International Journal of Computer Vision
https://doi.org/10.1007/s11263-025-02354-1

Learning with Enriched Inductive Biases for Vision-Language Models

Lingxiao Yang[1] · Ru-Yuan Zhang[2] · Qi Chen[3] · Xiaohua Xie[3,4,5]

Received: 24 April 2024 / Accepted: 6 January 2025

Abstract

Vision-Language Models, pre-trained on large-scale image-text pairs, serve as strong foundation models for transfer learning across a variety of downstream tasks. For few-shot generalization tasks, *i.e.*, when the model is trained on few-shot samples and then tested on unseen categories or datasets, there is a balance to be struck between generalization and discrimination when tweaking these models. Existing approaches typically rely on one or two strategies during training to learn task-specific knowledge, while preserving as much task-agnostic representation as possible. However, these methods overlook the importance of other useful inductive biases, thereby limiting their generalization capabilities. In this work, we propose a method – **L**earning **w**ith **E**nriched **I**nductive **B**iases (LwEIB) – to explore multiple inductive biases at the text, model, and optimization levels. Specifically, we first propose to enrich the handcrafted text prompt with Large Language Model generated descriptions for each category. To better capture structural cues in both linguistics and vision, we design two new adapters for text and image encoders, respectively. Additionally, we propose a slow-fast optimization method to explore different degrees of adaptation more efficiently, learning task-specific representations while maintaining task-agnostic ones. We empirically validate the effectiveness of LwEIB on three widely used benchmarks. Remarkably, our LwEIB outperforms numerous state-of-the-art methods across all evaluation metrics, demonstrating its efficacy and versatility. Our code is available at https://github.com/ZjjConan/VLM-LwEIB.

Keywords Vision language models · Inductive biases · Few-shot adaptation · Transformer

1 Introduction

Deep Neural Networks (DNNs) (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; He et al., 2016; Dosovitskiy et al., 2020) are powerful tools for image understanding. Since the introduction of AlexNet (Krizhevsky et al., 2012), the dominant paradigm for vision-related tasks has been to pre-train DNNs on large-scale datasets (Deng et al., 2009; Zhou et al., 2017) and then fine-tune them for specific tasks such as image classification (Hu et al., 2018; Woo et al., 2018; Yang et al., 2021; Tan and Le, 2019), object detection (Girshick et al., 2014; Ren et al., 2015; Liu et al., 2016; Redmon et al., 2016; Lin et al., 2017), semantic segmentation (Long et al., 2015; He et al., 2017; Chen et al., 2024; Ronneberger et al., 2015), person re-identification (Sun et al., 2017; Ye et al., 2021; Zhang et al., 2024a, b), *etc*. This success is largely attributed to the availability of large, crowd-labeled datasets like ImageNet (Deng et al., 2009), PLACES (Zhou et al., 2017) and MS-COCO (Lin et al., 2014). However, collecting these datasets and their high-quality labels is costly.

Communicated by Kaiyang Zhou.

✉ Xiaohua Xie
xiexiaoh6@mail.sysu.edu.cn

Lingxiao Yang
yanglx9@mail.sysu.edu.cn

Ru-Yuan Zhang
ruyuanzhang@sjtu.edu.cn

Qi Chen
chenq377@mail2.sysu.edu.cn

[1] School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou, China

[2] Brain Health Institute, National Center for Mental Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine and School of Psychology, Shanghai, China

[3] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

[4] Guangdong Province Key Laboratory of Information Security Technology, Guangzhou, China

[5] Pazhou Lab (HuangPu), Guangzhou, China

Published online: 28 January 2025

Recently, Vision-Language Models (VLMs) (Radford et al., 2021; Jia et al., 2021; Yao et al., 2021; Yuan et al., 2021; Zhai et al., 2022; Huang et al., 2023; Wang et al., 2021; Alayrac et al., 2022) have emerged to reduce the need of manually collecting such high-quality annotations. VLMs are usually pre-trained on massive web-searched data, such as the 400 million image-text pairs used in **C**ontrastive **L**anguage-**I**mage **P**retraining (CLIP) (Radford et al., 2021). These models often contain a text encoder and an image encoder and are pre-trained to construct a unified representation space where related images and texts are grouped, while unrelated ones are separated. This extensive pre-training allows VLMs to capture complex image-language relationships and exhibit good generalization across various tasks.

Although fine-tuning pre-trained VLMs is the most straightforward strategy, the massive number of parameters in VLMs poses challenges when fine-tuning them for different downstream tasks, especially in scenarios with limited data and labels (*i.e*., few-shot settings). To address this, prompt engineering (Radford et al., 2021) has emerged as a key technique. This approach involves strategically formulating input queries to guide VLMs toward desired outputs. For example, in CLIP (Radford et al., 2021), handcrafted text prompts like "a photo of a <CN>" are input into the text encoder, where "<CN>" is replaced by actual category names (*e.g*., "a photo of a dog") to generate category-specific features. These features are then compared with the visual features produced by the image encoder to predict the output class.

However, devising effective prompts requires substantial expert knowledge and time. To address this, researchers have recently proposed two types of parameter-efficient fine-tuning (PEFT) strategies: prompt-based and network-based methods. Prompt-based methods incorporates a small number of learnable prompts into either the text encoder (Zhou et al., 2022a, b; Li and Liang, 2021; Lester et al., 2021; Bulat and Tzimiropoulos, 2023), the image encoder (Chen et al., 2022b; Rao et al., 2022), or both (Khattak et al., 2023a, b; Lee et al., 2023; Wang et al., 2024), of pre-trained models. During fine-tuning, only these added prompts are optimized, while the rest of the model remains fixed. This approach allows researchers to manipulate prompts to specific contextual challenges. Another strategy is to construct lightweight networks called adapters (Chen et al., 2022b; Houlsby et al., 2019; Chen et al., 2022c; Gao et al., 2023; Zhang et al., 2022; Stickland and Murray, 2019; Hu et al., 2021) for downstream tasks. Unlike prompt-based methods, where task-specific cues are learned and stored via additional prompted tokens, adapters are shallow networks (*e.g*., MLPs) that enhance generalizability through feature fusion. Like prompt-based methods, optimization focuses only on the added adapters, reducing memory usage and overfitting. Adapters are also versatile, operating independently of network architecture and easily integrating into various models (He et al., 2016; Dosovitskiy et al., 2020; Liu et al., 2021). As a result, these PEFT methods have gained popularity for their practical utility in VLMs.

Despite the promising results achieved by these methods, they still face three challenges: (1) Images often contain diverse textures and complex environments. A simple text prompt (*i.e*., "a photo of a <CN>") used in most existing methods (Zhou et al., 2022a, b; Khattak et al., 2023a, b; Lee et al., 2023) lacks precise descriptions of target categories, making it difficult for VLMs to accurately align text and visual modalities; (2) Transformer-based models generally outperform Convolutional Neural Networks (CNNs) in many vision and language tasks, because they rely on more flexible self-attention layers to capture long-range dependencies. However, due to fewer hard priors assumed as CNNs (*e.g*., weight sharing for translation invariance), pure transformers (Dosovitskiy et al., 2020) are considered less efficient than CNNs (He et al., 2016) in data-limited scenarios; (3) Most existing methods are designed to fit well to training data distributions. This increases the risk of overfitting and reducing the model's ability to generalize to unseen samples. These issues arise mainly because current approaches do not well explore inductive biases, such as prior knowledge in the data, model structure, and during optimization. We argue that these inductive biases are crucial for effective adaptation, as they help VLMs make predictions about unseen data by learning from the limited training examples.

To this end, we propose a novel adaption framework – **L**earning **w**ith **E**nriched **I**nductive **B**iases (LwEIB) – that refers to the process of tuning VLMs by incorporating additional inductive biases. Specifically, our LwEIB incorporates three levels of inductive biases: text-level, model-level, and optimization-level. First, to bridge semantic gaps between language and vision modalities, we propose the text-level inductive bias by supplementing the prompt text with many LLM-generated descriptions (Pratt et al., 2023) to provide detailed information for each category. Second, to enable the model to well capture inductive biases, we propose two types of adapters for text and image encoders, respectively. Specifically, for the text encoder, we design a phrase adapter to explicitly explore connections between adjacent words. For the image encoder, we design a spatial adapter to enable the model to capture more local relationships and details (Liu et al., 2021; Wu et al., 2021; Sd'Ascoli et al., 2021). Third, based on our adapters, we further propose a simple optimization-level inductive bias to reduce overfitting. This is achieved by a new dynamic training strategy that allows the model to adjust to different degrees of fitting. We perform extensive experiments and show that our LwEIB can handle diverse few-shot generalization tasks. Compared to previous methods, LwEIB achieves an average HM of 81.21 on base-to-novel generalization, an average recogni-

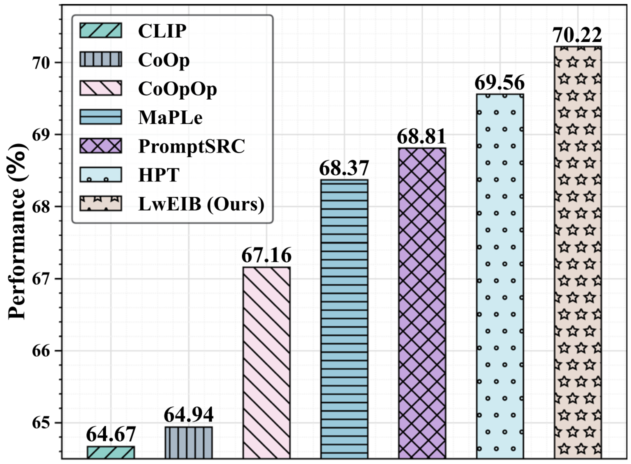


Fig. 1 Average performance over 3 benchmarks. Similar to previous studies, we mainly evaluate the proposed LwEIB on three few-shot image recognition benchmarks (see experiment section for more details). The average performance across 3 benchmarks reveals that our LwEIB yields the best performance outcomes when compared with many existing state-of-the-art methods

tion accuracy of 68.61 and 60.84 on cross-dataset and domain generalization evaluation, setting a new state-of-the-art average performance over these three benchmarks (see Fig. 1). In summary, the main contributions of our approach are three-folds:

1. We propose a novel parameter-efficient fine-tuning framework – Learning with Enriched Inductive Biases (LwEIB) – that can be trained end-to-end to leverage multiple inductive biases.
2. We propose three levels of inductive biases, *i.e.*, text-level, model-level and optimization-level, inductive biases, to increase the generalizability of VLMs in few-shot settings.
3. We evaluate LwEIB on three widely used and challenging few-shot generalization tasks. Experimental results show that LwEIB achieves leading performance among all compared methods in all evaluated benchmarks.

The paper is organized as follows. Sect. 2 provides an overview of the studies related to our research. In Sect. 3, we present the proposed LwEIB and include all experiment results in Sect. 4. Finally, we draw conclusions in Sect. 5.

2 Related Work

2.1 Vision-Language Models

Recent advances in Vision-Language Models (VLMs) (Radford et al., 2021; Jia et al., 2021; Yao et al., 2021; Yuan et al., 2021; Zhai et al., 2022; Huang et al., 2023; Wang et al., 2021) have substantially influenced the fields of computer vision and machine learning, particularly in efforts to integrate language understanding with image analysis. These models capitalize on the self-supervised training paradigm, utilizing large amounts of multi-modal data collected from the web for their pre-training. For instance, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) train their models in a contrastive manner with the InfoNCE loss (Oord et al., 2018), leveraging datasets comprising approximately 400 million and one billion image-text pairs, respectively. With the increasing amount of collected multi-modal data (*e.g.*, five billion image-text pairs) (Schuhmann et al., 2022), VLMs show commendable effectiveness in various applications (Ilharco et al., 2021). Despite acquiring good generalized representations, the efficient adaptation of these pre-trained VLMs to specific downstream tasks remains a formidable challenge, especially in scenarios with limited training data (*e.g.*, few-shot settings). To address this challenge, numerous studies have been proposed and have achieved good performance on a variety of tasks, such as few-shot image recognition (Gao et al., 2023; Kim et al., 2021; Zhang et al., 2022; Zhou et al., 2022a, b; Chen et al., 2022a), object detection (Feng et al., 2022; Gu et al., 2021; Zang et al., 2022a; Zhou et al., 2022d; Zhong et al., 2022), and segmentation (Ding et al., 2022; He et al., 2023; Zhou et al., 2022c). In contrast, the present work proposes incorporating several inductive biases to effectively facilitate the adaptation of VLMs in different few-shot generalization tasks.

2.2 Efficient Transfer Learning for VLMs

To transfer large-scale pre-trained models to downstream tasks, conventional approaches (Devlin et al., 2018; Brown et al., 2020a) fine-tune all parameters of such pre-trained models. However, as model size continues to increase, the conventional paradigm is inherently constrained by significant computational requirements. More importantly, fine-tuning such a large number of trainable parameters introduces the risk of overfitting, especially in the few-shot setting. Consequently, the NLP community has introduced several parameter-efficient methods (Li and Liang, 2021; Houlsby et al., 2019; Hu et al., 2021), which have been further extended to the fields of computer vision (Chen et al., 2022b; Jia et al., 2022) and visual language understanding (Zhou et al., 2022a, b). Since the aim of this work is to develop an efficient transfer learning method for VLMs, we mainly present two dominant lines of research in the following: token-based prompt learning and network-based adapters.

Prompt learning involves the initial provision of textual prompts to the language component of VLMs, with the aim of improving the models' adaptability in vision-language understanding. This approach often tunes the added tokens while the whole pre-trained model is frozen. For example, CoOp (Zhou et al., 2022b) improves the few-shot transfer capabilities of CLIP (Radford et al., 2021) by strategically optimizing a continuous set of prompt tokens within its language branch. This optimization contributes to a more effective use of textual instructions to guide model responses. CoCoOp (Zhou et al., 2022a) further extends CoOp by conditioning language prompts on specific image instances. This refinement allows language features to be more closely aligned with the associated visual content. Recently, several studies have made great strides in this area. These methods include multiple template-based prompt learning (Lu et al., 2022; Chen et al., 2023), improving the alignment of text and image features through optimal transport (Chen et al., 2022a), incorporating pre-trained CLIP as general knowledge (Yao et al., 2023; Bulat and Tzimiropoulos, 2023; Khattak et al., 2023b; Zhu et al., 2023) to address the problem of overfitting to seen examples, and adding prompt tokens in both image and text branches (Khattak et al., 2023a; Zang et al., 2022b; Lee et al., 2023).

Adapters are first proposed in the NLP community to adapt large-scale pre-trained language models (Houlsby et al., 2019; Hu et al., 2021; Stickland and Murray, 2019), and have recently been introduced in pure vision and vision-language models. They are lightweight networks, *e.g.* MLPs, that are inserted into the pre-trained VLMs. Similar to prompt learning, during fine-tuning, only the weights of the added adapters are optimized while the other parameters of the entire pre-trained model are frozen. With this strategy, task-specific information is learned while the general knowledge stored by the pre-training is retained. Recent representative studies include adding an adapter layer after either image (Gao et al., 2023; Zhang et al., 2022; Chen et al., 2022b, c) or text encoders (Yu et al., 2023). More recently, a cross-modal adapter (Jiang et al., 2022) has been developed for text-to-video retrieval.

2.3 Inductive Biases for VLMs

In machine learning, inductive biases refer to the assumptions baked into algorithms that guide them toward particular solutions or hypotheses. A typical example of inductive bias is convolutional constraints, such as weight sharing and translation invariance, which have been incorporated into foundation models (*e.g.,* Dosovitskiy et al. (2020), Liu et al. (2021), Sd'Ascoli et al. (2021) to enable efficient training in the relatively small data regime. Recently, several studies have introduced inductive biases into the tuning process for VLMs. For instance, PromptSRC (Khattak et al., 2023b) and KgCoOp (Yao et al., 2023) regularize prompted representations by the frozen pre-trained models. This helps to retain more task-agnostic general representations. ProDA (Lu et al., 2022) introduces multiple handcrafted prompt templates to enhance the representational capacity of the text encoder, further improved by the use of LLM-generated text prompts in HPT (Wang et al., 2024). Unlike these methods, which typically employ one or two inductive biases during training, we propose to systematically incorporate inductive biases at text, model, and optimization levels, leading to more effective VLM tuning and improved generalization.

3 Our Methodology

Following previous studies (Zhou et al., 2022a; Khattak et al., 2023a; Wang et al., 2024; Lee et al., 2023; Khattak et al., 2023b), our approach uses pre-trained transformer-based CLIP models (Radford et al., 2021), *i.e.*, using transformers in both text and visual encoders. In this section, we first provide some preliminary knowledge on CLIP and then elaborate on our proposed LwEIB.

3.1 Preliminaries

CLIP (Radford et al., 2021) represents a significant advancement in Vision-Language Models (VLMs), attracting considerable scientific attention in both natural language processing and computer vision. Roughly speaking, CLIP consists of a text encoder and an image encoder. In transformer-based CLIP, both the text and image encoders are transformers with identical network architectures, comprising an embedding layer, a series of transformer blocks (L), and a projection layer. Each transformer block includes a multi-head self-attention layer (MSA) and a feed-forward network (FFN). We can formulate a single transformer block as follows:

$$MSA: \quad z \leftarrow z + Attention(LN_{MSA}(z)) \tag{1}$$

$$FFN: \quad z \leftarrow z + FC2(FC1(LN_{FFN}(z))) \tag{2}$$

where z represents the text or image input, LN denotes the layer normalization, and $Attention$ is a standard softmax-based self-attention layer used to capture long-range dependencies. $FC1$ and $FC2$ are two fully-connected layers. For simplicity, we omit the $GELU$ activation function between the two FC layers. By jointly training the text and image encoders with a contrastive objective (Oord et al., 2018; Radford et al., 2021) on massive image-text pairs (Radford et al., 2021), CLIP aligns the representations of related image-text pairs and pushes those of unrelated pairs further apart. This extensive pre-training enables CLIP to simultaneously encode both images and text descriptions such that CLIP can be applied in a wide range of downstream tasks. Specifically, given an image $\boldsymbol{I}$ and a text description $\boldsymbol{T}$, CLIP first tokenizes each of them to generate N and M tokens, respectively. Visual features ($\boldsymbol{x}$) and text features ($\boldsymbol{w}$) are then extracted by

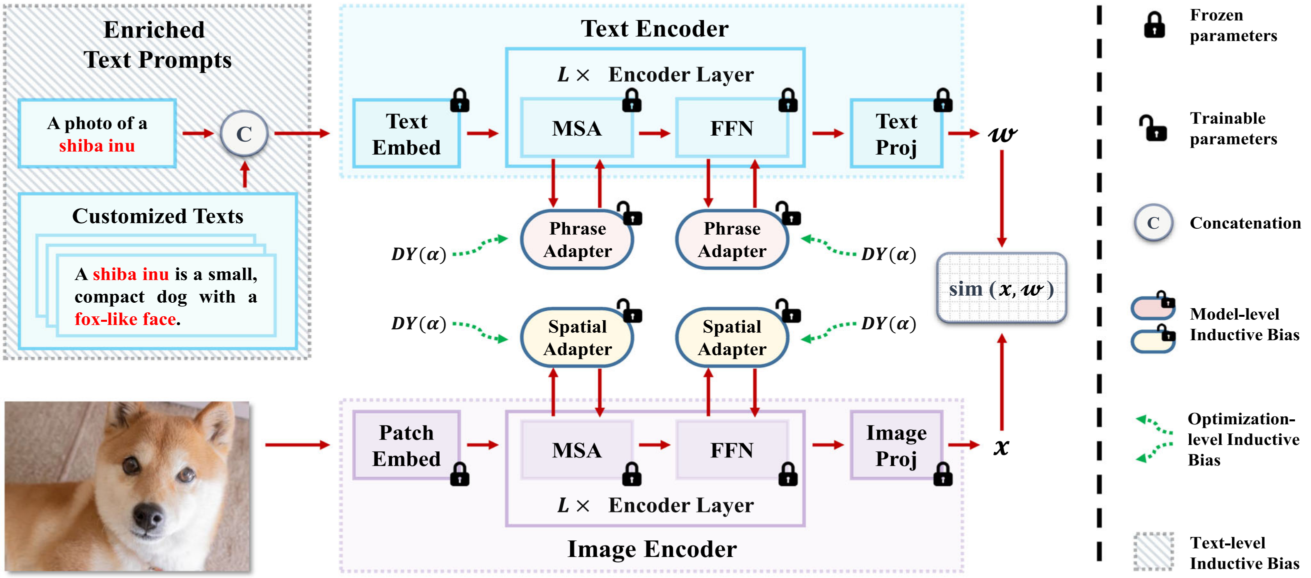


Fig. 2 Overview of our LwEIB. LwEIB is built on transformer-based CLIP models and incorporates three types of inductive biases: (1) *Text-level Inductive Bias*: Supplementing "a photo of a <CN>" with a few LLM-generated customized texts to provide more category-specific information; (2) *Model-level Inductive Bias*: Adapting the text and image encoders with specially designed adapters to learn hard inductive biases; (3) *Optimization-level Inductive Bias*: In the adapters, α is a hyper-parameter that controls the degree of task-specific knowledge being learned and used. During training, we dynamically adjust α ($DY(\alpha)$) to allow the model to learn different degrees of task-specific knowledge. This optimization strategy strikes a good balance in recognizing both seen and unseen scenarios

the pre-trained image and text encoders. Afterward, cosine similarity scores ($sim(\boldsymbol{x}, \boldsymbol{w})$) between these image and text features can be computed to facilitate task-specific predictions.

3.2 LwEIB

Following (Zhou et al., 2022a; Khattak et al., 2023a; Lee et al., 2023; Khattak et al., 2023b; Wang et al., 2024), we utilize the transformer-based CLIP model (Radford et al., 2021) and focus on few-shot generalization tasks. The goal is to tune CLIP using a limited number of training samples. The tuned model should also have a good generalization ability for unseen scenarios. To achieve this, we propose a novel approach called **L**earning **w**ith **E**nriched **I**nductive **B**iases (LwEIB). As illustrated in Fig. 2, LwEIB incorporates three types of inductive biases: text-level, model-level, and optimization-level inductive biases.

3.2.1 Text-level Inductive Bias

Given the widespread use of handcrafted prompts such as "a photo of a <CN>" in Zhou et al. (2022a, b); Khattak et al. (2023a); Lee et al. (2023); Khattak et al. (2023b); Bulat and Tzimiropoulos (2023), these approaches aim to learn a set of continuous prompt tokens to adapt CLIP for various downstream tasks. For example, replacing "a photo of a" with four learnable tokens, and optimizing these tokens in different downstream tasks to provide trained descriptions of the given "CN". Here, we argue that, in most existing studies, the learnable tokens are shared across all categories and cannot provide sufficient information to distinguish between different categories. Furthermore, a limited set of learnable tokens cannot adequately describe visual images in detail, as images often contain much richer textures and complex environments. These issues might be addressed by increasing the number of learnable prompt tokens. However, doing this greatly increases the number of trainable parameters and leads to overfitting in few-shot settings, as shown in many previous studies (Zhou et al., 2022a, b; Khattak et al., 2023a, b).

To overcome these problems, we propose to enrich the handcrafted prompt with some customized texts (Pratt et al., 2023) generated by a LLM (Brown et al., 2020b). Our process is shown in Fig. 2 and is formulated as follows:

$$\boldsymbol{T}(CN)_i \leftarrow [\boldsymbol{T}_{hc}(CN); \boldsymbol{T}_{ct}(CN)_i], \quad i = 1, ..., K \tag{3}$$

Where CN denotes the category name, $\boldsymbol{T}_{hc}$ represents the handcrafted prompt, *i.e.*, "a photo of a <CN>". We use LLM-generated descriptions (Brown et al., 2020b; Pratt et al., 2023) as custom texts and denote the i-th customized text for the category CN as $\boldsymbol{T}_{ct}(CN)_i$. Here, K is the total number of custom texts for each category, and $[\cdot;\cdot]$ indicates the concatenation operation. Through this process, the category CN is enriched with K detailed descriptions, each potentially containing useful cues. For example, as shown in Fig. 2, the category "shiba inu" is enriched with multiple attributes such as "small", "compact", or "fox-like face". We expect these detailed descriptions to bridge the semantic gap between language and vision, leading to a better generalization ability.

Currently, there are a few studies that leverage LLM-generated text to facilitate the transfer of CLIP. Menon and Vondrick (2022) and Pratt et al. (2023) use GPT-generated class-specific attributes to improve zero-shot performance. These methods are further enhanced by fine-tuning CLIP for fine-grained recognition (Saha et al., 2024), using self-attention to refine text features across different sentences (Maniparambil et al., 2023), and designing relationship-guided attention to capture pairwise correspondences among entities and attributes (Wang et al., 2024). Overall, these studies share a similar goal with ours: enhancing CLIP's transfer capability by better leveraging LLM-generated sentences. However, unlike them, we introduce a strong inductive bias into the text encoder via convolution (Sect. 3.2.2), which, unlike attention-based methods, captures structural information across sentences in a distinct manner. We further propose an optimization-level inductive bias for model training (Sect. 3.2.3).

3.2.2 Model-level Inductive Bias

Based on large-scale pre-trained transformers, several tuning methods (Hu et al., 2021; Khattak et al., 2023a, b; Lee et al., 2023; Wang et al., 2024) have been proposed and obtained good performance in few-shot generalization tasks. However, all of them are prompt learning based methods and thus cannot explicitly model some useful inductive biases such as text phrases, local spatial connections, and translation invariance in images. We are also interested in whether these inductive biases can improve the tuning of VLMs in extremely low-shot settings – a question that has not been thoroughly explored in current VLM literatures. To address this, we follow the network-based approach and propose two adapters for VLMs in few-shot tasks.

Our adapters are used in a manner similar to many previous approaches (Hu et al., 2021; Chen et al., 2022b; Houlsby et al., 2019), where they are integrated into each transformer block (see Fig. 2, *i.e.*, within MSA and FFN). A general view of our method is shown in Fig. 3 (left). It is possible to achieve better performance by integrating adapters at different layers or in different blocks for the text and image encoders. However, for simplicity, this work uses the same arrangement for both text and image encoders, because this approach greatly reduces the efforts for engineering tuning. We find such an identical integration method also achieves leading performance in different benchmarks. Next, we will present the details in the following.

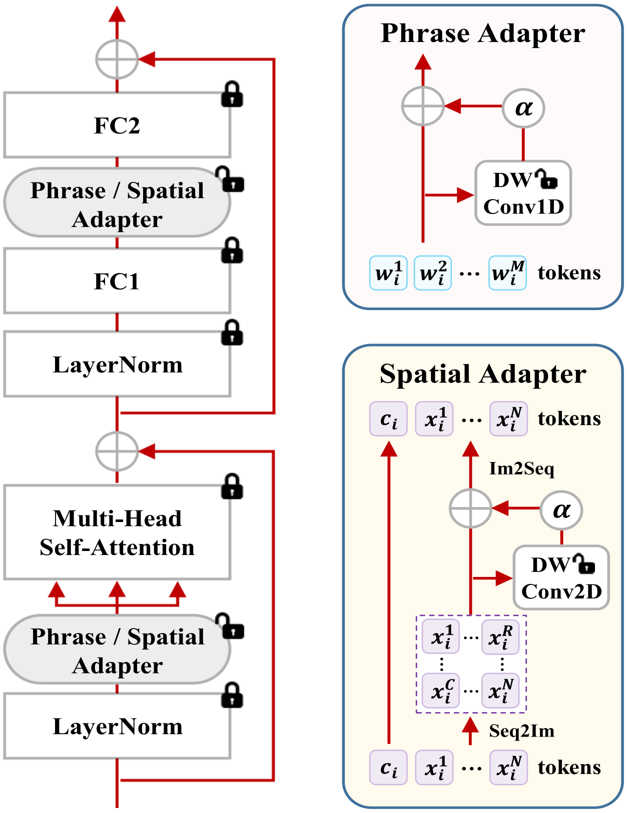


Fig. 3 Model-level Inductive Biases Integration. Left: The general view of our adapters for a transformer block. In detail, we enrich the model with inductive biases after the layer normalization and the first FC layer in MSA and FFN blocks respectively. Top-right: the structure of phrase adapter for text encoder. Bottom-right: the structure of spatial adapter for image encoder, where Seq2Im$(\cdot)$ reshapes 1D sequence to 2D feature map, and Im2Seq$(\cdot)$ reshapes it back. α is a scaling factor. Note that, Only DWConv1D and DWConv2D in adapters are tuned

Phrase Adapter (PA). The widely used self-attention mechanism is a powerful method to capture relationships from large training corpus, with an emphasis on longer-range dependencies between words across the entire sentence. Currently, the enriched prompts, *e.g.*, "A photo of shiba. A shiba inu is a small, compact dog with a fox-like face", contain strong connections between adjacent words. To tailor the few-shot generalization task where relationships between adjacent words is pivotal, we instead use a simpler convolution approach – "Phrase Adapter (PA)" to explicitly model such phrase cues. Let us use $\boldsymbol{W} = [\boldsymbol{w}^j]_{j=1}^{M}$, $\boldsymbol{W} \in \mathcal{R}^{M \times D}$ to denote M word tokens with D dimensions. $\boldsymbol{W}$ will be fed into the transformer block. In the following, we omit the layer index for simplicity. Our PA is shown in Fig. 3 (Top-right) and can be formulated as:

$$PA(\boldsymbol{W}) : \boldsymbol{W} \leftarrow \boldsymbol{W} + \alpha \cdot DWConv1D(\boldsymbol{W}) \tag{4}$$

where $DWConv1D$ is a 1D depthwise convolutional layer with a kernel size of $3 \times 1 \times D$. This layer convolves along M words to capture phrase relationships through adjacent words. The output shape of this layer is the same as $\boldsymbol{W}$ by zero padding. $DWConv$ is adopted in the PA because it introduces only a small number of trainable parameters, which helps reduce overfitting. Additionally, similar to other adapters (Hu et al., 2021; Chen et al., 2022b; Houlsby et al., 2019), we use a scaling factor α to control the degree of integration of specific knowledge into the pre-trained VLMs. Obviously, a larger α integrates more task-specific knowledge and vice versa. After defining PA, we add Eq. (4) to each transformer block as shown in Fig. 3 (Left), and modify Eqs. (1) and (2) for the text encoder as follows:

$$\boldsymbol{W} \leftarrow \boldsymbol{W} + Attention(\underline{PA}(LN_{MSA}(\boldsymbol{W})) \tag{5}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} + FC2(\underline{PA}(FC1(LN_{FFN}(\boldsymbol{W})))) \tag{6}$$

As shown in Fig. 3 (Top-right), only kernels of $DWConv1D$ in PA are optimized to capture task-specific knowledge, highlighted by underline in Eqs. (5) and (6).

Spatial Adapter (SA). It is well known that certain inductive biases, such as locality, translation equivariance, and translation invariance, are essential for learning effective representations of images (Krizhevsky et al., 2012; He et al., 2016; Simonyan and Zisserman, 2015). By integrating these inductive biases, models become more sample- and parameter-efficient. However, these inductive biases have not been leveraged in the adaptation of current VLMs, and their effectiveness in the extremely few-shot scenarios has not been demonstrated. Therefore, we propose a "Spatial Adapter (SA)" to capture these inductive biases. Let $\boldsymbol{Y} = [\boldsymbol{c}; \boldsymbol{X}]$, $\boldsymbol{Y} \in \mathcal{R}^{(N+1)\times D}$ represent the input to the transformer block of the image encoder, where $\boldsymbol{c}$ is the extra class token and $\boldsymbol{X} = [\boldsymbol{x}^j]_{j=1}^N$, $\boldsymbol{X} \in \mathcal{R}^{N\times D}$ consists of N patch tokens. Our SA is shown in Fig. 3 (Bottom-right) and formulated as follows:

$$\begin{aligned} SA(\boldsymbol{Y}): \quad & \boldsymbol{c}, \boldsymbol{X} \leftarrow Split(\boldsymbol{Y}) \\ & \boldsymbol{X} \leftarrow Seq2Im(\boldsymbol{X}) \\ & \boldsymbol{X} \leftarrow \boldsymbol{X} + \alpha \cdot DWConv2D(\boldsymbol{X}) \\ & \boldsymbol{Y} \leftarrow [\boldsymbol{c}; Im2Seq(\boldsymbol{X})] \end{aligned} \tag{7}$$

where $Seq2Im$ transforms the sequential patch tokens $\boldsymbol{X}$ into an R-row by C-column tensor $\mathcal{R}^{R\times C\times D}$ (Fig. 3 (Bottom-right)). A 2D depthwise convolutional layer – $DWConv2D$ with a kernel size of $3 \times 3 \times 1 \times D$ – is used to convolve along the 2D spatial dimensions to capture inductive biases embedded in the image. $Im2Seq$ then transforms the tensor back into sequential patch tokens. With SA, our formulation for each transformer block in the image encoder is shown as follows:

$$\boldsymbol{Y} \leftarrow \boldsymbol{Y} + Attention(\underline{SA}(LN_{MSA}(\boldsymbol{Y})) \tag{8}$$

$$\boldsymbol{Y} \leftarrow \boldsymbol{Y} + FC2(\underline{SA}(FC1(LN_{FFN}(\boldsymbol{Y})))) \tag{9}$$

Similar to PA, we only tune parameters in SA, which is denoted using underline.

3.2.3 Optimization-level Inductive Bias

Given a set of enriched text prompts, we can optimize the parameters in PA and SA to adapt VLMs in different downstream tasks. For the optimization process, a common approach to balance underfitting and overfitting is to monitor the training loss and accuracy on a held-out validation set. This strategy is widely used in many machine learning algorithms. Unfortunately, it is difficult to systematically balance between underfitting and overfitting in the few-shot scenario. To address this issue, we propose an inductive bias for the training step.

Let us recall our adapters. The hyper-parameter α plays an important role in learning task-specific knowledge. On one hand, if α is too small, VLMs cannot be well trained to fit current training data, *i.e.* in an underfitting state. This could be helpful for model generalization in unseen situations, as the general knowledge of VLMs is preserved. On the other hand, if α is too large, VLMs will quickly fit to the limited number of training samples, which can easily lead to an overfitting state. In summary, we should use a small α to improve the generalization ability of VLMs, while use a relative large α to learn useful task-specific knowledge for seen categories. Based on our observation, we propose a *slow-fast optimization* method, which is achieved by a dynamic scaling of α and is shown as follows:

$$DY(\alpha) \leftarrow \begin{cases} s \cdot \alpha, & prob > 0.5 \\ \alpha, & otherwise \end{cases} \tag{10}$$

s is a new introduced hyper-parameter, which is set ≥ 1 to scale the adapter's α in this study. We use the uniform distribution to generate a random number: $prob$, which lies in the range of [0, 1]. During the training phase, we replace the α in all PA and SA with the newly designed Eq. (10). In this case, when a random generated number > 0.5, $DY(\alpha)$ enlarges the α to make the model fit the current data faster, otherwise vice. After training, we directly use α for inference in both seen and unseen categories. The proposed slow-fast optimization dramatically improves the model's generalization ability in unseen scenarios.

4 Experiments

To demonstrate the effectiveness of our proposed LwEIB, the evaluation settings are the same as CoOpOp (Zhou et al. (2022a)), including *Generalization from Base-to-Novel Classes*, *Cross-Dataset Evaluation*, and *Domain Generalization*.

Generalization from Base-to-Novel Classes. Following the protocol used in Zhou et al. (2022a, b), our LwEIB is evaluated on 11 widely used image classification datasets. These include two datasets relevant to general object recognition: ImageNet (Deng et al., 2009) and Caltech101 (Fei-Fei et al., 2004), and five fine-grained image recognition datasets: Pets (Parkhi et al., 2012), Cars (Krause et al., 2013), Flowers (Nilsback and Zisserman, 2008), Food101 (Bossard et al., 2014), and Aircraft (Maji et al., 2013). Additionally, the evaluation covers a scene understanding dataset – SUN397 (Xiao et al., 2010), a texture dataset – DTD (Cimpoi et al., 2014), a satellite-image recognition dataset – EuroSAT (Helber et al., 2019), and an action classification dataset – UCF101 (Soomro et al., 2012). This comprehensive evaluation spans diverse recognition tasks, facilitating an assessment of the models' generalization capabilities. Similar to Zhou et al. (2022a); Khattak et al. (2023a); Lu et al. (2022); Yao et al. (2023); Bulat and Tzimiropoulos (2023), the model is trained on base classes with 16 shots, and then tested on both base and novel categories.

Cross-Dataset Evaluation. Similar to the Base-to-Novel experiments, the cross-dataset evaluation uses the 11 datasets mentioned above. Following the methodology advocated in CoCoOp (Zhou et al., 2022a), all models are trained on ImageNet with 1,000 categories, each category consisting of 16 training samples. After training, the models are directly evaluated on the other 10 datasets without any additional tuning.

Domain Generalization. To assess the robustness of models under large distribution shifts, Zhou et al. (2022a) propose to examine the ImageNet fine-tuned models on other four ImageNet variants, each characterized by different types of domain shifts. These datasets are ImageNetV2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021), and ImageNet-R (Hendrycks et al., 2021). We also adopt this approach for a more comprehensive study.

Implementation Details. Following previous studies (Zhou et al., 2022a, b; Khattak et al., 2023a; Lu et al., 2022; Yao et al., 2023; Bulat and Tzimiropoulos, 2023), our experiment operates only in a few-shot manner, specifically using 16 shots per category. We utilize the ViT-B/16 based CLIP model in all experiments. The standard template a photo of a <CN>", with <CN>" substituting class names, serves as the handcrafted text prompt. Furthermore, in our text-level inductive biases, we use CuPL (Pratt et al., 2023) as the additional category-wise description. The number of descriptions K for each category is set to 20. For PA and SA, we use a kernel size of 3 in both $DWConv1D$ and $DWConv2D$ to capture inductive biases at the model level. In the Base-to-Novel generalization, the scaling factor for the adapters, α, is set to 0.025, and the multiplier s in Eq. (10) is set to 2.5. We train our model for 30 epochs with a batch size of 16 and a learning rate of 0.25. In other two experimental settings, similar to MaPLe (Khattak et al., 2023a), which adjusts training configurations to avoid overfitting, we train our model for 10 epochs with a batch size of 64 and a learning rate of 0.2. The scaling factor α and the multiplier s are set to 0.05 and 10.0, respectively. All optimization is performed using an SGD solver with a momentum of 0.9 and a weight decay of 0.0005. All models are trained under a cosine learning rate schedule on a single GPU device with mixed-precision. We report Base and Novel class accuracies and their harmonic mean (HM) in the Base-to-Novel Generalization. For other two settings, we report class accuracy on each dataset. All results are averaged over three runs with three different seeds.

4.1 Main Results

Base-To-Novel Generalization. We conduct an analysis of our LwEIB in comparison with many state-of-the-art methods, including the zero-shot baseline – CLIP (Radford et al., 2021), text-driven prompt learners such as CoOp (Zhou et al., 2022b), CoOpOp (Zhou et al., 2022a), ProDA (Lu et al., 2022), KgCoOp (Yao et al., 2023), and LASP-V (Bulat and Tzimiropoulos, 2023), and multi-modal prompt learners such as RPO (Lee et al., 2023), MaPLe (Khattak et al., 2023a), PromptSRC (Khattak et al., 2023b), and HPT (Wang et al., 2024). The evaluation is based on the recognition accuracy on 11 datasets comprising both base (Base) and novel (Novel) classes, as well as their harmonic mean (HM) (Xian et al., 2017; Zhou et al., 2022a). All results are presented in Table 1.

Based on the reported results, two main conclusions can be made. First, the proposed LwEIB exhibits the highest overall performance across 11 datasets, as assessed by various evaluation metrics that include both base and novel class accuracies, as well as their HM scores. Specifically, among all the compared methods, HPT (Wang et al., 2024) provides the best results with an HM score of 80.23. This method adapts VLMs with structured linguistic knowledge such as category-wise descriptions, attributes, and their relationships, which serve as strong text-level inductive biases for different categories in different datasets. Instead of using only text-level inductive biases, our LwEIB extensively integrates text-level, model-level, and optimization-level inductive biases to adapt VLMs. By integrating such multi-level inductive biases, our method can better tune VLMs in most few-shot scenarios. As a result, our LwEIB achieves significantly better performance than HPT in base, novel, and HM scores respectively (see Table 1: Base: 84.45 *vs.* 84.32; Novel: 78.21 *vs.* 76.86; HM: 81.21 *vs.* 80.23)

Table 1 Comparison with state-of-the-art methods across diverse datasets in the Base-to-Novel Generalization setting

Methods	Entry	Avg			ImNet			Caltech101			Pets		
		Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	2021	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	2022	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	2022	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	2022	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	2023	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	2023	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
LASP-V	2023	83.18	76.11	79.48	76.25	71.17	73.62	98.17	94.33	96.21	95.73	**97.87**	**96.79**
RPO	2023	81.13	75.00	77.78	76.60	71.57	74.00	97.97	94.37	96.03	94.63	97.50	96.05
PromptSRC	2023	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
HPT	2024	84.32	76.86	80.23	**77.95**	70.74	**74.17**	98.37	94.98	96.65	**95.78**	97.65	96.71
LwEIB	Ours	**84.45**	**78.21**	**81.21**	76.64	**71.64**	74.06	**98.47**	**95.47**	**96.95**	95.70	97.40	96.54

Methods	Entry	Cars			Flowers			Food101			Aircraft		
		Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	2021	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	2022	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	2022	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	2022	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	2023	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	2023	72.94	74.00	73.47	95.92	72.46	82.56	90.71	**92.05**	91.38	37.44	35.61	36.50
LASP-V	2023	75.23	71.77	73.46	97.17	73.53	83.71	**91.20**	91.90	**91.54**	38.05	33.20	35.46
RPO	2023	73.87	75.53	74.69	94.13	76.67	84.50	90.33	90.83	90.58	37.33	34.20	35.70
PromptSRC	2023	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
HPT	2024	76.95	74.23	75.57	**98.17**	**78.37**	**87.16**	90.46	91.57	91.01	42.68	38.13	40.28
LwEIB	Ours	**80.07**	74.01	**76.92**	97.53	77.50	86.37	90.63	91.73	91.18	**45.11**	**42.60**	**43.82**

Methods	Entry	SUN397			DTD			EuroSAT			UCF101		
		Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	2021	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	2022	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	2022	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	2022	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	2023	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	2023	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
LASP-V	2023	80.70	79.30	80.00	81.10	62.57	70.64	**95.00**	**83.37**	**88.86**	85.53	78.20	81.70
RPO	2023	80.60	77.80	79.18	76.70	62.13	68.61	86.63	68.97	76.79	83.67	75.43	79.34
PromptSRC	2023	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	**87.10**	78.80	82.74
HPT	2024	82.57	79.26	**80.88**	**83.84**	63.33	72.16	94.24	77.12	84.82	86.52	80.06	83.16
LwEIB	Ours	81.10	**79.80**	80.44	82.87	**67.83**	**74.60**	**95.00**	80.01	86.86	85.73	**82.37**	**84.02**

Bold values indicate the best result for each evaluation setting among different methods

Herein, “Base” and “Novel” represent the recognition accuracies on base and novel classes, respectively. Furthermore, “HM” denotes the harmonic mean of base and novel accuracies, thereby encapsulating the balance between adaptation and generalization. All results of other methods are directly taken from their original papers. The proposed LwEIB demonstrates commendable adaptability alongside remarkable efficacy in the generalization of novel classes, thus achieving state-of-the-art performance in HM score.

Secondly, as shown in Table 1, none of the investigated methods achieves superior performance on all evaluation metrics across all 11 datasets. Our LwEIB demonstrates superior performance in novel classes in 6 out of 11 datasets, while also being competitive with other methods in base classes. HPT achieves the best overall performance in Flowers102, and LASP-V performs the best in EuroSAT. Additionally, the zero-shot classifier – CLIP (Radford et al., 2021) – also shows comparable performance to other methods in Caltech101, Pets, and Food101. These results underscore the persistent challenge of Base-to-Novel generalization and highlight the ability of LwEIB to provide the most favorable trade-off.

Cross-Dataset Evaluation. A comprehensive summary of the results is presented in Table 2. Significantly, our LwEIB achieves the highest average accuracy of 68.61, surpassing other state-of-the-art methods. A notable observation is that LwEIB consistently outperforms the second-ranked approach – HPT (Wang et al., 2024) in 6 out of 10 datasets. Additionally, LwEIB remains competitive when evaluated on the training source dataset – ImageNet. These results highlight the commendable zero-shot transferability of our LwEIB, indicating its potential for broader applications on diverse datasets.

Domain Generalization. We directly evaluate the ImageNet fine-tuned model (in the cross-dataset setting) on other four variants of ImageNet. All results are presented inTable 3. Remarkably, our LwEIB demonstrates better performance on 3 out of 4 out-of-distribution datasets. This finding suggests the inherent robustness of our LwEIB, particularly when faced with significant domain shifts. Such an effect underscores the model's ability to overcome limitations imposed by dataset boundaries, thereby enhancing its applicability and effectiveness in diverse real-world scenarios.

4.2 Ablative Analysis

We run a number of ablation experiments and show results in Tables 4, 5, 6, 7, Figs. 4, and 5. These results are all averaged over 11 datasets used in the Base-to-Novel Generalization setting. We report base accuracy (Base Acc), novel accuracy (Novel Acc), and their HM.

Effectiveness of Different Inductive Biases. Our goal is to effectively integrate multi-level inductive biases for tuning VLMs in different downstream tasks. In this study, we perform an in-depth analysis of each proposed inductive bias. The baseline is the original zero-shot CLIP classifier. We gradually add text-level, model-level, and optimization-level (Optim-level) inductive biases to CLIP and show all results in Table 4. By enriching custom texts, each category gains detailed descriptions, and such descriptions can

Table 2 Comparison with state-of-the-art methods in the Cross-Dataset Evaluation setting

Methods	Entry	ImNet	Caltech101	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Avg
CLIP	2021	66.72	92.98	89.13	65.29	71.30	86.11	24.90	62.59	44.56	47.84	66.83	65.15
CoOp	2022	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp	2022	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe	2023	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
PromptSRC	2023	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
HPT	2024	**71.72**	94.20	92.63	66.33	74.84	86.21	25.68	68.75	50.87	47.36	**70.50**	67.74
LwEIB	Ours	71.31	**94.51**	92.50	**66.58**	73.03	**86.37**	**27.70**	**69.33**	50.63	**55.37**	70.03	**68.61**

Bold values indicate the best result for each evaluation setting among different methods

All models are trained on the full 1000 categories in ImageNet (Deng et al., 2009) dataset under the 16-shot experiment setting and directly transferred to other datasets for evaluation. Following previous works (Zhou et al., 2022a, b), the average performance is calculated over the transferred 10 datasets. All results of other methods are directly from their original papers. Overall, our LwEIB performs the best in 6 out of 10 datasets, and also achieves the best average performance over 10 datasets, demonstrating strong zero-shot transferability.

Table 3 Comparison with state-of-the-art methods in the Domain Generalization setting

Methods	Entry	ImNet	-V2	-S	-A	-R	Avg
CLIP	2021	66.73	60.83	46.15	47.77	73.96	57.18
CoOp	2022	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp	2022	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe	2023	70.72	64.07	49.15	50.90	76.98	60.27
PromptSRC	2023	71.27	64.35	49.55	50.90	77.80	60.65
HPT	2024	**71.72**	65.25	49.36	50.85	77.38	60.71
LwEIB	Ours	71.31	64.47	**50.07**	**51.00**	**77.81**	**60.84**

Bold values indicate the best result for each evaluation setting among different methods

All models are trained on the full 1000 categories in ImageNet (Deng et al., 2009) dataset under the 16-shot experiment setting and directly used for evaluating on domain generalization. Following previous works (Zhou et al., 2022a, b), the average performance is calculated over the transferred 4 datasets. All results of other methods are directly from their original papers. Overall, our LwEIB obtains the best performance in 3 out of 4 out-of-distribution datasets, showing strong robustness to domain shifts.

Table 4 Effect of integration of our proposed inductive biases

Inductive Biases	Base Acc	Novel Acc	HM
1: Baseline (CLIP)	69.34	74.22	71.70
2: + Text-level	72.26	76.11	74.14
3: + Model-level	85.01	76.56	80.56
4: + Optim-level	84.45	78.21	81.21

We gradually add different levels of inductive biases into the baseline method – CLIP for tuning.

Table 5 Performance with different variants of trainable components

Component Variants	Base Acc	Novel Acc	HM
1: Only PA	80.71	76.01	78.29
2: Only SA	81.30	77.92	79.57
3: Only $\boldsymbol{T}_{ct}$	83.72	78.00	80.76
4: LwEIB	84.45	78.21	81.21

Integrating with all proposed components performs the best.

Table 6 Performance with different adapters in the text encoder

Adapters	Base Acc	Novel Acc	HM
1: MLP	84.17	74.42	78.99
2: $LoRA$	84.10	74.20	78.84
3: $Attention$	82.95	73.38	77.87
4: Our PA	84.45	78.21	81.21

We replace our PA $(DWConv1D)$ with other adapters.

Table 7 Integrating our adapters after different layers in Eqs. (1) and (2)

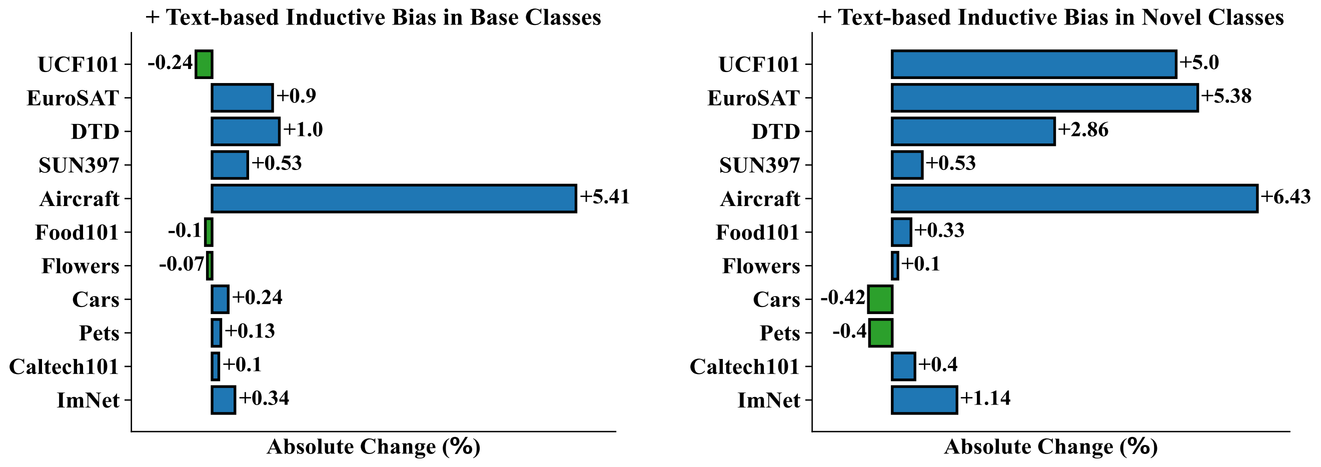
Different Layers	Base Acc	Novel Acc	HM
1: $Attention$ - $FC2$	82.37	77.85	80.05
2: $Attention$ - $FC1$	84.28	78.25	81.15
3: LN_{MSA} - $FC2$	82.94	76.96	79.84
4: LN_{MSA} - $FC1$	84.45	78.21	81.21

Our default strategy is adding adapters after LN_{MSA} and $FC1$ layers (denoted as LN_{MSA} - $FC1$).

provide more discriminative information between different categories. This approach enhances the model's performance in all evaluation metrics. Additionally, the performance is further improved by introducing adapters to learn more inductive biases within the model. Finally, when training with our optimization-level inductive bias (*i.e.*, the slow-fast method), the base class accuracy has a small performance decrement but the accuracy of the novel class is significantly improved. This indeed confirms that our slow-fast optimization method can provide a good trade-off between underfitting and overfitting, leading to better HM scores.

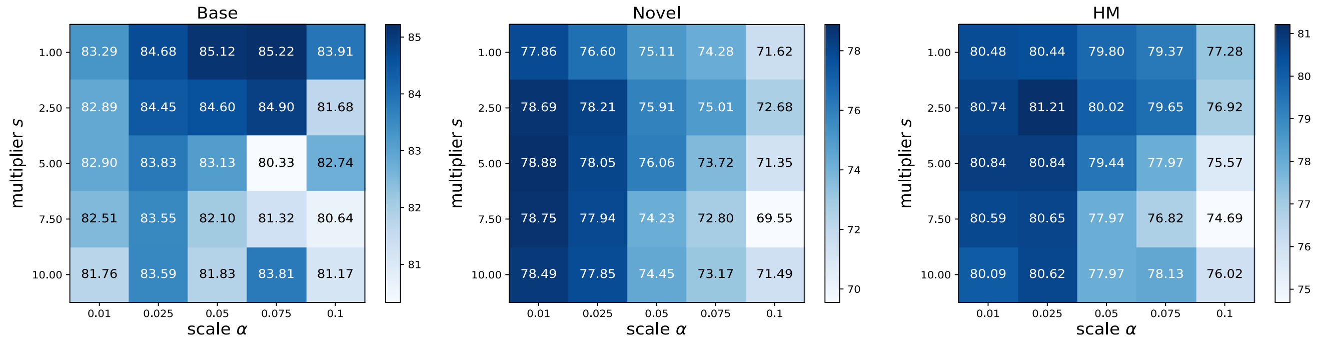
Previous studies (Pratt et al., 2023; Menon and Vondrick, 2022) have demonstrated that incorporating text-based inductive bias (*i.e.*, LLM-generated descriptions) can enhance zero-shot performance across diverse datasets. Similarly, we present per-dataset performance changes with and without the proposed text-level inductive bias to examine how this inductive bias, in combination with other two inductive biases, performs. Details are presented in Fig. 4. As shown, adding the text-level inductive bias generally leads to improved performance, particularly for novel classes. For instance, this approach results in an absolute improvement of 1.14% on the challenging ImageNet dataset. In addition, as presented in Fig. 4 and Table 2, the performance enhancements are particularly notable for more challenging datasets, such as EuroSAT, DTD and Aircraft. This phenomenon may be attributed to a significant distribution shift between these datasets and the original CLIP training data. The introduced text-level inductive bias combined with other two proposed mechanisms can effectively mitigate this shift.

Variants of Trainable Components. We evaluate the effectiveness of different design choices. These design alternatives relate to unimodal adapters introduced in either the test (Only



(a) Results on **base** Classes (b) Results on **novel** Classes

Fig. 4 Comprehensive comparisons of using text-based inductive bias or not in the Base-to-Novel Generalization setting. Adding text-level inductive biases gains improvements in most cases



(a) Results on **base** classes. (b) Results on **novel** classes. (c) Their **HM** scores.

Fig. 5 The effectiveness of different scaling factor α and multiplier s. The horizontal and vertical axis represent different values of α and s. The results in $s = 1.00$ means that we do not use the proposed slow-fast optimization method (Eq. (10)) during training phase. Best viewed in PDF form

PA) or the image (Only SA) encoder, with the exclusion of a handcrafted text prompt (Only $\boldsymbol{T}_{ct}$). In Table 5, we present the aggregated results averaged over 11 recognition datasets used in the Base-to-Novel setting. Our analysis reveals that using only PA or SA alone cannot incorporate inductive biases from dual domains and does not perform as well as our full LwEIB. Furthermore, the inclusion of the handcrafted prompt elevates the HM from 80.76 to 81.21, underscoring its significance.

Different Adapters in Text Encoder. It often uses MLP (Houlsby et al., 2019), LoRA (Hu et al., 2021), or self-attention (Devlin et al., 2018) to tune the text encoder. Here, we adopt 1D convolutional layers to explicitly extract information from text phrases. The key distinction between our method and others is that we intentionally add a strong inductive bias into the text encoder. To verify the effectiveness of our choice, we compare various adapters in Table 6. As shown, neither MLP nor LoRA effectively model relationships between words. Additionally, attention mechanisms, which can capture long-range dependencies, do not perform as well as ours. This may be due to two factors: first, self-attention without inductive biases requires substantial data to learn these relationships (Dosovitskiy et al., 2020); second, self-attention may overly focus on unrelated context (Ye et al., 2024), resulting in noise features. All these results demonstrate that our PA introduces a strong inductive bias for text phrases, leads to more stable training outcomes, and better generalization ability in few-shot settings.

Integrating Adapters after Different Layers. Currently, our adapters are applied after the layer normalization (LN_{MSA}) in MSA and after the first fully-connected layer ($FC1$) in FFN respectively. This raises the question of whether placing our adapters in different locations are more efficient. Therefore, we test three additional configurations: placing adapters after attention in MSA and after $FC2$ in FFN ($Attention$-$FC2$), after the attention in MSA and after $FC1$ in FFN

($Attention$-$FC1$), and after LN_{MSA} in MSA and after $FC2$ in FFN (LN_{MSA}-$FC2$). As shown in Table 7, our current design (LN_{MSA}-$FC1$) still performs the best among all configurations.

Analysis on Scale α and Multiplier s. We carefully explore the influence of different scaling factor α and multiplier s in Eq. (10) for our slow-fast optimization method. All results are shown in Fig. 5. A relatively large α helps our model to better recognize base classes but results in inferior performance for novel classes (*e.g.*, $\alpha = 0.01, s = 1.00$), while a smaller α makes the model hard to tune and usually results in a lower accuracy for base classes but a higher accuracy for novel classes (*e.g.*, $\alpha = 0.01, s = 1.00$). Also, adjusting s in our slow-fast optimization slightly decreases the base accuracy while significantly increases the novel accuracy, reaching the best HM (81.21) at ($\alpha = 0.025, s = 2.50$). This further confirms the effectiveness of our optimization-level inductive bias. In addition, when we train our model with very large values such as ($\alpha = 0.1, s = 10.0$), the model tends to overfit to the few-shot training examples, leading to inferior results in both base and novel classes. In addition, our slow-fast optimization is probabilistic, which suggests results may vary significantly on each run. We thus report mean and standard deviation across three random runs with different seeds here: Base: 84.45±0.30, Novel: 78.21±0.32, HM: 81.21±0.17. All results show that our method demonstrates a certain degree of robustness across different runs. Possible future extensions of this work will involve proposing a new slow-fast optimization method with adaptive scaling factor and multiplier, or further reducing the effects caused by randomness.

Comparison our $DY(\alpha)$ with Dropout Regularization. Our optimization-level inductive bias, $DY(\alpha)$, can be viewed as a probabilistic perturbation training mechanism that aims to balance the model between overfitting and underfitting. This approach shares the objective of commonly used dropout methods - namely, to reduce overfitting and enhance generalization through probabilistic perturbations. However, traditional dropout does not account for differences in fit between tasks and may not be well-suited for few-shot generation scenarios. To test this hypothesis, we conducted a series of experiments, with results shown in Table 8. From the table, it is evident that an appropriate dropout ratio can indeed improve generalization to the training task (Base Acc), but offers limited usage for novel tasks (Novel Acc). In contrast, our method achieves a balanced performance across both training and novel tasks.

Table 8 Comparison of the proposed optimization-level inductive bias – $DY(\alpha)$ with the dropout method

Dropout Method	Base Acc	Novel Acc	HM
1: Dropout−0.1	84.77	76.89	80.64
2: Dropout−0.2	84.65	76.88	80.58
3: Dropout−0.5	83.24	76.70	79.84
4: Our $DY(\alpha)$	84.45	78.21	81.21

Dropout regularization with different ratios is denoted as Dropout-X. Results with a dropout ratio greater than 0.5 are not shown, as large dropout rates can make the network unstable.

4.3 Computational Costs

Table 9 compares computational costs including additional parameters, and train/test FPS, of LwEIB with other approaches. First, our LwEIB adds significantly fewer additional parameters compared to MaPLe (507 K *vs.* 3555 K) but has more parameters than two recently state-of-the-art methods, HPT and PromptSRC. However, our method achieves the best HM score among all compared methods. Second, the training speed of our LwEIB is comparable to that of CoOp, MaPLe, and PromptSRC, while the testing speed surpasses that of CoCoOp and HPT, placing it at a moderate level. These comparative results indicate that efficiently utilizing parameters while maintaining the running speed still remains a big challenge in leveraging VLMs. Our method currently provides a favorable trade-off.

Table 9 Comparison of additional parameters (+Params) and running speed (FPS) among different methods using ImageNet dataset

Method	+Params	Train FPS	Test FPS	HM
CoOp	2 K	12	558	71.66
CoCoOp	35 K	3	8	75.83
MaPLe	3555 K	12	578	78.55
PromptSRC	46 K	11	577	79.97
HPT	296 K	7	150	80.23
LwEIB	507 K	11	264	81.21

Train FPS is tested with a batch size of 4 due to the high GPU memory usage in CoCoOp. Test FPS is tested with a batch size of 256. All speeds are tested on a single GTX 8000 GPU.

5 Conclusion

The integration of large-scale Vision-Language Models (VLMs) into downstream tasks poses a significant challenge, mainly due to the vast number of model parameters in contrast to the limited availability of training data. In this study, we propose a novel learning framework – **L**earning **w**ith **E**nriched **I**nductive **B**iases (LwEIB) – that can simultaneously incorporate additional inductive biases at the text, model, and optimization levels. Additionally, our LwEIB fine-tunes only the added adapters to capture inductive biases within the model, which is also a parameter-efficient fine-tuning method. To evaluate the effectiveness of our proposed LwEIB, we conduct a series of experiments on three challenging tasks: adaptation to novel classes, transfer to new target datasets, and accommodation of unseen domain shifts. Comparative evaluations with state-of-the-art methods underline the superior performance of our LwEIB framework across all three evaluation criteria.

Author Contributions Lingxiao Yang (First Author): Conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, writing—original draft, writing—review & editing; Ru-Yuan Zhang: Data curation, formal analysis, investigation, methodology, validation, visualization, writing—original draft; Qi Chen: Data curation, validation, writing—review & editing; Xiaohua Xie (Corresponding Author): Conceptualization, methodology, project administration, supervision, validation, writing—original draft, writing—review & editing.

Funding This project is partially supported by National Key Research and Development Program of China (2024YFA1011900), Key-Area Research and Development Program of Guangzhou (202206030003), National Natural Science Foundation of China (62206316, 32441102, 32100901), Natural Science Foundation of Guangdong (2022A1515011254), Project of Guangdong Provincial Key Laboratory of Information Security Technology (2023B1212060026), Open Research Fund of the State Key Laboratory of Brain-Machine Intelligence Zhejiang University (BMI2400021), and Key Laboratory of Brain-Machine Intelligence for Information Behavior-Ministry of Education (2023JYBKFKT005).

Declarations

Conflict of interest. The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al. (2022). Flamingo: A visual language model for few-shot learning. *Advances in neural information processing systems, 35*, 23716–23736.

Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101–mining discriminative components with random forests. *European conference on computer vision* (pp. 446–461).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877–1901.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Bulat, A., & Tzimiropoulos, G. (2023). Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23232–23241).

Chen, G., Yao, W., Song, X., Li, X., Rao, Y., & Zhang, K. (2022a). Plot: Prompt learning with optimal transport for vision-language models. *International conference on learning representations.*

Chen, Q., Chen, Y., Huang, Y., Xie, X., & Yang, L. (2024). Region-based online selective examination for weakly supervised semantic segmentation. *Information Fusion, 107*, 102311.

Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., & Luo, P. (2022). Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems, 35*, 16664–16678.

Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., & Qiao, Y. (2022c). Vision transformer adapter for dense predictions. *International conference on learning representations.*

Chen, Z., Huang, X., Guan, Q., Lin, L., & Luo, W. (2023). A retrospect to multi-prompt learning across vision and language. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22190–22201).

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3606–3613).

Deng, J., Dong, W., Socher, R., Li, L.- J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 248–255).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805,

Ding, J., Xue, N., Xia, G.-S., & Dai, D. (2022). Decoupling zero-shot semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11583–11592).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *International conference on learning representations.*

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 178–178).

Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., & Ma, L. (2022). Promptdet: Towards open-vocabulary detection using uncurated images. *European conference on computer vision* (pp. 701–717).

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., & Qiao, Y. (2023). Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision, 132*(2), 581–595.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 580–587).

Gu, X., Lin, T.-Y., Kuo, W., & Cui, Y. (2021). Open-vocabulary object detection via vision and language knowledge distillation. *International conference on learning representations.*

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2961–2969).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 770–778).

He, W., Jamonnak, S., Gou, L., & Ren, L. (2023). Clip-s4: Language-guided self-supervised semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11207–11216).

Helber, P., Bischke, B., Dengel, A., & 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. others (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8340–8349).

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural adversarial examples. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15262–15271).

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. *International conference on machine learning* (pp. 2790–2799).

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Wang, L. (2021). *others*. Lora: Low-rank adaptation of large language models. International conference on learning representations

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7132–7141).

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S. others (2023). Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., & Schmidt, L. (2021). Openclip. *Zenodo*. https://doi.org/10.5281/zenodo.5143773

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *International conference on machine learning* (pp. 4904–4916).

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S.-N. (2022). Visual prompt tuning. *European conference on computer vision* (pp. 709–727).

Jiang, H., Zhang, J., Huang, R., Ge, C., Ni, Z., Lu, J., & Huang, G. (2022). Cross-modal adapter for text-video retrieval. arXiv preprint arXiv:2211.09623

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., & Khan, F. S. (2023a). Maple: Multi-modal prompt learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19113–19122).

Khattak, M. U., Wasim, S. T., Naseer, M., Khan, S., Yang, M.-H., & Khan, F. S. (2023b). Self-regulating prompts: Foundational model adaptation without forgetting. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15190–15200).

Kim, K., Laskin, M., Mordatch, I., & Pathak, D. (2021). How to adapt your large-scale vision-and-language model.

Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 554–561).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Communications of the ACM, 60*, 84–90.

Lee, D., Song, S., Suh, J., Choi, J., Lee, S., & Kim, H. J. (2023). Read-only prompt optimization for vision-language few-shot learning. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1401–1411).

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. *Proceedings of the conference on empirical methods in natural language processing* (pp. 3045–3059).

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. arXiv:2101.00190 [cs.CL]

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2117–2125).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European conference on computer vision* (pp. 740–755).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. *European conference on computer vision* (pp. 21–37).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3431–3440).

Lu, Y., Liu, J., Zhang, Y., Liu, Y., & Tian, X. (2022). Prompt distribution learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5206–5215).

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

Maniparambil, M., Vorster, C., Molloy, D., Murphy, N., McGuinness, K., & O'Connor, N. E. (2023). Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 262–271).

Menon, S., & Vondrick, C. (2022). Visual classification via description from large language models. arXiv preprint arXiv:2210.07183

Nilsback, M.-E., & Zisserman, A. (2008). Automated flower classification over a large number of classes. *Sixth indian conference on computer vision, graphics & image processing* (pp. 722–729).

Oord, A.v.d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748

Parkhi, O. M., Vedaldi, A., Zisserman, A., & Jawahar, C. (2012). Cats and dogs. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3498–3505).

Pratt, S., Covert, I., Liu, R., & Farhadi, A. (2023). What does a platypus look like? generating customized prompts for zero-shot image classification. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15691–15701).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. others (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning* (pp. 8748–8763).

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., & Lu, J. (2022). Denseclip: Language-guided dense prediction with context-aware prompting. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18082–18091).

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? *International conference on machine learning* (pp. 5389–5400).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence, 39*(6), 1137–1149.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Saha, O., Van Horn, G., & Maji, S. (2024). Improved zero-shot classification by adapting vlms with text descriptions. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17542–17552).

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., & Jitsev, J. (2022). LAION-5b: An open large-scale dataset for training next generation image-text models. *Thirty-sixth conference on neural information processing systems datasets and benchmarks track*. https://openreview.net/forum?id=M3Y74vmsMcY

Sd'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., & Sagun, L. (2021). Convit: Improving vision transformers with soft convolutional inductive biases. *International conference on machine learning* (pp. 2286–2296).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International conference on learning representations.*

Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402

Stickland, A. C., & Murray, I. (2019). Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. *International conference on machine learning* (pp. 5986–5995).

Sun, Y., Zheng, L., Deng, W., & Wang, S. (2017). Svdnet for pedestrian retrieval. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3800–3808).

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning* (pp. 6105–6114).

Wang, H., Ge, S., Lipton, Z., & Xing, E. P. (2019). Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems, 32*, 10506–10518.

Wang, Y., Jiang, X., Cheng, D., Li, D., & Zhao, C. (2024). Learning hierarchical prompt with structured linguistic knowledge for vision-language models. *Proceedings of the AAAI conference on artificial intelligence.*

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., & Cao, Y. (2021). Simvlm: Simple visual language model pretraining with weak supervision. *International conference on learning representations.*

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. *European conference on computer vision* (pp. 3–19).

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22–31).

Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4582–4591).

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3485–3492).

Yang, L., Zhang, R.-Y., Li, L., & Xie, X. (2021). Simam: A simple, parameter-free attention module for convolutional neural networks. *International conference on machine learning* (pp. 11863–11874).

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

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., & Xu, C. (2021). Filip: Fine-grained interactive language-image pre-training. *International conference on learning representations.*

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(6), 2872–2893.

Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., & Wei, F. (2024). Differential transformer. arXiv preprint arXiv:2410.05258

Yu, T., Lu, Z., Jin, X., Chen, Z., & Wang, X. (2023). Task residual for tuning vision-language models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10899–10909).

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J. (2021). Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432

Zang, Y., Li, W., Zhou, K., Huang, C., & Loy, C. C. (2022a). Open-vocabulary detr with conditional matching. *European conference on computer vision* (pp. 106–122).

Zang, Y., Li, W., Zhou, K., Huang, C., & Loy, C. C. (2022b). Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., & Beyer, L. (2022). Lit: Zero-shot transfer with locked-image text tuning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18123–18133).

Zhang, Q., Lai, J., Feng, Z., & Xie, X. (2024). Uncertainty modeling for group re-identification. *International Journal of Computer Vision, 132*, 3046.

Zhang, Q., Lai, J., Xie, X., Jin, X., & Huang, S. (2024). Separable spatial-temporal residual graph for cloth-changing group re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*, 5791.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., & Li, H. (2022). Tip-adapter: Training-free adaption of clip for few-shot classification. *European conference on computer vision* (pp. 493–510).

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H. others (2022). Regionclip: Region-based language-image pretraining. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16793–16803).

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(6), 1452–1464.

Zhou, C., Loy, C. C., & Dai, B. (2022c). Extract free dense labels from clip. *European conference on computer vision* (pp. 696–712).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022a). Conditional prompt learning for vision-language models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16816–16825).

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

Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., & Misra, I. (2022d). Detecting twenty-thousand classes using image-level supervision. *European conference on computer vision* (pp. 350–368).

Zhu, B., Niu, Y., Han, Y., Wu, Y., & Zhang, H. (2023). Prompt-aligned gradient for prompt tuning. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15659–15669).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.