# Exploring the Better Multimodal Synergy Strategy for Vision-Language Models

**Xiaotian Yin*, Xin Liu*, Si Chen, Yuan Wang, Yuwen Pan, Tianzhu Zhang†**

Deep Space Exploration Laboratory/School of Information Science and Technology,
University of Science and Technology of China
{xiaotianyin, xinliu99, cs_fisha, wy2016, panyw}@mail.ustc.edu.cn, tzzhang@ustc.edu.cn

## Abstract

Vision-Language models (VLMs) have shown great potential in enhancing open-world visual concept comprehension. Recent researches focus on an optimum multimodal collaboration strategy that significantly advances CLIP-based few-shot tasks. However, existing prompt-based solutions suffer from unidirectional information flow and increased parameters since they explicitly condition the vision prompts on textual prompts across different transformer layers using non-shareable coupling functions. To address this issue, we propose a **D**ual-**s**hared mechanism based on Lo**RA** (**DsRA**) that addresses VLM adaptation in low-data regimes. The proposed DsRA enjoys several merits. First, we design an inter-modal shared coefficient that focuses on capturing visual and textual shared patterns, ensuring effective mutual synergy between image and text features. Second, an intra-modal shared matrix is proposed to achieve efficient parameter fine-tuning by combining the different coefficients to generate layer-wise adapters placed in encoder layers. Our extensive experiments demonstrate that DsRA improves the generalizability under few-shot classification, base-to-new generalization, and domain generalization settings. Our code will be released soon.

## Introduction

Pre-trained Vision-Language Models (VLMs) such as CLIP (Radford et al. 2021) have demonstrated promising generalization power and transferability on various downstream tasks, including image segmentation (Rao et al. 2022; Wang et al. 2022), object detection (Shi and Yang 2023; Zang et al. 2022), image caption (Wang et al. 2023; Mokady, Hertz, and Bermano 2021; Shen et al. 2021), and so on, opening up new possibilities in these fields. Through contrastive training on a web-scale dataset of image-text pairs, VLMs achieve a global alignment between images and textual descriptions with the help of rich supervision provided by natural language. This alignment greatly enhances the representations provided by VLMs, enabling compelling zero-shot inference even in the absence of training samples.

Although CLIP is highly effective for generalizing to new concepts, the large scale of VLMs and the limited availability of training data in many downstream tasks (*e.g.*, few-shot settings) make it difficult to fine-tune the entire model. Previous works (Zhou et al. 2022b; Gao et al. 2023; Shu et al. 2023) reveal that naively fine-tuning CLIP on downstream few-shot tasks may cause catastrophic forgetting and overfitting. In sample-scarce cross-modal scenarios, the inability to achieve mutual synergy between vision and language modalities increases the difficulty of aligning features, resulting in unsatisfactory recognition performance (Khattak et al. 2023a). Therefore, designing a strategy for effective multimodal collaboration in few-shot settings poses a significant challenge.

To address the challenges mentioned above, recent works such as MaPLe (Khattak et al. 2023a), ALIGN (Wang et al. 2024), and CoPrompt (Roy and Etemad 2024) adopt a viable prompt-based solution by explicitly conditioning the vision prompts on textual prompts across different transformer layers. This process involves mapping the textual prompts into visual prompts using MLPs, thereby aligning the features between different modalities. While this paradigm has

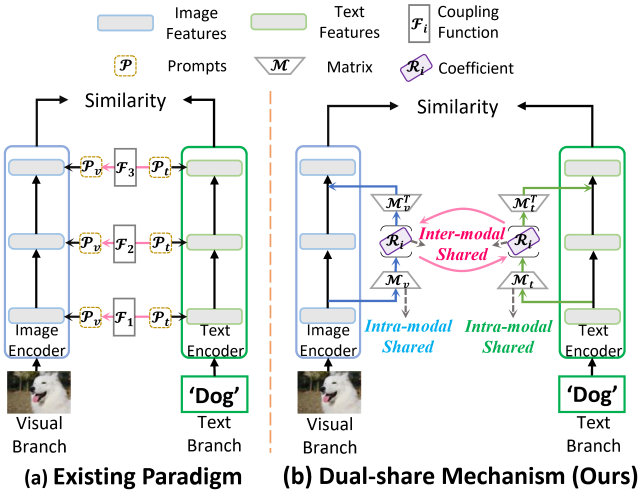


Figure 1: The existing paradigm aligns the features between different modalities by using unshareable coupling functions, which leads to insufficient modal interaction and increased complexity. Our DsRA effectively enhances the mutual perception between different features through a lightweight Dual-shared design.

*These authors contributed equally.
†Corresponding Author.

| Method | Paras | Fps | 1-shot | 2-shot | 8-shot |
|---|---|---|---|---|---|
| CoOp* | 0.03m | 505.0 | 68.80 | 71.42 | 75.68 |
| MaPLe* | 0.80m | 348.1 | 68.05 | 73.04 | 78.04 |
| MaPLe | 3.55m | 348.3 | 69.30 | 73.29 | 78.75 |
| ALIGN | 3.58m | 51.2 | 70.04 | 73.56 | 79.58 |
| CoPrompt | 4.74m | 285.8 | 67.58 | 70.99 | 77.67 |
| **Ours** | **0.17m** | **466.2** | **73.28** | **76.10** | **81.34** |

Table 1: Complexity analysis over various methods. We report the number of trainable parameters (Paras), frames per second (Fps), and few-shot classification results. CoOp* means our reproduced results with visual prompt tuning. MaPLe* denotes the MaPLe version with a shared coupling function across all layers.

achieved encouraging results, several factors make this explicit alignment of features potentially suboptimal. **(1) Unidirectional Information Flow:** When handling inputs from different modalities, using prompt-based mapping struggles to manage the synergy between the two modalities. As shown in Figure 1, existing works explicitly map textual prompts $\mathcal{P}_t$ from the text encoder into visual prompts $\mathcal{P}_v$ and feeds them into the vision encoder ($\mathcal{P}_t \rightarrow \mathcal{P}_v$). Although this allows textual information to be conveyed to the visual branch, this unidirectional flow cannot enhance the perception of text features to visual branches. The intermediate feature representation from the vision encoder cannot interact with the textual information, potentially leading to inadequate modal alignment. **(2) Layer-wise Computational Complexity:** To capture varying visual and textual patterns across different network layers, for example, MaPLe uses layer-wise coupling functions to encode features at various levels. Each layer of textual prompts in the encoder requires non-sharable linear layers for visual prompt mapping. The use of multiple MLPs significantly increases parameter overhead, specifically by 118 times compared to the original prompt learning method CoOp (Zhou et al. 2022b). This layer-wise mapping design undoubtedly adds to the model's complexity and computational load, leading to a higher risk of overfitting, especially in the extreme low-shot scenario (*e.g.*, comparison of 1-shot task results among MaPLe, CoPrompt, and CoOp in Table 1). To tackle the above challenge, exploring a more **effective** and **efficient** strategy to enable better information interaction at various levels across layers is imperative.

In this work, we focus on exploring a novel multimodal synergy strategy and proposing a **D**ual-**s**hared mechanism to enhance the reusability of module parameters, thereby improving efficiency. We propose a Lo**RA**-based approach called the **DsRA** for the adapter design using a bottleneck operation. As illustrated in Figure 1(b), our Dual-shared mechanism shares the up/down projection matrices $\mathcal{M}_v/\mathcal{M}_v^T$ or $\mathcal{M}_t/\mathcal{M}_t^T$ in low-rank design across different layers of the same encoder. Additionally, we learn inter-modal shared coefficients $\mathcal{R}$ to recombine these linear projections as layer-adaptive adapters within the encoder. **To achieve more effective interaction between different modalities,** we design inter-modal shared coefficients within the encoder. These shared coefficients encode varying visual and textual shared patterns while reconciling the inherent tension between textual and visual features. By placing different coefficients in various layers of the encoder, DsRA implicitly enhances the compactness of the feature space, leading to better multimodal alignment. **To achieve more efficient parameter fine-tuning,** each encoder shares the intra-modal shared matrix, which, combined with inter-modal shared coefficients, forms a layer-specific adapter. This clever Dual-shared design enhances the reusability of the module while reducing parameter overhead. Besides, our designed adapters can be seamlessly integrated into pre-trained networks without adding extra computation during inference due to their linear nature (Luo et al. 2023). We strive for our work to become a fundamental reference in CLIP-based transfer learning, establishing a new baseline for future research.

In summary, the contributions of this work include: (1) We propose a novel method, DsRA, with a Dual-shared mechanism by jointly exploring an efficient and effective strategy to fully boost CLIP's adaptability across layers and modalities for few-shot recognition. (2) DsRA incorporates the inter-modal shared coefficients to achieve interaction between different modalities and the intra-modal shared matrix to maintain parameters and computation efficiency. Additionally, Our DsRA is compatible with existing prompt tuning methods as a general plugin module. (3) Experiments show that DsRA consistently outperforms other state-of-the-art CLIP adaptation methods on 11 challenging benchmarks and three settings while maintaining the trade-off between accuracy and efficiency.

## Related Works

In this section, we introduce several lines of research in Vision-Language models and parameter-efficient transfer learning.

**Vision-Language Models (VLMs)** aim to connect visual and language aspects through extensive pre-training. CLIP achieves this by training on a large dataset of image-text pairs, creating a shared embedding space for images and textual descriptions. The cross-modal alignment between vision and language enables CLIP to recognize various visual contents. By inserting a class name into a text template (*e.g.*, "A photo of a [CLASS]"), CLIP can generate classification weights to classify new classes in the real world. Despite the effectiveness of VLMs, efficiently applying these pretrained models to downstream tasks is still crucial, especially in situations with limited data, such as the few-shot setting. In addition, the immense size of these models makes it difficult to fine-tune them without sacrificing generalization ability. Consequently, recent studies have focused on adapting large foundational models for downstream tasks while keeping the pretrained backbones frozen. To effectively adapt CLIP for zero-shot and few-shot recognition tasks, we propose a novel LoRA-based Dual-shared strategy to enhance the perception between vision and text features while improving the reusability of the module and reducing the number of parameters.

**Parameter-Efficient Transfer Learning (PETL)** aims to approach the fully tuned performance on downstream tasks by updating a few parameters or additional parameters, reducing the training and storage burdens. One of the popular techniques in PETL is prompt tuning, which has proven effective in various natural language processing tasks (Brown et al. 2020). Inspired by this success, prompt learning has also been explored in vision-language models. Specifically, CoOp introduces learnable prompts to enable efficient finetuning and adaptation of the CLIP text encoder. CoCoOp (Zhou et al. 2022a) addresses the inferior performance of CoOp on novel classes by explicitly conditioning prompts on image instances, alleviating the generalization issue. ProDA (Lu et al. 2022) extends CoOp by estimating the distribution of multiple text prompts. MaPLe (Khattak et al. 2023a) conditions visual prompts on textual prompts across transformer layers. PromptSRC (Khattak et al. 2023b) proposes a self-regulating loss to improve the generalization of prompts. ALIGN (Wang et al. 2024) leverages the optimal transportation to learn and align a set of prompt tokens across modalities. TCP (Yao, Zhang, and Xu 2024) proposes a textual-based class-aware prompt combined with class tokens to generate task-specific textual knowledge.

Another line of research adopts an adapter strategy for VLM-based transfer learning. CLIP-Adapter (Gao et al. 2023) impressively introduces a lightweight residual style adapter to efficiently fine-tune CLIP. Tip-Adapter (Zhang et al. 2022) leverages the visual prototypes obtained from the few-shot support samples to compute the similarity with the test image's visual embedding, which is later used to modify the CLIP visual embedding. CaFo (Zhang et al. 2023) proposes a cascade of foundation models, including DINO (Caron et al. 2021), DALL-E (Ramesh et al. 2021), and GPT3 (Brown et al. 2020), to create a powerful cache model for refining the Tip-adapter. CoPrompt (Roy and Etemad 2024) combines prompt tuning with adapters and proposes a consistency constraint loss to optimize them. Different from previous works, we introduce a novel LoRA-based approach called the DsRA, which enables VLMs to effectively exploit multi-level relations between visual and textual cues while ensuring parameter and compute efficiency.

## Method

In this section, we first revisit CLIP to provide an introduction to the notations. Then, we introduce the details of our proposed DsRA.

### Revisit CLIP

In our approach, we adopt CLIP as the pre-trained vision-language model. CLIP consists of a visual branch and a text branch. The model in each branch contains a transformer encoder with multiple layers.

**In the text branch,** the CLIP text encoder $g$ generates feature representations by projecting the tokenized words to input embeddings $\boldsymbol{t} = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_{L-1}, \boldsymbol{c}\} \in \mathbb{R}^{L \times d_t}$, where the last token $\boldsymbol{c}$ denotes the embedding of class name and $\{,\}$ is concatenation operation. $\boldsymbol{t}$ is then fed into several consecutive transformer layers, each consisting of a Multi-head Attention Block (MAB) and a Feed-forward Network Block (FNB). LayerNorm (LN) is applied before each block, and residual connections are applied thereafter. The process of each encoder layer is defined as:

$$\hat{\boldsymbol{t}}^{(i)} = \mathrm{MAB}\left(\mathrm{LN}\left(\boldsymbol{t}^{(i-1)}\right)\right) + \boldsymbol{t}^{(i-1)}, \tag{1}$$

$$\boldsymbol{t}^{(i)} = \mathrm{FNB}\left(\mathrm{LN}\left(\hat{\boldsymbol{t}}^{(i)}\right)\right) + \hat{\boldsymbol{t}}^{(i)}, \tag{2}$$

where $\boldsymbol{t}^{(i-1)}$ denotes input text embeddings in $i$-th transformer layer, $\hat{\boldsymbol{t}}^{(i)}$ indicates intermediate features produced by MAB, $\boldsymbol{t}^{(i)}$ is the output of $i$-th layer.

In MAB, self-attention module take $\boldsymbol{t}^{(i-1)}$ as queries $\mathbf{Q}$, keys $\mathbf{K}$ and values $\mathbf{V}$. The $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}$ triplets are generated by the independent linear projection layers:

$$\mathbf{Q} = \boldsymbol{t}^{(i-1)} \times \mathbf{W}_q, \tag{3}$$

$$\mathbf{K} = \boldsymbol{t}^{(i-1)} \times \mathbf{W}_k, \tag{4}$$

$$\mathbf{V} = \boldsymbol{t}^{(i-1)} \times \mathbf{W}_v, \tag{5}$$

where $\mathbf{W}_q, \mathbf{W}_k$, and $\mathbf{W}_v \in \mathbb{R}^{d_t \times d_t}$ are linear projection layers. The attention score can be obtained by:

$$\boldsymbol{attn} = Softmax(\frac{\mathbf{Q} \times \mathbf{K}^T}{\sqrt{d_k}}), \tag{6}$$

where $\sqrt{d_k}$ is a scaling factor. Then we can obtain the output of MAB by:

$$\hat{\boldsymbol{t}}^{(i)} = (\boldsymbol{attn} \times \boldsymbol{V}) \times \mathbf{W}_o + \boldsymbol{t}^{(i-1)}, \tag{7}$$

where $\mathbf{W}_o \in \mathbb{R}^{d_t \times d_t}$ denotes linear projection. The FNB block contains an MLP with the GELU activation function:

$$\boldsymbol{t}^{(i)} = \mathrm{GELU}\left(\hat{\boldsymbol{t}}^{(i)} \times \mathbf{W}_1\right) \times \mathbf{W}_2 + \hat{\boldsymbol{t}}^{(i)}, \tag{8}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_t \times 4d_t}$ and $\mathbf{W}_2 \in \mathbb{R}^{4d_t \times d_t}$ are linear projections. Given the prompt $\boldsymbol{t}$ with the class name $\boldsymbol{c}_n$, a text feature $\boldsymbol{g}_n = g(\boldsymbol{t}_n)$ is obtained by the text encoder $g$ of CLIP.

**In the visual branch,** the image encoder first splits the image into some fixed-size patches which are projected into patch embeddings $\boldsymbol{\mu} = \left\{\boldsymbol{\mu}_n \in \mathbb{R}^{d_v} \middle|_{k=1}^{K}\right\}$, where $K$ denotes the number of image patch tokens. Then, the input sequence for the image encoder can be formulated as $\boldsymbol{e} = \{\boldsymbol{\mu}, \boldsymbol{\mu}_{cls}\}$, which is sent into the visual encoder $f$ to encode the image feature. $\boldsymbol{\mu}_{cls}$ denotes the learnable $cls$ token. Similarly to the process in the text branch, we can derive the $i$-th image feature by:

$$\hat{\boldsymbol{e}}^{(i)} = \mathrm{MAB}\left(\mathrm{LN}\left(\boldsymbol{e}^{(i-1)}\right)\right) + \boldsymbol{e}^{(i-1)}, \tag{9}$$

$$\boldsymbol{e}^{(i)} = \mathrm{FNB}\left(\mathrm{LN}\left(\hat{\boldsymbol{e}}^{(i)}\right)\right) + \hat{\boldsymbol{e}}^{(i)}, \tag{10}$$

where $\boldsymbol{e}^{(i-1)}$ denotes image embeddings in $i$-th vision transformer layer, $\hat{\boldsymbol{e}}^{(i)}$ indicates intermediate features produced

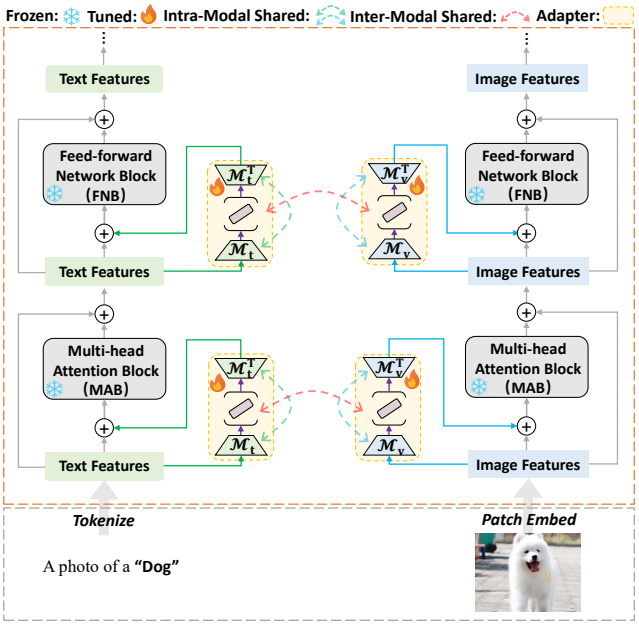


Figure 2: Illustration of the proposed DsRA in the first layer of the CLIP encoders. DsRA utilizes the inter-modal shared coefficient to effectively capture diverse visual and textual patterns and the intra-modal shared matrix to optimize parameter fine-tuning.

by MAB, $\boldsymbol{e}^{(i)}$ is the output of $i$-th layer. The final image representation $\boldsymbol{f}$ is can be obtained by: $\boldsymbol{f} = f(\boldsymbol{e})$. Notice that each encoders have a projection, which projects features of different dimensions into the same space to calculate similarity.

The zero-shot classification probability is computed based on the similarity between the image and text features:

$$p(y=n \mid \boldsymbol{x}) = \frac{\exp\left(\operatorname{sim}\left(\boldsymbol{f}, \boldsymbol{g}_n\right)/\tau\right)}{\sum_{n'=1}^{N} \exp\left(\operatorname{sim}\left(\boldsymbol{f}, \boldsymbol{g}_{n'}\right)/\tau\right)}, \tag{11}$$

where $\operatorname{sim}(,)$ is the similarity function and $\tau$ is a temperature parameter learned by CLIP.

## Our DsRA

Most existing methods (Khattak et al. 2023a; Wang et al. 2024; Roy and Etemad 2024) introduce prompts into various layers and learn layer-specific mapping functions to align text features with image features for fine-grained recognition. Despite the encouraging results, the unidirectional information flow and layer-specific coupling increase parameters make the explicit alignment of intermediate modalities potentially suboptimal. Drawing inspiration from these findings (Hu et al. 2021; Lian et al. 2022; Luo et al. 2023; Dong et al. 2024; Yin et al. 2024; Liu et al. 2024), we propose a new LoRA-based Dual-shared approach called DsRA to enhance the perception between vision and text features while increasing the module's reusability to reduce parameter size.

**Dual-shared mechanism.** The DsRA incorporates a Dual-share mechanism, which consists of two key components: the inter-modal shared coefficient for capturing varying visual and textual patterns and the intra-modal shared matrix for more efficient parameter fine-tuning.

As illustrated in Figure 2, to reconcile the inherent tension between textual and visual features, we design an inter-modal shared coefficient, which is then diagonalized into a specific diagonal matrix $\mathcal{R}^i \in \mathbb{R}^{d_r \times d_r}$ for each layer $i$. $d_r$ denotes the hidden dimensionality of the low-rank adapter, where $d_r \ll d_t < d_v$. This inter-modal shared diagonal matrix aims to enable efficient and effective adjustment of the adaptation parameters at each layer while enhancing the perception between text and visual features. To place different coefficients in various layers from different encoders, we design an intra-modal shared matrix to form a low-rank adapter using the bottleneck operation. Specifically, in the text branch, we use symmetric down-projection and up-projection within a single bottleneck structure:

$$\mathcal{M}_t = \left(\mathcal{M}_t^T\right)^T, \tag{12}$$

where $\mathcal{M}_t \in \mathbb{R}^{d_t \times d_r}$ is the up-projection and $\mathcal{M}_t^T \in \mathbb{R}^{d_r \times d_t}$ is the down-projection. Then, we can combine the inter-modal shared coefficients in a layer with the text shared matrix to generate a layer-specific adapter. Formally, given a text embeddings $\boldsymbol{t}_{in}^i$ output from the ($i$-1)-th layer, we can get the output as follows:

$$\boldsymbol{t}_{out}^i = \mathcal{A}\left(\boldsymbol{t}_{in}^i\right) = \boldsymbol{t}_{in}^i \mathcal{M}_t \mathcal{R}^i \mathcal{M}_t^T + \boldsymbol{t}_{in}^i, \tag{13}$$

where $\mathcal{A}$ denotes the composed adapter. Then, our DsRA proposes inserting the composed adapter sequentially before both the MAB and FNB blocks in the transformer. The text encoder incorporating our modules can be formulated as follows:

$$\hat{\boldsymbol{t}}^{(i)} = \mathrm{MAB}\left(\mathcal{A}_{MAB}\left(\mathrm{LN}\left(\boldsymbol{t}^{(i-1)}\right)\right)\right) + \boldsymbol{t}^{(i-1)}, \tag{14}$$

$$\boldsymbol{t}^{(i)} = \mathrm{FNB}\left(\mathcal{A}_{FNB}\left(\mathrm{LN}\left(\hat{\boldsymbol{t}}^{(i)}\right)\right)\right) + \hat{\boldsymbol{t}}^{(i)}, \tag{15}$$

where $\mathcal{A}_{MAB}$ and $\mathcal{A}_{FNB}$ represent the adapter $\mathcal{A}$ placed in positions before MAB and FNB, respectively. That is to say, the projection matrices of the two adapters are independent between of each other. Similarly, in CLIP's visual branch, we can recombine the inter-modal shared coefficient with image shared matrix to derive the image features:

$$\hat{\boldsymbol{e}}^{(i)} = \mathrm{MAB}\left(\mathcal{A}_{MAB}\left(\mathrm{LN}\left(\boldsymbol{e}^{(i-1)}\right)\right)\right) + \boldsymbol{e}^{(i-1)}, \tag{16}$$

$$\boldsymbol{e}^{(i)} = \mathrm{FNB}\left(\mathcal{A}_{FNB}\left(\mathrm{LN}\left(\hat{\boldsymbol{e}}^{(i)}\right)\right)\right) + \hat{\boldsymbol{e}}^{(i)}. \tag{17}$$

Compared to previous methods based on prompt-based cross-modal alignment solutions (Khattak et al. 2023a; Wang et al. 2024; Roy and Etemad 2024), our proposed DsRA enables the CLIP model to comprehensively exploit multi-level relations between visual and textual cues while decreasing the parameter size used in the adaptation process. Thus, it strengthens the efficiency and robustness of cross-modal alignment for few-shot recognition in open-set tasks.

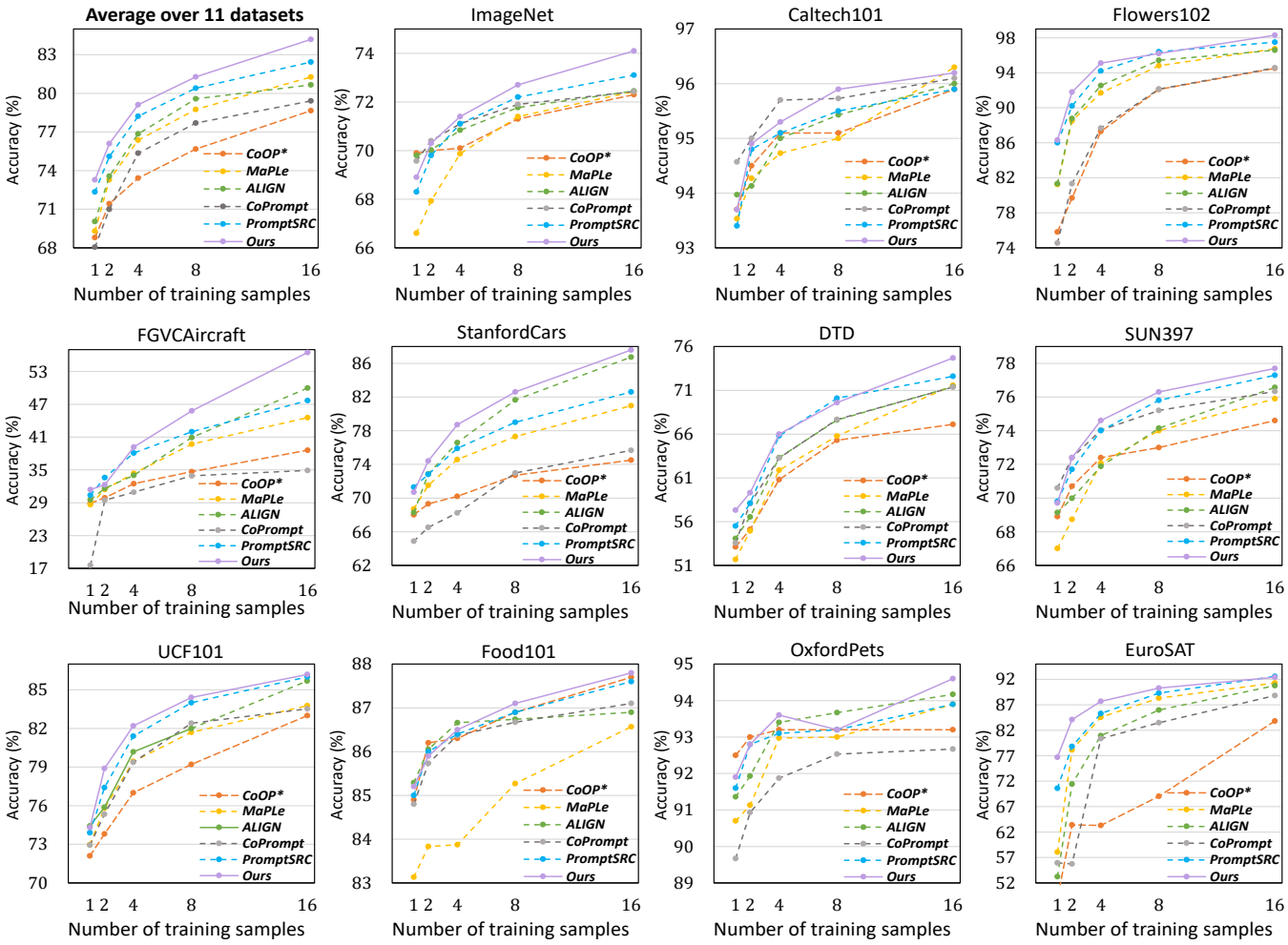


Figure 3: The accuracy performance on the few-shot classification setting. DsRA consistently outperforms the previously top-performing methods MaPLe, ALIGN, and PromptSRC, achieving higher average accuracy across 11 datasets. CoOp* denotes our reproduction version with additional visual prompt tuning.

## Experiments

In this section, we introduce the benchmark settings and implementation details and then present the experimental results comprehensively. We evaluate the adaptation and generalization capability of our method in three settings. (1) Few-Shot Classification (FSC). (2) Base-to-New Generalization (B2N). (3) Domain Generalization.

**Dataset.** In line with CLIP (Radford et al. 2021), we utilize 11 image benchmark datasets: ImageNet (Img) (Deng et al. 2009), Caltech101 (Cal) (Fei-Fei, Fergus, and Perona 2004), FGVCAircraft (FGV) (Maji et al. 2013), Flower102 (Flo) (Nilsback and Zisserman 2008), Food101 (Foo) (Bossard, Guillaumin, and Van Gool 2014), OxfordPets (Pet) (Parkhi et al. 2012), StanfordCars (Car) (Krause et al. 2013), EuroSAT (Eur) (Helber et al. 2019), DTD (Cimpoi et al. 2014), SUN397 (SUN) (Xiao et al. 2010), and UCF101 (UCF) (Soomro, Zamir, and Shah 2012). To ensure fairness, we follow the experimental methodologies outlined in CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a), including dataset splits, data augmentation, and backbones. Additionally, we conduct experiments to evaluate DsRA's domain generalization capabilities, leveraging ImageNet (Img) as the source dataset and considering its diverse domain variants, such as ImageNetV2 (V2) (Recht et al. 2019), ImageNetSketch (S) (Wang et al. 2019), ImageNet-A (A) (Hendrycks et al. 2021b), and ImageNet-R (R) (Hendrycks et al. 2021a), as the target datasets.

**Implementation Details.** All experiments are conducted using the CLIP model with a ViT-B/16 backbone. The hidden dimension $d_r$ is set to 50. SGD is used for optimization with a learning rate of 2.7e-3 and a batch size of 4. All experiments are conducted on a single RTX 3090 GPU. The main results are averaged over three runs.

### Few-shot Classification

We first evaluate our model on few-shot classification, where models are trained on 1, 2, 4, 8 and 16 shots and then applied to the test sets. Figure 3 summarizes the accuracy compari-

son between DsRA with other methods. Our method consistently outperforms previous pipelines in average accuracy across 11 datasets, demonstrating its effectiveness and generalization ability. Compared with prompt-based solutions such as MaPLe and ALIGN, DsRA significantly outperforms them in various shot settings. Specifically, it exceeds MaPLe by 3.98% and 2.75% in average accuracy for 1-shot and 4-shot settings, respectively. DsRA exhibits accuracy 3.24% and 1.71% improvements compared with ALIGN for 1-shot and 16-shot tasks. In addition, we find that our DsRA performs better on extremely low-shot settings. It achieves 73.28% and 76.10% average accuracy for 1-shot and 2-shot scenarios, respectively. The above improvements benefit from our Dual-share design, which effectively and efficiently enhances the mutual perception between vision and text features, even with extremely few samples. These outcomes illustrate DsRA's adaptability across diverse CLIP-based models and its effectiveness in enhancing few-shot adaptability.

## Base-to-New Generalization

To test the base-to-new generalization ability, we follow CoCoOp to train our model only on the base classes in a 16-shot setting and evaluate the model on base and new categories. We summarize the comparison between the proposed method and existing methods in Table 2.

Some key findings can be found: (1) Compared to CoOp and CoCoOp that do not involve deep multimodal interaction, DsRA significantly outperforms them by exploring a multimodal synergy strategy, achieving 2.67% improvements on average results in base classes (vs. CoOp) and an impressive improvement of 3.56% (vs. CoCoOp) in performance on new classes. The improvements highlight the effectiveness of our design in improving multimodal perception capabilities. (2) Based on prompt learning, methods like MaPLe, ALIGN, and CoPrompt condition the vision prompts on textual prompts to conduct deep modal interactions using non-sharable linear layers. Compared with the excessive parameters approach, our method outperforms them in base and new classes performance with fewer parameters. Specifically, on the harmonic mean (HM) results across 11 datasets, our method outperforms MaPLe and ALIGN by 2.55% and 1.85%, respectively. Additionally, our method outperforms CoPrompt by 1.36% on the base class while maintaining comparable performance on the new class. It is worth noting that DsRA has fewer trainable parameters than CoPrompt (0.17m vs. 4.74m). These results prove that our approach is more effective and efficient than the prompt-based solution for enabling better information interaction. (3) Compared to DePT (Zhang et al. 2024), which uses PromprSRC as the baseline, like our method does. DsRA outperforms DePT on base and new classes, reaching 85.36% and 77.25%, respectively. Our design can be plugged into more methods to improve performance further, and the results are in the section on ablation experiments.

| Method | Base | New | HM |
|---|---|---|---|
| CoOp[IJCV22] | 82.69 | 63.22 | 71.66 |
| CoCoOp[CVPR22] | 80.47 | 71.69 | 75.83 |
| MaPLe[CVPR23] | 82.28 | 75.14 | 78.55 |
| ALIGN[NeurIPS23] | 83.38 | 75.51 | 79.25 |
| PromptSRC[ICCV23] | 84.26 | 76.10 | 79.97 |
| CoPrompt[ICLR24] | 84.00 | 77.23 | 80.48 |
| DePT[CVPR24] | 85.19 | 76.17 | 80.43 |
| TCP[CVPR24] | 84.13 | 75.36 | 79.51 |
| Ours | **85.36** | **77.25** | **81.10** |

Table 2: The average results under the base-to-new generalization setting on 11 datasets. More detailed results can be found in the supplementary material.

| Method | Source | Target | | | | Avg. |
|---|---|---|---|---|---|---|
| | Img | V2 | S | A | R | |
| CLIP | 66.73 | 60.83 | 46.15 | 47.77 | 73.96 | 57.18 |
| CoOp | 71.51 | 64.20 | 47.99 | 49.71 | 75.21 | 59.28 |
| CoCoOp | 71.00 | 64.07 | 48.75 | 50.60 | 76.20 | 59.91 |
| MaPLe | 70.72 | 64.07 | 49.15 | 50.90 | 77.00 | 60.28 |
| PromptSRC | 71.27 | 64.35 | 49.55 | 50.90 | 77.80 | 60.65 |
| ALIGN | 72.03 | 64.64 | 49.96 | 50.94 | 76.16 | 60.43 |
| CoPrompt | 70.80 | 64.25 | 49.43 | 50.50 | 77.51 | 60.42 |
| Ours | 72.00 | 65.87 | 49.73 | 50.19 | 77.60 | **60.85** |

Table 3: Performance of DsRA on domain generalization setting and its comparison to existing methods.

## Domain Generalization

To evaluate the DsRA's robustness under domain shifts, we initially train it using the source dataset ImageNet and then evaluate its performance on target datasets. The overall results are summarized in Table 3. Specifically, DsRA achieves a performance comparable to the experimental results of the existing methods. Our method is slightly lower than the accuracy performance of previous methods on the ImageNet-A dataset. However, DsRA performs better on ImageNetV2, reaching 65.87%. In general, the above experiments demonstrate the generalization ability and robustness of our method to domain shifts.

## Ablation Experiments

**Impact of the hidden dimensions $d_r$ in DsRA.** Figure 4 (a) reveals the impact of different $d_r$. Even when the value is as small as 5, the performance remains comparable to previous methods, highlighting the effectiveness of our design. The accuracy of base and new classes increases when the dimension becomes larger, which could deliver richer context knowledge. However, the accuracy slightly decreases when the value exceeds 50, as an excessive amount of parameters may introduce redundancy and noise. We have not attempt a larger value as it would introduce more parameters and increase the risk of overfitting. In general, the accuracy changes relatively smoothly for different values of $d_r$.

**Shared vs. non-shared inter-modal coefficients.** The inter-modal shared coefficients are an essential design in

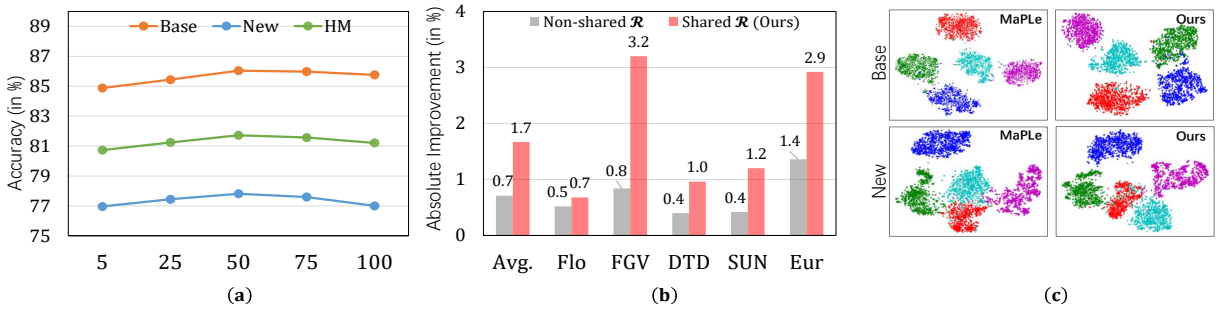


Figure 4: (a) The impact about the value of the inter-modal shared coefficients. We conduct experiments in the base-to-new generalization setting on ten datasets. (b) Ablation on shared and non-shared inter-modal coefficients. We report the absolute improvements compared with the baseline across five datasets in the few-shot classification. (c) The t-SNE visualization of MaPLe and DsRA on EuroSAT.

our Dual-share mechanism. In this section, we ablation on shared and non-shared inter-modal coefficients by conducting experiments with modal-shared (shared) or modal-specific (non-shared) coefficients, as depicted in Figure 4 (b). It is noted that by using non-shared coefficients, the performance surpasses the baseline model because each encoder can learn different coefficients to capture the specific patterns of various layers, thereby improving the recognition performance. With modal-shared coefficients, accuracy rises as shared $\mathcal{R}$ enables the CLIP to exploit multi-level relations between visual and textual context knowledge, strengthening the robustness of cross-modal alignment for few-shot tasks. The results prove that our method can promote collaboration between multiple modalities more effectively.

**Different placing strategies for adapters.** In DsRA, we implement CLIP's adaptation by positioning the composed adapter in MAB and FNB. We investigate the impact of various positioning strategies for the composed adapter in the few-shot classification. The results are presented in Table 4. Placing the adapters in the MAB (refer to MAB only) and FNB (refer to FNB only) reduces the number of learnable parameters, but it leads to performance degradation. However, using both adapters increases performance compared to using only one type of adapter for either MAB or FNB. This implies that our design allows for a more comprehensive adaptation of CLIP to the target task without significantly increasing the number of trainable parameters.

**DsRA on more methods.** To further evaluate the effectiveness of our proposed DsRA approach, we plug it into KgCoOp (Yao, Zhang, and Xu 2023) and ProGrad (Zhu et al. 2023) and present the results in Table 5. We observe consistent performance increases. Specifically, combining DsRA with KgCoOp results in a considerable improvement in accuracy for few-shot classification, contributing to an overall increase in the harmonic mean (from 77.00% to 79.08%) for base-to-new generalization. Compared to ProGrad, DsRA with the ProGrad attains significantly stronger performance across most shot settings, e.g., 1.11% on the 2-shot setting and 2.33% on the 16-shot setting. These outcomes illustrate DsRA's adaptability across diverse CLIP-based models and its effectiveness in enhancing few-shot learning.

| Strategy | Paras | FGV | Car | UCF | DTD | Eur | Avg. |
|---|---|---|---|---|---|---|---|
| MAB only | 0.11m | 40.12 | 78.36 | 80.90 | 64.30 | 83.82 | 69.50 |
| FNB only | 0.11m | 40.62 | 78.42 | 80.72 | 64.40 | 84.64 | 69.76 |
| Ours | 0.17m | **41.56** | **78.80** | **81.20** | **65.40** | **86.24** | **70.64** |

Table 4: Impact of different placing strategies for recomposed adapter in DsRA. We report the number of learnable parameters (Paras) and few-shot classification results.

| Method | FSC | | | B2N |
|---|---|---|---|---|
| | 1-shot | 4-shot | 16-shot | HM |
| KgCoOp | 68.73 | 73.42 | 75.68 | 77.00 |
| KgCoOp + DsRA | **70.54** | **77.35** | **82.87** | **79.08** |
| ProGrad | 70.57 | 74.57 | 81.98 | 76.16 |
| ProGrad + DsRA | **71.50** | **76.35** | **84.31** | **77.91** |

Table 5: Few-shot classification and base-to-new generation of other approaches w/ or w/o our DsRA.

**Visualization.** We further visualize the image embeddings of MaPLe and ours. As shown in Figure 4 (c), DsRA prefers to learn separable representations in base and new classes, which demonstrates that DsRA can effectively enable information interaction.

## Conclusion

This paper proposes a novel DsRA approach to adapt CLIP for few-shot tasks. DsRA contains a Dual-shared mechanism to improve synergy between vision and language modalities while maintaining parameters and computing efficiency. Additionally, DsRA is compatible with existing methods and consistently improves performance as a general plugin module. Experiments show the effectiveness.

## Acknowledgments

This work was supported by National Defense Science and Technology Foundation Strengthening Program Funding (No. 2023-JCJQ-JJ-0219), Basic Strengthening Program

Laboratory Fund ( No. NKLDSE2023A009), Youth Innovation Promotion Association.

## References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, 446–461. Springer.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision*, 9650–9660.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255. Ieee.

Dong, W.; Yan, D.; Lin, Z.; and Wang, P. 2024. Efficient adaptation of large vision transformer via adapter recomposing. *Advances in Neural Information Processing Systems*, 36.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 178–178. IEEE.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2023. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 1–15.

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

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *International Conference on Computer Vision*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15262–15271.

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

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

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

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 554–561.

Lian, D.; Zhou, D.; Feng, J.; and Wang, X. 2022. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35: 109–123.

Liu, X.; Wu, J.; Yang, W.; Zhou, X.; and Zhang, T. 2024. Multi-modal attribute prompting for vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5206–5215.

Luo, G.; Huang, M.; Zhou, Y.; Sun, X.; Jiang, G.; Wang, Z.; and Ji, R. 2023. Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106*.

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

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *International Conference on Computer Vision*, 722–729. IEEE.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3498–3505. IEEE.

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

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18082–18091.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 5389–5400. PMLR.

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

Shen, S.; Li, L. H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; and Keutzer, K. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Shi, C.; and Yang, S. 2023. Edadet: Open-vocabulary object detection using early dense alignment. In *International Conference on Computer Vision*, 15724–15734.

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

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Wang, D.; Li, M.; Liu, X.; Xu, M.; Chen, B.; and Zhang, H. 2024. Tuning multi-mode token-level prompt alignment across modalities. *Advances in Neural Information Processing Systems*, 36.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32.

Wang, L.; Qiu, H.; Qiu, B.; Meng, F.; Wu, Q.; and Li, H. 2023. TridentCap: Image-Fact-Style Trident Semantic Framework for Stylized Image Captioning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11686–11695.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *International Conference on Computer Vision*, 3485–3492. IEEE.

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

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

Yin, X.; Wu, J.; Yang, W.; Zhou, X.; Zhang, S.; and Zhang, T. 2024. Hierarchy-Aware Interactive Prompt Learning for Few-Shot Classification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, 106–122. Springer.

Zhang, J.; Wu, S.; Gao, L.; Shen, H. T.; and Song, J. 2024. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12924–12933.

Zhang, R.; Hu, X.; Li, B.; Huang, S.; Deng, H.; Qiao, Y.; Gao, P.; and Li, H. 2023. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15211–15222.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, 493–510. Springer.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

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

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