small [CLASS].". Nonetheless, manual prompt design is time-consuming and requires substantial expertise, yet it does not guarantee the discovery of optimal prompts. To address this limitation,

[*]Corresponding author

CoOp [56] introduces prompt learning[21] where prompts are modeled as continuous learnable vectors, optimized during training while keeping VLM parameters fixed, thereby enabling efficient dataset adaptation. Recently MaPLe [17] has identified that prompt learning solely within the text modality may be sub-optimal. In response, it proposes a multi-modal prompt learning approach, embedding deep prompts into the lower layers of both VLM encoders via a coupling function to enhance alignment between visual and textual representations.

In addition to prompt learning, adapter-style learning methods offer a different adaptation pathway: rather than modifying input prompts, lightweight modules (*e.g.*, multi-layer perceptrons, MLPs) are integrated within VLMs to adjust extracted features for downstream datasets. CLIP-Adapter [10] exemplifies this approach by maintaining the frozen VLM while fine-tuning features via an MLP adapter added to the image encoder, which incorporates residual connections for feature fusion. Similar to MaPLe, MMA [47] proposes a multimodal adapter that refines the alignment between text and vision representations by aggregating features from diverse branches into a unified feature space, allowing gradient flow across branches. Notably, MMA reveals that different layers within VLM encoders capture varying characteristics: higher layers encode discriminative, dataset-specific information, while lower layers retain more generalizable features.

However, the current multimodal deep prompt learning method [17], which applies prompt concatenation at shallow layers, may compromise generalizable knowledge. This approach map visual prompts from text prompts, incorporating visual information via gradient propagation but ultimately remaining text-centric, with updates focused mainly on text prompts. Moreover, both prompt learning and adapter-style methods solely optimize class token features using task-specific objectives, such as cross-entropy loss. As a result, these methods are vulnerable to overfitting to specific data distributions or task categories when training data is scarce (*e.g.*, few-shot setting), leading to a decline in the inherent generalization and zero-shot learning capabilities of VLMs.

To address these challenges, we propose a novel multi-modal representation learning framework that distinguishes itself from conventional prompt learning and adapter-style methods. Specifically, we introduce a shared, learnable representation space that is independent of any modality within the higher layers of the encoder. This space serves as a bridge for multimodal interaction, mapping tokens from this space to both image and text representation tokens, which are then concatenated with the original encoder tokens to enable effective multimodal interaction. Our representation tokens are designed to learn dataset-specific knowledge from downstream tasks while the original classification token is regularized to retain a significant amount of generalizable knowledge. MMRL offers three key advantages: (1) an unbiased shared representation space that promotes balanced multimodal learning; (2) preservation of original VLM generalization by avoiding prompt integration at shallow encoder layers; and (3) Unlike prompt learning or adapter-style methods that refine only the class token features through learnable prompts or adapters, our approach supports decoupled inference across classes. During training, we prioritize optimizing representation token features, with their projection layer trainable, while that of the original class token remain fixed. To further preserve the generalizability of the class token, a regularization term aligns its features with the zero-shot features from the frozen VLM. For inference, we utilize both representation and class token features for base classes, while for unseen classes or new datasets, only the class token features are employed.

Our main contributions are summarized as follows:

- We introduce the **M**ulti-**M**odal **R**epresentation **L**earning (MMRL) framework, which incorporates a shared, unbiased, learnable space that bridges image and text modalities, facilitating multimodal interaction at the high layers of the original encoder.
- A decoupling strategy preserves VLM generalization by adapting representation tokens for downstream tasks while regularizing the original class token for new tasks.
- Extensive experiments demonstrate that MMRL substantially improves downstream adaptation and generalization, achieving superior performance over baselines.

2. Related Work

2.1. Vision-Language Models

Vision-language models (VLMs) have emerged as powerful tools for capturing rich multimodal representations, standing apart from traditional models that rely exclusively on visual or textual supervision. Recent advances in VLMs, such as CLIP [34], ALIGN [16], FILIP [50], KOSMOS [15, 33], and VILA [24], have demonstrated remarkable performance across a variety of tasks. These models typically learn joint image-language representations through self-supervised learning, leveraging large-scale architectures and massive collections of image-text pairs. For instance, CLIP is trained on a collection of 400 million image-text pairs, while ALIGN leverages an impressive 1.8 billion pairs. Although these pre-trained models excel at learning generalized representations, efficiently adapting them to specific downstream tasks remains a challenge.

2.2. Efficient Transfer Learning

Prompt learning methods have proven effective for adapting VLMs. CoOp [56] pioneers prompt learning [21, 22, 25] by replacing fixed templates with learnable continuous vec-

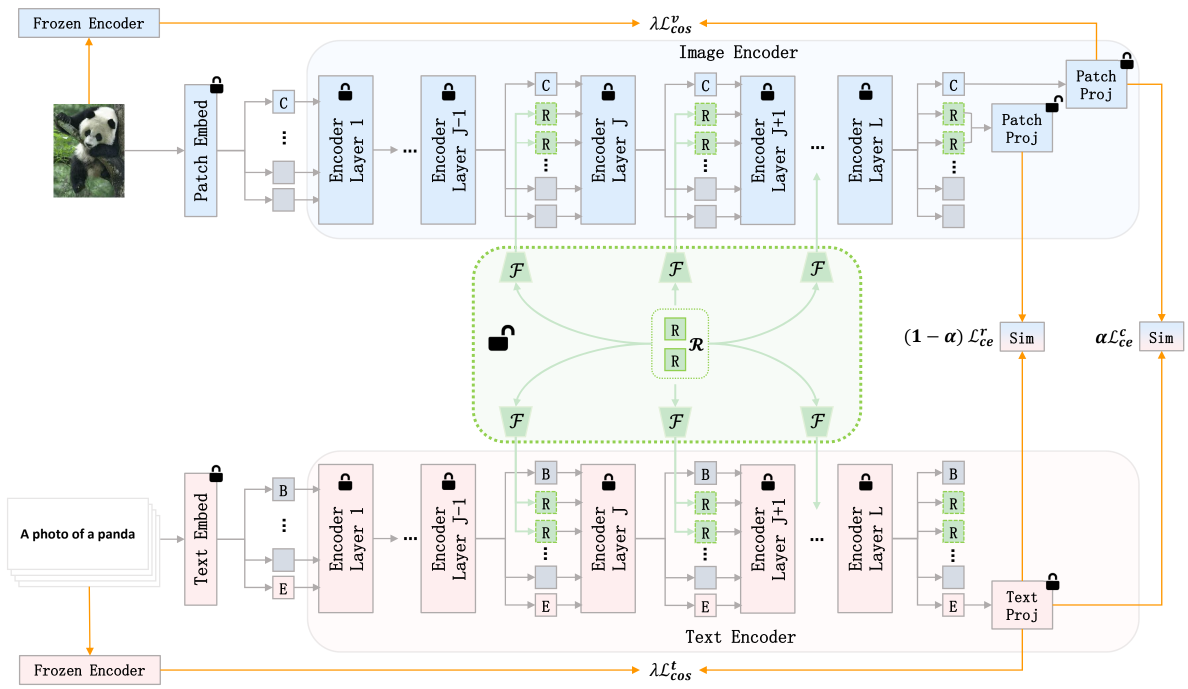

Figure 2. MMRL training framework. Here, 'C' denotes the class token, 'B' the BOT token, 'E' the EOT token, $\mathcal{R}$ our representation space, and 'R' the representation token. Only the representation space $\mathcal{R}$, mapping function $\mathcal{F}$, and the patch projection layer for the representation tokens are optimized, while the entire pre-trained CLIP model remains frozen. To preserve generalization knowledge, we integrate representation tokens in both encoders starting from layer J.

tors, enhancing flexibility but compromising CLIP's zero-shot and generalization capabilities. To address this, Co-CoOp [55] incorporates visual cues to generate instance-specific prompts, improving generalization to class distribution shifts, while ProDA [26] learns prompt distributions to enhance adaptability. PLOT [5] uses optimal transport to align the vision and text modalities. KgCoOp [48] retains general textual knowledge by minimizing divergence between learned and crafted prompts. ProGrad [57] selectively updates gradients aligned with general knowledge, and RPO [20] mitigates internal representation shifts using masked attention. Moving beyond text-focused approaches, MaPLe [17] integrates visual prompts mapped from text prompts through a coupling function, fostering cross-modal synergy. ProVP [46] employs single-modal visual prompts with contrastive feature re-formation to align prompted visual features with CLIP's distribution. Prompt-SRC [18] employs a self-regularization strategy to mitigate overfitting, while MetaPrompt [53] applies a meta-learning-based prompt tuning algorithm that encourages task-specific prompts to generalize across various domains or classes. TCP [49] adapts textual knowledge into class-aware tokens, enhancing generalization capabilities.

Adapter-style learning methods represent another efficient pathway for VLM adaptation. CLIP-Adapter [10] uses lightweight adapters, implemented as two-layer MLPs, to refine CLIP's feature representations through cross-entropy optimization. Building on this, Tip-Adapter [52] caches training features to facilitate efficient similarity calculations between test and training features. However, both methods process image and text representations independently before prediction. Addressing this separation, MMA [47] integrates features across branches into a shared space, allowing for cross-branch gradient flow and enhanced coherence between modalities.

In addition to the aforementioned methods, several approaches [23, 36, 38, 43] leverage large language models (LLMs) such as GPT-3 [4] for text augmentation or apply distillation over the entire dataset to improve performance. However, the increased computational requirements associated with these methods may place them beyond the intended scope of efficient transfer learning.

3. Method

Our approach, in line with previous methods, builds upon a pre-trained VLM, CLIP [34]. In this section, we detail the construction of our MMRL framework and the implementation specifics.

3.1. Preliminary

We begin by defining the notations used in our approach. CLIP comprises two encoders: an image encoder $\mathcal{V}$ and a text encoder $\mathcal{W}$.

Image Encoding: The image encoder $\mathcal{V}$ consists of L

transformer [40] layers, denoted $\{\mathcal{V}_i\}_{i=1}^{L}$. Given an input image $x \in \mathbb{R}^{H\times W\times 3}$, it is divided into M fixed-size patches, each projected into a patch embedding, resulting in $E_0 \in \mathbb{R}^{M\times d_v}$, where M represents the number of patches and d_v the embedding dimension. The initial patch embeddings E_0 are combined with a learnable class token c_0 and positional encodings, forming the input sequence for the transformer layers. Each layer processes this sequence as

$$[c_i, E_i] = \mathcal{V}_i([c_{i-1}, E_{i-1}]) \quad i = 1, 2, \ldots, L$$

After passing through all transformer layers, a patch projection layer, P_v^c, projects the output of the class token, c_L, into a shared V-L latent space,

$$f = P_v^c(c_L)$$

where $f \in \mathbb{R}^d$.

Text Encoding: For an input text, *e.g.*, "A photo of a [CLASS].", it is tokenized and converted into embeddings $T_0 \in \mathbb{R}^{N\times d_t}$, where N is the token length and d_t the embedding dimension. Beginning-of-text (BOT) and end-of-text (EOT) tokens, denoted b_0 and e_0, mark the sequence boundaries. These token embeddings, with positional encodings, are passed through the text encoder's L transformer layers, $\{\mathcal{W}_i\}_{i=1}^{L}$, as follows,

$$[b_i, T_i, e_i] = \mathcal{W}_i([b_{i-1}, T_{i-1}, e_{i-1}]) \quad i = 1, \ldots, L$$

After the final layer, the output of the EOT token, e_L, is projected into the shared V-L space using P_t,

$$w = P_t(e_L)$$

where $w \in \mathbb{R}^d$.

Classification with CLIP: With the image feature f and text features $\{w_c\}_{c=1}^{C}$ for C classes, CLIP calculates the cosine similarity between f and each w_c,

$$\text{sim}(f, w_c) = \frac{f \cdot w_c}{|f||w_c|},$$

where $|\cdot|$ represents the L_2 norm. Class probabilities are then computed using the softmax function,

$$p(y = c \mid f) = \frac{\exp(\text{sim}(f, w_c)/\tau)}{\sum_{i=1}^{C} \exp(\text{sim}(f, w_i)/\tau)}$$

where τ is a temperature parameter. The final predicted class is selected as the one with the highest probability score.

3.2. Multi-Modal Representation Learning (MMRL)

Our proposed MMRL aims to address the challenges of adapting pre-trained VLMs using few-shot data while maintaining generalization to new tasks. The training and inference frameworks of MMRL are shown in Fig. 2 and Fig. 3, respectively. In the following, we describe the specifics of the methodology.

3.2.1. Learnable Representation Space

MMRL establishes a shared, learnable representation space $\mathcal{R}$ to facilitate multimodal interactions, initialized through sampling from a Gaussian distribution. Using a learnable mapping function $\mathcal{F}(\cdot)$, implemented as a linear layer, we project the tokens $R \in \mathbb{R}^{K\times d_r}$ in this space—where K is the number of tokens and d_r is the dimension of the representation space—into both visual and textual modalities,

$$R^v = \{R_i^v\}_{i=J-1}^{L-1} \quad R_i^v = \mathcal{F}_i^v(R)$$
$$R^t = \{R_i^t\}_{i=J-1}^{L-1} \quad R_i^t = \mathcal{F}_i^t(R)$$

where $R_i^v \in \mathbb{R}^{K\times d_v}$ and $R_i^t \in \mathbb{R}^{K\times d_t}$ represent the representation tokens for visual and textual modalities, respectively, in the $(i+1)$-th transformer layer. The index J indicates the starting layer from which these representation tokens are integrated into the encoders.

3.2.2. Integration into Higher Encoder Layers

To preserve the generalized knowledge in the lower layers of the pre-trained CLIP model, the representation tokens $\mathcal{R}^v$ and $\mathcal{R}^t$ are integrated into the higher layers of the image encoder $\mathcal{V}$ and the text encoder $\mathcal{W}$, beginning from the J-th layer.

For the image encoder $\mathcal{V}$,

$$[c_i, E_i] = \mathcal{V}_i([c_{i-1}, E_{i-1}]) \quad i = 1, \ldots, J-1$$
$$[c_i, \_, E_i] = \mathcal{V}_i([c_{i-1}, R_{i-1}^v, E_{i-1}]) \quad i = J, \ldots, L-1$$
$$[c_i, R_i^v, E_i] = \mathcal{V}_i([c_{i-1}, R_{i-1}^v, E_{i-1}]) \quad i = L$$

For the text encoder $\mathcal{W}$, while previous prompt learning [17] involves replacing parts of T_i to incorporate deep prompts, we retain the entire T_i and insert R_i^t before it, aiming to preserve the original textual information,

$$[b_i, T_i, e_i] = \mathcal{W}_i([b_{i-1}, T_{i-1}, e_{i-1}]) \quad i = 1, \ldots, J-1$$
$$[b_i, \_, T_i, e_i] = \mathcal{W}_i([b_{i-1}, R_{i-1}^t, T_{i-1}, e_{i-1}]) \quad i = J, \ldots, L-1$$
$$[b_i, R_i^t, T_i, e_i] = \mathcal{W}_i([b_{i-1}, R_{i-1}^t, T_{i-1}, e_{i-1}]) \quad i = L$$

Note that due to the autoregressive nature of the text encoder, we adjust the attention mask matrix to accommodate the increased embedding length.

3.2.3. Representation Learning

Representation learning is designed to leverage representation tokens for dataset-specific adaptation, while the class token preserves the pre-trained knowledge of the original CLIP. Through a set of strategies aimed at retaining generalization during both training and inference, MMRL enables flexible inference for different tasks, as detailed below.

- **Training Phase:** We optimize the features of both the representation tokens and the original class token, with

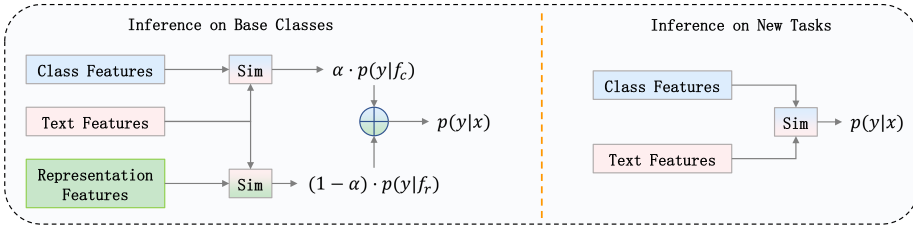

Figure 3. MMRL inference process, where different tasks utilize distinct features.

the primary focus on representation features to preserve pre-trained knowledge. Specifically, the projection layer for the representation tokens is trainable, while that for the class token remains fixed. For the image encoder $\mathcal{V}$, after passing through L transformer layers, we obtain the output $c_L \in \mathbb{R}^{d_v}$ for the class token and $R_L^v \in \mathbb{R}^{K \times d_v}$ for the K representation tokens. The final output of the representation tokens, r_L, is derived by averaging across the K tokens,

$$r_L = \text{Mean}(R_L^v)$$

where $r_L \in \mathbb{R}^{d_v}$. We then apply the patch projection layers to map the outputs of both the class and representation tokens into the common V-L latent space, yielding the class features f_c and representation features f_r.

$$f_c = P_v^c(c_L) \quad f_r = P_v^r(r_L)$$

Here, P_v^c is the original, frozen patch projection layer of CLIP for class features, while P_v^r for representation features is trainable.
For the text encoder $\mathcal{W}$, following the sequential nature of text, we map the EOT token e_L—as in the original CLIP model—after processing through L transformer layers into the common V-L space, yielding the text features.

$$w = P_t(e_L)$$

With the image features f_c, f_r, and the text classifiers $\{w_c\}_{c=1}^C$ for C classes, we apply cross-entropy loss to separately optimize the class and representation features,

$$\mathcal{L}_{ce}^c = -\sum_c^C y_c \log p(y = c \mid f_c)$$

$$\mathcal{L}_{ce}^r = -\sum_c^C y_c \log p(y = c \mid f_r)$$

where $y_c = 1$ if the image x belongs to class c, and $y_c = 0$ otherwise. To further preserve the generalization of class features, we maximize the cosine similarity between (f_c, w) and the frozen CLIP features (f_0, w_0), explicitly guiding the training trajectory,

$$\mathcal{L}_{cos}^v = 1 - \frac{f_c \cdot f_0}{|f_c||f_0|} \quad \mathcal{L}_{cos}^t = 1 - \frac{1}{C}\sum_c^C \frac{w^c \cdot w_0^c}{|w^c||w_0^c|},$$

The final MMRL loss function is

$$\mathcal{L}_{MMRL} = \alpha\mathcal{L}_{ce}^c + (1-\alpha)\mathcal{L}_{ce}^r + \lambda(\mathcal{L}_{cos}^v + \mathcal{L}_{cos}^t)$$

where α controls the balance between the features, and λ is the penalty coefficient.

- **Testing on Base Classes:** For in-distribution classes seen during training, we combine the dataset-specific representation features with the class features that preserve generalizability. The probability of an in-distribution test sample x belonging to the c-th class is

$$p(y = c \mid x) = \alpha \cdot p(y = c \mid f_c) + (1-\alpha) \cdot p(y = c \mid f_r)$$

where f_c and f_r are features extracted from the class token and representation tokens, respectively.

- **Testing on Novel Classes:** For classes unseen during training or for new datasets, we rely solely on the class tokens, which retain generalized knowledge.

$$p(y = c \mid x) = p(y = c \mid f_c)$$

4. Experiments

Details on implementation, datasets, and computational cost are provided in the Supplementary Materials.

4.1. Tasks and Datasets

We conduct four core evaluations to comprehensively assess MMRL's performance: base-to-novel generalization, cross-dataset evaluation, domain generalization, and few-shot learning. Except for few-shot learning, all experiments utilize a 16-shot setting, *i.e.*, only 16 training examples per category.

Base-to-Novel Generalization: In this evaluation, dataset classes are equally divided into base and novel classes. The model is trained exclusively on base classes and tested on both base and novel classes, allowing us to examine its transfer learning effectiveness on base classes as well as its ability to retain the inherent generalization or zero-shot capabilities of pre-trained VLMs for novel classes. We conduct this evaluation across 11 diverse image classification datasets: ImageNet [7], Caltech101 [9], OxfordPets [32], StanfordCars [19], Flowers102 [29], Food101 [3], FGVCAircraft [27], SUN397 [45], UCF101 [30], DTD [6], and EuroSAT [11].

Table 1. Comparison of MMRL with previous state-of-the-art methods on base-to-novel generalization across 11 datasets. Bold values indicate the best results. MMRL consistently enhances base class performance without compromising generalization.

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML2021)	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 (IJCV2022)	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoOpOp (CVPR2022)	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 (CVPR2022)	81.56	72.30	76.65	75.40	70.23	72.72	98.27	93.23	95.68	95.43	97.83	96.62
KgCoOp (CVPR2023)	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 (CVPR2023)	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
PromptSRC (ICCV2023)	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
ProVP (IJCV2024)	85.20	73.22	78.76	75.82	69.21	72.36	98.92	94.21	96.51	95.87	97.65	**96.75**
MetaPrompt (TIP2024)	83.65	75.48	79.09	77.52	70.83	74.02	98.13	94.58	96.32	95.53	97.00	96.26
TCP (CVPR2024)	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 (CVPR2024)	83.20	76.80	79.87	77.31	71.00	74.02	98.40	94.00	96.15	95.40	**98.07**	96.72
MMRL (Ours)	**85.68**	**77.16**	**81.20**	**77.90**	**71.30**	**74.45**	**98.97**	94.50	**96.68**	**95.90**	97.60	96.74

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML2021)	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 (IJCV2022)	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoOpOp (CVPR2022)	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 (CVPR2022)	74.70	71.20	72.91	97.70	68.68	80.66	90.30	88.57	89.43	36.90	34.13	35.46
KgCoOp (CVPR2023)	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 (CVPR2023)	72.94	74.00	73.47	95.92	72.46	82.56	90.71	**92.05**	**91.38**	37.44	35.61	36.50
PromptSRC (ICCV2023)	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	**37.87**	40.15
ProVP (IJCV2024)	80.43	67.96	73.67	98.42	72.06	83.20	90.32	90.91	90.61	**47.08**	29.87	36.55
MetaPrompt (TIP2024)	76.34	75.01	75.48	97.66	74.49	84.52	**90.74**	91.85	91.29	40.14	36.51	38.24
TCP (CVPR2024)	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 (CVPR2024)	78.50	73.10	75.70	97.77	75.93	85.48	90.13	91.30	90.71	40.57	36.33	38.33
MMRL (Ours)	**81.30**	**75.07**	**78.06**	**98.97**	**77.27**	**86.78**	90.57	91.50	91.03	46.30	37.03	**41.15**

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML2021)	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 (IJCV2022)	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoOpOp (CVPR2022)	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 (CVPR2022)	78.67	76.93	77.79	80.67	56.48	66.44	83.90	66.00	73.88	85.23	71.97	78.04
KgCoOp (CVPR2023)	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 (CVPR2023)	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
PromptSRC (ICCV2023)	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
ProVP (IJCV2024)	80.67	76.11	78.32	83.95	59.06	69.34	**97.12**	72.91	83.29	**88.56**	75.55	81.54
MetaPrompt (TIP2024)	82.26	79.04	80.62	83.10	58.05	68.35	93.53	75.21	83.38	85.33	77.72	81.35
TCP (CVPR2024)	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 (CVPR2024)	82.27	78.57	80.38	83.20	**65.63**	73.38	85.46	**82.34**	83.87	86.23	80.03	82.20
MMRL (Ours)	**83.20**	**79.30**	**81.20**	**85.67**	65.00	**73.82**	95.60	80.17	**87.21**	88.10	80.07	**83.89**

Cross-Dataset Evaluation: This evaluation measures the model's transferability to new, unseen datasets. Following CoCoOp [55], we train the model on all 1000 ImageNet classes in a few-shot setting and then directly apply it, without further fine-tuning, to other datasets to assess its cross-dataset generalization. We employ the same datasets as in the base-to-novel generalization task.

Domain Generalization: In this setting, we assess the resilience of the ImageNet-trained model to domain shifts and its generalization to out-of-distribution data. Specifically, we use ImageNet as the training dataset and evaluate on four variants—ImageNetV2 [35], ImageNet-Sketch [42], ImageNet-A [13], and ImageNet-R [12]—each introducing different types of domain variation.

Few-Shot Learning: This evaluation examines the model's transfer learning capability in limited-data scenarios, independent of its generalization performance. The model is trained on subsets of the training data with 1, 2, 4, 8, and 16 examples (shots) per class and subsequently evaluated on the full test sets.

4.2. Base-to-Novel Generalization

In this experiment, we compare MMRL with several models, including the zero-shot baseline CLIP and leading prompt learning approaches: CoOp [56], CoCoOp [55], ProDA [26], KgCoOp [48], MaPLe [17], PromptSRC [18], ProVP [46], MetaPrompt [53], TCP [49] and the multi-

Table 2. Comparison of MMRL with previous state-of-the-art methods on cross-dataset evaluation across 10 datasets.

	Source	Target										
	ImageNet	Average	Caltech101	OxfordPets	StanfordCars	Flowers101	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101
CoOp (IJCV2022)	71.51	63.88	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55
CoOpOp (CVPR2022)	71.02	65.74	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21
MaPLe (CVPR2023)	70.72	66.30	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69
PromptSRC (ICCV2023)	71.27	65.81	93.60	90.25	65.70	70.25	86.15	23.90	67.10	**46.87**	45.50	**68.75**
TCP (CVPR2024)	71.40	66.29	93.97	91.25	64.69	71.21	**86.69**	23.45	67.15	44.35	51.45	68.73
MMA (CVPR2024)	71.00	66.61	93.80	90.30	**66.13**	72.07	86.12	25.33	**68.17**	46.57	49.24	68.32
MMRL (Ours)	**72.03**	**67.25**	**94.67**	**91.43**	66.10	**72.77**	86.40	**26.30**	67.57	45.90	**53.10**	68.27

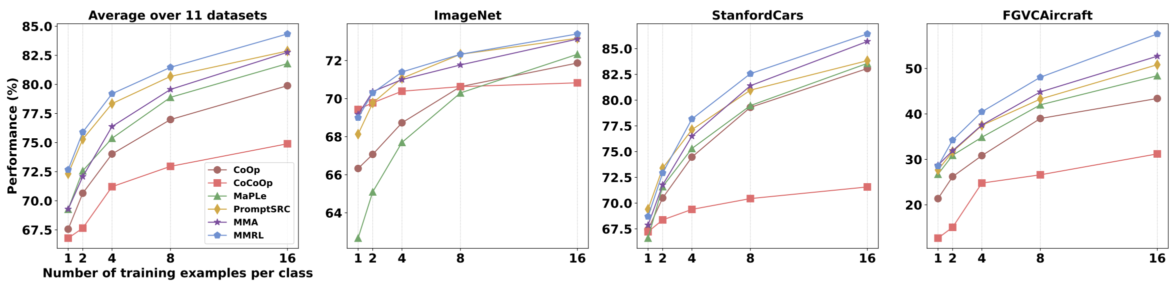

Figure 4. Comparison of MMRL with previous state-of-the-art methods on few-shot learning across 11 datasets. Detailed results on all 11 datasets are provided in the **Supplementary Material**.

modal adapter-style model MMA [47].

Tab. 1 provides detailed results for **Base** and **Novel** classes across 11 datasets, along with the balanced harmonic mean (**HM**) of their accuracies. Key findings include:

New SOTA Performance: Based on the average results across 11 datasets, MMRL achieves gains of 2.48%, 0.36%, and 1.33% in Base, Novel, and HM metrics, respectively, surpassing the previous best-performing model, MMA, and establishing a new state-of-the-art.

Strong Generalizability with Enhanced Transfer Learning: Notably, MMRL enhances generalizability while significantly boosting base accuracy, effectively improving transfer learning capabilities across downstream datasets such as ImageNet, StanfordCars, and SUN397. Although MMRL may not consistently achieve the highest novel accuracy on some datasets (e.g., UCF101, EuroSAT, DTD, and FGVCAircraft), it substantially outperforms other methods in the base category. For instance, on FGVCAircraft, MMRL's novel accuracy trails PromptSRC by 0.84%, yet it achieves a significant 3.57% gain in base accuracy. Similarly, on EuroSAT, MMRL underperforms MMA by 2.17% in the novel category but outperforms it by 10.14% in the base category!

4.3. Cross-Dataset Evaluation

As illustrated in Tab. 2, MMRL achieves a 1.03% accuracy improvement over the previous state-of-the-art method, MMA, on ImageNet. Beyond this, MMRL consistently exhibits superior performance across various target datasets, achieving the highest average accuracy, which underscores its strong cross-dataset generalization capability.

Table 3. Comparison of MMRL with previous state-of-the-art methods on domain generalization across 4 datasets.

	Source	Target			
	ImageNet	-V2	-S	-A	-R
CLIP (ICML2021)	66.73	60.83	46.15	47.77	73.96
CoOp (IJCV2022)	71.51	64.20	47.99	49.71	75.21
CoOpOp (CVPR2022)	71.02	64.07	48.75	50.63	76.18
MaPLe (CVPR2023)	70.72	64.07	49.15	50.90	76.98
PromptSRC (ICCV2023)	71.27	64.35	**49.55**	50.90	**77.80**
MMA (CVPR2024)	71.00	64.33	49.13	51.12	77.32
MMRL (Ours)	**72.03**	**64.47**	49.17	**51.20**	77.53

4.4. Domain Generalization

As summarized in Tab. 3, MMRL attains top performance on 2 out of the 4 domain-shifted datasets, showcasing its robust generalization capability across diverse domains.

4.5. Few-Shot Learning

As shown in Fig. 4, MMRL achieves the best average performance across 11 datasets under all shot settings, with performance margins increasing as the shot number rises. This trend confirms MMRL's strong transfer learning capability, even in data-scarce scenarios.

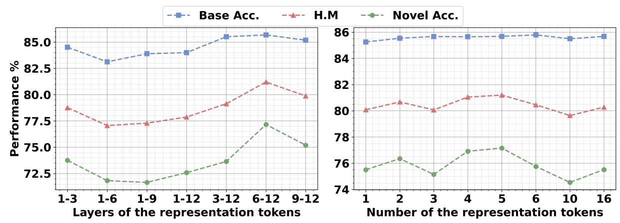

Figure 5. Ablation on layers (left) and K (right).

4.6. Ablation Analysis

All ablation experiments are conducted on base-to-novel generalization across 11 datasets, with results averaged, except for the analysis of λ on ImageNet; please refer to the **Supplementary Material** for the complete λ analysis across all datasets.

Variants of MMRL: The performance of MMRL variants, as shown in Tab. 4 (left), highlights the contributions of different components. In the variants 'w/o L' and 'w/o V', where only a single modality of representation tokens is utilized, we observe a performance decline, especially in 'w/o V', emphasizing the significance of multimodal interaction and representation features in MMRL. The 'w/o DS_1' variant, which omits the Decoupling Strategy and relies only on class features, degrades significantly on base classes, underscoring the role of representation features in capturing downstream knowledge. The 'w/o DS_2' variant, using both features for novel class evaluation, also drops notably, suggesting that the representation tokens primarily capture base class features, making transfer to new tasks challenging. The 'w/o RS' variant, which excludes the Representation Space, independently initializes textual and visual tokens without multimodal learning. This unimodal approach, while improving base class performance, severely limits generalization to novel classes, indicating the necessity of multimodal learning for effective generalization. Finally, MMRL† adopts a biased multimodal learning scheme similar to MaPLe [17], where text-side tokens are initialized randomly and mapped to the visual side. Results indicate that this biased approach underperforms compared to MMRL's balanced multimodal learning framework.

Dimension of Representation Space, d_r: As shown in Tab. 4 (right), adjusting d_r reveals that performance initially increases, followed by a decline as d_r continues to grow. This decline likely stems from overfitting caused by an overly complex representation space.

Table 4. Ablation on variants (left) and d_r (right).

Variants	Base	Novel	HM
w/o L	85.05	75.65	80.08
w/o V	82.83	75.03	78.74
w/o DS_1	83.59	77.16	80.25
w/o DS_2	85.68	73.80	79.30
w/o RS	**85.79**	75.55	80.34
MMRL†	85.60	76.02	80.55
MMRL	85.68	**77.16**	**81.20**

d_r	Base	Novel	HM
32	85.27	76.85	80.84
128	85.42	76.74	80.85
256	85.63	76.84	81.00
512	**85.68**	**77.16**	**81.20**
1024	85.57	76.97	81.04
2048	85.53	76.91	81.00

Table 5. Ablation on α (left) across 11 datasets and λ (right) on ImageNet.

α	Base	Novel	HM
0.0	82.96	72.34	77.29
0.3	84.57	75.45	79.75
0.5	85.42	76.11	80.50
0.7	**85.68**	**77.16**	**81.20**
1.0	83.79	75.49	79.42

λ	Base	Novel	HM
0.0	77.73	70.63	73.96
0.2	77.83	71.23	74.38
0.5	**77.90**	**71.30**	**74.45**
2.0	**77.90**	70.93	74.25
4.0	77.67	70.73	74.04

Layer for Representation Token Insertion: As depicted in Fig. 5 (left), model performance declines when representation tokens are introduced at lower encoder layers. This trend aligns with MMRL's design, as higher layers capture dataset-specific, discriminative features, while lower layers retain generalizable features. Furthermore, MMRL performs poorly on base classes when representation tokens are inserted into lower layers, suggesting that lower-layer features are less adaptable. Performance improves with insertion at higher layers but declines when placed too high, likely due to the reduced number of learnable parameters and limited capacity to influence CLIP's critical parameters.

Number of Representation Tokens, K: In Fig. 5 (right), increasing K slightly enhances base class accuracy due to additional learnable parameters. For novel classes, accuracy initially improves with K but eventually declines, indicating that an excessive number of tokens may lead to overfitting, reducing generalization capacity.

Balance Weight, α: The parameter α modulates reliance on representation token features versus class token features. Lower α values increase dependence on representation features, heightening overfitting risk due to the learnable projection layer, while higher values shift dependence to class token features, diminishing transferability. As shown in Tab. 5 (left), the optimal α is 0.7.

Penalty Coefficient, λ: The penalty coefficient λ regulates the regularization strength by aligning class token features with frozen CLIP's features. Higher λ generally enhances generalization but may restrict transfer flexibility. In Tab. 5 (right), the optimal λ for ImageNet is 0.5.

5. Conclusion

In this work, we introduce the MMRL framework to enhance the generalization of VLMs when adapting to diverse downstream datasets. MMRL establishes a shared, unbiased representation space that bridges image and text modalities, promoting balanced multimodal learning while preserving the pre-trained knowledge encapsulated in class tokens. By strategically decoupling representation tokens from class tokens during inference, MMRL effectively mitigates overfitting risks and reinforces adaptability. Extensive evaluations confirm MMRL's capacity for an optimal balance between task-specific adaptation and generalization, setting a new benchmark for efficient transfer learning.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China under Grant 62176062.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[2] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4662–4670, 2022. 1

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 5

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 3

[5] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models, 2023. 3

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[9] 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 *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 5

[10] 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. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 2, 3

[11] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5

[12] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 6

[13] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 6

[14] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11375–11385, 2024. 1

[15] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 1, 2

[16] 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 *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 2

[17] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2, 3, 4, 6, 8, 1

[18] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 3, 6, 1

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 5

[20] Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J Kim. Read-only prompt optimization for vision-language few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1401–1411, 2023. 3

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2

[22] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, 2021. Association for Computational Linguistics. 2

[23] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 3

[24] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 1, 2

[25] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland, 2022. Association for Computational Linguistics. 2

[26] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 3, 6

[27] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[28] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning, 2021. 1

[29] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 5

[30] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 5

[31] Övgü Özdemir and Erdem Akagündüz. Enhancing visual question answering through question-driven image captions as prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1562–1571, 2024. 1

[32] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 1, 5

[33] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. 1, 2

[34] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[35] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 6

[36] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *ICLR*, 2024. 3

[37] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862, 2021. 1

[38] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28578–28587, 2024. 3

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 1

[40] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[41] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. What makes for good visual tokenizers for large language models?, 2023. 1

[42] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[43] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5749–5757, 2024. 3

[44] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 1

[45] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5

[46] Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive visual prompt learning with contrastive feature re-formation. *International Journal of Computer Vision*, pages 1–16, 2024. 3, 6, 1

[47] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024. 2, 3, 7, 1

[48] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 3, 6

[49] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024. 3, 6, 1

[50] 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. 1, 2

[51] Shuhong Ye, Weikai Kong, Chenglin Yao, Jianfeng Ren, and Xudong Jiang. Video question answering using clip-guided visual-text attention. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 81–85. IEEE, 2023. 1

[52] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 3, 1

[53] Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. Learning domain invariant prompt for vision-language models. *IEEE Transactions on Image Processing*, 2024. 3, 6, 1

[54] Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, and Dinggang Shen. Clip in medical imaging: A comprehensive survey, 2024. 1

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 3, 6, 1

[56] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 6, 1

[57] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 3, 1