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 [3]KSoC, University of Utah

Motivation: Few-Shot Adaptation of Vision-Language Models

Vision-Language Models (VLMs) such as CLIP demonstrate strong zero-shot performance by learning aligned image–text representations from web-scale data. However, adapting these models to downstream tasks in **few-shot settings** remains challenging.

Why is adaptation difficult?

- Full fine-tuning is computationally expensive and prone to overfitting.
- Parameter-efficient tuning methods must balance:
  - learning task-specific knowledge
  - preserving CLIP's strong generalization ability.

Limitations of existing adapters:

- Many adapters process image and text streams independently.
- Multimodal adapters scale parameters with model depth.
- Layer-wise adapters rely heavily on frozen features.

Goal: Efficiently adapt CLIP while preserving generalization.

Our Idea: Recurrent Multi-Modal Adapter (R-MMA)

We propose **R-MMA**, a lightweight adapter that:

- Uses a **single shared adapter** across all transformer layers
- Aligns frozen CLIP features using **attention**
- Learns a **unified latent multimodal representation**

Key benefits:

- Parameter count independent of model depth
- Improved cross-modal consistency
- Strong few-shot and cross-domain generalization

Trainable parameters: only **0.48M**

Main Contributions

- Introduce a **Recurrent Multimodal Adapter** shared across all layers.
- Propose **attention-based alignment** with frozen CLIP features.
- Achieve **state-of-the-art** results on:
  - Base-to-Novel generalization
  - Cross-dataset evaluation
  - Domain generalization
- Provide strong **parameter efficiency** (7× fewer params than prior SOTA).

R-MMA Architecture

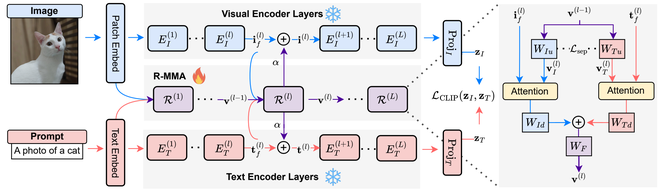

Figure 1. Frozen CLIP encoders (snowflake) with shared recurrent multimodal adapter (fire).

Core components of R-MMA:

- **Modality-Aware Routing (MAR):**

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

- **Attention alignment with frozen CLIP features:**

$$\tilde{\mathbf{v}}_I^{(l)} = \text{Attention}(\mathbf{i}_f^{(l)}, \mathbf{v}_I^{(l)}), \qquad \tilde{\mathbf{v}}_T^{(l)} = \text{Attention}(\mathbf{t}_f^{(l)}, \mathbf{v}_T^{(l)})$$

- **Modality fusion into unified latent token:**

$$\mathbf{v}^{(l)} = \mathbf{W}_F\left[\mathbf{W}_{Id}\tilde{\mathbf{v}}_I^{(l)}\ ;\ \mathbf{W}_{Td}\tilde{\mathbf{v}}_T^{(l)}\right]$$

- **Adapter injection into frozen CLIP layers:**

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

Key idea: A **single shared adapter** is reused across all transformer layers.

Training Objective

CLIP contrastive loss:

$$\mathcal{L}_{CLIP} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp\left(\langle \mathbf{i}_i, \mathbf{t}_i\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle \mathbf{i}_i, \mathbf{t}_j\rangle/\tau\right)} + \log\frac{\exp\left(\langle \mathbf{t}_i, \mathbf{i}_i\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle \mathbf{t}_i, \mathbf{i}_j\rangle/\tau\right)}\right]$$

Orthogonality regularization:

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

Total loss:

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

Encourages disentangled and non-redundant multimodal projections.

Datasets & Experimental Setup

Few-shot benchmarks: ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVC-Aircraft, SUN397, DTD, EuroSAT, UCF101

Domain Generalization: ImageNet-V2, ImageNet-Sketch, ImageNet-A, ImageNet-R

Training: CLIP ViT-B/16 (frozen), AdamW (LR 0.001), mixed precision, single RTX 4090

Overall Performance Summary

R-MMA achieves the strongest overall trade-off between adapting to base classes and preserving generalization to novel classes. It also transfers best to unseen datasets and remains robust under ImageNet domain shifts, while being highly parameter- and compute-efficient.

Evaluation	Metric (Test)	Prior SOTA	R-MMA
Few-Shot Generalization	Base Acc (seen) ↑	85.68 (MMRL)	85.27
Few-Shot Generalization	Novel Acc (unseen) ↑	77.16 (MMRL)	**77.72**
Few-Shot Generalization	HM (Base–Novel) ↑	81.20 (MMRL)	**81.32**
Cross-Dataset Transfer	Avg. Acc ↑	67.25 (MMRL)	**67.37**
Domain Generalization	Best IN variants ↑	2 / 4 (MMRL)	**3 / 4**
Training Cost	ms / image ↓	2.2 (MMA)	**1.6**
Model Size	Trainable Params (M) ↓	4.99 (MMRL)	**0.48**

Key Ablation Findings

Each component contributes to the final performance: removing MAR, attention alignment, fusion, or $\mathcal{L}_{sep}$ consistently reduces HM. Hyperparameter tuning identify an optimal configuration (d=64, α=0.1, λ=0.6). Weight sharing (recurrent design) improves the accuracy–cost trade-off.

Variant	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**

Design Choice	Best	Evidence
Hidden Dim d	64	Large d ↓ HM
Scaling α	0.1	Small underfit, large ↓ gen
Loss weight λ	0.6	Best Base/Novel trade-off
Recurrent sharing	✓	↑ HM + ↓ Params

Conclusion

R-MMA improves few-shot VLM adaptation while preserving generalization, achieving **state-of-the-art performance** with only **0.482M parameters** and the **fastest training**. Its recurrent weight-sharing design makes it both efficient and scalable for real-world adaptation.