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]*CCDS, Independent University, Bangladesh* [2]*Penta Global Limited* [3]*KSoC, University of Utah*

{fahimcse381,farhan.ishmam}@gmail.com

Abstract

*Pre-trained vision-language models (VLMs) such as CLIP exhibit strong generalization but struggle with few-shot adaptation due to the trade-off between gaining task-specific knowledge and preserving general performance. While multimodal adapters add trainable modules that improve alignment and excel in few-shot generality, they greatly increase the trainable parameter count while relying heavily on the prior layer's frozen representation. Addressing these limitations, we introduce **R**ecurrent **M**ulti-**M**odal **A**dapter (R-MMA), a lightweight and efficient adapter that uses self-attention to compute a unified latent representation with a single set of shared adapter weights. Our attention-based alignment harmonizes the adapter outputs with the frozen encoder features before fusing the modalities, ensuring better preservation of pre-trained representations and cross-modal consistency. Our experiments show that R-MMA achieves state-of-the-art performance on most datasets for base-to-novel generalization, cross-dataset evaluation, and domain generalization, under few-shot settings. Our approach also achieves one of the highest forms of parameter efficiency with only a few trainable weight matrices for the whole network, regardless of its depth. Our code is available at:* https://github.com/farhanishmam/R-MMA.

1. Introduction

Vision language models (VLMs), like CLIP [37], jointly learn image-text representations by aligning related pairs while distancing unrelated ones. 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 [53]. Parameter-efficient techniques have emerged to ensure transferability to downstream tasks and bridge the modality alignment [19, 20]. Adapters have been particularly attractive as they are small, lightweight modules placed with both visual and textual encoders to align multimodal features without updating the weights of the entire model [48].

Conventional adapters have two major limitations for multimodal models: (i) the visual and textual representations are independently processed before prediction [11, 53], (ii) the layer-wise nature of adapters scales trainable parameter count with network depth [41]. While multimodal adapters [12, 48] address (i) by using adapters that integrate representations from both modalities, they are heavily dependent on the previous layer's frozen encoding. To the best of our knowledge, limitation (ii) has not been addressed by any adapter-style work. Recurrent or weight-sharing adapters can potentially address (ii) by enforcing parameter sharing across layers, significantly reducing the number of trainable parameters [38, 41]. Yet, naïvely reusing the same adapter across layers can be detrimental to the model's performance as it greatly reduces the number 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) [48], 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 [48], which applies k independent adapters in the final k layers, resulting in increased computational cost, we reuse a single adapter across layers by sharing weights. While MMA uses a shared projection across modalities, potentially affecting the representational space and transferability [51], R-MMA employs an attention-based alignment module to harmonize the latent features before fusion. Our method achieved strong few-shot generalization in diverse downstream tasks by combining the parameter efficiency of weight-sharing with the adaptability of alignment via attention, marking a step toward highly scalable, general-purpose VLM adaptation.

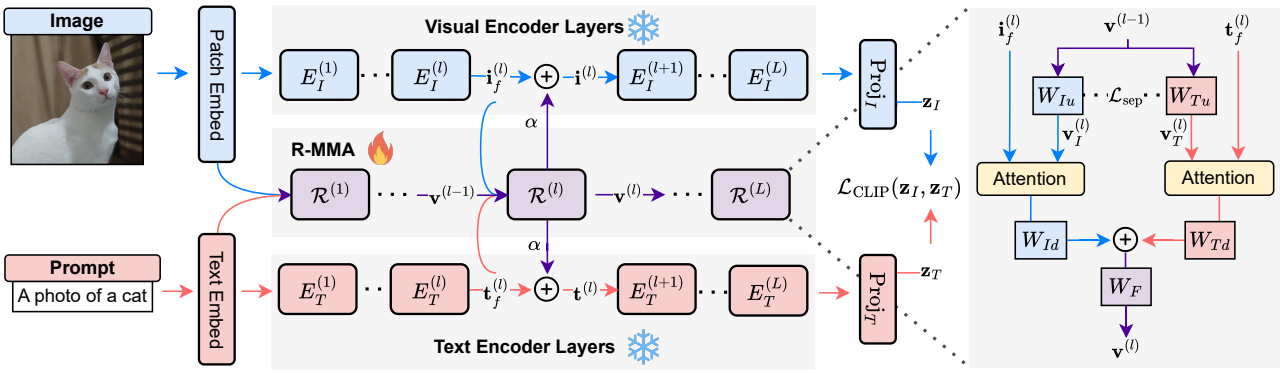

Figure 1. Model Architecture of our proposed R-MMA. Only the adapter modules are trained (fire symbol), while the entire pre-trained CLIP model is frozen (snowflake symbol). The frozen visual and textual embeddings are passed to the adapter module, which produces a unified adapter representation that is aggregated with the frozen representations. Inside the adapter module, attention is applied individually to the modalities and later fused to form the unified representation. The adapter weights are shared across layers.

2. Related Work

Parameter Efficient Transfer Learning (PETL) methods [4, 13, 19] 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 [22] or textual inputs [27], and low-rank adaptation (LoRA), which injects trainable low-rank matrices into the weight updates [20]. Adapters fall under the broader sphere of PETL, which were first introduced as lightweight modules inserted between transformer layers [19, 41]. Our work extensively focuses on adapters, as they remain a popular choice in Vision-Language tasks due to their modularity and efficiency.

Vision Language Adapters build on successful NLP adapters [19], extended to vision [5] and multimodal [48] 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 [11]. TIP-Adapter [53] builds on this using a training-free method. The aforementioned adapters process unimodal streams independently until the final prediction. MMA [48] 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. MMRL [12] extends MMA by inserting representation tokens at higher encoder layers. Despite the substantial performance gains of MMA and MMRL, its parameter footprint scales linearly with the architectural depth of the frozen model.

Adapter Weight Sharing further reduces the parameter overhead in knowledge transfer. Several works [29, 41] explore weight sharing among adapters across layers, tasks, or modalities. VL-Adapter [41] showed that a single shared adapter can match task-specific adapters in VL tasks. Adapter Re-Composing (ARC) [8] ties projection weights across all ViT layers [9]. UniAdapter [30] introduces partial sharing, *i.e.*, a subset of weights will be shared across modalities and tasks. Contrasting these works, our method shares the *full* trainable weights *layer-wise* for *vision-language* networks.

3. Preliminaries

3.1. CLIP

Contrastive Language-Image Pre-training (CLIP) [37] 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 [14], ViT [9]) and a text encoder (*e.g.* Transformer [43]). The input image and prompt are passed to the image and text embedding layers, respectively, to produce a sequence of *patch embeddings* and *textual embeddings*. 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 modalities. 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 $\mathbf{i}^{(l)} \in \mathbb{R}^{d_I}$ and $\mathbf{t}^{(l)} \in \mathbb{R}^{d_T}$ denote the representations of the visual and textual tokens at layer l, while d_I and d_T represent the dimension of visual and textual embeddings. The final representations are obtained by projecting the `[CLS]` tokens from the final L^{th} transformer layers using their respective projection layers, *i.e.*,

$$\mathbf{z}_I = \text{Proj}_I(\mathbf{i}_{[\text{CLS}]}^{(L)}), \quad \mathbf{z}_T = \text{Proj}_T(\mathbf{t}_{[\text{CLS}]}^{(L)}). \tag{2}$$

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:

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

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.

3.2. Multimodal Adapter

MMA [48] introduces a lightweight adapter that shares projections across modalities. The authors observed that in dataset-level recognition tasks, the lower transformer layers tend to learn more generalizable representations 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:

$$\mathbf{i}^{(l)} = E_I^{(l)}(\mathbf{i}^{(l-1)}) + \alpha \cdot \mathcal{A}_I^{(l)}(\mathbf{i}^{(l-1)}), \tag{4}$$

$$\mathbf{t}^{(l)} = E_T^{(l)}(\mathbf{t}^{(l-1)}) + \alpha \cdot \mathcal{A}_T^{(l)}(\mathbf{t}^{(l-1)}), \tag{5}$$

where $\mathcal{A}_I^{(l)}(\cdot)$ and $\mathcal{A}_T^{(l)}(\cdot)$ are learnable visual and textual adapter modules, and α is a hyperparameter that 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 joint latent space via modality-specific down-projection layers, then applies a shared projection matrix, and finally restores the original feature dimensions using up-projection layers, *i.e.*,

$$\mathcal{A}_I^{(l)}(\mathbf{i}^{(l-1)}) = \mathbf{W}_{Iu}^{(l)} \cdot \delta\left(\mathbf{W}_S^{(l)} \cdot \delta\left(\mathbf{W}_{Id}^{(l)} \cdot \mathbf{i}^{(l-1)}\right)\right), \tag{6}$$

$$\mathcal{A}_T^{(l)}(\mathbf{t}^{(l-1)}) = \mathbf{W}_{Tu}^{(l)} \cdot \delta\left(\mathbf{W}_S^{(l)} \cdot \delta\left(\mathbf{W}_{Td}^{(l)} \cdot \mathbf{t}^{(l-1)}\right)\right), \tag{7}$$

where $\mathbf{W}_{M,p}^{(l)}$ defines the l^{th} layer projection 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 projections allow gradients to flow between modalities during training, improving feature alignment and multimodal representation.

4. Methodology

At its core, R-MMA also computes a single unified latent representation through modality-aware decomposition and fusion. However, we depart from MMA's [48] layer-wise adapter with shared projections, by reusing the adapter weights across all L layers. This recursive strategy significantly reduces parameter and computational costs while preserving the frozen encoder's semantic space.

The R-MMA architecture consists of three core components: Modality-Aware Routing, Attention, and Modality Fusion. MAR first factorizes the latent tokens into modality-specific projections, which then attend independently to their respective modality streams via attention. This enables separate semantic alignment paths, similar in spirit to dual-stream architectures like LXMERT [42] and ViLBERT [31]. Finally, the contextualized outputs are projected and fused into a unified latent representation for efficient cross-modal interaction.

Modality-Aware Routing (MAR). 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, and d is the latent embedding dimension, such that $d < d_I$ and $d < d_T$, where d_I and d_T are the visual and textual embedding dimensions. The representation is projected into two modality-specific subspaces, *i.e.*,

$$\mathbf{v}_I^{(l)} = \mathbf{W}_{Iu} \cdot \mathbf{v}^{(l-1)}, \quad \mathbf{v}_T^{(l)} = \mathbf{W}_{Tu} \cdot \mathbf{v}^{(l-1)}, \tag{8}$$

where $\mathbf{W}_{Iu} \in \mathbb{R}^{d_I \times d}$ and $\mathbf{W}_{Tu} \in \mathbb{R}^{d_T \times d}$ are learnable up projection matrices for the image and text streams.

Attention. In R-MMA, the intermediate representations of visual and textual tokens, $\mathbf{i}^{(l)}$ and $\mathbf{t}^{(l)}$, differ from the frozen CLIP features, $\mathbf{i}_f^{(l)}$ and $\mathbf{t}_f^{(l)}$, following Eq. (1). The current layer's frozen features attend to the projections via scaled dot-product attention [43]:

$$\tilde{\mathbf{v}}_I^{(l)} = \text{Attention}(\mathbf{i}_f^{(l)}, \mathbf{v}_I^{(l)}), \tag{9}$$

$$\tilde{\mathbf{v}}_T^{(l)} = \text{Attention}(\mathbf{t}_f^{(l)}, \mathbf{v}_T^{(l)}), \tag{10}$$

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

$$\mathbf{v}^{(l)} = \mathbf{W}_F \cdot [\mathbf{W}_{Id} \cdot \tilde{\mathbf{v}}_I^{(l)}; \mathbf{W}_{Td} \cdot \tilde{\mathbf{v}}_T^{(l)}], \tag{11}$$

where $\mathbf{W}_{Id} \in \mathbb{R}^{d \times d_I}$ and $\mathbf{W}_{Td} \in \mathbb{R}^{d \times d_T}$ are the learnable down-projection matrices, and $\mathbf{W}_F \in \mathbb{R}^{d \times d}$ is the learnable fusion matrix. The sequence-wise concatenation increases the token length of the latent representation to $M + N$, where M and N represent the number of visual and textual tokens, respectively. We will sequentially slice

tokens $[1 : M]$ and $[M + 1 : M + N]$ for the visual and textual representations when the latent representation is projected back to their respective modalities.

R-MMA Module. We obtain adapter outputs by projecting the latent representation to their respective modalities using the following equations:

$$\mathcal{R}_I^{(l)}(\mathbf{v}^{(l-1)}, \mathbf{i}_f^{(l)}, \mathbf{t}_f^{(l)}) = \text{Proj}_{Iu}(\mathbf{v}^{(l)}), \tag{12}$$

$$\mathcal{R}_T^{(l)}(\mathbf{v}^{(l-1)}, \mathbf{i}_f^{(l)}, \mathbf{t}_f^{(l)}) = \text{Proj}_{Tu}(\mathbf{v}^{(l)}). \tag{13}$$

Finally, we combine the adapter representation with the frozen CLIP representation using a weighted sum, similar to Eqs. (6) and (7), but for each layer $l \in \{1, \ldots, L\}$:

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

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

It should be noted that all projection and fusion matrices are shared across layers to reduce parameter overhead, unlike previous methods with weight matrices for each layer (Eqs. (6) and (7)).

Initial Latent Representation. We construct the initial latent representation $\mathbf{v}_0 \in \mathbb{R}^d$ by projecting the embeddings of each modality to the latent dimension d and sequentially concatenating the projected embeddings. Unlike the adapter projections from Eqs. (12) and (13), the initial projection is a down-projection to the latent dimension $d < d_I$ and $d < d_T$, where d_I and d_T are the dimensions of the image and text embeddings. Formally, we represent the initial latent representation as:

$$\mathbf{v}^{(0)} = [\text{Proj}_I(\mathbf{i}^{(0)}); \text{Proj}_T(\mathbf{t}^{(0)})], \tag{16}$$

where $\mathbf{i}^0$ represents the patch embedding of the visual token and $\mathbf{t}^0$ represents the text embedding of the textual token.

Orthogonality Regularization. Following [28], to encourage disentangled and non-redundant projections, we apply an orthogonality regularization between the projection matrices alongside standard CLIP contrastive loss:

$$\mathcal{L}_{sep} = \left\|\mathbf{W}_{Iu}^\top \cdot \mathbf{W}_{Tu}\right\| + \left\|\mathbf{W}_{Id}^\top \cdot \mathbf{W}_{Td}\right\|. \tag{17}$$

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

5.1. Base-to-Novel Generalization

Tab. 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 prior works [23, 55, 56]. We evaluate our method on 11 diverse image classification datasets: ImageNet [7] and Caltech101 [10] for general object recognition; OxfordPets [36], StanfordCars [25], Flowers102 [34], Food101 [2], and FGVC-Aircraft [33] for fine-grained classification; SUN397 [35] for scene recognition; DTD [6] for texture classification; EuroSAT [15] for satellite image recognition; and UCF101 [40] 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. R-MMA is compared against several strong baselines (Tab. 1) where it consistently demonstrates superior performance, evidenced by the following observations:

Generalization and Overall Performance: R-MMA achieves the best average harmonic mean (HM) of 81.32% across all 11 datasets, surpassing the previous state-of-the-art method, MMRL [12] (81.20%), by 0.12%. R-MMA also achieves a leading average novel accuracy of 77.72%, approximately 0.56% higher than MMRL (77.16%). These results highlight R-MMA's strong generalization to unseen classes while maintaining a good balance between adaptation and generalization. The performance is consistent across all datasets, with R-MMA achieving the best or close to the best results.

Base Class Performance: R-MMA maintains a competitive base class accuracy, averaging 85.27% across the 11 datasets, which is marginally lower than the highest average of 85.68% by MMRL. R-MMA also achieves leading or near-leading base accuracy on several datasets, *e.g.*, achieving state-of-the-art 78.06% base accuracy on the popular ImageNet. This shows that our method preserves performance on seen classes while substantially improving generalization capabilities to unseen ones.

Performance Nuances: Despite R-MMA's strong overall performance, MMRL [12] slightly outperforms our method across all metrics in Flowers102 and SUN397. For Flowers102, MMRL achieves higher accuracy by 0.16% base, 0.03% novel, and 0.08% 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,

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [37]	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 [56]	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 [55]	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 [32]	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 [49]	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 [23]	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 [3]	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 [26]	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 [24]	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 [47]	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 [54]	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 [50]	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 [48]	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 [12]	**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	85.27	**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 [37]	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 [56]	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 [55]	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 [32]	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 [49]	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 [23]	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 [3]	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 [26]	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 [24]	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 [47]	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 [54]	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 [50]	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 [48]	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 [12]	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	**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 [37]	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 [56]	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 [55]	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 [32]	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 [49]	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 [23]	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 [3]	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 [26]	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 [24]	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 [47]	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 [54]	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 [50]	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 [48]	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 [12]	**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	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 Base-to-Novel Generalization setting. The base and novel class classification accuracies and their harmonic mean (HM) have been provided. HM quantifies the trade-off between adaptation and generalization.

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 performance.

Methods	ImageNet	**Average**	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVC-Aircraft	SUN397	DTD	EuroSAT	UCF101
CoOp [56]	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 [55]	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 [23]	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 [24]	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 [50]	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 [48]	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 [12]	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	**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 with state-of-the-art methods on the Cross-Dataset Evaluation setting. Here, ImageNet is the source dataset, and the rightmost 10 columns are the target datasets. The average is taken across the target datasets.

Method	Source	Target			
	ImageNet	**-V2**	**-S**	**-A**	**-R**
CLIP [37]	66.73	60.83	46.15	47.77	73.96
CoOp [56]	71.51	64.20	47.99	49.71	75.21
CoCoOp [55]	71.02	64.07	48.75	50.63	76.18
MaPLe [23]	70.72	64.07	49.15	50.90	76.98
PromptSRC [24]	71.27	64.35	49.55	50.90	**77.80**
MMA [48]	71.00	64.33	49.13	51.12	77.32
MMRL [12]	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 with state-of-the-art methods on domain generalization across 4 ImageNet variants: V2, Sketch (S), A, and R.

Method	Mod	#Trainable Params	t_{Train} *ms/image*	t_{Train} *min/all*	FPS *100 BS*	HM
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**	894.6	**81.32**

Table 4. Computation cost comparison of different methods using MMRL's setup [12] on ImageNet. 'V-L' denotes vision-language interaction, 'V, L' indicates separate fine-tuning, and 'L' represents language-only fine-tuning. The training time t_{Train} is given for each image in milliseconds (ms) and for the whole dataset (16 shots) in minutes (min). FPS represents frames per second at a 100 batch size during inference.

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 [55], all models are initially trained on ImageNet using §5.1's few-shot setting and then evaluated on the remaining 10 datasets. This setup assesses the model's dataset generalization without fine-tuning or adaptation. Tab. 2 shows R-MMA achieving 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 [12], and a 0.76% improvement over MMA [48]. Specifically, R-MMA obtains the highest accuracy on the source dataset, ImageNet (72.47%), and leads on 4 out of 10 target datasets (OxfordPets, StanfordCars, FGVCAircraft, and EuroSAT). These results validate R-MMA's ability to transfer knowledge effectively and generalize to new, unseen domains.

5.3. Domain Generalization

We evaluate the resilience of models to domain shifts and generalization to out-of-distribution (OOD) data. Following CoCoOp [55], the models, initially trained on ImageNet, are evaluated directly on four ImageNet variants, ImageNetV2 [39], ImageNet-Sketch [44], ImageNet-A [17], and ImageNet-R [16], each introducing a distinct domain variation. Tab. 3 summarizes the domain generalization results where R-MMA achieves the best performance on 3 out of 4 datasets, highlighting its robustness and adaptability to training data distribution shifts.

5.4. Comparison of Computational Cost

To ensure a fair comparison, we follow the configuration used for computational cost estimation by MMRL [12]. All methods are trained on ImageNet, based on the publicly available code and default hyperparameters of the respective papers. Training time is reported as both the average time per image and the total duration required to train the full dataset using 16 shots. Inference speed is measured in frames per second (FPS) with a batch size of 100, indicating the amount of data processed by the model in a unit time.

Following Tab. 4, our proposed method, R-MMA, achieves the highest HM score while using only 0.482M parameters, with the fastest training time (1.6ms/image). Although the inference speed of R-MMA may not be the

(a) Different Model Variants

Model Variants	Base	Novel	HM
w/o MAR	83.04	75.80	79.25
w/o $\mathcal{L}_{sep}$	84.07	75.91	79.78
w/o Attention	83.24	76.78	79.88
w/o Fusion	83.97	76.46	80.40
R-MMA	**85.27**	**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	**85.27**	**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	**85.27**	**77.72**	**81.32**
0.2	83.78	75.92	79.66

Table 5. Ablations of our modules on the 11 datasets used in the Base-to-Novel Generalization setting.

Design Choice	Accuracy			Training Cost	
	Base	Novel	HM	Params (M)	Train Time (ms/img)
w/o Recurrent					
Concatenation	77.92	71.25	74.44	0.563	1.78
Co-Attention	77.64	71.34	74.36	1.538	2.35
R-MMA	**78.32**	71.18	74.58	0.728	1.84
with Recurrent					
Concatenation	77.28	70.07	73.49	**0.394**	**1.18**
Co-Attention	77.59	70.87	74.09	0.927	1.92
R-MMA	78.06	**71.64**	**74.71**	0.482	1.21

Table 6. Comparison of R-MMA with other design approaches on the ImageNet dataset. Here **Bold** indicates the best performance, and Underline refers to the second best.

fastest (491.3 FPS), it outperforms its competitors, MMA [48] and MMRL [12]. Compared to the previous state-of-the-art, MMRL, R-MMA offers a marginal performance gain while reducing parameter count by 7x and improving inference speed by over 1.5×.

5.5. Ablation Study

Model Variants. Tab. 5a validates R-MMA's design choices by showing the performance on 11 datasets used in Base-to-Novel generalizations. The scores drop across all three ablation settings: MAR module (81.32 → 79.75 HM), orthogonality regularization, $\mathcal{L}_{sep}$ (81.32 → 79.78 HM), attention (81.32 → 79.88 HM), and modality fusion (81.32 → 80.40 HM). Tab. 6 shows that other forms of attention, such as concatenation and co-attention, drop performance on ImageNet (details in §C.2).

Dimensions of Hidden Layers. The adapter's hidden layer dimension influences its expressiveness and ability to capture cross-modal relationships. From Tab. 5b, we observe the best performance at the hidden dimension, $d = 64$. Smaller dimensions underfit the model, limiting its ability to learn complex multimodal interactions. Conversely, while increasing the dimension initially improved performance, it eventually degraded, likely due to overfitting from the increased parameter count in few-shot settings.

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 7. Ablation of the loss factor λ on ImageNet.

Model	Base	Novel	HM
Base			
CLIP	72.43	68.14	70.22
SigLIP	74.28	70.46	72.32
R-MMA			
CLIP	78.06	71.64	74.71
SigLIP	**78.75**	**72.13**	**75.34**

Table 8. Ablation of CLIP and SigLIP backbones on ImageNet.

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. In Tab. 5c, we found that a $\alpha = 0.1$ yielded the best results. A smaller α resulted in undermining the adapter features, leading to lower performance. Conversely, larger values undermine the pre-trained knowledge of the frozen features, leading to overfitting and reduced generalization.

Loss Factor λ: Tab. 7 shows that the loss factor, $\lambda = 0.6$, produces optimal balance for R-MMA on ImageNet, while higher and lower values slowly degrade the performance.

CLIP Variations We analyze the impact of alternative CLIP backbones, such as SigLIP [52]. Base SigLIP and SigLIP with R-MMA outperform their CLIP counterparts in both base and novel class performance and their harmonic mean, indicating strong potential for other backbones in adapter-style works. However, for fair comparison, we used the CLIP backbone across all our experiments.

Weight-sharing Ablation. Unlike prior works that use independent adapters at the last k encoder layers [12, 48], our design shares the same adapter weights across all layers. As shown in Tab. 6, weight sharing improves novel class performance at the cost of a minor drop in base class performance, while drastically reducing both parameter count and training cost. The weight-sharing design has two key intuitions. First, our attention-based alignment iteratively refines multimodal representations, enabling a single adapter to capture both generic and dataset-specific

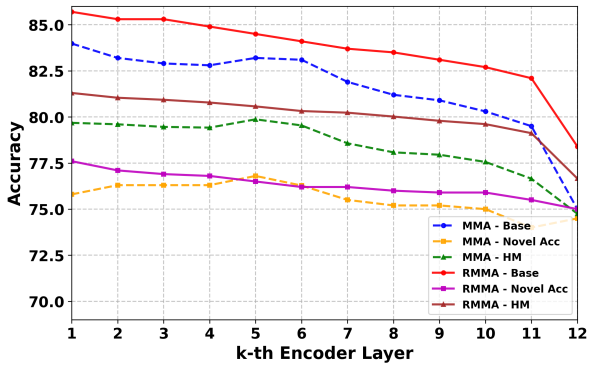

Figure 2. Comparison between R-MMA and MMA on *k-to-12* settings where the adapters are used from layer $l \geq k$. We report average scores of Base, Novel, and HM over 11 datasets from the Base-to-Novel generalization setting.

patterns, similar to how transformers progressively refine features [43]. Secondly, enforcing the same adapter across layers encourages learning a unified representation rather than multiple independent ones. This acts as an effective regularizer in few-shot scenarios by reducing overfitting and leading to stronger generalization on novel classes.

R-MMA in the Higher Layers. We evaluate R-MMA against MMA in the k-to-12 setting [48], where adapters are applied from layer k onward. Following Fig. 2, R-MMA consistently outperforms MMA across all configurations, achieving a 1.62 to 1.91 point improvement in Harmonic Mean (HM) accuracy over 11 datasets. This superiority is primarily driven by a significant 1.7 to 3.4 point gain in base class accuracy, while novel class performance remains competitive. The high efficacy of R-MMA even when adapting only the final few layers demonstrates that our recurrent multimodal adaptation effectively refines the high-level semantic representations in CLIP's higher layers, improving generalization. Details of the experimental setting are provided in §C.1.

Layer Fine-tuning. Tab. 9 presents a comparison of R-MMA, MMA[48], and baselines that fine-tune the final k CLIP layers on all 11 datasets. Fine-tuning more layers typically improves base class accuracy but often degrades novel class generalization, as it impairs the pre-trained model's general knowledge. While some fine-tuning approaches outperform MMA in base accuracy, R-MMA remains at the top in terms of all three metrics, showing better dataset-centric specialization and generalization against more parameter-heavy approaches.

Ablation on Loss $\mathcal{L}_{sep}$. Tab. 10 ablates our orthogonal projection loss $\mathcal{L}_{sep}$ against (i) Cosine loss, encouraging orthogonality by minimizing cosine similarity, (ii) L1 and L2 losses, measuring element-wise absolute and squared differences, respectively, and (iii) MSE loss, penalizing mean squared error between features. Our orthogonal projection loss outperforms the other formulations in all three metrics.

Layer	12	10→12	8→12	5→12	MMA	R-MMA
Base	80.77	83.02	83.77	83.21	83.20	**85.27**
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 9. HM of R-MMA with baseline fine-tuning of the last layers on 11 datasets from the Base-to-Novel generalization setting. "10→12" refers to fine-tuning the last 3 layers.

Loss	Base	Novel	HM
Cosine	85.22	77.26	81.05
L1	85.24	77.12	80.91
L2	85.18	77.05	80.88
MSE	85.22	77.37	81.05
Orthogonal Proj	**85.27**	**77.72**	**81.32**

Table 10. Ablation on the choices of loss $\mathcal{L}_{sep}$, on the 11 datasets used in the Base-to-Novel generalization setting.

6. Discussion: Adapters in the era of LVLMs

Large Vision-Language Models (LVLMs) like GPT-4V, Gemini Vision, and Qwen2.5-VL [1] have raised questions about the need for domain or task-specific adaptation due to their strong zero-shot capabilities. While these capabilities are often demonstrated in multimodal tasks such as VQA [21], Image Captioning [18], and Visual Grounding [46], they still lag in zero-shot and few-shot image classification. For example, GPT-4V achieves a zero-shot performance of only 63.1% on ImageNet [45], which is significantly lower than any of our baseline methods. This suggests that current LVLMs require further improvement to match the performance of fine-tuned approaches.

7. Conclusion

We present R-MMA, a recurrent adapter framework designed to enhance the generalization capabilities of pre-trained VLMs in few-shot settings. By sharing the adapter weights across layers, R-MMA captures rich cross-modal interactions with high parameter efficiency. It uses attention to learn important features in the latent dimension and align the modality streams. Extensive evaluation across 15 datasets on diverse tasks demonstrates R-MMA's effectiveness in generalizing while preserving the rich pre-trained knowledge of the frozen backbone. We envision a potential class of weight-sharing adapters that enhance the transfer learning capabilities of VLMs.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 8

[2] 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, 1

[3] 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. 5

[4] 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

[5] 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

[6] 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, 1

[7] 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. 4, 1

[8] 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

[9] 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

[10] 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, 1

[11] 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

[12] 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, 2, 4, 5, 6, 7

[13] 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

[14] 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

[15] 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, 1

[16] 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. 6, 1

[17] 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. 6, 1

[18] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 8

[19] 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

[20] 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

[21] Md Farhan Ishmam, Md Sakib Hossain Shovon, Muhammad Firoz Mridha, and Nilanjan Dey. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, 106:102270, 2024. 8

[22] 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

[23] 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, 1

[24] 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. 5, 6

[25] 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, 1

[26] 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. 5

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

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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. 5

[33] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. 4, 1

[34] 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, 1

[35] 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, 1

[36] 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, 1

[37] 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, 6

[38] Thomas Rolland and Alberto Abad. Exploring shared-weight mechanisms in transformer and conformer architectures for automatic speech recognition. In *Proc. Interspeech 2025*, pages 2885–2889, 2025. 1

[39] 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. 6, 1

[40] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. 4, 1

[41] 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

[42] 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

[43] 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, 8

[44] 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. 6, 1

[45] Wenhao Wu, Huanjin Yao, Mengxi Zhang, Yuxin Song, Wanli Ouyang, and Jingdong Wang. Gpt4vis: What can gpt-4 do for zero-shot visual recognition? *arXiv preprint arXiv:2311.15732*, 2023. 8

[46] Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. *arXiv preprint arXiv:2412.20206*, 2024. 8

[47] 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. 5

[48] 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, 5, 6, 7, 8

[49] 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. 5

[50] 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. 5, 6, 1

[51] 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. 1

[52] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 7, 1

[53] 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

[54] 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. 5

[55] 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. 4, 5, 6, 1

[56] 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, 1