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

Md Fahim[1,2], Md Farhan Ishmam[3], Mir Sazzat Hossain[1], M Ashraful Amin[1],
Amin Ahsan Ali[1], AKM Mahbubur Rahman[1]
[1]*Center for Computational & Data Sciences, Independent University, Bangladesh*
[2]*Penta Global Limited, Dhaka, Bangladesh*
[3]*University of Utah, USA*
{fahimcse381}@gmail.com

Abstract

Despite the strong generalization capabilities of pre-trained vision-language models (VLMs) such as CLIP, adapting them to few-shot generalization tasks remains a fundamental challenge. Stemming from the discrimination–generalization dilemma, the models are required to gain task-specific knowledge while simultaneously preserving their general performance. Lightweight adapters, like Multimodal Adapter (MMA), improve alignment by adding learnable multimodal modules, but the use of multiple independent adapters greatly increases the trainable parameter overhead. Additionally, these adapters depend on the previous layer's frozen features while disregarding the current layer's representation during training. Addressing these limitations, we introduce ***R****ecurrent* ***M****ulti-****M****odal* ***A****dapter (R-MMA), a lightweight and efficient extension of MMA that enhances both performance and generalization while reducing the trainable parameter count.* R-MMA *employs a recurrent adapter module that shares weights layer-wise and integrates the current layer representation in the adapter blocks. We integrate an attention-based alignment mechanism to harmonize the adapter outputs with the frozen encoder features before fusion to ensure better preservation of the pre-trained representations and cross-modal consistency. Extensive experiments across 15 datasets on diverse tasks, including few-shot learning, generalization to novel classes and domains, and dataset transfer scenarios, demonstrate that R-MMA consistently surpasses state-of-the-art baselines, achieving strong performance with improved efficiency and a better balance between adaptation and generalization. Our work achieves one of the highest forms of parameter efficiency with only three trainable weight matrices for the whole network, regardless of the network depth.*

1. Introduction

Vision language models (VLMs) often jointly learn image-text representations by aligning related pairs while distancing unrelated ones [35]. Training on the large-scale web data endows these models with strong zero and few-shot performance. However, these capabilities come at a cost; the size makes full fine-tuning impractical for downstream tasks, particularly in resource-scarce domains.

Parameter-efficient techniques have emerged to ensure transferability to downstream gaps and bridge the modality alignment [16, 17]. Adapters have been particularly attractive as they are small, lightweight modules that are placed with both visual and textual encoders to align multimodal features without updating the weights of the entire model [43].

Conventional adapters have two major limitations for multimodal models: (i) the visual and textual representations are independently processed before prediction [9, 47], (ii) adapters are used at each layer, which scales trainable parameter count with network depth Multimodal adapters [10, 43] address (i) by using adapters that integrate both unimodal representations. However, these adapters rely on the previous layer's frozen encoding, disregarding the current layer's representation.

Limitation (ii) has not been addressed by any current adapter-style work. Recurrent or weight-sharing adapters [38] can potentially address this by enforcing shared parameters across adapter modules in multiple layers, significantly reducing the number of trainable parameters. However, the recurrence of the same adapter module can be detrimental to the model's performance as it greatly reduces the amount of learned features. Designing effective recurrent multimodal adapters remains challenging due to the heterogeneous nature of visual and textual modalities.

Inspired by the Multimodal Adapter (MMA) proposed in [43], we introduce the *Recurrent Multi-Modal Adapter*

(R-MMA), a lightweight yet expressive module applied recurrently across multiple layers of both the image and text encoders. Unlike MMA, which applies k independent adapters in the final layers—resulting in increased computational cost—our method reuses a single adapter across layers with shared weights, reducing parameter overhead and improving efficiency. MMA also directly fuses the learnable adapter representation with the frozen encoder outputs, which can negatively affect the representational space and transferability [46].

In contrast, R-MMA employs an attention-based alignment module to harmonize these representations prior to fusion. Our approach achieved strong few-shot generalization in diverse downstream tasks by combining the parameter efficiency of weight-tied designs with the adaptability of multimodal conditioning, marking a step toward highly scalable, general-purpose VLM adaptation.

2. Related Work

Parameter Efficient Transfer Learning (PETL) methods [3, 11, 16] aim to adapt large pretrained models by training only a smaller subset of existing or new parameters while keeping the bulk of the weights frozen. Methods include prompt tuning, which appends a learnable prompting token to the visual [18] or textual inputs [25], and low-rank adaptation (LoRA), which injects trainable low-rank matrices into the weight updates [17]. Adapters fall under the broader sphere of PETL, which were first introduced as lightweight modules inserted between transformer layers [16, 38]. Our work extensively focuses on adapters, as it remains a popular choice in Vision-Language tasks due to its modularity and efficiency.

Vision Language Adapters. Building on successful NLP adapters [16], they were soon applied to vision [4] and multimodal [43] tasks. Early adapters create a low-dimensional bottleneck to learn at a lower feature space, often improving VL alignment, *e.g.*, CLIP adapters added a bottleneck MLP on CLIP's frozen features and improved performance at a fraction of trainable parameters [9]. TIP-Adapter [47] builds on this using a training-free method.

The aforementioned adapters process unimodal streams independently until the final prediction. MMA [43] introduces a new class of multimodal adapters to address the modality gap by jointly attending to intermediate representations from both modalities in each adapter block. Despite the substantial performance gains of MMA, its parameter footprint scales linearly with the architectural depth of the frozen model.

Adapter Weight Sharing. To further reduce the parameter overhead in knowledge transfer, several works [27, 38] explore weight sharing among adapters across layers, tasks, or modalities. VL-Adapter [38] showed that a single shared adapter can match task-specific adapters in VL tasks. Adapter Re-Composing (ARC) [6] ties projection weights across all ViT layers [7]. UniAdapter [28] introduces partial sharing, *i.e.*, a subset of weights will be shared across modalities and tasks. Our work draws contrast from these by sharing the *full* trainable weights *layer-wise* for *vision-language* networks.

3. Preliminaries

3.1. CLIP

Contrastive Language-Image Pre-training (CLIP) [35] has demonstrated strong performance in open-set visual recognition tasks by aligning visual and textual embeddings into a unified representation space. CLIP consists of two encoders: an image encoder (*e.g.*, ResNet [12], ViT [7]) and a text encoder (*e.g.* Transformer [40]). For the ViT-based image encoder, the input image x_I is passed to the image embedding layer Embed_I to produce a sequence of *patch embeddings*. Similarly, the textual input x_T is tokenized through Embed_T. The embeddings are then processed by a stack of L transformer layers, $\{E_I^{(l)}\}_{l=1}^L$ and $\{E_T^{(l)}\}_{l=1}^L$, for the visual and textual modality. The intermediate representation at the l^{th} layer is defined by:

$$\mathbf{I}^{(l)} = E_I^{(l)}(\mathbf{I}^{(l-1)}), \quad \mathbf{T}^{(l)} = E_T^{(l)}(\mathbf{T}^{(l-1)}) \tag{1}$$

where $\mathbf{I}^{(l)}$ and $\mathbf{T}^{(l)}$ represent the visual and textual representations at layer l. The final representations are obtained by projecting the `[CLS]` tokens from the last transformer layers using their respective projection layers, *i.e.*,

$$\mathbf{z}_I = \text{Proj}_I(\mathbf{I}^{(L)}), \quad \mathbf{z}_T = \text{Proj}_T(\mathbf{T}^{(L)}). \tag{2}$$

CLIP is trained using contrastive loss that aligns visual and textual representations by maximizing the cosine similarity between matched image-text pairs within a training batch of N samples, while minimizing it for mismatched pairs. The training objective is defined as:

$$\begin{aligned}\mathcal{L}_{\text{CLIP}} = \frac{1}{2N}\sum_{i=1}^{N}\Bigg[&-\log\frac{\exp\left(\cos(\mathbf{z}_I^i,\mathbf{z}_T^i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\cos(\mathbf{z}_I^i,\mathbf{z}_T^j)/\tau\right)} \\ &-\log\frac{\exp\left(\cos(\mathbf{z}_T^i,\mathbf{z}_I^i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\cos(\mathbf{z}_T^i,\mathbf{z}_I^j)/\tau\right)}\Bigg]\end{aligned} \tag{3}$$

where $\mathbf{z}^i$ is a representation of sample i, $\cos(\cdot,\cdot)$ denotes cosine similarity, and τ is the temperature hyperparameter. After training, CLIP enables zero-shot image recognition. Let $\mathbf{z_I}$ be an image representation and $\{\mathbf{z_T}^i\}_{i=1}^K$ be a set of textual class prompt representations, (*e.g.*, "A photo of a ⟨class⟩."), where K is the number of classes. The most probable class is selected based on the classification probability of each class y, computed as:

$$p(y|\mathbf{z}_I) = \frac{\exp\left(\cos(\mathbf{z}_I, \mathbf{z}_T^y)/\tau\right)}{\sum_{i=1}^{K} \exp\left(\cos(\mathbf{z}_I, \mathbf{z}_T^i)/\tau\right)}, \tag{4}$$

3.2. Mulitmodal Adapter

MMA [43] introduces a lightweight multimodal adapter for CLIP-like VLMs. The authors observed that in dataset-level recognition tasks, the lower transformer layers tend to learn more generalizable representations transferable across datasets, while higher layers capture dataset-specific semantics. This suggests that fine-tuning higher layers is particularly effective for adapting to new tasks. MMA exploits this by using adapter modules from the k^{th} transformer layer. The update equations for layers $l \geq k$ are:

$$\mathbf{I}^{(l)} = E_I^{(l)}(\mathbf{I}^{(l-1)}) + \alpha \cdot A_I^{(l)}(\mathbf{I}^{(l-1)}) \tag{5}$$

$$\mathbf{T}^{(l)} = E_T^{(l)}(\mathbf{T}^{(l-1)}) + \alpha \cdot A_T^{(l)}(\mathbf{T}^{(l-1)}) \tag{6}$$

where $A_I^{(l)}(\cdot)$ and $A_T^{(l)}(\cdot)$ are learnable visual and textual adapter modules, and α balances between pre-trained and task-specific features. To bridge the cross-modal semantic gap, each adapter includes an additional multimodal block. The block projects modality-specific representations into a joining latent space via modality-specific up-projection layers, then applies a shared projection matrix, and finally restores the original feature dimensions using down-projection layers. The adapters are defined as:

$$\mathcal{A}_I^{(l)}(\mathbf{I}^{(l-1)}) = W_{Iu}^{(l)} \cdot \delta\left(W_S^{(l)} \cdot \delta\left(W_{Id}^{(l)} \cdot \mathbf{I}^{(l-1)}\right)\right) \tag{7}$$

$$\mathcal{A}_T^{(l)}(\mathbf{T}^{(l-1)}) = W_{Tu}^{(l)} \cdot \delta\left(W_S^{(l)} \cdot \delta\left(W_{Td}^{(l)} \cdot \mathbf{T}^{(l-1)}\right)\right) \tag{8}$$

where $W_{M,p}^{(l)}$ defines the l^{th} layer project matrix for the image, text, or shared modality, *i.e.*, $M \in \{I, T, S\}$, and the up/down project type, *i.e.*, $p \in \{u, d\}$. $\delta(\cdot)$ denotes a non-linear activation function. The shared weights allow gradients to flow between modalities during training, improving feature alignment and multimodal representation.

4. Methodology

The architecture of R-MMA is grounded in three key principles. First, the latent token is factorized into modality-specific projections, inspired by Mixture-of-Experts [19] and the factorized attention mechanism. Second, each projection attends independently to its modality stream, enabling separate semantic alignment paths, similar in spirit to dual-stream architectures such as LXMERT [39] and ViLBERT [29]. Finally, the fusion stage merges the contextualized representations into a unified latent token, allowing efficient cross-modal interaction and better multimodal reasoning.

To explicitly reduce the semantic gap between modalities, we depart from MMA's [43] unified adapter strategy by computing a *single unified latent representation* for both modalities through a modality-aware decomposition and fusion framework. R-MMA reuses the multimodal adapter module across layers rather than instantiating independently at each of the final k layers. By sharing all the adapter weights across all L layers, R-MMA cuts parameter count and computational cost compared to MMA, while preserving the frozen encoder's semantic space. The R-MMA architecture consists of three core components: i) Modality-Aware Routing (MAR), ii) Attention, and iii) Modality Fusion modules.

Modality-Aware Routing (MAR) Module. Let $\mathbf{v}^{(l-1)} \in \mathbb{R}^d$ be the unified latent token entering the l^{th} layer from the $(l-1)^{th}$ layer. This representation is projected into two modality-specific subspaces:

$$\mathbf{v}_I^{(l)} = W_I \mathbf{v}^{(l-1)}, \quad \mathbf{v}_T^{(l)} = W_T \mathbf{v}^{(l-1)} \tag{9}$$

where $W_I, W_T \in \mathbb{R}^{d \times d}$ are learnable projection matrices for the image and text streams, and d is the CLIP-embedding dimension.

Attention Module. In R-MMA , the intermediate visual and textual representations, $\mathbf{I}^{(l)}$ and $\mathbf{T}^{(l)}$, differ from the frozen CLIP features, $\mathbf{I}_f^{(l)}$ and $\mathbf{T}_f^{(l)}$, following Eq. (1). Each projections attend to its respective modality's current layer's frozen features via dot-product attention [40]:

$$\tilde{\mathbf{v}}_I^{(l)} = \mathbf{I}_f^{(l)\top} \cdot \mathbf{v}_I^{(l)}, \quad \tilde{\mathbf{v}}_T^{(l)} = \mathbf{T}_f^{(l)\top} \cdot \mathbf{v}_T^{(l)}. \tag{10}$$

Modality Fusion Module. The attended outputs are then concatenated and passed through a linear transformation layer to form the updated latent representation. The output of modality fusion for layer l is defined as:

$$\mathbf{v}^{(l)} = W_F \cdot [\tilde{\mathbf{v}}_I^{(l)}; \tilde{\mathbf{v}}_T^{(l)}] \tag{11}$$

where $W_F \in \mathbb{R}^{2d \times d}$ is the learnable fusion matrix. It should be noted that the projection and fusion matrices, W_I, W_T, and W_F, are shared across layers to reduce parameter overhead, unlike previous methods with weight matrices for each layer (Eqs. (7) and (8)).

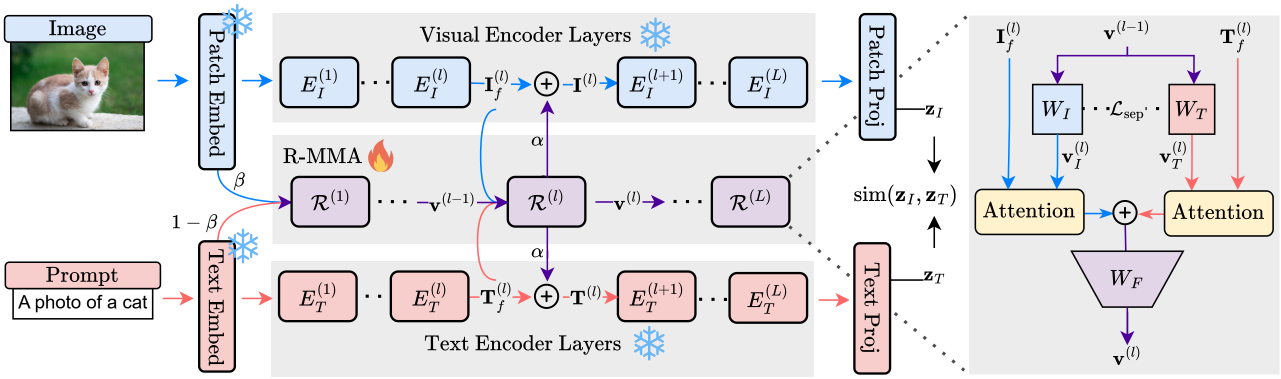

Figure 1. Model Architecture of our proposed R-MMA . Only the extra adapters are optimized, while the whole pre-trained CLIP models are frozen. The adapters are shared the weights and work in recurrent manner. R-MMA creates a unified representation for both image and text encoder for a particular layer. As R-MMA works in recurrent manner, it not only resource effective but also captures more fine grained information.

Recurrent Multimodal Adapter (R-MMA)

Combining Eqs. (9) to (11), we get the overall formulation of the recurrent adapter, $\mathcal{R}^{(l)}$:

$$\mathbf{v}^{(l)} = \mathcal{R}^{(l)}(\mathbf{v}^{(l-1)}, \mathbf{I}_f^{(l)}, \mathbf{T}_f^{(l)}) \tag{12}$$

$$= W_{\text{F}} \cdot \left[\mathbf{I}_f^{(l)\top} \cdot W_I \mathbf{v}^{(l-1)} \, ; \mathbf{T}_f^{(l)\top} \cdot W_T \mathbf{v}^{(l-1)} \right]. \tag{13}$$

Finally, we combine the adapter representation with the frozen CLIP representation using a weighted sum, similar to Eqs. (7) and (8), but for each layer $l \in \{1, \dots, L\}$:

$$\mathbf{I}^{(l)} = \mathbf{I}_f^{(l)} + \alpha \cdot \mathcal{R}^{(l)}(\mathbf{v}^{(l-1)}, \mathbf{I}_f^{(l)}, \mathbf{T}_f^{(l)}), \tag{14}$$

$$\mathbf{T}^{(l)} = \mathbf{T}_f^{(l)} + \alpha \cdot \mathcal{R}^{(l)}(\mathbf{v}^{(l-1)}, \mathbf{I}_f^{(l)}, \mathbf{T}_f^{(l)}). \tag{15}$$

Initial Latent Representation. The initial unified latent representation in our framework is constructed using a weighted sum of CLIP embeddings from both modalities:

$$\mathbf{v}^{(0)} = \beta \cdot \text{Embed}_I(x_I) + (1-\beta) \cdot \text{Embed}_T(x_T), \tag{16}$$

where x_i and x_t denote the input image and text, respectively, $\text{Embed}_I(\cdot)$ and $\text{Embed}_T(\cdot)$ denote the patch and textual embedding, respectively, and β balances between the embeddings.

Orthogonality Regularization. Following [26], to encourage disentangled and non-redundant projections, we apply an orthogonality regularization between the projection matrices alongside standard CLIP contrastive loss:

$$\mathcal{L}_{sep} = \left\| W_I^\top \cdot W_T \right\|. \tag{17}$$

The total loss $\mathcal{L}_{\text{Total}}$ combines the standard CLIP loss $\mathcal{L}_{\text{CLIP}}$ with an additional separation loss $\mathcal{L}_{\text{sep}}$, scaled by a weighting factor λ:

$$\mathcal{L}_{Total} = \mathcal{L}_{CLIP} + \lambda \, \mathcal{L}_{sep}$$

5. Result and Analysis

To evaluate the effectiveness of R-MMA , we conduct extensive experiments across a spectrum of tasks, including base-to-novel generalization, cross-dataset evaluation, and domain generalization. A 16-shot setting is used for all experiments, where only 16 samples per class are utilized for training. Full implementation details are provided in the Supplementary Material.

5.1. Base-to-Novel Generalization

Table 1 presents the results of R-MMA on the base-to-novel generalization task, where the model is trained on a set of base classes and tested on both base and novel classes, following the setup in [21, 49, 50]. We evaluate our method on 11 diverse image classification datasets: ImageNet [20] and Caltech101 [8] for general object recognition; OxfordPets [34], StanfordCars [23], Flowers102 [32], Food101 [1], and FGVC-Aircraft [31] for fine-grained classification; SUN397 [33] for scene recognition; DTD [5] for texture classification; EuroSAT [13] for satellite image recognition; and UCF101 [37] for action recognition. This task allows us to assess R-MMA 's transfer learning effectiveness on base classes and its ability to preserve the inherent generalization and zero-shot capabilities of pre-trained VLMs on novel classes.

In this experiment, we compare R-MMA against several strong baselines, including CoOp [50], CoCoOp [49], ProDA [30], KgCoOp [44], MaPLe [21], LASP [2], RPO [24], PromptSRC [22], ProVP [42], MetaPrompt [48], TCP [45], MMA [43], and MMRL [10] and the zero-shot CLIP baseline. Our method, R-MMA , consistently

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [35]	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 [50]	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp [49]	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 [30]	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 [44]	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 [21]	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 [2]	82.70	74.90	78.61	76.20	70.95	73.48	98.10	94.24	96.16	95.90	97.93	96.90
RPO [24]	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 [22]	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
ProVP [42]	85.20	73.22	78.76	75.82	69.21	72.36	98.92	94.21	96.51	95.87	97.65	96.75
MetaPrompt [48]	83.65	75.48	79.09	77.52	70.83	74.02	98.13	94.58	96.32	95.53	97.00	96.26
TCP [45]	84.13	75.36	79.51	77.27	69.87	73.38	98.23	94.67	96.42	94.67	97.20	95.92
MMA [43]	83.20	76.80	79.87	77.31	71.00	74.02	98.40	94.00	96.15	95.40	98.07	96.72
MMRL [10]	**85.68**	77.16	81.20	77.90	71.30	74.45	**98.97**	94.50	96.68	95.90	97.60	96.74
R-MMA (Ours)	**85.67**	**77.72**	**81.32**	**78.06**	**71.64**	**74.71**	98.82	**94.77**	**96.75**	**96.10**	**98.17**	**97.12**

Method	StanfordCars			Flowers102			Food101			FGVC-Aircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [35]	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 [50]	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp [49]	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 [30]	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 [44]	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 [21]	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 [2]	75.17	71.60	73.34	97.00	74.00	83.95	**91.20**	91.70	**91.44**	34.53	30.57	32.43
RPO [24]	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 [22]	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
ProVP [42]	80.43	67.96	73.67	98.42	72.06	83.20	90.32	90.91	90.61	47.08	29.87	36.55
MetaPrompt [48]	76.34	75.01	75.48	97.66	74.49	84.52	90.74	91.85	91.29	40.14	36.51	38.24
TCP [45]	80.80	74.13	77.32	97.73	75.57	85.23	90.57	91.37	90.97	41.97	34.43	37.83
MMA [43]	78.50	73.10	75.70	97.77	75.93	85.48	90.13	91.30	90.71	40.57	36.33	38.33
MMRL [10]	81.30	75.07	78.06	**98.97**	**77.27**	**86.78**	90.57	91.50	91.03	46.30	37.03	41.15
R-MMA (Ours)	**81.90**	**75.48**	**78.56**	98.81	77.24	86.70	90.27	**92.64**	**91.44**	**47.28**	**38.17**	**42.24**

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [35]	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 [50]	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp [49]	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 [30]	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 [44]	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 [21]	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 [2]	80.70	78.60	79.63	81.40	58.60	68.14	94.60	77.78	85.36	84.77	78.03	81.26
RPO [24]	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 [22]	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
ProVP [42]	80.67	76.11	78.32	83.95	59.06	69.34	97.12	72.91	83.29	88.56	75.55	81.54
MetaPrompt [48]	82.26	79.04	80.62	83.10	58.05	68.35	93.53	75.21	83.38	85.33	77.72	81.35
TCP [45]	82.63	78.20	80.35	82.77	58.07	68.25	91.63	74.73	82.32	87.13	80.77	83.83
MMA [43]	82.27	78.57	80.38	83.20	65.63	73.38	85.46	82.34	83.87	86.23	80.03	82.20
MMRL [10]	**83.20**	**79.30**	**81.20**	**85.67**	65.00	73.82	**95.60**	80.17	87.21	88.10	80.07	83.89
R-MMA (Ours)	82.63	78.92	80.73	84.96	**65.88**	**74.21**	94.77	**81.25**	**87.49**	**88.79**	**80.71**	**84.56**

Table 1. Comparison with state-of-the-art methods on different datasets in the Base-to-Novel Generalization setting. “Base” and “Novel” are the recognition accuracies on base and novel classes respectively. “HM” is the harmonic mean of base and new accuracy, providing the trade-off between adaption and generalization. The proposed R-MMA shows a good adaptation ability, while being highly effective in novel class generalization.

demonstrates superior performance, particularly in novel class generalization and overall trade-off, as evidenced by the following key observations:

Novel Class Generalization and Overall Performance: Our proposed R-MMA achieves the best average harmonic mean (HM) of 81.32% across all 11 datasets, setting a new

Methods	ImageNet	**Average**	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVC-Aircraft	SUN397	DTD	EuroSAT	UCF101
CoOp [50]	71.51	63.88	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55
CoCoOp [49]	71.02	65.74	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21
MaPLe [21]	70.72	66.30	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69
PromptSRC [22]	71.27	65.81	93.60	90.25	65.70	70.25	86.15	23.90	67.10	**46.87**	45.50	**68.75**
TCP [45]	71.40	66.29	93.97	91.25	64.69	71.21	**86.69**	23.45	67.15	44.35	51.45	68.73
MMA [43]	71.00	66.61	93.80	90.30	66.13	72.07	86.12	25.33	**68.17**	46.57	49.24	68.32
MMRL [10]	72.03	67.25	**94.67**	91.43	66.10	**72.77**	86.40	26.30	67.57	45.90	53.10	68.27
R-MMA (Ours)	**72.47**	**67.37**	94.52	**91.74**	**66.25**	72.60	86.55	**26.41**	67.47	46.19	**53.28**	68.64

Table 2. Comparison of R-MMA with state-of-the-art methods in the Cross-Dataset Evaluation setting. Overall, our R-MMA obtains leading average performance over 10 datasets, demonstrating the good zero-shot transferable ability.

state-of-the-art. This significantly surpasses the previous leading method, MMRL [10], by 0.12% HM (81.32% vs. 81.20%). More critically, R-MMA also achieves a leading average novel accuracy of 77.72%, which is 0.56% higher than MMRL (77.16%). These results highlight R-MMA 's exceptional ability to generalize to unseen categories without novel class information during training, alongside providing a superior overall balance between adaptation and generalization.

This leading performance is consistently reflected in individual dataset results across a wide range of scenarios. R-MMA achieves the best novel accuracy on 9 out of 11 datasets (ImageNet, Caltech101, OxfordPets, StanfordCars, FGVC-Aircraft, DTD, EuroSAT, and UCF101). For example, on OxfordPets, R-MMA reaches an impressive 98.17% novel accuracy, surpassing all baselines. Similarly, for FGVC-Aircraft and UCF101, R-MMA achieves significant gains in novel accuracy, reaching 38.17% and 80.71% respectively, which are the highest among all methods. Furthermore, R-MMA consistently achieves the highest harmonic mean on 10 out of 11 datasets. Notable gains include 74.71% HM on ImageNet (vs. MMRL's 74.45%), 97.12% HM on OxfordPets (vs. MMRL's 96.74%), 42.24% HM on FGVC-Aircraft (vs. MMRL's 41.15%), and 84.56% HM on UCF101 (vs. MMRL's 83.89%).

Competitive Base Class Performance: While R-MMA primarily focuses on improving novel class generalization, it maintains highly competitive base class accuracy, with an average of 85.67%. This is only marginally lower than the top average of 85.68% by MMRL. On several individual datasets, R-MMA achieves leading or near-leading base accuracy. For instance, on ImageNet, it obtains 78.06% base accuracy, which is the best among all methods. On OxfordPets, it achieves 96.10% base accuracy, and on FGVC-Aircraft, it reaches 47.28% base accuracy, both leading. This indicates that our method does not significantly sacrifice performance on seen classes while substantially boosting generalization capabilities to unseen ones.

Performance Nuances on Flowers102 and SUN397: Despite R-MMA 's strong overall performance, we observe that on Flowers102 and SUN397, MMRL [10] slightly outperforms our method across base accuracy, novel accuracy, and harmonic mean. For Flowers102, MMRL achieves 0.16% higher base, 0.03% higher novel, and 0.08% higher HM. Similarly, on SUN397, MMRL leads by 0.57% in base, 0.38% in novel, and 0.47% in HM. These datasets are known for presenting unique challenges: Flowers102 features fine-grained classification with high inter-class similarity and intra-class variation, while SUN397 is a complex scene recognition dataset requiring robust understanding of global contextual information. This indicates that while R-MMA excels in broad generalization, there might be specific dataset characteristics where MMRL's inductive biases offer a marginal advantage. However, these specific cases do not detract from R-MMA 's overall state-of-the-art average performance.

5.2. Cross-dataset Evaluation

To evaluate the transferability of R-MMA to entirely new domains (unseen datasets), we conduct cross-dataset evaluation. Following CoCoOp [49], all models are initially trained on the ImageNet dataset using a few-shot setting described in Section 5.1. Subsequently, these trained models are directly evaluated on the remaining 10 datasets (i.e., all other datasets used in Section 5.1 excluding ImageNet). This setup allows us to assess the model's ability to generalize to new datasets without any further fine-tuning or adaptation to the target datasets.

Table 2 presents the results of R-MMA in this cross-dataset evaluation setting, comparing it with state-of-the-art baselines. Overall, R-MMA achieves a new state-of-the-art average accuracy of 67.37%, demonstrating superior zero-shot transferability across diverse domains. This marks a 0.12% improvement over the previous leading method, MMRL [10], and a 0.76% improvement over MMA [43]. Specifically, R-MMA obtains the highest accuracy on the source dataset, ImageNet (72.47%), and leads on 5 out of 10 target datasets (including OxfordPets, StanfordCars, FGVCAircraft, and EuroSAT). These results underscore R-MMA 's ability to transfer knowledge effectively and generalize to new, unseen domains without requiring any dataset-specific fine-tuning, thus setting a new benchmark for cross-dataset evaluation.

Source	**ImageNet**	**Target**			
Methods		**-V2**	**-S**	**-A**	**-R**
CLIP [35]	66.73	60.83	46.15	47.77	73.96
CoOp [50]	71.51	64.20	47.99	49.71	75.21
CoCoOp [49]	71.02	64.07	48.75	50.63	76.18
MaPLe [21]	70.72	64.07	49.15	50.90	76.98
PromptSRC [22]	71.27	64.35	49.55	50.90	**77.80**
MMA [43]	71.00	64.33	49.13	51.12	77.32
MMRL [10]	72.03	64.47	49.17	51.20	77.53
R-MMA	**72.47**	**64.58**	**49.63**	**51.49**	77.44

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

5.3. Domain Generalization

To evaluate the resilience of models to various domain shifts and their generalization to out-of-distribution (OOD) data, we conduct these experiments. Following CoCoOp [49], model initially trained on ImageNet are evaluated directly on four distinct ImageNet variants, each introducing a different type of domain variation. These datasets include ImageNetV2 [36], ImageNet-Sketch [41], ImageNet-A [15], and ImageNet-R [14]. This evaluation specifically assesses the model's robustness against data that deviates significantly from the training distribution.

Table 3 summarizes the results of R-MMA in this domain generalization task. R-MMA achieves best performance on 3 out of 4 datasets. These results highlight R-MMA 's robustness and adaptability to unseen data distributions.

5.4. Comparison of Computational Cost

To ensure a fair comparison, we adopt the exact configuration used for computational cost estimation by MMRL [10]. All methods are trained on a single NVIDIA RTX 4090 GPU using the ImageNet dataset. Each model is implemented using publicly available code and default settings as specified in their respective papers. Training time is reported both as the average time per image and the total duration required to train the full 16-shot dataset. Inference speed is measured in frames per second (FPS) with a batch size of 100.

Table 5 summarizes the performance of various fine-tuning approaches across several key metrics, including the number of learnable parameters, training efficiency, inference throughput, and HM score. Our proposed method, R-MMA , achieves the highest HM score while using fewer parameters (0.482M), with the fastest training time (1.6ms/image) and inference speed (491.3FPS). Compared to MMRL—the strongest baseline—R-MMA offers a marginal performance gain while reducing parameter count by a factor of 7 and improving inference speed by over 1.5×.

5.5. Ablation Study

Ablation on Model Variants: Table 4a lays out the performance of different R-MMA model configurations, clearly showing how effective our design choices are. When we stripped out the MAR module, performance took a significant hit. The HM dropped from 81.32 to 79.25. Dropping the orthogonality regularization term $\mathcal{L}_{\text{sep}}$ leads to a noticeable dip to 79.78 HM. And removing the Attention Module also results in a performance drop to 79.88 HM. We also explored a variant without the Modality Fusion Module (w/o Fusion). This module is responsible for taking the modality-specific, attention-guided outputs and fusing them into a single, unified latent token. Its removal resulted in a lower 80.4 HM.

Dimensions of Hidden Layers: The dimension of the hidden layers in the adapter is a critical hyperparameter that influences its expressiveness and ability to capture relationships between features across modalities and layers. We performed an ablation study by varying this dimension, and the results are presented in Table 4b. We observed that a dimension of 64 yielded the optimal performance. Smaller dimensions limited the adapter's expressiveness, hindering its ability to learn complex multimodal interactions. Conversely, while increasing the dimension initially improved performance, excessively large dimensions led to a slight decline.

Scaling Factor α: The scaling factor α in the adapter is crucial for balancing the influence of the task-specific adapter features with the task-agnostic frozen CLIP features. Table 4c shows the impact of varying α on performance. We found that a value of 0.1 yielded the best results. A smaller α resulted in underutilization of the adapter features, leading to lower performance. Conversely, larger values caused the adapter features to dominate, which can lead to overfitting and reduced generalization.

Initial Modality Balancing Factor β: The initial modality balancing factor β is a critical hyperparameter that controls the initial composition of the unified latent representation ($\mathbf{v}^{(0)}$) by balancing the contributions of the image and text input embeddings. Table 4d details the impact of varying β on performance. An optimal β value of 0.7 yielded

(a) Performance with Different Model Variants

Model Variants	Base	Novel	HM
w/o MAR	83.04	75.80	79.25
w/o $\mathcal{L}_{\text{sep}}$	84.07	75.91	79.78
w/o Attn	83.24	76.78	79.88
w/o Fusion	83.97	76.46	80.4
R-MMA	**85.67**	**77.72**	**81.32**

(b) Dimensions of Hidden Layers

Dims	Base	Novel	HM
8	81.36	75.66	78.41
16	83.82	75.87	79.65
32	84.08	76.58	80.15
64	**85.67**	**77.72**	**81.32**
128	84.77	76.97	80.68

(c) Scaling Factor α

α	Base	Novel	HM
0.001	75.92	74.98	75.45
0.005	78.64	75.26	76.91
0.01	82.37	75.53	78.80
0.1	**85.67**	**77.72**	**81.32**
0.2	83.78	75.92	79.66

(d) Scaling Factor β

β	Base	Novel	HM
0.5	83.92	76.04	79.79
0.6	84.76	76.62	80.48
0.7	**85.67**	**77.72**	**81.32**
0.8	85.28	77.03	80.95
0.9	84.87	76.71	80.58

Table 4. Ablation experiments over 11 datasets used in Base-to-Novel Generalization setting. Here, *Attn* means the Attention Module and *Fusion* indicates the Modality Fusion Module of R-MMA respectively.

Method	Modality	Params	Train time	Train time	FPS	HM
		(learnable)	(ms/image)	(minute/all)	(100 BS)	
MaPLe	V-L	3.555M	39.5	26.4	1757.6	78.55
PromptSRC	V, L	0.046M	40.0	106.8	1764.2	79.97
ProVP	V	0.147M	4.4	107.2	928.9	78.76
MetaPrompt	V, L	**0.031M**	30.7	32.8	659.8	79.09
TCP	L	0.332M	5.3	17.7	950.6	79.51
MMA	V-L	0.675M	2.2	1.5	688.5	79.87
MMRL	V-L	4.992M	5.3	3.6	762.4	81.20
R-MMA	V-L	0.482M	**1.6**	**1.2**	**491.3**	**81.32**

Table 5. Comparison of Computational Cost for Different Methods. 'V-L' denotes vision-language interaction, 'V, L' indicates separate fine-tuning, and 'L' represents textual-only fine-tuning.

λ	Base	Novel	HM
0.1	76.56	68.82	72.48
0.3	77.40	70.61	73.85
0.6	**78.06**	**71.64**	**74.71**
1.2	76.02	69.28	72.49
3.0	74.87	68.05	71.30

Table 6. Ablation on λ on ImageNet Dataset.

the highest HM. This indicates that for our architecture, a slightly stronger initial emphasis on the visual embedding than the textual embedding provides a more effective starting point for the recurrent adapter's learning process.

Impact of Loss Factor λ: We also perform an ablation study on the loss factor λ on the ImageNet dataset. As shown in the Table 6, the best results across all metrics are achieved at $\lambda = 0.6$, indicating an optimal balance. Both lower and higher values of λ lead to a drop in model performance.

Applying R-MMA in the Higher Layers. We evaluate R-MMA against MMA in the k-to-12 setting, where adapters are applied from layer k onwards. Figure 2 shows average performance of both adapters across 11 datasets. R-MMA consistently outperforms MMA in HM accuracy across all configurations, with improvements ranging from 1.62 to 1.91 points. The superiority stems primarily from enhanced Base class performance (1.7-3.4 point gains), while Novel class accuracy remains competitive. Notably, R-MMA maintains strong performance even when applied only to the final layers, demonstrating that the recurrent multimodal adaptation effectively leverages pre-trained knowledge in CLIP's higher layers for improved generalization.

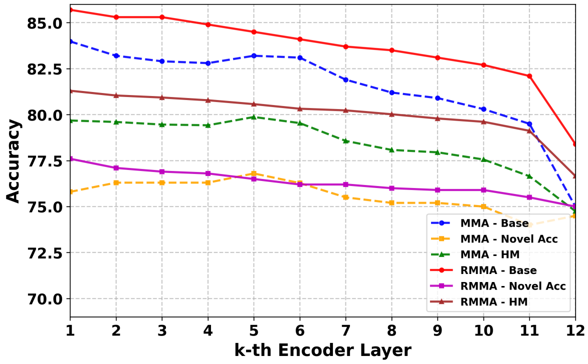

Figure 2. Comparison between R-MMA and MMA on *k-to-12* settings where the adapters are used from layers $\geq$ k. We report average scores of Base, Novel and HM over 11 datasets. R-MMA consistently outperforms MMA in terms of HM accuracy.

6. Conclusion

We presented R-MMA , a recurrent adapter framework designed to enhance the generalization capabilities of pretrained VLMs in few-shot settings. By using a recurrent architecture, R-MMA captures rich cross-modal interactions with high parameter efficiency. It incorporates a Modality-Aware Routing (MAR) module that learns modality-specific pathways, an orthogonality regularization term to ensure disentangled feature learning, and an attention mechanism to guide the adapter's outputs. These components work synergistically to adapt the VLMs to new tasks while preserving the rich pre-trained knowledge of CLIP. Extensive evaluation across 15 datasets on diverse tasks demonstrates R-MMA 's effectiveness in balancing adaptation and generalization, making it a promising approach for enhancing the transfer learning capabilities of VLMs.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, pages 446–461, Cham, 2014. Springer International Publishing. 4

[2] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23232–23241, 2023. 4, 5

[3] Hao Chen, Ran Tao, Han Zhang, Yidong Wang, Xiang Li, Wei Ye, Jindong Wang, Guosheng Hu, and Marios Savvides. Conv-adapter: Exploring parameter efficient transfer learning for convnets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1551–1561, 2024. 2

[4] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 2

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 4

[6] Wei Dong, Dawei Yan, Zhijun Lin, and Peng Wang. Efficient adaptation of large vision transformer via adapter recomposing. *Advances in Neural Information Processing Systems*, 36:52548–52567, 2023. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Li Fei-Fei, R. Fergus, and P. 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, 2004. 4

[9] 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. 1, 2

[10] Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25015–25025, 2025. 1, 4, 5, 6, 7

[11] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[13] 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. 4

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2021. 7

[15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15257–15266, 2021. 7

[16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 1, 2

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1, 2

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022. 2

[19] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 3

[20] 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. 4

[21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, 2023. 4, 5, 6, 7, 1

[22] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023. 4, 5, 6, 7

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 4

[24] 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 *Pro-*

ceedings of the IEEE/CVF international conference on computer vision, pages 1401–1411, 2023. 4, 5

[25] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 2

[26] S.A. Liu, Y. Zhang, Z. Qiu, H. Xie, Y. Zhang, and T. Yao. Learning orthogonal prototypes for generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11319–11328. IEEE, 2023. 4

[27] Ting Liu, Xuyang Liu, Siteng Huang, Honggang Chen, Quanjun Yin, Long Qin, Donglin Wang, and Yue Hu. Dara: Domain-and relation-aware adapters make parameter-efficient tuning for visual grounding. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 2

[28] Haoyu Lu, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. *arXiv preprint arXiv:2302.06605*, 2023. 2

[29] J. Lu, Z. Yang, L. Yu, Y. Hong, D. Joo, E. Choi, D. Batra, and D. Parikh. Vilbert: Pretraining task-agnostic visual-linguistic representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1328–1338, 2019. 3

[30] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5206–5215, 2022. 4, 5

[31] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. 4

[32] 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, 2008. 4

[33] Aude Oliva, Krista A. Ehinger, Antonio Torralba, James Hays, and Jianxiong Xiao. SUN database: Large-scale scene recognition from abbey to zoo . In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, Los Alamitos, CA, USA, 2010. IEEE Computer Society. 4

[34] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012. 4

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 5, 7

[36] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9641–9649, 2021. 7

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. 4

[38] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237, 2022. 1, 2

[39] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1553–1562, 2019. 3

[40] 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. 2, 3

[41] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. *Learning robust global representations by penalizing local predictive power*. Curran Associates Inc., Red Hook, NY, USA, 2019. 7

[42] Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive visual prompt learning with contrastive feature re-formation. *International Journal of Computer Vision*, 133(2):511–526, 2025. 4, 5

[43] 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. 1, 2, 3, 4, 5, 6, 7

[44] 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. 4, 5

[45] 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. 4, 5, 6, 1

[46] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 2

[47] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 1, 2

[48] Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. Learning domain invariant prompt for vision-language models. *IEEE Transactions on Image Processing*, 33:1348–1360, 2024. 4, 5

[49] 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 com-*

puter vision and pattern recognition, pages 16816–16825, 2022. 4, 5, 6, 7, 1

[50] 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. 4, 5, 6, 7, 1

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

Supplementary Material

A. Implementation Details

Experimental Setup. Following established protocols in prompt learning research [10, 21, 43, 45, 49, 50], we adopt CLIP with ViT-B/16 architecture [35] as our visual backbone across all experiments. Text prompts are manually designed following standard practices [35, 49, 50], with complete templates provided in Table A.1.

Training Configuration. We employ 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 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, and 5/50 epochs for few-shot learning on ImageNet/other datasets respectively. All results represent averages over three independent runs on a single NVIDIA RTX 4090 GPU.

Hyperparameter Settings. Based on the ablation study in Table 4, we configure R-MMA with hidden dimension 64, $\alpha = 0.1$, $\beta = 0.7$, and $\lambda = 0.6$. To ensure fair comparison, these hyperparameters remain fixed across all datasets rather than dataset-specific optimization.

B. Dataset Description

Following the prior work, we also 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, Caltech101, and SUN397, which use straightforward prompts like “a photo of a [CLASS].” Fine-grained classification is addressed through OxfordPets, StanfordCars, Flowers102, Food101, and FGVCAircraft, each focusing on specific object categories with prompts adapted to their domains (e.g., “a photo of a [CLASS], a type of flower”). DTD targets texture classification with prompts such as “[CLASS] texture,” while EuroSAT involves satellite images and uses prompts like “a centered satellite photo of [CLASS].” The UCF101 dataset is used for action recognition, where prompts describe actions (e.g., “a photo of a person doing [CLASS]”).

The remaining three datasets—ImageNetV2, ImageNet-Sketch, and ImageNet-A/R—serve to 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 the accompanying table.

C. R-MMA in Higher Layers

In the *k-to-12* experimental setting, we aim to explore the effectiveness of applying our R-MMA adapter only in the higher layers of the CLIP model, rather than across all layers. This design choice is motivated by the hypothesis that later layers in the transformer encoder encode more task-specific and semantically rich representations, making them more suitable for adaptation. Accordingly, the adapter is inserted starting from the k^{th} layer up to the 12^{th} (final) layer.

Unlike the regular setting where the initial input $\mathbf{v}^{(0)}$ is taken directly from the projection layer, here we construct $\mathbf{v}^{(0)}$ by fusing the frozen intermediate representations from the $(k-1)^{\text{th}}$ layer of both the image and text encoders. Specifically, we define:

$$\mathbf{v}^{(0)} = \beta \cdot I_f^{(k-1)} + (1-\beta) \cdot T_f^{(k-1)} \tag{18}$$

where $I_f^{(k-1)}$ and $T_f^{(k-1)}$ represent the frozen image and text features from the $(k-1)^{\text{th}}$ layer of CLIP, respectively. The scalar $\beta \in [0,1]$ balances the contribution from the visual and textual modalities. This fused representation serves as the starting point for subsequent adapter-based transformation, which is applied from layer k through layer 12 to produce the final output.

D. R-MMA Further Ablation

Impact of Recurrent Adaptation Across Layers: Table A.2 presents a comparison of our R-MMA with baselines that fine-tune only the final CLIP layers, alongside MMA. Fine-tuning more layers typically improves base class accuracy but often degrades novel class generalization, as this can impair the pre-trained model’s general

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.

Layer	12	10→12	8→12	5→12	MMA	R-MMA
Base	80.77	83.02	83.77	83.21	83.20	**85.67**
Novel	74.08	74.55	73.77	70.95	76.80	**77.72**
HM	77.28	78.56	78.45	76.59	79.87	**81.32**

Table A.2. Comparing our MMA with the baseline by fine-tuning last few layers on 11 datasets in Base-to-Novel Generalization setting. “10→12” refers to fine-tune the last 3 layers in both branches.

knowledge. Notably, R-MMA noteably outperforms these fine-tuning variants, as well as MMA, achieving an HM of 81.32. This highlights the superior parameter efficiency and generalization of our recurrent, weight-shared adapter, which effectively fosters cross-modal interactions while preserving the frozen CLIP backbone.

Loss	Base	Novel	HM
Cosine	85.52	77.61	81.36
L1	85.24	77.12	80.91
L2	85.18	77.05	80.88
MSE	85.22	77.37	81.05
Orthogonal Proj	**85.67**	**77.72**	**81.32**

Table A.3. Ablation on the Design Choice of $\mathcal{L}_{sep}$

Ablation on the Design Choice of $\mathcal{L}_{sep}$ Table A.3 presents an ablation study on different loss functions used for the separation regularization term, $\mathcal{L}_{sep}$. The objective of $\mathcal{L}_{sep}$ is to enforce dissimilarity between the feature representations of base and novel classes, encouraging the model to learn distinct and less overlapping embeddings. This separation is crucial for improving generalization, particularly on novel categories, without degrading performance on the base classes.

To achieve this, each loss function measures the discrepancy between the feature distributions or embeddings of base and novel classes in different ways:

- *Cosine loss* encourages orthogonality by minimizing cosine similarity.
- *L1* and *L2 losses* measure the element-wise absolute and squared differences, respectively.
- *MSE* penalizes the mean squared error between features.
- *Orthogonal Projection loss* explicitly enforces orthogonality between subspaces representing base and novel features.

These losses are incorporated into the training objective as a regularization term, guiding the model to separate the learned feature spaces. Among them, the Orthogonal Projection loss leads to the best trade-off, achieving the highest accuracy on both base (85.67%) and novel (77.72%) classes, as well as a strong harmonic mean score of 81.32%. This demonstrates its effectiveness in dissimilarity enforcement and overall performance enhancement.