A_3B_2: Adaptive Asymmetric Adapter for Alleviating Branch Bias in Vision-Language Image Classification with Few-Shot Learning

Yiyun Zhou[1], **Zhonghua Jiang**[1], **Wenkang Han**[1], **Kunxi Li**[1], **Mingjing Xu**[2], **Chang Yao**[1] and **Jingyuan Chen**[1*]
[1]Zhejiang University
[2]Swansea University
{yiyunzhou, jiangzhonghua, wenkangh, kunxili, changy, jingyuanchen}@zju.edu.cn,
mingjing.xu@swansea.ac.uk

Abstract

Efficient transfer learning methods for large-scale vision–language models (*e.g.*, CLIP) enable strong few-shot transfer, yet existing adaptation methods follow a fixed fine-tuning paradigm that implicitly assumes a uniform importance of the image and text branches, which has not been systematically studied in image classification. Through extensive analysis, we reveal a *Branch Bias* issue in vision–language image classification: adapting the image encoder does not always improve performance under out-of-distribution settings. Motivated by this observation, we propose $\boldsymbol{A_3B_2}$, an **A**daptive **A**symmetric **A**dapter that alleviates **B**ranch **B**ias in few-shot learning. A_3B_2 introduces *Uncertainty-Aware Adapter Dampening* (*UAAD*), which automatically suppresses image-branch adaptation when prediction uncertainty is high, enabling soft and data-driven control without manual intervention. Architecturally, A_3B_2 adopts a lightweight asymmetric design inspired by mixture-of-experts with *Load Balancing Regularization*. Extensive experiments demonstrate that A_3B_2 consistently outperforms 11 competitive prompt- and adapter-based baselines.

Extended version: https://arxiv.org/abs/2605.13161

1 Introduction

Large-scale Vision-Language Models (VLMs) with dual branches are foundational for downstream transfer learning tasks due to their strong generalization. CLIP [Radford *et al.*, 2021], trained on 400 million image-text pairs, aligns visual and textual modalities in a shared feature space via separate encoders and contrastive learning. However, full fine-tuning VLMs for specific tasks is computationally expensive and prone to overfitting in few-shot settings.

To address the aforementioned issues, efficient transfer learning methods have been proposed [Zhou *et al.*, 2022b; Yang *et al.*, 2024; Guo and Gu, 2025a; Guo and Gu, 2025b], enabling VLMs to adapt to different downstream tasks with minimal additional parameters. Previous approaches have mainly focused on prompt tuning and adapter-based methods. Prompt tuning aims to address the limitation of VLMs that require manually designed text prompts, which requires expert knowledge and is time-consuming. For instance, in the OxfordPets [Parkhi *et al.*, 2012] dataset, CLIP's text encoder computes class embeddings using the prompt "`A photo of a [CLASS], a type of pet.`". These embeddings are then matched with the outputs by the image encoder, and the most similar one is selected to predict the image category. Most prior prompt tuning methods [Zhou *et al.*, 2022b; Guo and Gu, 2025a] introduce learnable continuous vectors into the text encoder to dynamically model prompts, optimizing them during training while freezing other parameters, thereby enabling adaptation to specific datasets. In contrast, adapter-based methods [Yang *et al.*, 2024] add lightweight, architecture-agnostic modules, making them easily integrable into different VLMs and allowing the models to adapt to potential downstream task spaces.

Although prior methods have achieved impressive few-shot performance in data and model adaptation, most CLIP-based methods follow a fixed fine-tuning paradigm: prompt tuning methods optimize the text encoder, while adapter-based methods are attached to the image encoder [Li *et al.*, 2024; Zhang *et al.*, 2024]. This raises a compelling research question: ***which matters more in different few-shot tasks, CLIP's text or image encoder?*** (We refer to this as the CLIP's *Branch Bias* issue) Recent studies have sparked interest in this issue. Yang *et al.* train only the image encoder while keeping the text encoder fixed, achieving surprising results on compositional understanding and long-text description tasks. Gong *et al.* propose a kernel-based method, significantly improving the image encoder's performance on zero-shot target recognition, fine-grained spatial reasoning, and localization tasks. Fu *et al.* find that in vision-centered benchmarks (*e.g.*, depth estimation, correspondence), VLMs perform worse than their image encoders, with performance nearing random levels. In open-vocabulary semantic segmentation tasks, most past methods [Li *et al.*, 2022; Liang *et al.*, 2023; Xu *et al.*, 2023] freeze the text encoder to preserve generalization, while recent studies [Peng *et al.*, 2025] show that jointly fine-tuning both encoders leads to better segmentation results. However, **there is no consensus on the *Branch Bias* issue in the general tasks of image classification**.

*Corresponding author.

To address gaps in prior work, we conduct extensive exploratory experiments on three visual tasks: base-to-novel generalization, cross-dataset evaluation, and domain generalization. Crucially, our insights reveal a significant phenomenon: **fine-tuning the image encoder does not always yield performance gains and may even degrade robustness on out-of-distribution (OOD) tasks**. While a straightforward response to this discovery would be to rigidly lock the image encoder for OOD scenarios, such a hard-coded approach is cumbersome and impractical, as it requires prior knowledge of the target data distribution to manually toggle the adapters. To resolve this dilemma, we propose a novel ***A**daptive **A**symmetric **A**dapter for alleviating **B**ranch **B**ias* ($\boldsymbol{A_3B_2}$) in vision-language image classification. Unlike rigid switching mechanisms, A_3B_2 achieves adaptability through a uncertainty-aware optimization mechanism. Specifically, we introduce *Uncertainty-Aware Adapter Dampening* (*UAAD*). This mechanism dynamically monitors prediction confidence; when **the model encounters high-uncertainty samples (indicative of potential OOD data), it automatically suppresses the magnitude of the image adapter's contribution** via a regularization penalty. This soft-constraint approach effectively operationalizes our insights without requiring manual intervention. Structurally, inspired by the Mixture of Experts (MoE) design [Fedus *et al.*, 2022; Xue *et al.*, 2022; Tian *et al.*, 2024; Gao *et al.*, 2024a; Zhou *et al.*, 2025], A_3B_2 consists of a single dimensionality-reduction matrix and multiple dimensionality-expansion matrices, further enhanced by a *Load Balancing Regularization* to ensure diverse expert utilization [Shazeer *et al.*, 2017].

Our contributions are summarized as follows:

- We systematically explore the *Branch Bias* issue in image classification for **transformer-based CLIP models**. Three insights discovered help build the optimal adapter structure for vision-language image classification tasks.
- We propose a task-adaptive, structure-asymmetric adapter to alleviate *Branch Bias* in vision-language image classification. **Empirical evidence, theoretical support for the asymmetric structure, and theoretical guarantees of the method** are provided in the extended version, respectively.
- Evaluations on three few-shot tasks show that, across 11 image classification datasets, A_3B_2 consistently achieves leading performance compared to 11 competitive baselines. The experiments also confirm the generality of our insights.

2 Related Work

2.1 CLIP's Branch Bias in Various Tasks

Recent studies have sparked significant interest in the CLIP's *Branch Bias* issue, particularly in tasks requiring fine-grained visual discrimination. For compositional and spatial tasks, the image branch often requires more adaptation. Yang *et al.* demonstrate that training only the image encoder while keeping the text encoder fixed yields superior performance in compositional understanding and long-text description. Similarly, Gong *et al.* propose a kernel-based alignment method that focuses on the image encoder, significantly improving zero-shot target recognition and localization. Fu *et al.* further reveal that in strictly vision-centered benchmarks (*e.g.*, depth estimation), full VLMs can paradoxically perform worse than their frozen image encoders, suggesting textual interference. In contrast, for open-vocabulary semantic segmentation, the consensus has shifted. While earlier methods [Li *et al.*, 2022; Liang *et al.*, 2023; Xu *et al.*, 2023] froze the text encoder to preserve generalization, recent work by Peng *et al.* argues that jointly fine-tuning both encoders is essential for aligning pixel-level features with semantic concepts. Despite these diverse findings in structured prediction and reasoning tasks, the manifestation of this *Branch Bias* in general image classification remains largely underexplored.

2.2 Efficient Transfer Learning in Vision-Language Image Classification

Prompt learning methods enhance the adaptability of Vision-Language Models (VLMs) by dynamically adjusting pretrained representations. CoOp [Radford *et al.*, 2021] introduces the learnable prompt vector, replacing fixed templates with optimized contexts to improve flexibility, though this sacrifices CLIP's zero-shot generalization. CoCoOp [Zhou *et al.*, 2022a] addresses this limitation by incorporating dynamic instance-conditioned prompts, improving generalization to unseen categories. KgCoOp [Yao *et al.*, 2023] further constrains the semantic gap between learned and handcrafted prompts, preserving general textual knowledge and enhancing model generalization. RPO [Lee *et al.*, 2023] employs a masked attention mechanism to freeze the pretrained representations, reducing internal representation changes and improving model stability, particularly in data-scarce scenarios. MaPLe [Khattak *et al.*, 2023] achieves cross-modal alignment by coupling visual-text prompts, enhancing adaptability to new categories and unseen domains. TCP [Yao *et al.*, 2024] innovatively embeds category semantics into prompt tokens, strengthening generalization for new categories and few-shot settings. MMRL [Guo and Gu, 2025a] proposes a shared learnable representation space to facilitate efficient visual-text interaction, with regularization to prevent overfitting, especially suited for few-shot learning and task adaptation. MMRL++ [Guo and Gu, 2025b] further optimizes this framework by employing a parameter-efficient low-rank alignment and progressive feature composition, which jointly facilitate robust generalization and training stability with minimal overhead.

Recent adapter-style methods provide another effective approach for model adaptation, typically integrating lightweight modules into VLMs for downstream task. CLIP-Adapter [Gao *et al.*, 2024b] introduces a residual MLP layer at the image encoder's end, significantly improving performance in visual tasks. MMA [Yang *et al.*, 2024] proposes a multimodal adapter method, aggregating image and text features into a shared space, enhancing consistency between modalities and allowing cross-modal gradient flow, thereby improving the model's overall adaptability and task transfer capability. However, none of these methods have systematically investigated the *Branch Bias* in vision-language image classification, typically adhering to static adaptation paradigms that fail to dynamically balance the dual encoders. Consistent with previous research [Zhou *et al.*, 2022b; Zhou

et al., 2022a; Guo and Gu, 2025b], we use CLIP (transformer-based ViT-B/16 backbone) as the base model for our method.

3 Methodology

3.1 Preliminary

The CLIP model consists of an image encoder $\mathcal{V}$ and a text encoder $\mathcal{T}$, encoding images and corresponding texts.

Image Encoder The image encoder $\mathcal{V}$ consists of L Transformer blocks [Vaswani *et al.*, 2017], denoted as $\{\mathcal{V}_i\}_{i=1}^{L}$. The input image $I \in \mathbb{R}^{H\times W\times 3}$ is divided into M patches, which are then embedded to obtain the initial features x_0:

$$x_0 = \texttt{Patch Embedding}(I), \tag{1}$$

where $x_0 \in \mathbb{R}^{M\times d_v}$ (d_v is the image embedding dimension).

The patch embeddings x_{i-1}, along with a learnable class (CLS) token c_{i-1}, serve as inputs to the next Transformer block $\mathcal{V}_i$ and are processed sequentially:

$$[c_i, x_i] = \mathcal{V}_i([c_{i-1}, x_{i-1}]), \quad i = 1, 2, \ldots, L. \tag{2}$$

Finally, a patch projection layer maps the CLS token c_L from the last Transformer block $\mathcal{V}_L$ into a shared Vision-Language (V-L) latent space to obtain the final image representation x:

$$x = \texttt{Patch Projection}(c_L), \tag{3}$$

where $x \in \mathbb{R}^d$, and d is the feature dimension of the shared V-L space.

Text Encoder For a given text T (*e.g.*, "`A photo of a [CLASS].`"), it is first tokenized and converted into word embeddings w_0 using a Text Embedding layer:

$$[w_0^j]_{j=1}^{N} = \texttt{Text Embedding}(T), \tag{4}$$

where $w_0 \in \mathbb{R}^{N\times d_t}$, with N as the token length and d_t as the text embedding dimension.

At each layer, w_{i-1} is sequentially processed by the next Transformer block $\mathcal{T}_i$ to extract text features w_i:

$$[w_i^j]_{j=1}^{N} = \mathcal{T}_i([w_{i-1}^j]_{j=1}^{N}), \quad i = 1, 2, \ldots, L. \tag{5}$$

The final text representation w is obtained by projecting the last token output from the final Transformer block $\mathcal{T}_L$ into the shared V-L space via a Text Projection layer:

$$w = \texttt{Text Projection}(w_L^N), \tag{6}$$

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

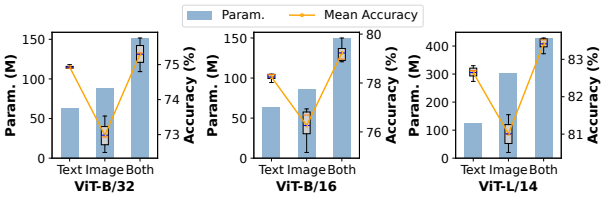
Zero-shot Classification Given the image representation x and text representations $\{w_c\}_{c=1}^{N_c}$, where N_c is the number of classes, CLIP computes the cosine similarity between them:

$$\cos(x, w_c) = \frac{x \cdot w_c}{|x||w_c|}, \tag{7}$$

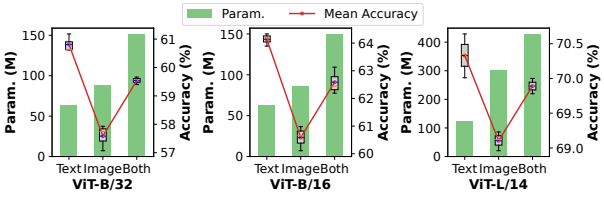
where $|\cdot|$ denotes the L_2 norm. The classification probability for each class c is then predicted as:

$$\mathbf{p}(\hat{y} = c \mid x) = \frac{\exp(\cos(x, w_c)/\tau)}{\sum_{i=1}^{N_c} \exp(\cos(x, w_i)/\tau)}, \tag{8}$$

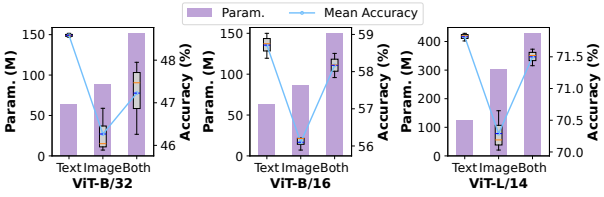
where τ is a learnable temperature parameter in CLIP. The class with the highest probability is selected as the final prediction.



(a) Base-to-Novel Generalization



(b) Cross-Dataset Evaluation



(c) Domain Generalization

Figure 1: The average performance of text or image adapters on three few-shot image classification tasks with three pretrained transformer-based CLIP models (empirical evidence from Branch Bias). The results on each dataset can be found in the extended version.

3.2 A_3B_2: Adaptive Asymmetric Adapter

Experimental Observations on Branch Bias

We conduct extensive exploratory experiments to investigate CLIP's *Branch Bias* in image classification tasks. Specifically, we apply 11 diverse image datasets across three widely used image classification tasks (base-to-novel generalization, cross-dataset evaluation, and domain generalization). By fine-tuning a basic adapter (a pair of dimensionality reduction and expansion matrices) on each layer of the image or text encoders in three different CLIP variants (ViT-B/32, ViT-B/16, ViT-L/14), we evaluate performance across different tasks. Fig. 1 shows the average performance across these tasks, along with the parameter sizes of the image and text encoders in various CLIP variants. Further details (experimental setup and additional results) are available in the extended version. We summarize the following three insights:

`Insight I` In various tasks, fine-tuning additional parameters on the image encoder results in lower performance gains compared to the text encoder.

`Insight II` In in-distribution tasks (*e.g.*, base-to-novel generalization), fine-tuning adapters on both encoders achieves the best performance.

`Insight III` In out-of-distribution tasks (*e.g.*, cross-dataset evaluation and domain generalization), fine-tuning

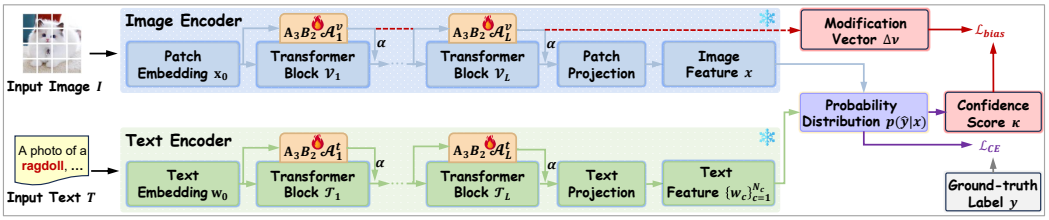


Figure 2: Overview of the proposed A_3B_2 architecture. The asymmetric adapters are integrated into each Transformer layer of the CLIP.

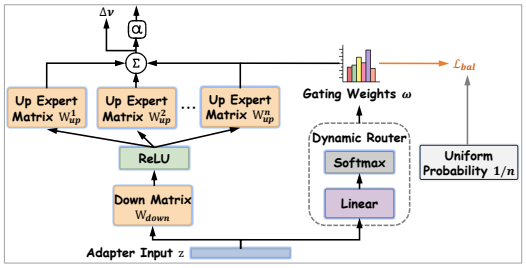


Figure 3: Detailed structure of the A_3B_2 adapter.

additional parameters on the image encoder may harm the transferability of VLMs on non-distribution data.

Task-adaptive and Structure-asymmetric Adapter

Based on the insights above, we propose an asymmetric architecture where adapters are always active on the text encoder (`Insight I`), while the image encoder's adaptation is dynamically regulated (`Insight II & III`).

As shown in Fig. 2, we integrate the A_3B_2 adapter (details in Fig. 3) into the CLIP backbone. For the text encoder, we modify Eq. 5 by adding the adapter $\mathcal{A}^t$ to each layer:

$$\begin{cases} [w_i^j]_{j=1}^N = \mathcal{T}_i([w_{i-1}^j]_{j=1}^N) + \alpha \underline{\mathcal{A}_i^t([w_{i-1}^j]_{j=1}^N)}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, L, \\ \mathcal{A}_i^t(z = [w_{i-1}^j]_{j=1}^N) = \sum_k^n \overset{t}{\mathfrak{w}}_i^k W_{up_i}^{t\ k} \\ \qquad\qquad\qquad\qquad \big(\texttt{ReLU}(W_{down\,i}^t(z))\big) \end{cases} \tag{9}$$

where underline indicates trainable modules, and α controls the adaptation strength. $\texttt{ReLU}(\cdot)$ is the activation function. $W_{down\,i}^t \in \mathbb{R}^{d_{ht} \times r}$ and $W_{up_i}^{t\ k} \in \mathbb{R}^{r \times d_{ht}}$, where d_{ht} represents the hidden layer dimension of the text encoder and rank $r \ll d_{ht}$, represent the "Down" matrix and k-th "Up" expert matrix in text encoder layer $\mathcal{T}_i$, respectively. Additionally, a dynamic router computes the expert weights:

$$\overset{t}{\mathfrak{w}}_i = \texttt{softmax}(\overset{t}{W}_i^{\,g} z), \tag{10}$$

where $\overset{t}{W}_i^{\,g} \in \mathbb{R}^{d_{ht} \times n}$, n is the number of experts.

Similarly, for the image encoder, we introduce an adapter $\mathcal{A}^v$ at each layer to modify Eq. 2:

$$\begin{cases} [c_i, x_i] = \mathcal{V}_i([c_{i-1}, x_{i-1}]) + \alpha \underline{\mathcal{A}_i^v([c_{i-1}, x_{i-1}])}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, 2, \ldots, L, \\ \mathcal{A}_i^v(z = [c_{i-1}, x_{i-1}]) = \sum_k^n \overset{v}{\mathfrak{w}}_i^k W_{up_i}^{v\ k} \\ \qquad\qquad\qquad\qquad \big(\texttt{ReLU}(W_{down\,i}^v(z))\big) \end{cases} \tag{11}$$

where $W_{down\,i}^v \in \mathbb{R}^{d_{hv} \times r}$ and $W_{up_i}^{v\ k} \in \mathbb{R}^{r \times d_{hv}}$ are the "Down" expert matrix and k-th "Up" matrix of image encoder layer $\mathcal{V}_i$, respectively, with d_{hv} as the hidden dimension of the image encoder. $\overset{v}{\mathfrak{w}}_i \in \mathbb{R}^n$ represents the gating weights from the router:

$$\overset{v}{\mathfrak{w}}_i = \texttt{softmax}(\overset{v}{W}_i^{\,g} z), \tag{12}$$

where $\overset{v}{W}_i^{\,g} \in \mathbb{R}^{d_{hv} \times n}$.

Rather than rigidly locking the image encoder, we fine-tune the image adapters while constraining their optimization via a novel uncertainty-aware mechanism, described below.

Optimization Objectives

To effectively alleviate *Branch Bias* and ensure structural diversity within the adapters [Mu and Lin, 2025], we introduce two auxiliary objectives: *Uncertainty-Aware Adapter Dampening* (*UAAD*) and *Load Balancing Regularization*.

Uncertainty-Aware Adapter Dampening `Insight III` suggests that aggressive modification of image features on out-of-distribution (OOD) data degrades performance. To address this without manual locking, we propose *UAAD*. This mechanism leverages the model's prediction confidence to gauge whether a sample lies within the distribution. **Low confidence implies potential OOD data, triggering a penalty on the image adapter's activation magnitude** [Hendrycks and Gimpel, 2016].

Let $\mathbf{p}(\hat{y}|x)$ be the softmax probability distribution for the image feature x. We define the confidence score as $\kappa = \max \mathbf{p}(\hat{y}|x)$. The bias alleviation loss $\mathcal{L}_{bias}$ is formulated as:

$$\mathcal{L}_{bias} = \frac{1}{B} \sum_{j=1}^{B} (1 - \kappa_j) \cdot \sum_{i=1}^{L} \|\Delta \mathbf{v}_{i,j}\|_2, \tag{13}$$

where B is the batch size, and $\Delta \mathbf{v}_{i,j} = \mathcal{A}_i^v(z_j)$ represents the feature modification vector produced by the image adapter at layer i. When the model is uncertain (low κ), $\mathcal{L}_{bias}$ forces the adapter output $\Delta \mathbf{v}$ towards zero, effectively reverting the model to the robust pre-trained CLIP features.

Load Balancing Regularization Standard MoE routing often suffer from expert collapse. To prevent this, we enforce a uniform distribution of expert usage across the batch. Unlike methods relying on selection frequency [Fedus *et al.*, 2022], we minimize the variance of the routing probabilities directly. Let $\bar{\mathfrak{w}}_{l,j,k}$ (derived from Eq. 10 and Eq. 12) denote the gating probability of the k-th expert for the j-th sample in

Dataset	Setting	CLIP	CoOp	CoCoOp	KgCoOp	RPO	MaPLe	CLIP-Adapter	TCP	MMA	MMRL	MMRL++	A_3B_2
ImageNet	Base	72.40	76.18	75.63	76.08	74.53	76.73	74.10	75.39	77.19	77.00	76.87	**77.37**
	Novel	68.09	70.47	70.45	70.59	70.26	70.49	68.89	70.21	69.63	70.70	70.72	**70.83**
	HM	70.18	73.21	72.95	73.23	72.33	73.48	71.40	72.08	73.22	73.72	73.67	**73.96**
Caltech101	Base	97.16	98.36	97.59	98.15	97.14	98.26	97.29	97.61	98.54	98.19	98.26	**98.73**
	Novel	94.21	94.18	94.69	94.65	94.32	94.14	94.10	94.54	93.63	94.07	94.36	**95.17**
	HM	95.66	96.22	96.12	96.37	95.71	82.54	95.67	96.05	96.02	96.09	96.27	**96.92**
OxfordPets	Base	91.26	95.27	95.39	94.84	88.76	95.21	94.22	94.79	95.20	95.41	95.16	**95.52**
	Novel	97.15	97.78	97.32	97.43	96.66	**97.93**	97.05	97.59	97.59	97.26	97.60	97.83
	HM	94.11	96.51	96.35	96.12	92.54	96.55	95.61	96.17	96.38	96.33	96.36	**96.66**
StanfordCars	Base	63.77	74.89	69.28	71.93	66.18	76.17	65.93	70.47	78.95	75.55	75.60	**81.96**
	Novel	74.96	71.70	73.83	73.35	**75.65**	73.33	75.44	75.40	72.53	74.99	74.83	75.58
	HM	68.91	73.26	71.48	72.63	70.60	74.72	70.37	72.85	75.60	75.27	75.21	**78.64**
Flowers102	Base	71.73	97.12	86.29	96.20	73.82	97.09	75.18	93.41	97.79	97.79	97.68	**98.14**
	Novel	77.45	71.94	74.21	72.88	**78.44**	72.03	75.58	75.25	75.96	75.13	76.59	77.39
	HM	74.48	82.65	79.80	82.93	76.06	82.70	75.38	83.35	85.50	84.98	85.86	**86.54**
Food101	Base	90.07	90.26	90.66	90.24	88.92	90.29	90.29	90.57	88.88	90.52	**90.73**	89.21
	Novel	91.15	91.25	91.58	91.50	89.73	91.54	91.12	91.53	90.51	91.82	**92.35**	91.75
	HM	90.61	90.75	91.12	90.87	89.32	90.91	90.70	91.05	89.69	91.17	**91.53**	90.46
FGVCAircraft	Base	27.63	37.82	33.53	36.85	28.15	39.22	30.69	35.79	43.18	38.70	39.53	**47.59**
	Novel	35.95	33.31	32.27	33.29	34.25	34.35	36.35	35.19	35.39	**37.51**	37.16	37.48
	HM	31.25	35.42	32.89	34.98	30.90	36.62	33.28	35.49	38.90	38.10	38.31	**41.93**
SUN397	Base	69.35	81.23	78.39	80.82	74.40	81.85	74.57	79.96	81.50	81.87	81.55	**82.36**
	Novel	75.56	75.43	77.37	74.97	77.67	76.93	77.40	78.11	77.75	**79.16**	79.05	78.72
	HM	72.32	78.22	77.88	77.79	76.00	79.31	75.96	79.02	79.58	80.49	80.28	**80.50**
DTD	Base	53.24	78.63	67.75	78.32	56.02	80.71	57.72	78.20	84.03	83.22	83.18	**84.18**
	Novel	60.87	50.48	55.64	56.88	61.31	55.63	60.79	53.90	**65.05**	62.04	62.57	64.71
	HM	56.80	61.49	61.10	65.90	58.55	65.86	59.22	63.81	**73.33**	71.09	71.42	73.17
EuroSAT	Base	57.02	88.81	78.41	85.82	53.52	92.63	62.80	84.25	94.22	90.39	90.64	**95.86**
	Novel	64.03	59.38	66.90	58.89	65.53	70.13	62.76	**70.17**	62.94	68.44	68.73	69.75
	HM	60.32	71.17	72.20	69.85	58.92	79.82	62.78	76.57	75.47	77.90	78.18	**80.75**
UCF101	Base	70.89	83.71	78.71	83.54	71.53	84.89	75.52	81.76	86.20	85.52	85.47	**86.49**
	Novel	78.42	74.17	74.56	74.36	73.12	78.06	77.54	77.93	77.73	78.35	79.57	**79.58**
	HM	74.47	78.65	76.58	78.68	72.32	81.33	76.52	79.80	81.75	81.78	82.42	**82.89**
Average	Base	69.50	82.03	77.42	81.16	70.27	83.00	72.57	80.20	84.15	83.11	83.15	**85.22**
	Novel	74.35	71.83	73.53	72.62	74.27	74.05	74.27	74.53	74.43	75.41	75.78	**76.25**
	HM	71.84	76.59	75.42	76.65	72.21	78.27	73.41	77.26	78.99	79.07	79.29	**80.49**

Table 1: Comparison of A_3B_2 with 11 methods in base-to-novel generalization across 11 datasets.The averages with 3 different seeds are reported. **Best results** in bold, next best underlined.

the l-th layer, averaged over the batch. We adopt a coefficient of variation [Shazeer *et al.*, 2017] based loss:

$$\mathcal{L}_{bal} = n^2 \cdot \frac{1}{L \cdot n} \sum_{l=1}^{L} \sum_{k=1}^{n} \left(\left(\frac{1}{B} \sum_{j=1}^{B} \bar{\mathfrak{w}}_{l,j,k} \right) - \frac{1}{n} \right)^2, \tag{14}$$

where $1/n$ represents the target uniform probability. Minimizing this objective penalizes deviations from a balanced distribution, ensuring diverse expert utilization.

Total Objective The model is primarily optimized using the cross-entropy loss, derived from the classification probabilities defined in Sec. 3.1. For a batch of size B, given the ground-truth label y_j for the image future x^j, this loss is defined as:

$$\mathcal{L}_{CE} = -\frac{1}{B} \sum_{j=1}^{B} \log \mathbf{p}(\hat{y} = y_j \mid x^j). \tag{15}$$

The final training objective is a weighted sum of the cross-entropy loss and the auxiliary regularizations:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_{bias}\mathcal{L}_{bias} + \lambda_{bal}\mathcal{L}_{bal}, \tag{16}$$

where λ_{bias} governs the suppression strength for uncertain samples, and λ_{bal} enforces the diversity of expert utilization. This formulation allows A_3B_2 to adaptively balance between general linguistic knowledge and task-specific visual adaptation, dynamically alleviating branch bias during training.

4 Experiments

We comprehensively evaluate A_3B_2's performance around three core image classification tasks in few-shot setting and validate the effectiveness of the discovered insights.

4.1 Experimental Setup

Base-to-Novel Generalization Consistent with previous studies [Brown *et al.*, 2020; Guo and Gu, 2025a], this evaluation divides the dataset into base and novel categories. The model is trained on base categories under a few-shot setting and tested separately on both base and novel categories. This evaluation includes 11 visual classification benchmark datasets: ImageNet [Deng *et al.*, 2009], Caltech101 [Fei-Fei *et al.*, 2004], OxfordPets [Parkhi *et al.*, 2012], StanfordCars [Krause *et al.*, 2013], Flowers102 [Nilsback and

Method	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	Average
CoOp	71.30	94.20	89.91	64.20	70.66	85.82	22.34	66.36	43.36	48.50	68.90	65.43
KgCoOp	71.12	94.24	90.29	**65.39**	71.23	86.24	21.99	67.03	44.39	43.89	68.67	65.34
MaPLe	71.91	93.39	90.41	64.37	70.28	85.63	22.88	66.04	44.60	39.87	67.91	64.54
TCP	70.36	94.21	90.16	65.23	71.46	86.73	23.74	67.13	45.06	43.00	68.91	65.56
MMA	72.45	92.39	89.59	61.69	68.52	81.94	24.19	66.32	44.29	39.70	67.86	63.65
MMRL	72.10	93.90	91.31	65.07	72.22	85.68	**25.44**	67.24	44.78	**48.66**	68.07	66.24
MMRL++	71.96	93.97	**91.43**	65.38	**72.57**	85.72	25.28	**67.65**	44.83	48.49	68.52	66.38
A_3B_2	**72.57**	**94.52**	91.36	65.32	72.15	**86.75**	24.53	67.15	**45.30**	47.42	**69.43**	**66.39**

Table 2: Comparison of A_3B_2 and 7 leading methods in cross-dataset evaluation, with more results provided in the extended version.

Method	ImageNet	-V2	-S	-A	-R	Average
CoOp	71.30	64.52	49.02	51.29	76.69	60.38
KgCoOp	71.12	64.33	48.97	51.41	76.60	60.33
MaPLe	71.91	64.69	48.91	49.26	76.47	59.83
TCP	70.36	63.60	48.74	50.50	76.70	59.89
MMA	72.45	65.23	48.09	48.07	75.43	59.21
MMRL	72.10	64.35	48.70	48.82	76.40	59.57
MMRL++	71.97	64.61	48.95	48.69	76.52	59.69
A_3B_2	**72.52**	**65.47**	**49.13**	**51.75**	**76.91**	**60.81**

Table 3: Comparison of A_3B_2 and 7 leading methods in domain generalization, with more results provided in the extended version.

Zisserman, 2008], Food101 [Bossard *et al.*, 2014], FGVCAircraft [Maji *et al.*, 2013], SUN397 [Xiao *et al.*, 2010], DTD [Cimpoi *et al.*, 2014], EuroSAT [Helber *et al.*, 2019], and UCF101 [Soomro *et al.*, 2012].

Cross-Dataset Evaluation This evaluation measures a model's discriminative ability on unseen datasets and its grasp of general knowledge. Following CoCoOp [Zhou *et al.*, 2022a], the model is trained on ImageNet with 1k classes and 16 shots per class, then directly evaluated on 10 other datasets mentioned in the base-to-novel generalization task.

Domain Generalization This task evaluates the model's robustness to domain shifts and its discriminative ability to out-of-distribution data. Following the setup of CoCoOp [Zhou *et al.*, 2022a], we fine-tune the model on ImageNet and test it on four of its variants: ImageNetV2 [Recht *et al.*, 2019], ImageNet-Sketch [Wang *et al.*, 2019], ImageNet-A [Hendrycks *et al.*, 2021b], and ImageNet-R [Hendrycks *et al.*, 2021a].

Baselines We compare the proposed A_3B_2 with 11 competitive methods, including: CLIP (zero-shot) [Radford *et al.*, 2021], CoOp [Radford *et al.*, 2021], CoCoOp [Zhou *et al.*, 2022a], KgCoOp [Yao *et al.*, 2023], RPO [Lee *et al.*, 2023], MaPLe [Khattak *et al.*, 2023], CLIP-Adapter [Gao *et al.*, 2024b], TCP [Yao *et al.*, 2024], MMA [Yang *et al.*, 2024], MMRL [Guo and Gu, 2025a], and MMRL++ [Guo and Gu, 2025b].

Implementation Details Consistent with previous studies [Yang *et al.*, 2024; Guo and Gu, 2025b], all experiments are conducted using a pre-trained transformer-based ViT-B/16 backbone of the CLIP model under a 16-shot setting (*i.e.*, 16 training samples per class). To ensure consistency with prior work [Radford *et al.*, 2021; Zhou *et al.*, 2022a; Yang *et al.*, 2024] as much as possible, we set the batch size to 8, with 5 epochs for ImageNet and 10 epochs for other datasets. The same image augmentation methods are used, and the SGD optimizer is employed with an initial learning rate of 0.0015, a momentum of 0.9, and a weight decay of 0.0005, combined with a cosine annealing scheduler with 1 epoch of warm-up. The text templates for the proposed A_3B_2 follow [Gao *et al.*, 2024b], with the default number of experts $n = 3$, low-rank $r = 32$, coefficient $\alpha = 0.001$, suppression strength $\lambda_{bias} = 0.3$ and balance parameter $\lambda_{bal} = 0.1$. The optimization strategies for other baselines follow the settings of the original papers. **For fairness, all baselines are re-implemented under the same experimental setup** and trained with mixed-precision acceleration on a single NVIDIA A800 GPU. Final performance is reported as the average over three random seeds (seed=1,2,3).

4.2 Base-to-Novel Generalization

In this task, we report the model's accuracy on both base and novel classes, along with their harmonic mean (HM), as shown in Table 1. Overall, our method consistently outperforms all baselines across all metrics and datasets. The key findings are as follows:

- Among all baselines, the recently proposed MMA and MMRL++ achieve the best average performance. MMRL++ leverages its decoupled inference strategy and progressive inter-layer composition to achieve leading novel-class accuracy with high parameter efficiency. MMA, on the other hand, integrates features from different branches into a shared space, enhancing gradient flow across branches, thus improving consistency between modalities and achieving optimal base-class recognition.
- The proposed A_3B_2 consistently leads all metrics across 11 datasets. Compared to MMA, which performs best on base-class recognition, and MMRL++, which excels on the novel and HM metrics, $\boldsymbol{A_3B_2}$ **shows a significant advantage in average performance**. These impressive results validate the effectiveness of our insights.
- In this in-distribution task, methods with dual-branch optimization (*e.g.*, MaPLe, MMA, and MMRL++) generally outperform single-branch optimization methods (*e.g.*, CoOp, CoCoOp, KgCoOp, and TCP), **further confirming the universality and broad applicability of our `Insight II`**.

4.3 Cross-Dataset Evaluation

We have compared the top 7 methods in the base-to-novel generalization task with the proposed A_3B_2 in the cross-

n	Base	Novel	HM	r	Base	Novel	HM	α	Base	Novel	HM	λ_{bias}	Base	Novel	HM	λ_{bal}	Base	Novel	HM
1	82.75	73.95	78.10	8	82.93	74.69	78.59	0.0001	85.13	75.62	80.09	0.1	84.82	75.69	80.00	0.01	85.15	75.93	80.28
2	84.89	75.83	80.10	16	84.24	75.26	79.50	0.0005	85.07	75.98	80.27	0.2	84.93	75.91	80.17	0.05	**85.26**	76.13	80.44
3	**85.22**	**76.25**	**80.49**	32	85.22	**76.25**	**80.49**	0.001	85.22	**76.25**	**80.49**	0.3	**85.22**	**76.25**	**80.49**	0.1	85.22	**76.25**	**80.49**
4	84.92	75.52	79.94	64	**85.37**	75.63	80.21	0.005	85.25	76.15	80.44	0.4	85.08	75.62	80.07	0.15	85.23	76.17	80.45
5	83.41	74.73	78.83	128	84.57	73.82	78.83	0.01	**85.38**	76.09	80.47	0.5	84.67	75.59	79.87	0.2	85.03	75.82	80.16

Table 4: Analysis experiments on the A_3B_2 hyperparameters.

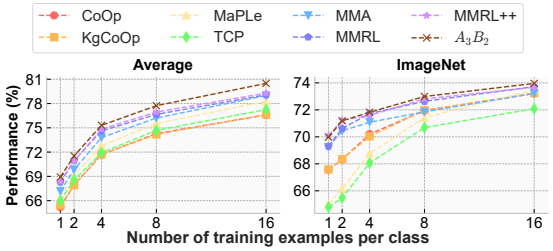


Figure 4: Comparison (HM) of A_3B_2 and 7 leading methods on few-shot learning, with more results provided in the extended version.

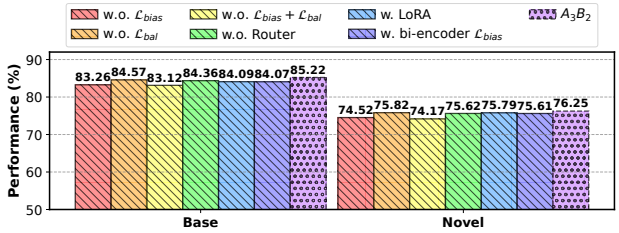


Figure 5: The performance of different variants in A_3B_2.

dataset evaluation task, as shown in Table 2. We observe that MMA performs poorly in our experimental setup, which we attribute to MMA simultaneously adapting the image encoder in out-of-distribution tasks, contrary to our `Insight III`. And this experiment validates the broad applicability of the `Insight III`: Text-only encoder optimization methods (*e.g.*, CoOp, CoCoOp, KgCoOp) generally outperform methods with image encoder optimization (*e.g.*, RPO, MaPLe, MMA). Notably, the MMRL and MMRL++ with image encoder optimization even outperform the proposed A_3B_2 in some cases. We believe these unexpected results stem from their use of a loss-function–driven model exploration strategy during optimization, differing from other methods, whereas we adopt a more general and adaptive structural design based on exploratory observations.

4.4 Domain Generalization

As shown in Table 3, the text-only encoder optimization methods significantly outperform the methods with image encoder optimization, **further validating the effectiveness of the `Insight III`**. Meanwhile, we also observe that the MMRL and MMRL++ methods perform poorly on this task, while our A_3B_2 method consistently achieves leading performance. This suggests that the A_3B_2, which focuses more on model structure optimization, is more stable, providing strong support for its future application and the insights derived for related research.

4.5 In-depth Analysis

We conduct analytical experiments on 11 datasets for the base-to-novel task; detailed results are in the extended version.

Few-Shot Learning As shown in Fig. 4, A_3B_2 consistently achieves the best performance across 11 datasets under all shot settings, with its advantage increasing as the number of shots grows. This demonstrates its strong transfer capability and robustness, even in data-scarce few-shot scenarios.

Sensitivity Analysis We analyze the sensitivity of the A_3B_2 hyperparameters, as shown in Table 4.

Number of Experts n. Too few experts limit model expressiveness, while too many may lead to overfitting. The optimal balance is achieved when $n = 3$.

Low-rank r. The low-rank r faces a similar trade-off to the number of experts. Performance is best when $r = 32$.

Balance Coefficient α. The coefficient α controls the relationship between pre-trained and learnable weights, which are more effective for base and novel settings, respectively. A good balance is achieved when $\alpha = 0.001$.

Suppression Strength λ_{bias}. This term controls the strength of uncertainty-aware dampening on image adapters. Moderate suppression ($\lambda_{bias} = 0.3$) best balances robustness and adaptability, while overly large values over-constrain visual adaptation.

Balance Parameter λ_{bal}. This parameter regulates expert utilization diversity. A moderate value ($\lambda_{bal} = 0.1$) effectively prevents expert collapse and yields the most stable performance across base and novel classes.

Ablation Study As shown in Fig. 5, removing any component degrades performance, confirming their effectiveness. Notably, removing $\mathcal{L}_{bias} + \mathcal{L}_{bal}$ leads to the largest drop, highlighting their synergy. Applying $\mathcal{L}_{bias}$ to the text encoder ("w. bi-encoder $\mathcal{L}_{bias}$") also causes a significant decline, validating the applicability of `Insight II & III`.

We provide a cost analysis of A_3B_2 and additional experiments on ResNet-based CLIP in the extended version.

5 Conclusion

This paper identifies the Branch Bias issue in few-shot vision-language classification, showing that fine-tuning the image encoder can be detrimental to out-of-distribution robustness. We propose A_3B_2, an adaptive asymmetric adapter that modulates the image branch based on prediction uncertainty. Extensive experiments demonstrate that A_3B_2 consistently improves performance across various few-shot settings, providing an effective solution for branch-aware adaptation.

Acknowledgments

This research was supported by grants from the "Pioneer" and "Leading Goose" R&D Program of Zhejiang under Grant No. 2025C02022 and the National Natural Science Foundation of China (No.62307032).

References

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

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

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

[Fedus *et al.*, 2022] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[Fei-Fei *et al.*, 2004] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[Fu *et al.*, 2025] Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

[Gao *et al.*, 2024a] Chongyang Gao, Kezhen Chen, Jinmeng Rao, Baochen Sun, Ruibo Liu, Daiyi Peng, Yawen Zhang, Xiaoyuan Guo, Jie Yang, and VS Subrahmanian. Higher layers need more lora experts. *arXiv preprint arXiv:2402.08562*, 2024.

[Gao *et al.*, 2024b] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[Gong *et al.*, 2025] Shizhan Gong, Yankai Jiang, Qi Dou, and Farzan Farnia. Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models. *arXiv preprint arXiv:2506.02557*, 2025.

[Guo and Gu, 2025a] Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. *arXiv preprint arXiv:2503.08497*, 2025.

[Guo and Gu, 2025b] Yuncheng Guo and Xiaodong Gu. Mmrl++: Parameter-efficient and interaction-aware representation learning for vision-language models. *arXiv preprint arXiv:2505.10088*, 2025.

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

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[Hendrycks *et al.*, 2021a] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

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

[Khattak *et al.*, 2023] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.

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

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

[Li *et al.*, 2022] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

[Li *et al.*, 2024] Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. *arXiv preprint arXiv:2409.16718*, 2024.

[Liang *et al.*, 2023] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023.

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

[Mu and Lin, 2025] Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025.

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

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

[Peng *et al.*, 2025] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Changsong Wen, Yu Huang, Menglin Yang, Feilong Tang, and Wei Shen. Understanding fine-tuning clip for open-vocabulary semantic segmentation in hyperbolic space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4562–4572, 2025.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

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

[Shazeer *et al.*, 2017] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

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

[Tian *et al.*, 2024] Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng-Zhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Advances in Neural Information Processing Systems*, 37:9565–9584, 2024.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2019] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

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

[Xu *et al.*, 2023] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2945–2954, 2023.

[Xue *et al.*, 2022] Fuzhao Xue, Ziji Shi, Futao Wei, Yuxuan Lou, Yong Liu, and Yang You. Go wider instead of deeper. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8779–8787, 2022.

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

[Yang *et al.*, 2025] Jingfeng Yang, Ziyang Wu, Yue Zhao, and Yi Ma. Language-image alignment with fixed text encoders. *arXiv preprint arXiv:2506.04209*, 2025.

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

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

[Zhang *et al.*, 2024] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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

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

[Zhou *et al.*, 2025] Yiyun Zhou, Chang Yao, and Jingyuan Chen. Cola: Collaborative low-rank adaptation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14115–14130, 2025.