R-MMA: Enhancing Vision-Language Models with Recurrent Adapters for Few-Shot and Cross-Domain Generalization

Supplementary Material

A. Implementation Details

Experimental Setup. Following standard in adapter tuning research [12, 23, 48, 50, 55, 56], we adopt CLIP with ViT-B/16 architecture [37] as our visual backbone across all experiments. Text prompts are manually designed following standard practices [37, 55, 56], with templates provided in Tab. A.1. We use the same setting for the SigLIP [52] ablation.

Training Configuration. We employ the AdamW optimizer with a learning rate of 0.001 and mixed-precision training for computational efficiency. Dataset-specific batch sizes are used: 32 for ImageNet and 4 for the remaining datasets. Training epochs vary by task: 5 epochs for ImageNet base-to-novel evaluation, 1 epoch for cross-dataset and domain generalization on ImageNet, 5 epochs for few-shot learning on ImageNet, and 50 epochs for few-shot learning on the remaining datasets. All results represent averages over three independent runs. All the experiments were run on a single NVIDIA RTX 4090 GPU.

Hyperparameter Settings. Based on the ablation study in Tab. 5, we configure R-MMAwith hidden dimension $d = 64$, $\alpha = 0.1$, and $\lambda = 0.6$. For fairness in comparison, the hyperparameters remain fixed across all datasets.

B. Dataset Description

Following prior works [12, 48], we evaluate our approach on 14 datasets spanning various recognition tasks, including generic object classification, fine-grained recognition, scene understanding, texture classification, satellite imagery interpretation, and action recognition. This set comprises 11 distinct datasets and 3 variants of ImageNet designed for robustness evaluation. Generic object recognition tasks are covered by datasets such as ImageNet [7], Caltech101 [10], and SUN397 [35], which use straightforward prompts like "a photo of a [CLASS]." Fine-grained classification is addressed through OxfordPets [36], StanfordCars [25], Flowers102 [34], Food101 [2], and FGVCAircraft [33], each focusing on specific object categories with prompts adapted to their domains (e.g., "a photo of a [CLASS], a type of flower"). DTD [6] targets texture classification with prompts such as "[CLASS] texture," while EuroSAT [15] involves satellite images and uses prompts like "a centered satellite photo of [CLASS]." The UCF101 [40] dataset is used for action recognition, where prompts describe actions (e.g., "a photo of a person doing [CLASS]").

The remaining three datasets, ImageNetV2 [39], ImageNet-Sketch [44], ImageNet-A [17], and ImageNet-R [16], assess model robustness and generalization. ImageNetV2 offers a newly curated test set maintaining ImageNet's class structure. ImageNet-Sketch provides sketch-based representations, while ImageNet-A and ImageNet-R include natural adversarial and artistic renditions of ImageNet classes, respectively. All prompt templates follow established conventions from prior work and are designed to be both descriptive and adaptable across tasks. Dataset statistics, including class counts and split sizes, are summarized in Tab. A.1.

C. Ablation Design

In this section, we detail a series of experiments conducted to explore the impact of different relevant architectural design choices for R-MMA.

C.1. R-MMA in Higher Layers

In the *k-to-12* experimental setting, we aim to explore the effectiveness of applying R-MMA only in the higher layers of CLIP. This design choice is motivated by the hypothesis that later layers in the transformer encoder capture more task-specific and semantically rich representations, making them more suitable for adaptation [48]. Accordingly, the adapter is inserted starting from the k^{th} layer up to the 12^{th}, *i.e.*, final layer. Unlike the regular setting where the initial input $\mathbf{v}^{(0)}$ is taken directly from the projection layer, we construct $\mathbf{v}^{(0)}$ by fusing the frozen intermediate representations from the $(k-1)^{\text{th}}$ layer of both the vision and text encoders. This fused representation serves as the starting point for subsequent adapter-based transformation, which is applied from layer k to 12 to produce the final output.

C.2. Design Choices of R-MMA

Concatenation. In this experimental setting, we sequentially concatenate the visual representation $\mathbf{i}_f^{(l)}$ and the textual representation $\mathbf{t}_f^{(l)}$, both extracted from a frozen CLIP model at layer l. This concatenated feature vector is passed through their respective multi-layer perceptron (MLP) layers that apply the weight matrices $\mathbf{W}_I$ and $\mathbf{W}_T$, to produce

Dataset	Classes	Train	Val	Test	Description	Prompt
ImageNet	1000	1.28M	∼	50000	Recognition of generic objects	"a photo of a [CLASS]."
Caltech101	101	4128	1649	2465	Recognition of generic objects	"a photo of a [CLASS]."
OxfordPets	37	2944	736	3669	Fine-grained classification of pets	"a photo of a [CLASS], a type of pet."
StanfordCars	196	6509	1635	8041	Fine-grained classification of cars	"a photo of a [CLASS]."
Flowers102	102	4093	1633	2463	Fine-grained classification of flowers	"a photo of a [CLASS], a type of flower."
Food101	101	50500	20200	30300	Fine-grained classification of foods	"a photo of [CLASS], a type of food."
FGVCAircraft	100	3334	3333	3333	Fine-grained classification of aircraft	"a photo of a [CLASS], a type of aircraft."
SUN397	397	15880	3970	19850	Scene classification	"a photo of a [CLASS]."
DTD	47	2820	1128	1692	Texture classification	"[CLASS] texture."
EuroSAT	10	13500	5400	8100	Land use & cover classification with satellite images	"a centered satellite photo of [CLASS]."
UCF101	101	7639	1898	3783	Action recognition	"a photo of a person doing [CLASS]."
ImageNetV2	1000	∼	∼	10,000	New test data for ImageNet	"a photo of a [CLASS]."
ImageNet-Sketch	1,000	∼	∼	50,889	Sketch-style images of ImageNet classes	"a photo of a [CLASS]."
ImageNet-A	200	∼	∼	7,500	Natural adversarial examples of 200 ImageNet classes	"a photo of a [CLASS]."
ImageNet-R	200	∼	∼	30,000	Renditions of 200 ImageNet classes	"a photo of a [CLASS]."

Table A.1. Summary of all 14 datasets used in this work, including 11 distinct datasets and 3 variants of ImageNet.

the updated representations:

$$\mathbf{v}_I^{(l)} = \mathbf{W}_I \cdot [\mathbf{i}_f^{(l)}; \mathbf{t}_f^{(l)}] \tag{18}$$

$$\mathbf{v}_T^{(l)} = \mathbf{W}_T \cdot [\mathbf{i}_f^{(l)}; \mathbf{t}_f^{(l)}] \tag{19}$$

Finally, we combine the updated representation with the original frozen feature to obtain the final representation:

$$\mathbf{i}^{(l)} = \mathbf{i}_f^{(l)} + \alpha \cdot \mathbf{v}_I^{(l)}, \quad \mathbf{t}^{(l)} = \mathbf{t}_f^{(l)} + \alpha \cdot \mathbf{v}_T^{(l)}. \tag{20}$$

Co-Attention. In this setup, we use a co-attention-based aggregation mechanism within the adapter. We replace Eqs. (9) and (10) by the following two equations:

$$\tilde{\mathbf{v}}_I^{(l)} = \text{Cross-Attention}(\mathbf{i}_f^{(l)}, \mathbf{v}_I^{(l)}), \tag{21}$$

$$\tilde{\mathbf{v}}_T^{(l)} = \text{Cross-Attention}(\mathbf{t}_f^{(l)}, \mathbf{v}_T^{(l)}), \tag{22}$$

Non-recurrent Setting. In our original recurrent setting, a single set of weights is shared across all layers. However, in the non-recurrent variant, we use layer-specific weights. In this case, a shared weight matrix $\mathbf{W}_i$ will be replaced by a layer-specific weight matrix $\mathbf{W}_i^{(l)}$, analogous to MMA's [48] adapter equations: Eqs. (6) and (7).