

Sustaining Cross-Modal Dependency for Generalizable Vision-Language Model Tuning

Chenhao Ding¹, Jizhou Han², Songlin Dong³, Qiang Wang²,
Xinyuan Gao⁴, Lixin Xiu¹, Wei Ke², Yihong Gong², Hideki Nakayama^{1,*}

¹The University of Tokyo ²Xi'an Jiaotong University
³Shenzhen University of Advanced Technology ⁴Kuaishou
ding225739@eccc.u-tokyo.ac.jp, nakayama@ci.i.u-tokyo.ac.jp

Abstract

Recent parameter-efficient fine-tuning (PEFT) methods effectively adapt Vision-Language Models (VLMs) to seen tasks, yet often exhibit limited generalization to unseen novel classes. To better understand this phenomenon, we analyze adaptation dynamics from the perspective of cross-modal mutual information (MI). Our empirical analyses reveal a consistent pattern across existing methods: during the adaptation process, cross-modal MI for novel classes frequently encounters early plateaus or declines, indicating weakened cross-modal semantic dependency. Motivated by this observation, we propose **Retrieval-Augmented Dynamic Adaptation (RADA)**, a retrieval-enhanced adaptation framework that continuously injects rich instance-specific semantic priors through an LLM-enhanced dual-modality memory bank and a lightweight modality-agnostic token generator. By facilitating fine-grained cross-modal alignment while preserving pretrained zero-shot ability, RADA achieves more transferable adaptation across diverse scenarios. Extensive experiments across multiple widely used benchmarks demonstrate that RADA consistently achieves sota performance while maintaining strong zero-shot capability. In addition, our empirical analyses show that RADA preserves more stable cross-modal dependency trajectories throughout adaptation, offering further insights into the relationship between adaptation dynamics and generalization in VLMs.

1 Introduction

Vision-Language Models (VLMs), such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and SigLIP (Zhai et al., 2023), achieve remarkable open-vocabulary recognition and zero-shot transfer by pretraining on large-scale image-text pairs and aligning multimodal representations in a shared semantic space. To enhance downstream adaptability,

recent studies (Zhou et al., 2022b; Khattak et al., 2023a; Yang et al., 2024a; Ding et al., 2024; Han et al., 2026b; Dong et al., 2025; Li et al., 2026b) have widely adopted the paradigm of Parameter-Efficient Fine-Tuning (PEFT) (Hu et al., 2022; Jia et al., 2022; Li et al., 2026a; Zhao et al., 2026), which freezes the backbone while optimizing only a small set of learnable parameters. However, despite their strong performance improvements on seen distributions, a fundamental yet largely overlooked paradox remains: as training progresses, the generalization gains on unseen tasks often quickly saturate or even deteriorate. In other words, improvements on seen distributions do not necessarily translate into stable and stable cross-task generalization gains.

To better understand this phenomenon, we pose a more fundamental question: is cross-modal semantic dependency between vision and language genuinely strengthened during adaptation? To investigate this, we continuously track the dynamic evolution of cross-modal mutual information (MI) between visual and textual modalities throughout the entire training process across a range of representative PEFT methods. Surprisingly, this diagnosis reveals two highly consistent patterns across all evaluated methods (Fig. 1(b)): (i) **MI remains largely unchanged or even decreases relative to zero-shot initialization** (e.g., MaPLe (Khattak et al., 2023a)); and (ii) **MI exhibits only a brief early increase before quickly saturating** (e.g., MMA (Yang et al., 2024a)), failing to sustain cross-modal dependency over time. More importantly, we further observe that this MI stagnation or decline is highly synchronized with generalization degradation on unseen classes (Fig. 1(c)), suggesting that **cross-modal MI dynamics are not merely a by-product of training, but a meaningful indicator of generalization capability**. This observation motivates us to investigate whether reinforcing such dependency can improve transferability.

* Corresponding author.

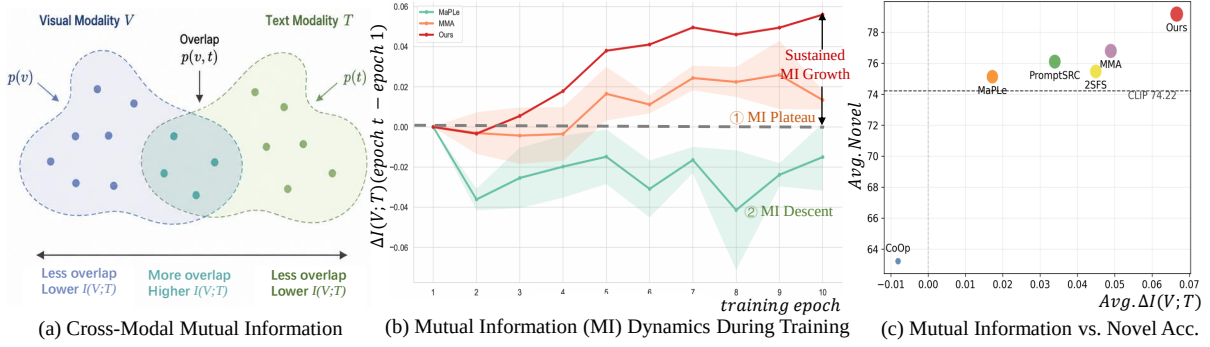


Figure 1: **The impact of cross-modal Mutual Information (MI) on generalization.** (a) MI measures cross-modal semantic overlap. (b-c) Unlike baselines that suffer from MI plateau or descent during training, our method achieves sustained MI growth. This persistent cross-modal dependency directly translates to superior generalization.

We attribute this generalization issue to the **insufficient continual modeling of cross-modal semantic dependency** in existing PEFT methods. Standard adaptation procedures are primarily driven by supervision from base classes, gradually biasing the model toward learning base-specific alignment patterns, while struggling to preserve the transferable cross-modal semantic associations established during large-scale pretraining. As a result, when cross-modal dependency is not stably and continuously reinforced throughout adaptation, the model acquires only task-specific specialization rather than genuinely transferable generalization capability. This analysis further reveals two key requirements for an effective solution: (i) continuously introducing cross-modal semantic priors that complement the limitations of base-class supervision during tuning, and (ii) dynamically modeling vision–language dependency in an instance-specific manner, thereby preserving the inherent zero-shot transferability of VLMs.

Motivated by the above observations, we propose **RADA (Retrieval-Augmented Dynamic Adaptation)**, a framework designed to naturally restore and continuously strengthen cross-modal consistency through retrieval-augmented semantic priors and instance-adaptive modulation. Specifically, to address the limited expressiveness of hand-crafted prompt templates (e.g., “a photo of a [class]”), RADA first constructs a large-scale dual-modal memory bank enhanced by LLMs, providing rich and informative cross-modal semantic prototypes. Given an input image, the framework performs efficient retrieval to obtain semantically relevant visual–textual priors, which are then softly aggregated to form contextualized cross-modal representations. On top of this, a lightweight modality-agnostic token generator dynamically produces

instance-conditioned prompt tokens based on the retrieved context, which are subsequently injected into both the visual and textual encoders. Through this structured design, RADA preserves the inherent zero-shot transferability of VLMs while continuously reinforcing cross-modal dependency, thereby effectively alleviating the generalization saturation problem.

Contributions. Our main contributions are three-fold: **(1)** We analyze the generalization bottleneck in parameter-efficient fine-tuning of VLMs through cross-modal mutual information dynamics, revealing a close connection between performance saturation and the stagnation of cross-modal semantic dependency. **(2)** We propose RADA, a retrieval-augmented dynamic adaptation framework that enhances vision–language consistency via cross-modal semantic priors and instance-adaptive modulation. **(3)** We validate RADA across base-to-novel, few-shot, cross-dataset, and domain generalization settings, showing consistent improvements and interpretable MI dynamics.

2 Related Works

Parameter-efficient fine-tuning. PEFT methods adapt pretrained models by freezing the backbone and updating only a small set of learnable parameters, generally falling into three categories. Important-parameter tuning (Agarwal et al., 2025; Xu et al., 2025; Liu et al., 2025) selectively updates the most task-critical parameters via Fisher information or gradient sensitivity. Adapter-based methods (He et al., 2021; Hu et al., 2022; Zhu et al., 2024; Tian et al., 2024; Ding et al., 2025; Fedus et al., 2022) insert lightweight trainable modules into frozen backbones, with LoRA (Hu et al., 2022) being a representative example. Prompt tuning (Jia et al., 2022) adapts models via learnable tokens in

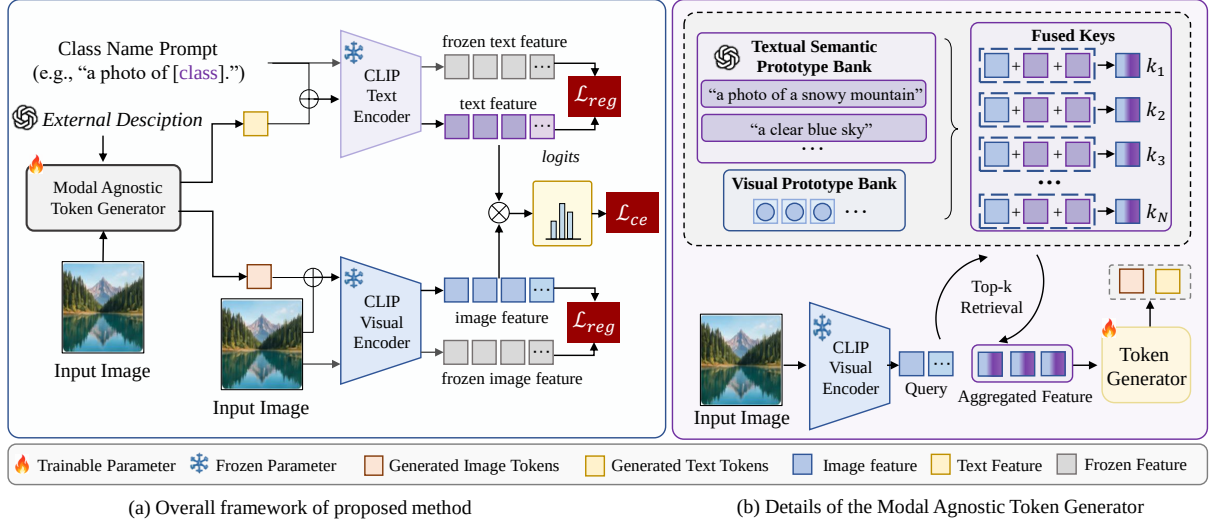


Figure 2: **Overview of the RADA Framework.** (a) Instance-specific tokens are dynamically generated and injected into frozen CLIP encoders, optimized by $\mathcal{L}_{ce}$ and $\mathcal{L}_{reg}$. (b) The token generator retrieves priors from a multimodal memory bank to output adaptive visual and textual tokens.

the input or intermediate layers, and has been extended to vision-language tasks (Zhou et al., 2022b; Khattak et al., 2023a) with strong transferability in models like CLIP.

Information-Theoretic Approaches in Adaptation. In recent years, information-theoretic objectives have been widely applied to visual-language adaptation (Shu et al., 2022; Du et al., 2024; Han et al., 2026a; Xiu et al., 2026; Ding et al., 2026). In the context of test-time adaptation (Feng et al., 2023), entropy minimization is commonly employed to reduce prediction uncertainty and enable unsupervised self-training on target domains. Additionally, some approaches (Karmanov et al., 2024; Zhang et al.) leverage entropy selection to assess sample confidence, further enhancing test-time generalization.

Base-to-Novel of VLMs. While existing PEFT methods have adapted VLMs to downstream tasks, they often overfit base classes, limiting intrinsic generalization to novel classes, a challenge known as the Base-to-Novel (B2N) problem (Zhou et al., 2022a). Early approaches like CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) use static or image-conditioned soft prompts to reduce overfitting. Later methods such as MaPLe (Khattak et al., 2023a), APEX (Yang et al., 2024b), MMA (Yang et al., 2024a), and RMAdapter (Lin et al., 2026) leverage multimodal prompts or adapters to jointly optimize visual and textual representations. Recent works (KgCoOp (Yao et al., 2023), PromptSRC (Khattak et al., 2023b), ProLog (Park et al., 2026)) explicitly preserve pre-

trained knowledge via regularization or distillation, further alleviating catastrophic forgetting.

Despite these advances, generalization remains limited. Unlike prior methods focusing on prompts or knowledge preservation, we revisit B2N from an information-theoretic perspective, noting that existing approaches often fail to continuously strengthen cross-modal semantic dependencies. Motivated by this, we propose a mutual information-based adaptation framework to enhance transferable alignment and improve generalization on unseen tasks while adapting to downstream distributions.

3 Methodology

Our approach builds upon the pre-trained vision-language model CLIP (Radford et al., 2021), following prior work (Zhou et al., 2022b). Here, we provide an overview of the RADA framework and its implementation details.

3.1 Preliminaries

Problem Setting. In this work, we consider a downstream training dataset denoted as $\mathcal{D}_{base} = \{(x_i, y_i)\}_{i=1}^N$, where x_i represents an input image and $y_i \in \mathcal{Y}$ is its corresponding class label. Our primary objective is to optimize a minimal set of learnable parameters on $\mathcal{D}_{base}$ such that the model excels at downstream tasks during inference. Depending on the evaluation protocols, this setup evaluates both in-distribution adaptation (e.g., few-shot learning) and out-of-distribution generalization capabilities (e.g., across unseen classes or tasks).

Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021). CLIP aligns visual and textual modalities in a shared embedding space via contrastive pre-training. Given an image x and a handcrafted textual prompt $p_i = T(w_i)$ for class name w_i , the zero-shot classification probability is derived via the cosine similarity of their encoded features $z_v = E_v(x)$ and $z_t^i = E_t(p_i)$, formulated as $p(y = i|x) = \exp(\cos(z_v, z_t^i)/\tau) / \sum_{j=1}^C \exp(\cos(z_v, z_t^j)/\tau)$, where τ is the learnable temperature.

3.2 Probing the Generalization Bottleneck in B2N: An Information-Theoretic View

To rigorously investigate whether cross-modal semantic dependency is strengthened during adaptation, we analyze the B2N generalization bottleneck through the lens of cross-modal MI dynamics.

Rényi MI Computation. Given a batch of N samples $(x_j, y_j)_{j=1}^N$ from $\mathcal{D}_{\text{novel}}$, we extract visual features $z_v^j = E_v(x_j)$ and label-matched textual features $z_t^{y_j} = E_t(p_{y_j})$, where $p_{y_j} = T(w_{y_j})$ denotes the constructed class prompt. The L_2 -norm features are stacked into matrices $V = [z_v^1, \dots, z_v^N]^\top$ and $T = [z_t^{y_1}, \dots, z_t^{y_N}]^\top$. The cross-modal alignment is then measured using a matrix Rényi mutual information estimator:

$$\hat{I}_\alpha(V; T) = \hat{H}_\alpha(K_V) + \hat{H}_\alpha(K_T) - \hat{H}_\alpha(K_{VT}), \quad (1)$$

where K_V and K_T denote the Gaussian kernel matrices for the visual and textual features, respectively. The joint kernel matrix K_{VT} is obtained via the Hadamard product of the trace-normalized matrices, i.e., $K_{VT} = A \odot B$, where $A = K_V / \text{tr}(K_V)$ and $B = K_T / \text{tr}(K_T)$.

To enhance numerical stability and adapt to different feature spaces, the Gaussian kernel bandwidth σ is dynamically computed using a median heuristic. The matrix Rényi entropy $\hat{H}_\alpha(K)$ is defined based on the normalized eigenvalues $\{\lambda_i\}$ of the kernel matrix.

$$\hat{H}_\alpha(K) = \frac{1}{1-\alpha} \log_2 \left(\sum_i \lambda_i^\alpha \right), \quad (2)$$

where $\{\lambda_i\}$ are the normalized eigenvalues of the kernel matrix. We set $\alpha = 1.01$ to ensure a robust estimation of the training dynamics, which is used to measure the average cross-modal information. As shown in Fig. 1(a) and 7, higher MI reflects stronger, transferable semantic alignment rather than unilateral feature collapse.

Analysis. We investigate the dynamic evolution of cross-modal MI to understand the differences between the zero-shot state and models trained via the standard tuning paradigm. To this end, we use MaPLE and MMA as baseline methods and plot their MI trajectories over the training epochs in Fig. 1(b). By tracking the MI dynamics during training, we observe two patterns: (i) **MI plateau or decline**: MI remains largely unchanged or decreases relative to zero-shot initialization (e.g., MaPLE (Khattak et al., 2023a)); and (ii) **early saturation**: MI briefly increases before quickly stabilizing (e.g., MMA (Yang et al., 2024a)), failing to sustain cross-modal dependency over time.

More importantly, this MI stagnation or decline is **highly synchronized with generalization degradation** on novel classes (Fig. 1(c)). These observations suggest that adaptation strategies focused solely on fitting base-class decision boundaries may produce predominantly base-specific gains, while failing to preserve transferable cross-modal dependencies essential for robust generalization. Consequently, existing PEFT methods appear susceptible to overfitting seen-class semantics, contributing to the B2N generalization bottleneck. Motivated by this observation, we further probe whether sustaining cross-modal dependency plays a functional role in generalization through targeted representation-freezing interventions (Appendix B).

We attribute this to the insufficient continual modeling of cross-modal semantic dependency in standard PEFT: base-class supervision gradually biases the model toward base-specific alignment, eroding the transferable cross-modal associations built during pretraining. This motivates two concrete design requirements for **RADA**: (i) continuously supplying cross-modal semantic priors to complement base-class supervision, and (ii) dynamically modeling vision-language dependency in an instance-specific manner. We therefore pose the following question:

Can we continuously reinforce cross-modal semantic dependency during fine-tuning to overcome the MI stagnation bottleneck, while preserving the VLM’s zero-shot transferability?

3.3 Retrieval-Augmented Dynamic Adaptation (RADA)

To satisfy the two requirements identified above, we propose **RADA** (Retrieval-Augmented Dynamic Adaptation), which addresses require-

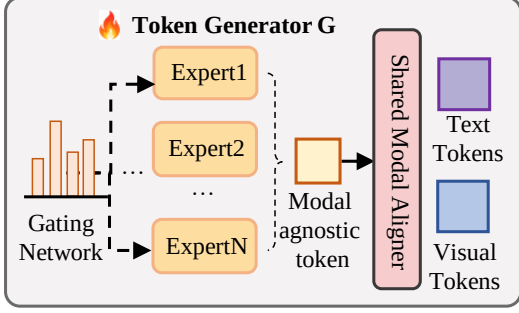


Figure 3: Detailed illustration of our token generator.

ment (i) via a non-parametric dual-modality memory bank supplying LLM-enhanced cross-modal priors, and requirement (ii) via a lightweight modality-agnostic token generator producing instance-conditioned tokens injected into both encoders. An overview is shown in Fig. 2.

RADA consists of a non-parametric multimodal memory bank, a Top- K retrieval module for retrieval-augmented feature construction, and a lightweight token generator with modality-specific projections injected across CLIP layers.

Non-Parametric Multimodal Memory Bank. Conventional handcrafted prompts often provide sparse semantic guidance. To supply richer and more stable semantic priors, RADA constructs a frozen multimodal memory bank that stores text and visual prototypes for each class.

i) Text Bank. For C classes $\{w_c\}_{c=1}^C$, we first leverage an external LLM to generate multi-perspective and fine-grained textual descriptions for each class. Specifically, given M diverse descriptions $\{d_{c,j}\}_{j=1}^M$ generated for class c , we encode them using the text encoder E_t to obtain the category-level enhanced textual representation $\tilde{z}_t^c \in \mathbb{R}^d$ and the text memory bank $\mathcal{B}_t = \{\tilde{z}_t^c\}_{c=1}^C$:

$$\tilde{z}_t^c = \text{Norm} \left(\frac{1}{M} \sum_{j=1}^M E_t(d_j^c) \right). \quad (3)$$

ii) Visual Prototypes. Meanwhile, we extract visual prototypes from available training images using E_v , computed as:

$$m_v^c = \text{Norm} \left(\frac{1}{|\mathcal{D}_c|} \sum_{(x,y) \in \mathcal{D}_c} \mathbb{I}[y=c] \cdot E_v(x) \right), \quad (4)$$

where $\mathcal{D}_c$ denotes the set of training samples belonging to class c . This thereby constructs the visual memory bank $\mathcal{B}_v = \{m_v^c\}_{c=1}^C$.

Joint Retrieval and Aggregation. To alleviate the generalization bottleneck caused by MI stagnation, RADA adopts a Top- K retrieval mechanism to dynamically retrieve relevant joint distributions. As illustrated in Fig. 2(b), to unify bimodal priors in a shared retrieval space, we fuse the enhanced textual representations $\tilde{z}_{t,c}$ with the visual prototypes m_v^c to construct modality-aligned fused keys:

$$k_c = \text{Norm}(m_v^c + \tilde{z}_t^c), \quad c = 1, \dots, C. \quad (5)$$

Given an input image x , we extract its query representation as $z_v = E_v(x)$. We then compute cosine similarities between z_v and all fused keys, i.e., $s_c = \cos(z_v, k_c)$, and select the top- K nearest semantic neighbors $\mathcal{N}_K(x)$.

Next, we apply exponential normalization over the retrieved similarity scores to obtain aggregation weights a_i , and derive the retrieval-enhanced representation r as:

$$a_i = \frac{\exp(s_i)}{\sum_{j \in \mathcal{N}_K(x)} \exp(s_j)}, \quad i \in \mathcal{N}_K(x), \quad (6)$$

$$r = \text{Norm} \left(\sum_{i \in \mathcal{N}_K(x)} a_i k_i \right). \quad (7)$$

Modality-Agnostic Token Generation. Given the aggregated representation r , a lightweight MoE-based generator $\mathcal{G}$ transforms it into dynamic, instance-conditioned prompt tokens. Unlike prior methods that design separate generators per modality, RADA introduces a **Layer Shared Modal Aligner** that projects a single modality-agnostic token P_{ma} into both the visual and textual feature spaces, explicitly enforcing cross-modal consistency at the feature level.

Specifically, r is fed into a gating network to compute soft routing weights:

$$g = \text{Softmax}(W_g r). \quad (8)$$

The representation r is then dispatched to parallel experts, whose outputs are aggregated according to g to produce a modality-agnostic token:

$$P_{ma} = \sum_{j=1}^N g_j \cdot \text{Expert}_j(r). \quad (9)$$

The Shared Modal Aligner then decouples P_{ma} via projection matrix A_t and A_v , mapping it into the CLIP text and visual feature spaces:

$$P_t^{(\ell)} = A_t(P_{ma}), \quad P_v^{(\ell)} = A_v(P_{ma}). \quad (10)$$

Method	LLM Usage	Average across datasets			ImageNet			EuroSAT		
		Base	New	HM	Base	New	HM	Base	New	HM
CLIP (Radford et al., 2021)	✗	69.34	74.22	71.70	72.43	68.14	70.22	56.48	64.05	60.03
CoOp (Zhou et al., 2022b)	✗	82.69	63.22	71.66	76.47	67.88	71.92	92.19	54.74	68.69
MaPLE (Khattak et al., 2023a)	✗	82.28	75.14	78.55	76.66	70.54	73.47	94.07	73.23	82.35
PromptSRC (Khattak et al., 2023b)	✗	84.26	76.10	79.97	77.60	70.73	74.01	92.90	73.90	82.32
MMA (Yang et al., 2024a)	✗	83.20	76.80	79.87	77.31	71.00	74.02	85.46	<u>82.34</u>	83.87
2SFS (Farina et al., 2025)	✗	85.55	75.48	80.20	77.71	70.99	74.20	96.91	67.09	79.29
RMAdapter (Lin et al., 2026)	✗	84.52	<u>77.36</u>	80.62	77.87	<u>71.50</u>	<u>74.52</u>	92.17	78.33	84.69
CuPL (Pratt et al., 2023)	✓	74.31	75.25	74.78	75.05	68.43	71.59	65.38	69.97	67.60
HPT (Wang et al., 2024b)	✓	84.32	76.86	80.23	<u>77.95</u>	70.74	74.17	94.24	77.12	84.82
ProLoG (Park et al., 2026)	✓	84.77	76.89	80.64	78.25	70.51	74.18	93.87	78.86	85.67
RADA (ours)	✓	85.75	78.50	81.97	77.60	71.70	74.53	<u>96.83</u>	86.57	91.41

Table 1: **Comparison of RADA with previous SOTA methods on Base-to-new generalization benchmark.** The best results are **bolded** and the second-best are underlined.

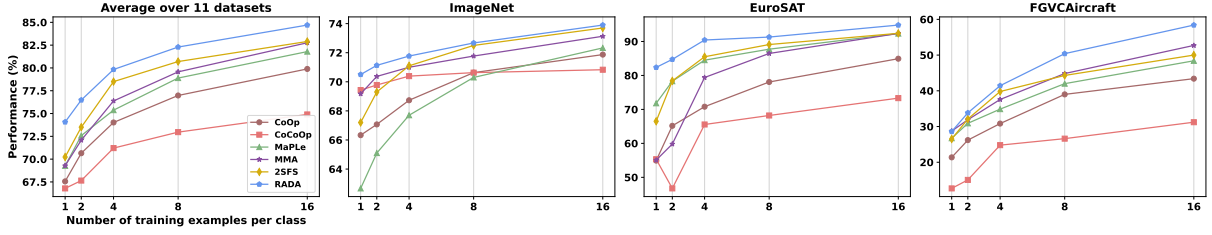


Figure 4: **Comparison of RADA with previous SOTA methods on Few-shot learning benchmark.**

The generated tokens are injected across multiple layers of the CLIP encoders, producing enhanced visual representations $\tilde{z}_v$ and adapted class-specific textual representations $\{\tilde{z}_t^c\}_{c=1}^C$.

Training Objective. During the fine-tuning stage, we jointly optimize a cross-entropy loss and a feature regularization loss to adapt to downstream tasks while mitigating catastrophic forgetting of pretrained knowledge. Given a training sample with ground-truth label $y \in 1, \dots, C$, we compute the logit using the enhanced representations:

$$\tilde{p}(y = c|x) = \frac{\exp(\cos(\tilde{z}_v, \tilde{z}_t^c)/\tau)}{\sum_{j=1}^C \exp(\cos(\tilde{z}_v, \tilde{z}_t^j)/\tau)}, \quad (11)$$

where τ denotes the learnable temperature parameter in CLIP and j serves as the class index.

The cross-entropy loss is then formulated as:

$$\mathcal{L}_{ce} = -\log \tilde{p}(y|x). \quad (12)$$

To prevent the injected tokens from distorting the pretrained feature space, we introduce a regularization loss $\mathcal{L}_{reg}$ that penalizes deviation of the adapted features $\tilde{z}_v$ and $\tilde{z}_t^c$ from their frozen zero-shot counterparts:

$$\mathcal{L}_{reg} = \left(1 - \frac{1}{B} \sum_{i=1}^B \cos(z_{v,i}, \tilde{z}_{v,i})\right) + \left(1 - \frac{1}{C} \sum_{c=1}^C \cos(z_t^c, \tilde{z}_t^c)\right), \quad (13)$$

where B is batch-size. The overall optimization objective is defined as:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{ce} + (1 - \lambda) \mathcal{L}_{reg}, \quad (14)$$

where λ is a hyperparameter that controls loss weight.

Test-time Knowledge Fusion. To satisfy requirement (ii) at inference time, the visual encoder simultaneously produces an enhanced representation $\tilde{z}_v$ and an unenhanced representation z_v from the original pathway. We compute a base logit $p_b(y = c|x)$ from z_v paired with the adapted textual feature $\tilde{z}_t^c$:

$$p_b(y = c|x) = \frac{\exp(\cos(z_v, \tilde{z}_t^c)/\tau)}{\sum_{j=1}^C \exp(\cos(z_v, \tilde{z}_t^j)/\tau)}. \quad (15)$$

The final prediction is obtained by combining p_b with the enhanced logit $\tilde{p}$ defined in Eq. 11:

$$p_{fus} = (p_b + \tilde{p}) / 2. \quad (16)$$

For zero-shot scenarios where no visual training samples are available, we set $m_v^c = \tilde{z}_t^c$, allowing RADA to degrade gracefully to a text-driven retrieval mode.

4 Experiments

4.1 Experimental Setup

Benchmark. To comprehensively evaluate the generalization capability of the proposed method and

Method	Source	Target										
	ImNet	Ave.	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	UCF
CoOp	71.51	63.88	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55
MaPLe	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	71.27	65.81	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75
MMA	71.00	66.61	93.80	90.30	66.13	72.07	86.12	25.33	68.17	46.57	49.24	68.32
TCP	71.40	66.29	93.97	91.25	64.69	71.21	86.69	23.45	67.15	44.35	51.45	68.73
HPT	71.72	67.74	94.20	92.63	66.33	74.84	86.21	25.68	68.75	50.87	47.36	70.50
RMAdapter	71.37	67.56	94.10	90.85	65.80	72.65	86.33	24.90	68.36	47.33	51.50	69.96
ProLoG	71.06	66.43	94.35	90.44	66.21	72.41	86.46	23.69	67.51	45.39	48.16	69.67
RADA (ours)	72.17	68.08	94.83	91.67	66.70	73.30	86.87	26.57	68.33	47.67	54.90	70.00

Table 2: Comparison with previous SOTA methods on cross-dataset transfer across 11 datasets.

its effectiveness on downstream tasks, we follow the experimental protocols of prior works (Zhou et al., 2022b,a; Yang et al., 2024a) and conduct experiments on 4 widely adopted benchmarks: *Base-to-Novel Generalization*, *Few-shot Learning*, *Cross-Dataset Transfer*, and *Domain Generalization*. For fair comparison, all methods are evaluated using the same backbone architecture. Detailed descriptions of the implementation details, datasets, baselines, and benchmark settings are provided in the Appendix A.

4.2 Base-to-Novel Generalization

Tables 1 and 17 report results across 11 datasets. RADA achieves average improvements of +0.98% on base, +1.61% on novel, and +1.33% in HM over the previous best method ProLoG, establishing SOTA performance. Notably, on EuroSAT and FGVC Aircraft, datasets with large domain discrepancy, RADA achieves HM scores of 91.41% and 42.27%, respectively, substantially outperforming all baselines. These gains are directly supported by RADA’s sustained MI growth throughout adaptation. Our intervention studies (Appendix B) consistently demonstrate a strong association between sustained cross-modal dependency and improved novel-class generalization.

Observation 1: Sustained cross-modal MI growth is consistently associated with stronger B2N generalization.

4.3 Domain Generalization

As summarized in Table 3, RADA achieves the best performance across almost all datasets subject to domain shift (ImageNet-V2, ImageNet-Sketch, ImageNet-A, and ImageNet-R). By attaining a highest average accuracy of 61.09%, RADA demonstrates robust cross-domain generalization and strong resilience to distributional shifts.

Method	Source	Target				
	ImNet	-V2	-S	-A	-R	Ave.
CLIP	66.73	60.83	46.15	47.77	73.96	57.17
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
MaPLe	70.72	64.07	49.15	50.90	76.98	60.26
PromptSRC	71.27	64.35	49.55	50.90	77.80	60.65
MMA	71.00	64.33	49.13	51.12	77.32	60.48
RMAdapter	71.37	64.45	49.50	51.21	77.70	60.71
ProLoG	71.06	64.36	49.64	51.05	77.50	60.64
HPT	71.72	65.25	49.36	50.85	77.38	60.71
RADA (ours)	72.17	64.90	49.67	51.87	77.93	61.09

Table 3: Comparison with previous SOTA methods on domain generalization across 4 datasets.

4.4 Cross-Dataset Transfer

As shown in Table 2, RADA achieves a 72.17% accuracy on the source ImageNet dataset, surpassing previous SOTA methods such as ProLoG and RMAdapter. Beyond this, RADA consistently exhibits superior performance across all 10 various target datasets, achieving the highest average accuracy of 68.08%, which underscores its strong cross-dataset generalization capability.

Observation 2: Continuously reinforced cross-modal dependency yields robust generalization under distributional shifts across domains even under domain shift.

4.5 Few-Shot Learning

As shown in Fig. 4, RADA achieves the best average performance across all shot settings on 11 datasets. Notably, even under extremely data-scarce conditions (e.g., 1-shot and 2-shot settings), RADA still demonstrates strong learning capability and stable performance advantages. This observation further indicates that RADA maintains strong transfer learning and generalization capabilities in low-resource scenarios. **More detailed results are provided in the Appendix C.2.**

Method	Base	Novel	HM
CLIP	69.34	74.22	71.70
+(1) Retrieval Bank	75.43	74.67	75.05
+(1)+(2) Token Generator	86.51	76.83	81.38
+(1)+(2)+(3) $\mathcal{L}_{\text{reg}}$	85.75	77.94	81.54
+ (1) + (2) + (3) + (4) Logit Fusion	85.75	78.50	81.97

Table 4: **Component-wise ablation study of RADA on Base-to-Novel Generalization.**

Observation 3: Dynamic instance-conditioned adaptation maintains strong generalization even in low-shot regimes.

4.6 Ablation Analysis

All ablation experiments are conducted on Base-to-Novel generalization across 11 datasets, with the results reported as averages.

The Effectiveness of Each Component. To verify the effectiveness of each component in our method, we conducted ablation experiments on the B2N task, as summarized in Table 4 and Fig. 7 in Appendix B.3. The results support our theoretical motivation regarding cross-modal MI stagnation and the B2N bottleneck. Starting from the Zs-CLIP baseline, introducing the non-parametric retrieval bank (1) **significantly improves Base from 69.34% to 75.43%**. Adding the dynamic Token Generator (2) **further enhances cross-modal alignment**, raising Base to 86.51% and Novel to 76.83%, indicating that the generated tokens contribute to downstream task adaptation and provide an initial improvement on Novel performance. Incorporating $\mathcal{L}_{\text{reg}}$ (3) maintains consistency between fine-tuned features and Zero-shot features, which **increases Novel to 77.94%**, with Base slightly decreasing to 85.75%. Finally, integrating Logit Fusion (4) at inference combines the robustness of the pretrained model, **achieving a balanced performance between Base and Novel classes, resulting in Novel of 78.50% and HM of 81.97%**.

Observation 4: Each component of RADA actively contributes to preventing mutual information stagnation during adaptation.

The Influence of Retrieval Quantity (Top-K). As shown in Table 5, RADA’s performance varies with the number of retrieved neighbors K . The HM peaks at $K = 4$ (**81.97%**), successfully balancing rich semantic priors with contextual relevance. Smaller K restricts semantic coverage, while larger

K	Base	Novel	HM	L	Base	Novel	HM
1	84.31	76.43	80.18	1	84.03	75.76	79.68
2	85.12	77.81	81.30	3	85.06	77.58	81.15
4	85.75	78.50	81.97	5	85.75	78.50	81.97
8	85.83	78.14	81.81	10	85.91	77.37	81.42
16	85.24	77.62	81.25	16	86.13	76.62	81.10

Table 5: **Sensitivity analysis of the retrieval quantity (Top- K) and dynamic token length (L) on Base-to-Novel generalization.**

K introduces irrelevant noise, leading to a gradual decline in novel-class accuracy.

The Influence of Token Length (L). The dynamic token length L governs the cross-modal adaptation capacity. Very short sequences ($L = 1$) prove insufficient for effective alignment. Conversely, while longer sequences ($L \geq 10$) further elevate base accuracy (up to **86.13%**), they risk overfitting to the base distribution, thereby degrading novel-class generalization. Consequently, we adopt $K = 4$ and $L = 5$ as our default configuration to achieve an optimal trade-off.

Beyond the main results, we provide extensive supplementary analyses in Appendix C and Appendix B. These include ablations on the memory bank design (modality choices and prompt types) and token generator architecture (aligner designs and MoE configuration), as well as hyperparameter sensitivity analyses, expert usage frequency, and t-SNE visualizations. Notably, we further conduct an epoch-wise intervention study examining whether sustained MI growth contributes to generalization improvements across diverse domains (Appendix B), and an information-theoretic analysis exploring why explicitly optimizing MI does not necessarily improve transferability.

5 Conclusion

In this paper, we observe that cross-modal MI stagnation correlates with the B2N generalization gap during VLM adaptation. To address this, we propose the **RADA** framework, which incorporates instance-specific semantic priors into the adaptation process through a multimodal memory bank and a modality-agnostic dynamic token generator. This method encourages stronger cross-modal consistency while better retaining the model’s zero-shot capabilities. Extensive experiments across benchmarks demonstrate that RADA consistently maintains higher estimated MI values during adaptation and achieves SOTA performance. Our find-

ings suggest that sustaining transferable cross-modal dependency may be an important ingredient for robust generalization in VLM adaptation.

Limitations

While RADA effectively mitigates the Base-to-Noel generalization bottleneck, our current evaluation is primarily concentrated on image classification, leaving its applicability to broader vision-language tasks (e.g., image-text retrieval or multi-modal generation) for future exploration. Regarding the architectural design, the proposed MoE-based token generator and dual-modality memory bank represent a specific instantiation of our theoretical framework, suggesting that alternative retrieval mechanisms could yield further optimization. Moreover, due to computational resource constraints, our empirical validation is currently limited to foundational VLMs like CLIP. Scaling this framework to assess its efficacy on Large Vision-Language Models (LVLMs) such as Qwen-VL (Wang et al., 2024a) remains a critical next step for our future research.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. U21B2048, No. 62302382, and No. 62576224, the Shenzhen Key Technical Projects under Grant CJGJZD20220517141605013, the China Postdoctoral Science Foundation No. 2024M752584, and the Guangdong Provincial Key Laboratory of Computability Microelectronics with Grant 2024B1212010007.

References

- Aradhye Agarwal, Suhas Kamasetty Ramesh, Ayan Sen-gupta, and Tanmoy Chakraborty. 2025. Step-by-step unmasking for parameter-efficient fine-tuning of large language models. *Transactions of the Association for Computational Linguistics*, 13:1767–1788.
- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101—mining discriminative components with random forests. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Chenhao Ding, Xinyuan Gao, Songlin Dong, Jizhou Han, Qiang Wang, Zhengdong Zhou, Yuhang He, and Yihong Gong. 2024. Lpt: Less-overfitting prompt tuning for vision-language model. *arXiv preprint arXiv:2410.10247*.
- Chenhao Ding, Jizhou Han, Songlin Dong, Xiang Song, Xinyuan Gao, Qiang Wang, Yuhang He, Zhiheng Ma, and Yihong Gong. 2026. Reorienting the frozen space: Training-free test-time adaptation by geometric transformation. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Chenhao Ding, Jiangyang Li, Songlin Dong, Xinyuan Gao, Yuhang He, and Yihong Gong. 2025. Sulora: Subspace low-rank adaptation for parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5334–5349.
- Songlin Dong, Chenhao Ding, Jiangyang Li, Jizhou Han, Qiang Wang, Yuhang He, and Yihong Gong. 2025. Beyond clip generalization: Against forward&backward forgetting adapter for continual learning of vision-language models. *arXiv preprint arXiv:2505.07690*.
- Xin Du, Lixin Xiu, and Kumiko Tanaka-Ishii. 2024. Bottleneck-minimal indexing for generative document retrieval. *arXiv preprint arXiv:2405.10974*.
- Matteo Farina, Massimiliano Mancini, Giovanni Iacca, and Elisa Ricci. 2025. Rethinking few-shot adaptation of vision-language models in two stages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29989–29998.
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.
- Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE.
- Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. 2023. Diverse data augmentation with diffusions for effective test-time prompt tuning.

- In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714.
- Jizhou Han, Chenhao Ding, Songlin Dong, Xinyuan Gao, Qiang Wang, Yuhang He, and Yihong Gong. 2026a. Unleashing the potential of all test samples: Mean-shift guided test-time adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Jizhou Han, Chenhao Ding, Yuhang He, Qiang Wang, Shaokun Wang, SongLin Dong, and Yihong Gong. 2026b. Learning like humans: Analogical concept learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28734–28743.
- Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164*.
- Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. 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.
- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, and 1 others. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349.
- Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.
- Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer.
- Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. 2024. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171.
- Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023a. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122.
- Muhammad Uzair Khattak, Syed Talal Wasim, Muza-mmam Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023b. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200.
- Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561.
- Jiangyang Li, Chenhao Ding, SongLin Dong, Qiang Wang, Jianchao Zhao, Yuhang He, and Yihong Gong. 2026a. Is parameter isolation better for prompt-based continual learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3887–3897.
- Jiangyang Li, Cong Wan, SongLin Dong, Chenhao Ding, Qiang Wang, Zhiheng Ma, and Yihong Gong. 2026b. Trajectory-diversity-driven robust vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 9143–9154.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Xiang Lin, Weixin Li, Shu Guo, Lihong Wang, and Di Huang. 2026. Rmadapter: Reconstruction-based multi-modal adapter for vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23594–23602.
- Xinxin Liu, Aaron Thomas, Cheng Zhang, Jianyi Cheng, Yiren Zhao, and Xitong Gao. 2025. Refining salience-aware sparse fine-tuning strategies for language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31932–31945.
- Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.
- Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.
- Jungwuk Park, Dong-Jun Han, and Jaekyun Moon. 2026. Prolog: Hybrid prompt and lora based adaptation of vision-language models for ood generalization. In

- Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 24782–24791.
- Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.
- Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15691–15701.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.
- Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289.
- Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11):1–7.
- Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. 2024. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *arXiv preprint arXiv:2404.19245*.
- Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. 2024b. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 5749–5757.
- Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.
- Lixin Xiu, Xufang Luo, and Hideki Nakayama. 2026. A comprehensive information-decomposition analysis of large vision-language models. *arXiv preprint arXiv:2603.29676*.
- Haoyun Xu, Runzhe Zhan, Yingpeng Ma, Derek F Wong, and Lidia S Chao. 2025. Let’s focus on neuron: Neuron-level supervised fine-tuning for large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9393–9406.
- Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. 2024a. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23826–23837.
- Yongjin Yang, Jongwoo Ko, and Se-Young Yun. 2024b. Towards difficulty-agnostic efficient transfer learning for vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2066–2085.
- Hantao Yao, Rui Zhang, and Changsheng Xu. 2023. 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.
- Hantao Yao, Rui Zhang, and Changsheng Xu. 2024. 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.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.
- Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping.
- Jianchao Zhao, Chenhao Ding, SongLin Dong, Jiangyang Li, Qiang Wang, Yuhang He, and Yihong Gong. 2026. Shared & domain self-adaptive experts with frequency-aware discrimination for continual test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 28697–28705.
- Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348.

Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez De Ocáriz Borde, Rickard Brüel Gabrielson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. 2024. Asymmetry in low-rank adapters of foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62369–62385.

A Additional Experimental Details

A.1 Implementation Details.

Unless otherwise specified, we employ the AdamW optimizer for all experiments. The proposed MoE generator contains 4 experts, and the number of generated prompt tokens is set to 5. Across all benchmarks, each expert within the MoE adopts LoRA with a rank of 8. These tokens are injected into layers $\{6, 7, 8, 9, 10, 11, 12\}$ of the CLIP encoder. Additionally, we generate $M = 8$ fine-grained descriptions for each class to provide semantic priors. We empirically set the loss weight λ for different benchmarks. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU, and we report the average results across three random seeds $\{1, 2, 3\}$. We follow prior work (Khattak et al., 2023b; Park et al., 2026) by injecting prompt tokens into layers $\{6, 7, 8, 9, 10, 11, 12\}$ of CLIP.

The computation of cross-modal MI on the novel dataset $\mathcal{D}_{\text{novel}}$ is conducted strictly for diagnostic and analytical purposes. These MI values are neither utilized as an optimization objective nor used to guide architectural choices, hyperparameter tuning, or early stopping decisions. This protocol ensures that no information from the novel distribution leaks into the training process, thereby adhering to the strict requirements of zero-shot evaluation.

For **Base-to-Novel Generalization**, the model is trained with a batch size of 4 for 10 epochs using an initial learning rate of 0.001. The regularization weight is set to $\lambda = 0.8$.

For **Cross-Dataset Transfer** and **Domain Generalization**, we follow a lightweight adaptation setting and train the model with a batch size of 32 for a single epoch. The learning rate is set to 0.001, while the remaining hyperparameter is kept identical to B2N, i.e., $\lambda = 0.8$. For **Few-Shot Learning**, the model is trained with a batch size of 32 for 100 epochs using a learning rate of 0.002. We set $\lambda = 0.7$ to better balance adaptation and generalization.

For semantic prior construction, we leverage GPT-generated fine-grained textual descriptions for each class. Specifically, GPT-3.5Turbo (Brown et al., 2020) is prompted to generate multi-perspective semantic descriptions, including class attributes, visual appearance, functional characteristics, and contextual cues, which are subsequently encoded by the frozen CLIP text encoder to con-

Table 6: Statistics of datasets used in experiments.

Dataset	Classes	Test Samples
ImageNet	1,000	50,000
ImageNet-V2	1,000	10,000
ImageNet-Sketch	1,000	50,000
ImageNet-A	200	7,500
ImageNet-R	200	30,000
FGVC Aircraft	100	3,333
Caltech101	101	2,465
Stanford Cars	196	8,041
DTD	47	1,692
EuroSAT	10	8,100
Flowers102	102	2,463
Food101	101	30,300
OxfordPets	37	3,669
SUN397	397	19,850
UCF101	101	3,783

struct the text memory bank. The details of the prompt are provided in Table 16.

A.2 Benchmark Protocols

Base-to-Novel Generalization. In this setting, each dataset is split into disjoint base and novel classes. The model is trained exclusively on the base classes and evaluated on both base and novel classes, enabling us to assess both transfer performance and zero-shot generalization capability. We conduct evaluations across 11 diverse image classification datasets, including ImageNet, Caltech101 (Fei-Fei et al., 2004), OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback and Zisserman, 2008), Food101 (Bossard et al., 2014), FGVC Aircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), UCF101 (Soomro et al., 2012), DTD (Cimpoi et al., 2014), and EuroSAT (Helber et al., 2019). The details of all datasets are shown in Table 6. For this setting, we follow recent works (Zhou et al., 2022a; Khattak et al., 2023a; Yang et al., 2024b) and adopt the ViT-B/16 backbone with $k = 16$ shots. The evaluation metrics are threefold: (i) top-1 accuracy on base classes, (ii) top-1 accuracy on novel classes, and (iii) the harmonic mean between them.

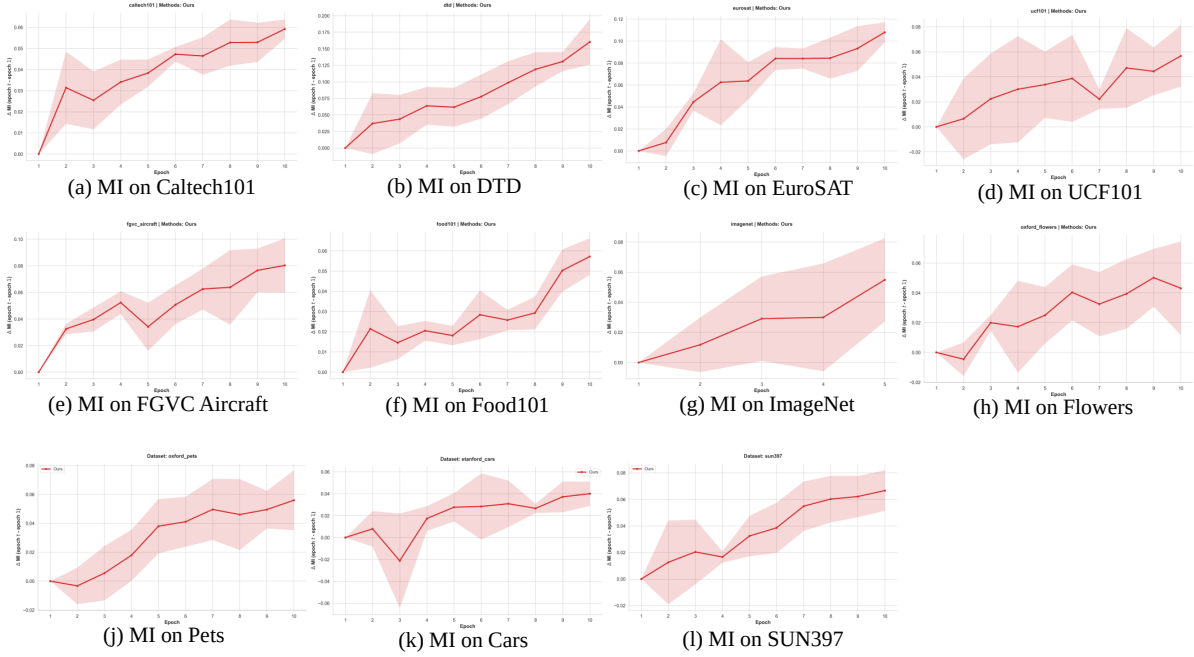


Figure 5: **Detailed Visualization of Mutual Information Trajectories Across All Evaluated Datasets.** The MI values across all datasets are averaged over three random seeds $\{1, 2, 3\}$, with the shaded regions indicating the 95% confidence intervals.

Cross-Dataset Transfer. This setting evaluates the transferability of the model to unseen target datasets. Following a standard evaluation protocol, the model is trained on all 1000 ImageNet (Deng et al., 2009) classes in a few-shot learning scenario and then directly evaluated on other datasets without any further fine-tuning. This allows us to measure the model’s cross-dataset generalization ability under domain shift conditions. The same set of datasets used in the Base-to-Novel setting is adopted for evaluation.

Domain Generalization. This setting assesses the robustness of the model under distribution shifts and its ability to generalize to out-of-distribution data. Specifically, the model is trained on ImageNet (Deng et al., 2009) and evaluated on multiple variants that exhibit different types of domain shift, including ImageNetV2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-R (Hendrycks et al., 2021a).

Few-Shot Learning. This setting evaluates the model’s transfer capability under extremely limited labeled data. The model is trained using a small number of labeled examples per class, specifically 1, 2, 4, 8, and 16 shots, and is subsequently evaluated on the full test sets. This protocol focuses

on assessing data efficiency and adaptation performance in low-resource scenarios.

A.3 Baselines

In the Base-to-Novel Generalization setting, we compare a total of eight methods, including CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), and CuPL (Pratt et al., 2023) as classical baselines, MaPLe (Khattak et al., 2023a), APEX (Yang et al., 2024b), HPT (Wang et al., 2024b), and MMA (Yang et al., 2024a) as representative modern improvements, and 2SFS (Farina et al., 2025), RMAdapter (Lin et al., 2026), and ProLoG (Park et al., 2026) as recent strong-performing methods. In addition, we report the zero-shot performance of hand-crafted prompt-based CLIP as a reference baseline.

For the Cross-Dataset Transfer setting, we evaluate the same set of eight methods. Since 2SFS does not report results on this benchmark, we additionally include TCP (Yao et al., 2024) for a fair comparison.

In the Domain Generalization setting, we also consider eight competing methods. However, since neither 2SFS nor TCP provides reported results under this setting, TCP is excluded and we instead include alternative comparable baselines to ensure a consistent evaluation protocol.

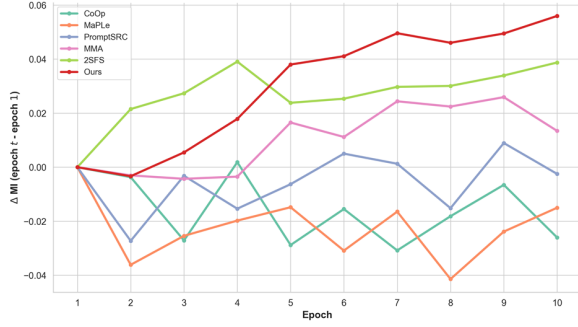


Figure 6: **Detailed Mutual Information (MI) Dynamics During Training.**

For the Few-Shot Learning setting, we compare against CoOp, CoCoOp, MaPLe, MMA, and 2SFS. We exclude APEX, RMAAdapter, and ProLoG since their few-shot results are not reported in the original papers.

B Validating the Role of Mutual Information

In this section, we provide a step-by-step analysis to examine the role of cross-modal MI, progressing from correlational observations to targeted interventions and counterfactual analysis. We first examine the natural trajectories of MI, followed by a direct causal intervention study, component-level dynamic tracking, and a counterfactual explicit optimization analysis.

B.1 Correlational Observations: MI Trajectories and Generalization

Impact of Mutual Information on Generalization. As discussed in Section 3.2, cross-modal MI quantifies the semantic dependency and alignment quality between visual and textual features. During fine-tuning, **stagnation or decline in MI indicates the occurrence of constrained generalization**, where the model overfits to base-class decision boundaries without learning transferable visual-language alignment. Maintaining continuous MI growth throughout training is therefore **critical for breaking the generalization bottleneck**.

Analysis of RADA’s MI Trajectories in Fig. 5. Fig. 5 depicts the MI changes (ΔI) of RADA across all evaluated datasets over training iterations. We observe several patterns that directly contribute to RADA’s superior generalization:

1. Sustained and stable growth: On multi-source datasets such as ImageNet, Caltech101, Eu-

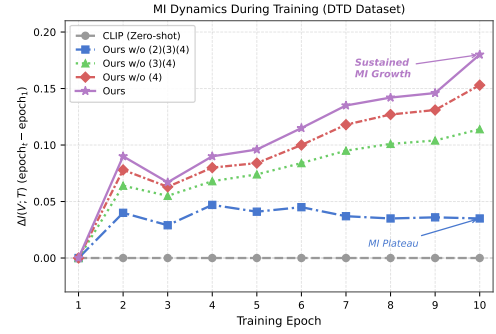


Figure 7: **Cross-modal MI Dynamics of Each RADA Component During Training on DTD.**

roSAT, and SUN397, MI exhibits a consistent upward trend without premature saturation. This sustained enhancement of cross-modal dependencies explains why RADA achieves substantial improvements on novel-class generalization (e.g., as shown in Table 6, EuroSAT novel-class accuracy reaches 96.83%).

2. Robust recovery and stability: On fine-grained or distribution-specific datasets like Stanford Cars, Flowers102, and UCF101, MI may initially fluctuate due to domain shifts. However, RADA consistently drives MI back up and stabilizes it at a high positive level, demonstrating strong robustness against base-class overfitting in challenging distributions.

More Comparison on MI. As shown in Fig. 6, we further provide a detailed comparison of the visualization in Fig. 1(b) by incorporating additional representative methods, including CoOp, PromptSRC, and 2SFS. From Fig. 6, we can clearly observe that existing methods exhibit two distinct patterns in the evolution of cross-modal MI, namely **MI Plateau** and **MI Descent**. Combined with the performance trends illustrated in Fig. 1(c), these findings suggest that **both MI stagnation and degradation constrain the model’s generalization capability on unseen classes, ultimately leading to performance saturation under the Base-to-Novel setting**.

Observation 5: Our analysis reveals that the stagnation or decline of cross-modal mutual information is a primary indicator of constrained generalization, whereas sustained MI growth is pivotal for achieving robust and transferable generalization in VLMs.

Intervention Strategy	Generator Status	DTD		EuroSAT		Oxford Pets	
		Final ΔI	HM (%)	Final ΔI	HM (%)	Final ΔI	HM (%)
Zero-shot CLIP	-	0.000	56.37	0.000	60.03	0.000	94.12
Freeze at Epoch 2	Frozen early	+0.037	64.09	+0.008	72.35	-0.003	94.55
Freeze at Epoch 5	Frozen mid-training	+0.061	68.22	+0.064	83.12	+0.038	95.88
Freeze at Epoch 8	Frozen late-training	+0.119	71.65	+0.084	88.54	+0.046	96.34
RADA (No Freeze)	Active continuously	+0.160	73.62	+0.110	91.41	+0.056	96.82

Table 7: **Impact of Epoch-wise Generator Intervention Across Diverse Domains.** Freezing the dynamic token generator prematurely halts MI growth and strictly upper-bounds the HM across distinct data distributions. This establishes a robust causal link between the duration of active dynamic adaptation and out-of-distribution transferability.

B.2 Intervention: Cross-Domain Validation of MI as a Generalization Helper

To rigorously address the question of whether cross-modal MI growth is a causal driver of generalization rather than a mere correlational byproduct, we conduct a comprehensive epoch-wise intervention study. We select three datasets representing distinct distribution shifts: DTD (unstructured textures), EuroSAT (satellite imagery), and Oxford Pets (fine-grained objects). By manually freezing the parameter updates of the dynamic token generator at different training stages (Epochs 2, 5, and 8), we artificially truncate the mechanism responsible for cross-modal alignment while keeping the overall architecture and optimization target fixed.

As presented in Table 7, the intervention reveals a consistent, monotonic causal dependence across all domains. For instance, on the EuroSAT dataset, when the generator is frozen early at Epoch 2, the MI trajectory prematurely plateaus at a marginal $+0.008 \Delta I$, which strictly upper-bounds the corresponding HM) to 72.35%. Conversely, allowing continuous adaptation yields $+0.110 \Delta I$ and 91.41% HM.

More compellingly, on the Oxford Pets dataset, freezing at Epoch 2 captures a transient negative fluctuation in MI ($-0.003 \Delta I$), which is directly reflected by an almost stagnant generalization gain (94.55% vs. zero-shot 94.12%). Once the generator continues to actively align the domains (Epoch 5 and beyond), both MI and HM recover and grow synchronously.

This multi-dataset intervention robustly rules out alternative explanations such as dataset-specific overfitting or catastrophic forgetting. If generalization saturation were solely caused by base-class overfitting, freezing the generator late (Epoch 8) would have preserved base accuracy while allowing novel performance to drop randomly. Instead, we

observe that the ceiling of the HM score is tightly associated with the level of MI achieved at the moment of intervention, suggesting that sustained cross-modal alignment is a primary functional contributor to novel-class transferability.

The fact that this causal link holds across unstructured textures, aerial perspectives, and fine-grained visual classes strongly supports the hypothesis that the continuous reinforcement of instance-level cross-modal dependencies serves as a key driving mechanism for breaking the Base-to-Novel generalization bottleneck.

Observation 6: The epoch-wise intervention provides strong evidence that mutual information growth is not a mere correlational byproduct, but a key functional driver of out-of-distribution generalization: prematurely truncating the dynamic alignment process establishes a strict upper bound on novel-class transferability.

B.3 Component-Level Dynamics: Disentangling the Source of MI Growth

Fig. 7 plots the $\Delta I(V; T)$ trajectory for each ablation variant on DTD, providing MI-level evidence connecting Table 4 to our theoretical motivation in Section 3.2. The retrieval bank (1) lifts ΔI above the zero-shot reference but plateaus at $\approx +0.035$ after epoch 4, explaining the marginal novel-class gain ($+0.45\%$): static priors alone cannot sustain cross-modal dependency. The token generator (2) breaks this plateau and drives continuous growth to $\Delta I = +0.114$ ($3.3\times$ higher), confirming it as the primary architectural driver of generalization. $\mathcal{L}_{\text{reg}}$ (3) further stabilizes late-epoch growth (epochs 7–10) by preventing feature drift from the zero-shot manifold, reaching $\Delta I = +0.153$ and producing the predicted Base–Novel tradeoff (-0.76% Base, $+1.11\%$ Novel).

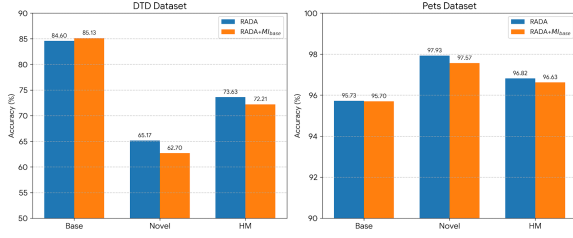


Figure 8: **Performance comparison between RADA and RADA + MI_{base} on DTD and Pets datasets.** Explicitly maximizing MI_{base} degrades novel-class generalization.

The complete RADA achieves the strongest sustained growth of $\Delta I = +0.180$ ($5.1\times$ over retrieval-only), consistently dominating all variants at every epoch and corroborating Appendix D: sustained MI growth tightens the theoretical generalization bound and directly translates into superior novel-class performance.

The consistency of this pattern across unstructured textures, aerial perspectives, and fine-grained visual categories provides converging empirical support that continuously reinforcing instance-level cross-modal dependencies plays a meaningful functional role in alleviating the Base-to-Novel generalization bottleneck.

Mechanism Behind RADA’s Improved Generalization. Unlike baselines that rely on static prompts or unimodal adapters, RADA mechanistically sustains $I(V; T)$ through two complementary pathways. **First**, the LLM-enhanced memory bank enriches the effective support of $H(T)$ by providing multi-perspective textual prototypes that go beyond the sparse, handcrafted class descriptions used in standard PEFT. This prevents the textual marginal distribution from collapsing onto a narrow base-class manifold, thereby preserving the upper bound of $I(V; T)$ throughout training. **Second**, the modality-agnostic token generator injects instance-conditioned perturbations *symmetrically* into both the visual and textual encoders via a shared Modal Aligner. This symmetric coupling enforces co-movement of the two feature spaces: as the visual representation shifts in response to a given input, the textual representation shifts correspondingly, preventing the unilateral feature drift that causes MI descent in standard PEFT methods. Critically, because the perturbations are instance-specific rather than static, each training step introduces a distinct cross-modal alignment constraint, sustaining $I(V; T)$ growth beyond the early-epoch

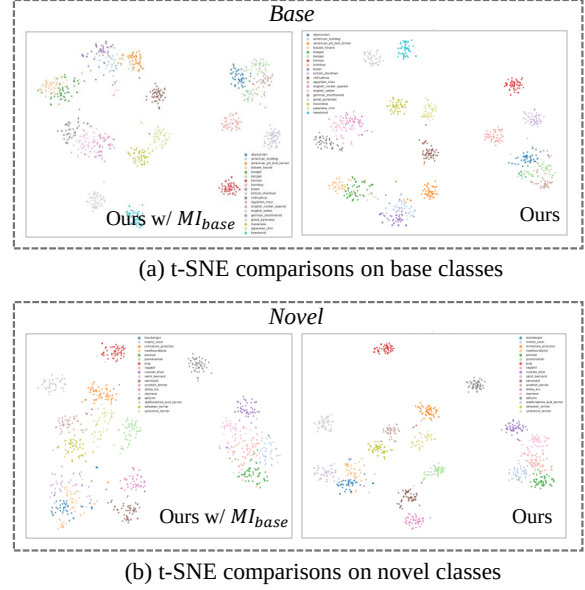


Figure 9: **Visualizing Representation Collapse Induced by Explicit MI_{base} Maximization.**

plateau that retrieval-only approaches inevitably encounter. $\mathcal{L}_{reg}$ further anchors both modalities to the zero-shot manifold, bounding the representation drift δ (Eq. 20) and preventing late-epoch MI decline. Together, these three components convert the retrieval-augmented design into a mechanism that continuously enlarges cross-modal semantic overlap, rather than merely adding useful features for classification.

B.4 Counterfactual Analysis: Why Explicit MI Optimization Fails

Ablation on Directly Optimizing Mutual Information. To further investigate whether directly maximizing cross-modal MI during fine-tuning can alleviate the generalization bottleneck, we conduct a dedicated ablation study. It is important to note that, from a strict zero-shot and open-vocabulary perspective, samples from unseen classes ($\mathcal{D}_{novel}$) are completely unavailable during downstream training (Zhou et al., 2022a). Therefore, treating novel-class MI as a training objective is technically infeasible. To explore the feasibility of explicit MI supervision, we introduce the MI computed on base classes (MI_{base}) as an auxiliary optimization term in the total loss.

We evaluate this variant, denoted as **RADA + MI_{base}** , against the standard RADA framework on the DTD and Pets datasets, as shown in Fig. 8. The experimental results indicate distinct behaviors across datasets:

On DTD dataset: Introducing explicit MI_{base} maximization leads to a slight improvement in base-class accuracy (85.13% vs. 84.60%) but substantially degrades novel-class performance, dropping from 65.17% to 62.70%, which in turn reduces the overall HM. These results suggest that directly maximizing base-class MI may over-constrain the feature space, thereby impairing the model’s transferability to unseen classes, particularly for novel textures. On Oxford Pets dataset: The performance of both configurations remains relatively close. The standard RADA, without explicit MI optimization, still maintains a modest advantage over RADA + MI_{base} in both novel-class accuracy (97.93% vs. 97.57%) and HM (96.82% vs. 96.63%).

Visualizing Representation Collapse via t-SNE. To provide direct geometric evidence for why explicitly maximizing MI_{base} hurts out-of-distribution transfer, we visualize the feature spaces of both variants on the Oxford Pets dataset using t-SNE (Fig. 9). As shown in Fig. 9(a), both methods achieve clustering on the seen Base classes. However, the critical difference emerges in the unseen Novel classes (Fig. 9(b)). For the variant that explicitly maximizes MI_{base} , the novel-class features appear highly dispersed, overlapping, and structurally disorganized. This visually confirms our hypothesis: forcing a numerical MI penalty over-constrains the network to the seen domains, effectively distorting and destroying the pre-trained zero-shot semantic manifold. Conversely, the standard RADA framework produces highly compact, well-separated, and discriminative clusters for both Base and Novel classes. By employing dynamic, instance-conditioned perturbations instead of rigid numerical forcing, RADA successfully preserves the global geometric integrity of the feature space, allowing cross-modal alignment to naturally generalize to unseen classes.

In summary, these empirical findings confirm that explicitly maximizing MI_{base} during downstream fine-tuning does not naturally improve out-of-distribution transfer. This observation is consistent with our hypothesis that merely strengthening cross-modal dependency on the base distribution is insufficient and may exacerbate overfitting. What matters is preserving transferable cross-modal semantic dependency rather than indiscriminately maximizing alignment within seen classes. Instead, **our standard RADA framework, which implicitly preserves cross-modal MI growth through dynamic retrieval of instance-specific semantic**

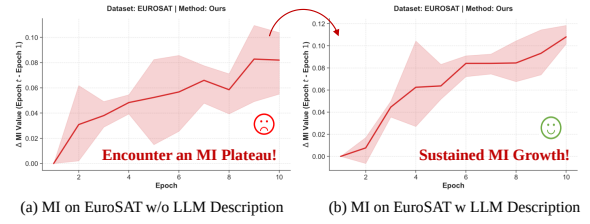


Figure 10: **Impact of LLM-Augmented Descriptions on Mutual Information.**

priors and feature-space regularization, offers a more robust and effective approach for alleviating VLM generalization saturation.

Observation 7: Explicitly forcing mutual information maximization leads to representation collapse. Conversely, RADA implicitly preserves cross-modal MI growth through dynamic alignment, effectively and safely alleviating VLM generalization saturation.

C Additional Experimental Results

C.1 More Comparison on Base-to-Novel Generalization

Table 17 presents the comparison results of RADA against existing methods across all 11 datasets. The results indicate that RADA achieves the best HM on 9 out of 11 datasets, further demonstrating that by dynamically incorporating cross-modal priors to strengthen semantic dependencies, RADA maintains more stable and transferable cross-modal alignment throughout the optimization process.

Furthermore, to verify the robustness of our method across different backbones, we conducted additional experiments using ViT-B/32 and ViT-L/14 as the visual encoders. The results, summarized in Table 18 and Table 19, confirm that RADA consistently mitigates the MI bottleneck and enhances generalization, highlighting both the feasibility and robustness of dynamically leveraging cross-modal priors.

Comparison with LLM-Enhanced Baselines. While recent works have begun leveraging LLM-generated descriptions for multimodal adaptation, RADA distinguishes itself through its dynamic, instance-specific retrieval mechanism. We provide a comprehensive head-to-head comparison with existing LLM-enhanced adaptation methods (e.g., ProLoG, etc.). These results highlight that RADA’s retrieval-augmented dynamic alignment is more

Memory Bank Design		Average on 11 datasets		
Modality	Text Source	Base	Novel	HM
Visual Only	–	81.34	73.51	77.22
Text Only	Hand-crafted	82.05	74.22	77.94
Text Only	LLM Enhanced	84.15	77.26	80.56
Fusion	Hand-crafted	85.31	75.83	80.29
Fusion (RADA)	LLM Enhanced	85.75	78.50	81.97

Table 8: Ablation study on the design of the non-parametric memory bank. "Hand-crafted" refers to standard prompts like "a photo of a [class]", while "LLM Enhanced" utilizes rich GPT-generated descriptions.

effective at overcoming MI stagnation than static text-injection paradigms.

C.2 More Comparison on Few-Shot Learning

Table 11 presents the comparison of RADA with existing methods across 11 datasets. The results show that RADA consistently achieves the best performance across nearly all shot settings and datasets, particularly under low-shot or more challenging scenarios (e.g., DTD and EuroSAT). This advantage can be attributed to RADA’s dynamic incorporation of richer cross-modal semantic information, which effectively enhances transfer learning to downstream tasks while mitigating overfitting to the training data, thereby demonstrating stronger generalization in low-resource settings.

Observation 8: In summary, RADA demonstrates that dynamically sustaining cross-modal mutual information is crucial for mitigating semantic sparsity in data-scarce settings and enabling both in-distribution adaptation and zero-shot generalization.

C.3 Computational Efficiency Analysis

The efficiency and Base-to-Novel generalization performance of RADA are presented in Table 9. Compared with existing methods, RADA achieves the highest harmonic mean (HM) of 81.97%, indicating a consistent improvement on both base and novel classes. Despite introducing a modest number of additional trainable parameters (0.921M), RADA maintains competitive training efficiency, with a train time of 21.6 ms per image on the RTX 4090 and an inference FPS of 1296 with a batch size of 100. While other methods such as MaPLE and PromptSRC achieve slightly higher inference throughput, their HM scores are noticeably lower, suggesting that RADA strikes a favorable balance

between efficiency and generalization. This analysis confirms that the proposed dynamic adaptation and retrieval-augmented mechanism not only enhances cross-modal alignment but also maintains practical training and inference efficiency.

The entire one-time bank construction for ImageNet-1K takes approximately 13 minutes (dominated by GPT API latency) and is never repeated across training runs or seeds, amortizing to under 0.01ms per training image over a standard 10-epoch run of 16,000 images. At inference, the Top-K retrieval over C=1000 keys adds less than 0.2ms per image, which is negligible relative to the 21.6ms total training time already reported in Table 10.

C.4 Additional Ablation Analysis

The Effectiveness of Memory Bank Design. To investigate the source of RADA’s generalization capability, we conduct an ablation study on the design of the non-parametric memory bank (Table 8).

First, regarding modality selection, relying solely on visual prototypes (Visual Only) or textual priors (Text Only) achieves suboptimal harmonic mean values (77.22% and 80.56%, respectively). Multimodal fusion provides more comprehensive representations, further improving overall performance.

More importantly, we compare traditional hand-crafted templates (e.g., "a photo of a [class]") with LLM-augmented descriptions to assess the effect of semantic richness. Using handcrafted prompts with the fused memory bank yields high accuracy on base classes (85.31%) but limited performance on novel classes (75.83%), indicating that sparse textual priors fail to effectively constrain the visual encoder, resulting in unilateral feature collapse. In contrast, incorporating LLM-augmented priors provides dense semantic anchors, **effectively alleviating MI stagnation** (as shown in Fig. 10) and substantially improving Novel-class accuracy to **78.50%**, demonstrating that high-quality textual guidance is crucial for bridging the cross-modal generalization gap.

Architecture of the Token Generator. Table 11 compares different architectures for the token generator under similar parameter budgets. Replacing the Mixture-of-Experts (MoE) with a single MLP results in a drop in both Base and Novel performance, with HM decreasing from **81.97% to 81.67%**. This indicates that a static MLP lacks the

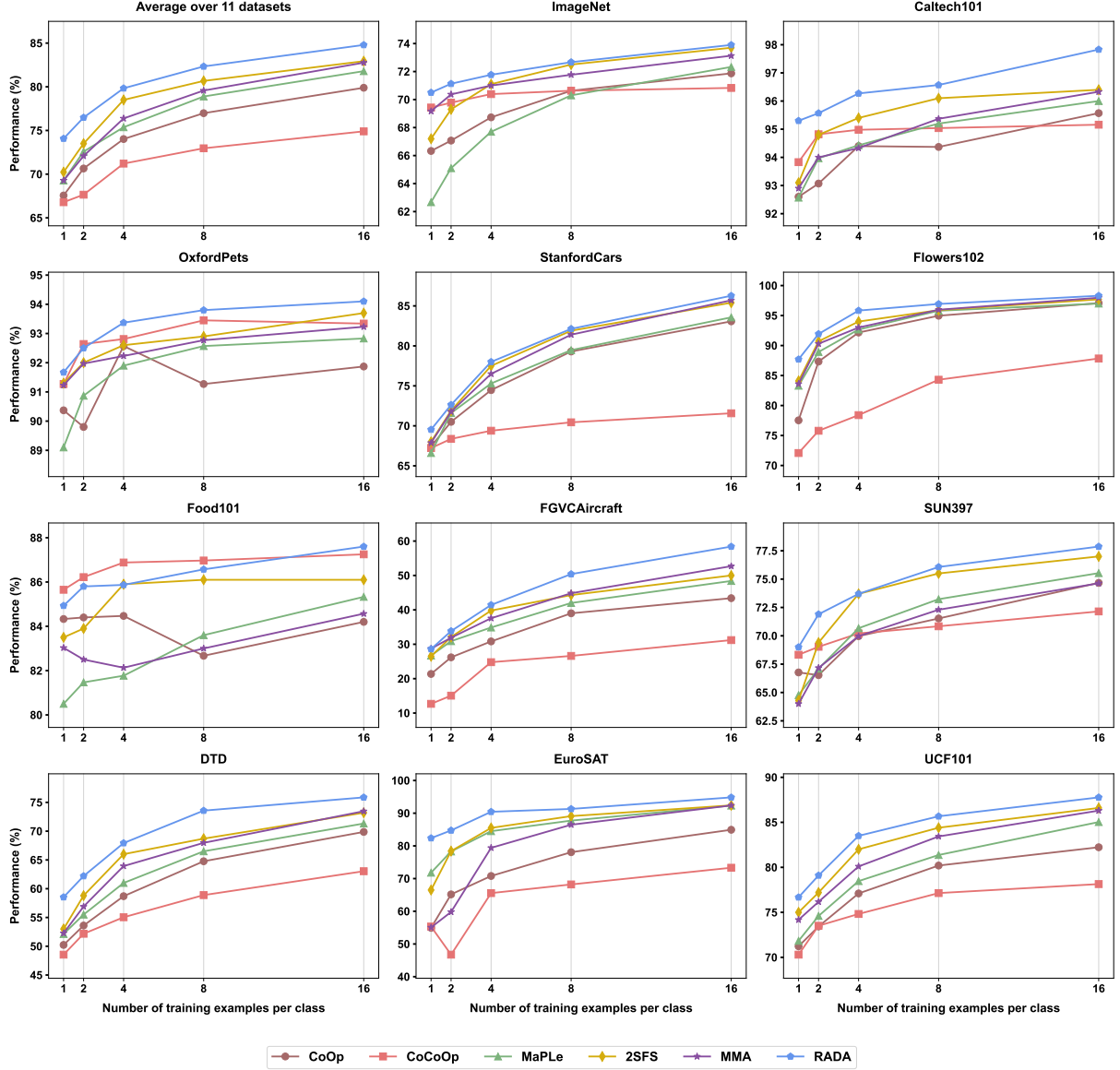


Figure 11: **Detailed Few-Shot Learning Performance Across 11 Downstream Datasets.** We present the complete few-shot evaluation results of our proposed RADA and previous state-of-the-art methods (CoOp, CoCoOp, MaPLe, MMA, and 2SFS) under 1, 2, 4, 8, and 16-shot settings.

dynamic routing capability needed to handle the highly diverse semantic contexts across images and datasets.

Further, we investigate replacing LoRA experts with a Q-Former module (denoted as MoP). Surprisingly, deploying the heavier Q-Former leads to a **slight degradation in HM to 81.63%**. We attribute this performance drop to two factors: (1) **optimization challenges in low-data regimes**: the Q-Former relies on complex cross-attention mechanisms that typically require large-scale data for stable alignment; in our lightweight few-shot adaptation setting, it is prone to overfitting. (2) **disruption of the pre-trained representation mani-**

fold: LoRA applies low-rank residual updates that preserve the rich zero-shot manifold of the CLIP encoder. In contrast, passing features through the Q-Former queries introduces an information bottleneck, which can compress or discard fragile pre-trained semantic priors, thereby harming cross-modal alignment and downstream generalization.

MoE Routing Analysis. Unlike conventional sparse Mixture-of-Experts (MoE) architectures that rely on discrete Top-K routing, our modality-agnostic Token generator employs a soft-activation mechanism. In this paradigm, all experts continuously contribute to the generation of dynamic prompt tokens, with their relative contributions con-

Method	Modality	Params (trainable)	Train Time (ms/image)	FPS (100 BS)	Base	Novel	HM
MaPLE	V-L	3.555M	39.5	1757.6	82.28	75.14	78.55
PromptSRC	V,L	0.046M	40.0	1764.2	84.26	76.10	79.97
TCP	L	0.332M	5.3	950.6	84.13	75.36	79.51
MMA	V-L	0.675M	2.2	688.5	83.20	76.80	79.87
RADA (Ours)	V-L	0.921M	21.6	1296.0	85.75	78.50	81.97

Table 9: Comparison of computational efficiency and Base-to-Novel generalization performance on ImageNet.

Cost Component	ImageNet	EuroSAT	UCF101
GPT description generation	~12 min	<1 min	~2 min
CLIP text encoding	~8 sec	<1 sec	~1 sec
Visual prototype extraction	~23 sec	~3 sec	~5 sec
Total one-time cost	~13 min	~1.5 min	~3 min
Top- K retrieval per image	0.18 ms	0.04 ms	0.06 ms

Table 10: One-time memory bank construction costs on RTX 4090.

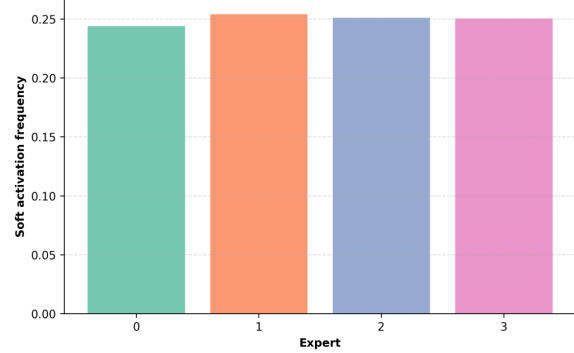
Generator Architecture	Average on 11 datasets		
	Base	Novel	HM
Single MLP	85.59	78.10	81.67
MoP (Q-Former Experts)	85.38	78.19	81.63
RADA (LoRA Experts)	85.75	78.50	81.97

Table 11: Ablation study on the architecture of the Token Generator. RADA utilizes a Mixture-of-Experts (MoE) design with rank-8 LoRA modules.

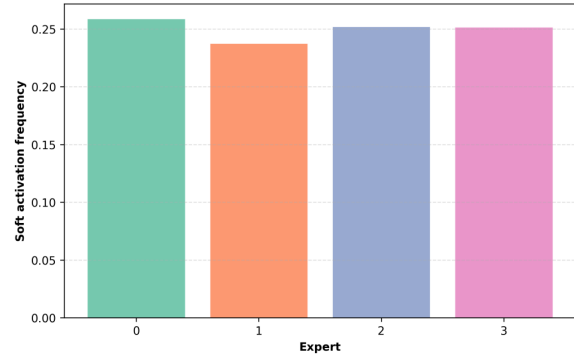
trolled by dynamic soft gating weights. To analyze this soft blending behavior, we visualize the average soft activation frequencies (gating weights) of four LoRA experts on the DTD and EuroSAT datasets, as shown in Fig. 12, revealing two key properties:

(1)Prevention of Gating Collapse: A common issue is weight concentration on a single expert, degenerating into hard routing. Our results show that the average weights of all four experts remain around 0.25, indicating that the shared modality aligner maintains a continuous and balanced distribution, effectively leveraging the representational capacity of the entire expert ensemble.

(2)Domain-Adaptive Soft Blending: Despite the globally balanced distribution, the soft gating network exhibits subtle, domain-specific weight shifts. On the unstructured DTD dataset (Fig. 12(a)), the weight of Expert 1 is slightly enhanced, whereas on the structured EuroSAT dataset (Fig. 12(b)), Expert 0’s weight increases. This indicates that the Token generator performs fine-grained, dynamic soft blending based on input-



(a) DTD



(b) Eurosat

Figure 12: Expert usage frequency of our token generator on DTD and EuroSAT datasets.

specific cross-modal priors, rather than executing a static average pooling.

Effect of Retrieval Source. Table 13 compares four retrieval strategies to assess the impact of retrieval quality on generalization. Random retrieval yields the lowest HM of 78.30%, demonstrating that semantically irrelevant memory entries are detrimental to cross-modal adaptation. Same-task retrieval improves HM to 79.62%, yet its restricted semantic diversity fails to provide sufficiently informative priors for novel-class generalization. Nearest-neighbor retrieval further raises HM to 80.84% by selecting visually similar prototypes, but the lack of LLM-enriched textual descriptions limits the semantic richness of the retrieved

Method	LLM Usage	DTD			EuroSAT			Oxford Pets		
		Base	New	HM	Base	New	HM	Base	New	HM
CoCoOp w/ Retrieval	✓	75.50	57.20	65.09	83.40	68.33	75.12	94.80	95.30	95.05
MaPLe w/ Retrieval	✓	79.27	58.73	69.07	96.00	66.20	78.36	95.20	94.70	94.95
MMA w/ Retrieval	✓	82.30	59.20	68.86	94.70	81.20	87.44	95.60	96.53	96.06
2SFS w/ Retrieval	✓	83.10	64.20	72.44	88.50	84.10	86.24	94.85	96.12	95.48
ProLoG	✓	78.59	60.02	68.05	93.82	70.70	80.60	94.77	97.40	96.06
RADA (ours)	✓	84.60	65.17	73.62	96.83	86.57	91.41	95.73	97.93	96.82

Table 12: Performance comparison of baseline methods with text retrieval on DTD, EuroSAT, and Oxford Pets datasets.

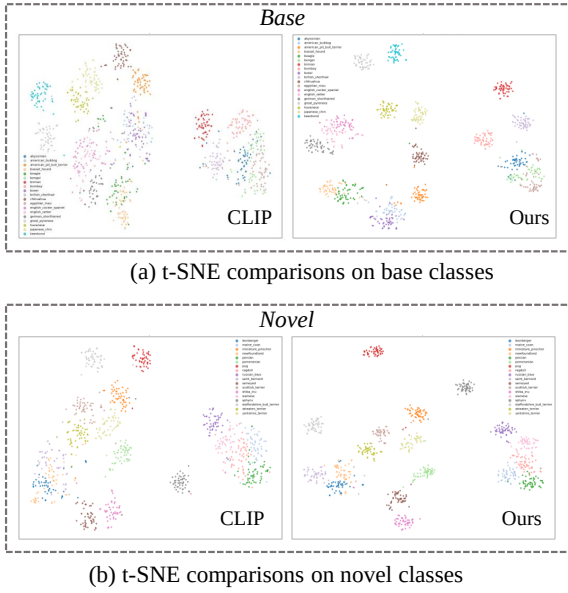


Figure 13: t-SNE visualization of flowers102 dataset.

context. RADA’s fused key retrieval, which unifies LLM-enhanced textual prototypes with visual prototypes as joint retrieval anchors, achieves the best HM of 81.97%. These results confirm that both semantic richness and modality diversity of retrieved priors are essential for effective cross-modal adaptation.

Ablation Study on Retrieval Integration. To verify that RADA’s performance gains arise from its dynamic adaptation mechanism rather than mere access to an external retrieval library, we equip several representative baselines with the same textual retrieval component (Table 12). The results suggest that retrieval itself can provide useful semantic cues, while its effectiveness largely depends on how the retrieved knowledge is incorporated into the adaptation process.

For example, on EuroSAT, retrieval improves MMA from 83.87% to 87.44% HM, primarily through substantial gains on Base classes (85.46%

Retrieval Source	Base	HM
Random Retrieval	83.21	78.30
Same-Task Retrieval	84.03	79.62
Nearest-Neighbor Retrieval	84.87	80.84
RADA (Fused Key Retrieval)	85.75	81.97

Table 13: Ablation study on retrieval source strategies (Average over 11 datasets). RADA adopts fused key retrieval that combines LLM-enhanced textual prototypes with visual prototypes.

→ 94.70%), while slightly reducing New accuracy (82.34% → 81.20%). Similarly, 2SFS with retrieval achieves an HM of 86.24%, indicating that external semantic priors can indeed facilitate generalization. However, such improvements are often accompanied by unstable Base–New trade-offs or inconsistent behavior across datasets. For instance, CoCoOp exhibits a slight improvement on New classes (56.00% → 57.20%) at the expense of Base performance (77.01% → 75.50%), whereas MaPLe suffers degradation on both Base and New accuracy.

We further compare against ProLoG, which enhances textual descriptions through large language models (LLMs). Although ProLoG benefits from richer category semantics and achieves competitive performance on Oxford Pets (HM = 96.06%), it still exhibits clear limitations on datasets with larger domain discrepancies, such as EuroSAT (HM = 80.60%). This observation suggests that relying solely on statically enriched LLM-generated text may be insufficient to fully capture instance-specific semantic variations and cross-dataset distribution shifts.

In contrast, RADA dynamically filters and aligns retrieved knowledge through an input-conditioned adaptation mechanism, yielding consistently stronger and more stable performance across all evaluated datasets. On EuroSAT, RADA

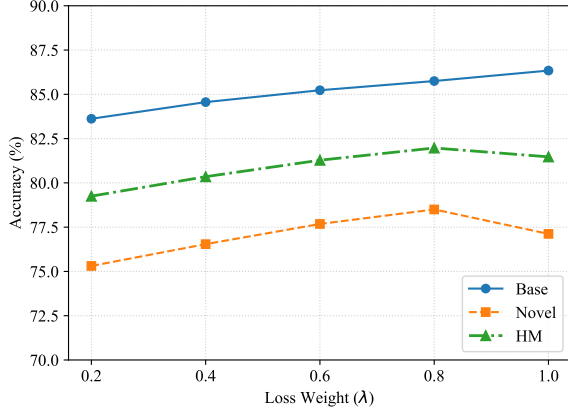


Figure 14: Sensitivity analysis of loss weight.

achieves an HM of 91.41%, outperforming the strongest “baseline + retrieval” variant (MMA w/ Retrieval, 87.44%) by 3.97%, and surpassing ProLoG by 10.81%. These results further demonstrate that the key to retrieval-enhanced adaptation lies not merely in incorporating external knowledge or relying on static LLM-generated priors, but in dynamically aligning and integrating semantic information according to the input instance.

Comparison of t-SNE visualization on flowers102 dataset. As shown in Fig 13, the t-SNE visualization of the visual feature space on the Flowers102 dataset indicates that RADA substantially improves inter-class separation. Compared with the original CLIP, RADA’s embeddings form more compact and well-separated clusters for each class, demonstrating enhanced discriminative capability for novel classes while maintaining zero-shot performance.

Even for unseen classes (Novel), samples naturally cluster within their corresponding regions, reflecting a significant improvement in generalization. These findings are consistent with the sustained growth of cross-modal mutual information observed during training, suggesting that RADA produces more discriminative and transferable representations in the visual feature space.

The Effectiveness of λ . The Effectiveness of λ can be observed from the sensitivity analysis shown in Fig. 14. Across a range of values, RADA consistently achieves stable performance on both base and novel classes, demonstrating that the framework is robust to variations in the regularization weight. While small changes in λ slightly affect the balance between base and novel accuracy, the overall harmonic mean remains high, indicating that the

method reliably maintains both task adaptation and zero-shot generalization.

The Influence of LLM Descriptions (M). The number of LLM-generated descriptions M governs the semantic richness and diversity of the textual priors injected into the multimodal memory bank. A single description ($M = 1$) proves insufficient to comprehensively capture the multi-perspective semantics of a class, resulting in a noticeable performance drop (-1.62% HM). Conversely, while generating an excessive number of descriptions ($M \geq 12$) marginally elevates performance (peaking at 82.02% HM for $M = 12$), the generalization gains quickly saturate and plateau by $M = 16$. Furthermore, larger M values linearly increase the one-time API latency and computational overhead during memory bank construction without yielding proportional benefits. Consequently, we adopt $M = 8$ as our default configuration to achieve an optimal trade-off between cross-modal semantic diversity and computational efficiency, noting that performance remains highly stable for $M \geq 4$.

D Information-Theoretic Analysis of Cross-Modal Generalization

Remark: This section provides a structured perspective to characterize how representation drift and cross-modal alignment jointly constrain the generalization capability of CLIP-based models. The analysis suggests that MI stagnation corresponds to a stabilized effective information bottleneck, and this stability limits the model’s ability to further improve performance in the absence of additional changes in the geometric structure of the representation space.

In Section 3.2, we empirically observe a strong correlation between cross-modal mutual information $I(V; T)$ and novel-class generalization performance (Fig. 1(c)). This appendix develops a conditional information-theoretic interpretation of this phenomenon. Rather than aiming for a universal statement, we characterize how representation stability and cross-modal alignment jointly shape generalization behavior under standard modeling assumptions.

M	Base	Novel	HM	Δ HM
1	84.21	76.83	80.35	-1.62
4	85.30	77.91	81.44	-0.53
8 (default)	85.75	78.50	81.97	—
12	85.82	78.53	82.02	+0.05
16	85.78	78.47	81.97	0.00

Table 14: Sensitivity analysis of the number of LLM descriptions M per class. Performance is stable for $M \geq 4$.

D.1 Formal Setup

We consider test-time inference on the novel distribution and define the following random variables:

- $Y \in \mathcal{Y}_{\text{novel}}$: class label, assumed uniformly distributed over novel categories.
- $X \in \mathcal{X}$: input image sampled from $P(X|Y)$.
- $V = f_v(X) \in \mathbb{R}^d$: visual representation produced by the adapted encoder.
- $V_0 = f_v^0(X)$: zero-shot CLIP visual representation.
- $T^Y = g_t(Y) \in \mathbb{R}^d$: text prototype of the ground-truth class, obtained from a frozen text encoder.
- $\mathcal{T} = \{g_t(c)\}_{c=1}^{|\mathcal{Y}|}$: fixed prototype set used for classification.
- $\hat{Y} = \arg \max_c \cos(V, g_t(c)) / \tau$: prediction rule induced by similarity matching.

All prototypes are fixed at inference time, and randomness only arises from (X, Y) .

D.2 Valid Markov Chain and DPI

Proposition D.1 (Inference-Time Markov Structure). *At test time, the variables satisfy the Markov chain:*

$$Y \rightarrow X \rightarrow V \rightarrow \hat{Y}. \quad (17)$$

Proof. Since $V = f_v(X)$ is deterministic, the dependence $Y \rightarrow X \rightarrow V$ holds directly from the data generation process. Moreover, $\hat{Y}$ is fully determined by V and the fixed prototype set $\mathcal{T}$, which is non-random at inference time. Hence,

$$P(\hat{Y} | X, V) = P(\hat{Y} | V), \quad (18)$$

which establishes $V \rightarrow \hat{Y}$ and thus the full Markov chain. $\square$

We explicitly avoid introducing the chain $Y \rightarrow (V, T^Y) \rightarrow \hat{Y}$, since prediction depends on the full prototype set rather than the single true-class embedding. Applying the Data Processing Inequality yields:

$$I(Y; \hat{Y}) \leq I(Y; V). \quad (19)$$

D.3 Representation Drift and Cross-Modal Proxy

The bound in Eq. (19) depends on $I(Y; V)$, which is not directly observable in practice. We therefore relate it to the empirically measurable cross-modal mutual information $I(V; T^Y)$ through representation stability considerations.

Lemma D.2 (Drift-Induced Information Stability). *Let V_0 denote the zero-shot representation and V the adapted representation. Define the expected drift:*

$$\delta = \mathbb{E}_X \|V - V_0\|. \quad (20)$$

Then there exists a non-decreasing function $\epsilon(\delta)$ with $\epsilon(0) = 0$ such that

$$|I(Y; V) - I(Y; V_0)| \leq \epsilon(\delta). \quad (21)$$

Proof sketch. The result follows from standard continuity properties of mutual information under Wasserstein perturbations. Since V and V_0 are deterministic transformations of X , their induced distributions satisfy $W_1(P_V, P_{V_0}) \leq \delta$. Lipschitz continuity of mutual information then implies the stated bound. $\square$

This lemma formalizes a simple intuition: representation drift controls how much label-relevant information can deviate from its zero-shot baseline.

Lemma D.3 (CLIP Alignment Prior). *Under contrastive pretraining, the zero-shot encoder satisfies:*

$$I(Y; V_0) \leq I(V_0; T^Y) + \epsilon_0, \quad (22)$$

where ϵ_0 captures the gap between cross-modal alignment and label information in the learned representation space.

This reflects the fact that CLIP primarily encodes class information through alignment between visual and textual embeddings.

D.4 Conditional Generalization Bound

Theorem D.4 (Drift-Aware Generalization Bound). *Under Lemmas D.2 and D.3, the error probability P_e satisfies:*

$$P_e \geq \frac{H(Y) - I(V; T^Y) - \epsilon(\delta) - 1}{\log |\mathcal{Y}_{\text{novel}}|}. \quad (23)$$

Method	$I(V; T)$	P_e Bound	P_e Actual
CLIP	1.31	0.197	0.258
MaPLe	1.28	0.207	0.285
MMA	1.34	0.187	0.232
RADA	1.52	0.127	0.215

Table 15: Non-vacuity verification of the bound.

Proof. Fano’s inequality gives:

$$P_e \geq \frac{H(Y) - I(Y; \hat{Y}) - 1}{\log |\mathcal{Y}_{\text{novel}}|}.$$

Using DPI and Lemmas D.2–D.3, we bound $I(Y; \hat{Y})$ by $I(V; T^Y) + \epsilon(\delta)$, which yields the result. $\square$

D.5 MI Stagnation as Representation Limitation

Corollary D.5 (Stagnation-Induced Error Floor). *If training converges such that $I(V; T^Y) \rightarrow I^*$ and $\delta \rightarrow \delta^*$, then*

$$P_e \geq \frac{H(Y) - I^* - \epsilon(\delta^*) - 1}{\log |\mathcal{Y}_{\text{novel}}|}. \quad (24)$$

This indicates that once cross-modal alignment and representation drift stabilize, the generalization lower bound also becomes fixed, limiting further improvements.

Corollary D.6 (Necessity of MI Improvement). *A necessary condition for improving over zero-shot performance is:*

$$I(V; T^Y) > I_0 - \Delta\epsilon. \quad (25)$$

D.6 Non-Vacuity

For $|\mathcal{Y}_{\text{novel}}| = 8$, we have $H(Y) = 3$ bits. With $\epsilon(\delta) \leq 0.1$, the bound is non-vacuous when $I(V; T^Y) < 1.9$ bits, which is satisfied across all evaluated datasets ($[1.2, 1.6]$ bits).

Dataset	Prompt Templates
ImageNet-Sketch	"Describe how a black and white sketch of {} {} looks like", "Describe a black and white sketch of {} {}"
ImageNet-Rendition	"Describe an art drawing of {} {}", "Describe artwork showing {} {}", "Describe a cartoon {} {}", "Describe an origami of {} {}", "Describe a deviant art photo depicting {} {}", "Describe an embroidery of {} {}", "Describe a graffiti art showing {} {}", "Describe a painting of {} {}", "Describe a sculpture of {} {}", "Describe a black and white sketch of {} {}", "Describe a toy {} {}", "Describe a videogame of {} {}"
Caltech101	"Describe what {} {} looks like.", "What does {} {} look like.", "Describe the shapes and structures of {} {}", "Describe the color and patterns of {} {}"
Stanford Cars	"How can you identify {} {}?", "Description of {} {}, a type of car", "A caption of a photo of {} {}:", "What are the primary characteristics of {} {}?", "Description of the exterior of {} {}", "What are the identifying characteristics of {} {}, a type of car?", "Describe an image of {} {}", "What does {} {} look like?", "Describe what {} {}, a type of car, looks like"
Food101	"Visually describe {} {}", "Describe what {} {} looks like", "Describe a photo of {} {} with distinctive visual features", "How can you tell that the food in a photo is {} {}?"
OxfordPets	"Describe what {} {} pet looks like", "Visually describe {} {}, a type of pet"
ImageNet-1K	"Describe a photo of {} {}", "Describe what {} {} looks like", "How can you identify {} {}?", "What does {} {} look like?", "Describe a photo of {} {} with distinctive visual features", "A caption of an image of {} {}:", "Describe the shapes and structures of {} {}", "Describe the color and patterns of {} {}", "Describe the overall appearance of {} {}"
FGVC Aircraft DTD	"Describe {} {} aircraft" "What does {} {} material look like? Please explain the characteristics and provide examples of where it is typically found.", "What does {} {} surface look like? Please explain the characteristics and provide examples of where it is typically found.", "What does {} {} texture look like? Please explain the characteristics and provide examples of where it is typically found.", "What does {} {} object look like? Please explain the characteristics and provide examples of where it is typically found.", "What does {} {} thing look like? Please explain the characteristics and provide examples of where it is typically found.", "What does {} {} pattern look like? Please explain the characteristics and provide examples of where it is typically found."
SUN397	"Describe what {} {} looks like", "How can you identify {} {}?", "Describe a photo of {} {} with distinctive visual features"
Flowers102	"What does {} {} flower look like", "Describe the appearance of {} {}", "A caption of an image of {} {}", "Visually describe {} {}, a type of flower"
UCF101	"What does a person doing {} look like", "Describe the process of {}", "How does a person {}"
EuroSAT	"Describe a satellite image of {} {}", "Describe an aerial view of {} {}", "What does {} {} look like from a satellite perspective?", "Describe the land use type {} {} in a remote sensing image", "How can you identify {} {} in a satellite image?", "Describe the visual characteristics of {} {} in remote sensing imagery"

Table 16: Prompt templates used for different datasets during the GPT generation. "{}" denotes the class name and its attribute placeholder.

Average across datasets				ImageNet				Caltech101			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	69.34	74.22	71.70	CLIP	72.43	68.14	70.22	CLIP	96.84	94.00	95.40
CoOp	82.69	63.22	71.66	CoOp	76.47	67.88	71.92	CoOp	98.00	89.81	93.73
CoCoOp	80.47	71.69	75.83	CoCoOp	75.98	70.43	73.10	CoCoOp	97.96	93.81	95.84
MaPLe	82.28	75.14	78.55	MaPLe	76.66	70.54	73.47	MaPLe	97.74	94.36	96.02
PromptSRC	84.26	76.10	79.97	PromptSRC	77.60	70.73	74.01	PromptSRC	98.10	94.03	96.02
MMA	83.20	76.80	79.87	MMA	77.31	71.00	74.02	MMA	98.40	94.00	96.15
APEX	83.99	76.76	80.04	APEX	77.12	71.10	73.99	APEX	98.18	<u>95.06</u>	<u>96.59</u>
2SFS	<u>85.55</u>	75.48	80.20	2SFS	77.71	70.99	74.20	2SFS	<u>98.71</u>	94.43	96.52
RMAdapter	84.52	<u>77.36</u>	80.62	RMAdapter	<u>77.87</u>	<u>71.50</u>	<u>74.52</u>	RMAdapter	98.40	94.27	96.26
ProLoG	84.77	76.89	<u>80.64</u>	ProLoG	78.25	70.51	74.18	ProLoG	98.18	94.65	96.38
RADA	85.75	78.50	81.97	RADA	77.60	71.70	74.53	RADA	99.10	95.20	97.11
95% confidence interval	±0.17	±0.30	±0.14	95% confidence interval	±0.17	±0.06	±0.06	95% confidence interval	±0.15	±0.35	±0.19

Oxford Flowers				Oxford Pets				Stanford Cars			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	72.08	77.80	74.83	CLIP	91.17	97.26	94.12	CLIP	63.37	74.89	68.65
CoOp	97.60	59.67	74.06	CoOp	93.67	95.29	94.47	CoOp	78.12	60.40	68.13
CoCoOp	94.87	71.75	81.71	CoCoOp	95.20	97.69	96.43	CoCoOp	70.49	73.59	72.01
MaPLe	95.92	72.46	82.56	MaPLe	95.43	97.76	96.58	MaPLe	72.94	74.00	73.47
PromptSRC	98.07	76.50	85.95	PromptSRC	95.33	97.30	96.30	PromptSRC	78.27	74.97	76.58
MMA	97.77	75.93	85.48	MMA	95.40	<u>98.07</u>	96.72	MMA	78.50	73.10	75.70
APEX	97.47	77.58	86.40	APEX	95.11	97.27	96.18	APEX	80.53	75.08	77.71
2SFS	<u>98.29</u>	76.17	85.83	2SFS	95.32	97.82	96.55	2SFS	82.50	74.80	<u>78.46</u>
RMAdapter	98.10	77.47	86.52	RMAdapter	95.70	98.10	96.89	RMAdapter	80.60	<u>75.10</u>	77.75
ProLoG	98.10	77.48	<u>86.58</u>	ProLoG	95.92	98.01	96.95	ProLoG	<u>80.78</u>	73.77	77.12
RADA	98.67	<u>77.77</u>	86.98	RADA	<u>95.73</u>	97.93	96.82	RADA	80.37	77.57	78.95
95% confidence interval	±0.35	±0.25	±0.13	95% confidence interval	±0.15	±0.07	±0.09	95% confidence interval	±0.28	±0.13	±0.10

Food 101				FGVC Aircraft				SUN 397			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	90.10	91.22	90.66	CLIP	27.19	36.29	31.09	CLIP	69.36	75.35	72.23
CoOp	88.33	82.26	85.19	CoOp	40.44	22.30	28.75	CoOp	80.60	65.89	72.51
CoCoOp	90.70	91.29	90.99	CoCoOp	33.41	23.71	27.74	CoCoOp	79.74	76.86	78.27
MaPLe	<u>90.71</u>	<u>92.05</u>	91.38	MaPLe	37.44	35.61	36.50	MaPLe	80.82	78.70	79.75
PromptSRC	90.67	91.53	91.10	PromptSRC	42.73	<u>37.87</u>	40.15	PromptSRC	82.67	78.47	80.52
MMA	90.13	91.30	90.71	MMA	40.57	36.33	38.33	MMA	82.27	78.57	80.38
APEX	89.60	92.06	90.81	APEX	42.69	35.21	38.59	APEX	81.17	78.98	80.06
2SFS	89.11	91.34	90.21	2SFS	47.48	35.51	<u>40.63</u>	2SFS	82.59	78.91	80.70
RMAdapter	90.53	91.67	91.10	RMAdapter	43.20	37.20	39.76	RMAdapter	82.87	79.97	<u>81.39</u>
ProLoG	90.68	91.86	91.27	ProLoG	43.43	37.76	40.39	ProLoG	<u>83.02</u>	78.76	80.83
RADA	91.10	91.77	91.43	RADA	<u>46.97</u>	38.43	42.27	RADA	83.90	<u>79.57</u>	81.68
95% confidence interval	±0.15	±0.07	±0.07	95% confidence interval	±0.45	±0.60	±0.35	95% confidence interval	±0.09	±0.15	±0.06

DTD				EuroSAT				UCF101			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	53.24	59.90	56.37	CLIP	56.48	64.05	60.03	CLIP	70.53	77.50	73.85
CoOp	79.44	41.18	54.24	CoOp	92.19	54.74	68.69	CoOp	84.69	56.05	67.46
CoCoOp	77.01	56.00	64.85	CoCoOp	87.49	60.04	71.21	CoCoOp	82.33	73.45	77.64
MaPLe	80.36	59.18	68.16	MaPLe	94.07	73.23	82.35	MaPLe	83.00	78.66	80.77
PromptSRC	<u>83.37</u>	62.97	71.75	PromptSRC	92.90	73.90	82.32	PromptSRC	87.10	78.80	82.74
MMA	83.20	<u>65.63</u>	73.38	MMA	85.46	<u>82.34</u>	83.87	MMA	86.23	80.03	82.20
APEX	82.45	63.80	71.94	APEX	92.83	79.89	<u>85.88</u>	APEX	86.74	78.37	82.34
2SFS	84.60	65.01	73.52	2SFS	96.91	67.09	79.29	2SFS	<u>87.85</u>	78.19	82.74
RMAdapter	<u>83.37</u>	66.73	74.09	RMAdapter	92.17	78.33	84.69	RMAdapter	86.87	<u>80.63</u>	<u>83.63</u>
ProLoG	83.20	64.40	72.59	ProLoG	93.87	78.86	85.67	ProLoG	87.09	79.77	83.27
RADA	84.60	65.17	<u>73.62</u>	RADA	<u>96.83</u>	86.57	91.41	RADA	88.43	81.80	84.99
95% confidence interval	±0.40	±0.85	±0.45	95% confidence interval	±0.65	±2.10	±1.10	95% confidence interval	±0.40	±0.65	±0.42

Table 17: **Base-to-novel generalization performance on 11 datasets with ViT-B/16 visual backbone.** HM indicates Harmonic Mean. The best results are **bolded**, the second-best are underlined, and our proposed method RADA is highlighted.

Average across datasets				ImageNet				Caltech101			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	67.27	71.68	69.41	CLIP	67.49	64.06	65.73	CLIP	94.06	94.00	94.03
MMA	78.69	71.04	74.67	MMA	72.53	65.77	68.98	MMA	97.20	92.63	94.86
2SFS	82.32	71.23	76.37	2SFS	72.52	66.62	69.44	2SFS	97.83	93.30	95.51
RADA	82.52	74.25	78.17	RADA	72.41	67.33	69.78	RADA	98.22	94.07	96.10

Oxford Flowers				Oxford Pets				Stanford Cars			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	72.36	73.69	73.02	CLIP	90.64	96.87	93.65	CLIP	60.72	69.74	64.92
MMA	95.50	71.57	81.82	MMA	93.77	96.30	95.02	MMA	73.73	69.27	71.43
2SFS	96.64	70.02	81.20	2SFS	93.18	95.56	94.35	2SFS	78.21	70.30	74.04
RADA	97.02	71.62	82.41	RADA	93.59	95.67	94.62	RADA	76.08	73.07	74.54

Food 101				FGVC Aircraft				SUN 397			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	85.30	86.89	86.09	CLIP	21.25	29.27	24.62	CLIP	69.80	73.01	71.37
MMA	85.77	87.13	86.44	MMA	31.77	28.73	30.17	MMA	80.27	76.57	78.38
2SFS	84.75	87.37	86.04	2SFS	39.12	30.85	34.50	2SFS	81.11	78.02	79.53
RADA	86.74	87.80	87.27	RADA	38.61	33.77	36.03	RADA	82.42	78.68	80.51

DTD				EuroSAT				UCF101			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	54.17	58.21	56.12	CLIP	55.14	69.77	61.60	CLIP	69.08	72.96	70.97
MMA	79.50	57.00	66.40	MMA	71.83	62.97	67.11	MMA	83.77	73.47	78.28
2SFS	80.09	54.63	64.95	2SFS	96.80	62.68	76.09	2SFS	85.25	74.15	79.31
RADA	80.09	54.79	65.07	RADA	96.72	82.16	88.84	RADA	85.83	77.76	81.60

Table 18: **Base-to-novel generalization performance on 11 datasets with ViT-B/32 visual backbone.** HM indicates Harmonic Mean. The best results are **bolded**, and our proposed method RADA is highlighted.

Average across datasets				ImageNet				Caltech101			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	76.18	80.08	78.08	CLIP	79.18	74.04	76.53	CLIP	95.61	95.41	95.51
MMA	85.70	79.06	82.25	MMA	83.17	76.73	79.82	MMA	98.60	95.97	97.27
2SFS	89.05	79.64	84.08	2SFS	83.11	76.98	79.93	2SFS	98.82	96.69	97.74
RADA	89.26	82.66	85.83	RADA	83.00	77.69	80.23	RADA	99.21	97.46	98.33

Oxford Flowers				Oxford Pets				Stanford Cars			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	80.34	83.05	81.67	CLIP	93.78	96.53	95.14	CLIP	74.56	84.65	79.29
MMA	99.00	80.20	88.61	MMA	96.23	98.70	97.45	MMA	85.27	83.80	84.53
2SFS	98.99	80.73	88.93	2SFS	96.74	98.64	97.68	2SFS	87.46	84.56	85.99
RADA	99.37	82.33	90.05	RADA	97.15	98.75	97.94	RADA	85.33	87.33	86.32

Food 101				FGVC Aircraft				SUN 397			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	93.75	94.82	94.28	CLIP	37.52	44.21	40.59	CLIP	73.23	77.71	75.40
MMA	94.23	95.10	94.66	MMA	50.00	42.47	45.93	MMA	85.03	81.77	83.37
2SFS	93.59	94.93	94.26	2SFS	59.76	43.59	50.41	2SFS	85.57	82.24	83.87
RADA	95.58	95.36	95.47	RADA	59.25	46.51	52.11	RADA	86.88	82.90	84.84

DTD				EuroSAT				UCF101			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	59.14	67.87	63.21	CLIP	70.93	82.90	76.45	CLIP	79.94	79.66	79.80
MMA	85.23	70.77	77.33	MMA	77.33	62.77	69.29	MMA	88.60	81.37	84.83
2SFS	87.35	70.73	78.17	2SFS	98.41	64.69	78.06	2SFS	89.80	82.24	85.85
RADA	87.35	70.89	78.26	RADA	98.33	84.17	90.70	RADA	90.38	85.85	88.06

Table 19: **Base-to-novel generalization performance on 11 datasets with ViT-L/14 visual backbone.** HM indicates Harmonic Mean. The best results are **bolded**, and our proposed method RADA is highlighted.