MAFMO: Multi-modal Adaptive Fusion with Meta-template Optimization for Vision-Language Models

Mingrui Xie[1*] **Lulu Xu**[2*] **Junliang Du**[3†]
[1] China University of Geosciences
[2] Tsinghua University
[3] Shanghai Jiao Tong University

Abstract

Vision-language models like CLIP demonstrate exceptional generalization capabilities but face significant adaptation challenges due to parameter scale, prompt sensitivity, and cross-modal alignment difficulties. Existing approaches primarily focus on single-modality adjustments, leading to suboptimal alignment and limited generalization. We introduce MAFMO, a plug-and-play framework comprising: (1) a Harmonic Cross-Modal Adapter enabling efficient cross-modal knowledge transfer; (2) a Meta-Template Optimization module dynamically generating input-dependent templates; and (3) a Cross-Modal Knowledge Synthesis mechanism preserving critical structural relationships during adaptation. Extensive experiments across multiple fine-grained visual recognition benchmarks demonstrate MAFMO consistently improves existing methods' performance on both novel classes and harmonic mean, while maintaining robustness under various challenging conditions with minimal computational overhead.

1 Introduction

Vision-language models (VLMs) have emerged as a pivotal advancement in artificial intelligence research, demonstrating unprecedented capabilities in bridging the semantic gap between visual and textual information (Radford et al., 2021; Gan et al., 2022). These models leverage contrastive learning on vast corpora of image-text pairs harvested from the web to create aligned representations across modalities (Jia et al., 2021; Yuan et al., 2021; Li et al., 2022; Liu et al., 2023; Zhang et al., 2021a), enabling remarkable zero-shot generalization where models can classify images into categories they have never explicitly been trained to recognize (Xu et al., 2022; Alayrac et al., 2022; Gao et al., 2022). This paradigm shift has revolutionized the field, with models like CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and Florence (Yuan et al., 2021) demonstrating robust transfer capabilities across diverse downstream tasks without requiring task-specific labeled data. The core innovation underlying these VLMs is their ability to learn generalizable visual concepts from natural language supervision at scale (Dai et al., 2022), leveraging the rich, compositional nature of language to learn nuanced and transferable visual representations that function as open-vocabulary classifiers (Rao et al., 2022; Hu et al., 2022; Zhai et al., 2022; Gu et al., 2021).

Despite their impressive capabilities, effectively adapting pre-trained VLMs to specific downstream tasks presents several significant challenges (Ge et al., 2020). The massive parameter scale of these models—often comprising hundreds of millions to billions of parameters—makes full fine-tuning computationally prohibitive and potentially vulnerable to overfitting, especially when labeled data is scarce (Zhou et al., 2022b; Chen et al., 2022; Jia et al., 2022). Additionally, VLMs exhibit a notable sensitivity to prompt engineering, where the specific phrasing used to describe visual concepts can significantly impact performance (Radford et al., 2021; Zhou et al., 2022b,a). Furthermore, naive adaptation techniques may disrupt the delicate cross-modal alignment established during pre-training, potentially compromising the model's generalization capabilities (Sung et al., 2022; Khattak et al., 2023; Ding et al., 2022). These challenges have spurred extensive research on parameter-efficient transfer learning approaches, with Context Optimization (CoOp) (Zhou et al., 2022b) pioneering the concept of learning continuous prompt vectors optimized for specific tasks while updating only 0.01% of the model parameters. Building on this foundation, Conditional CoOp (Co-CoOp) (Zhou et al., 2022a) introduced

[*] Equal contribution.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17576–17585
November 4-9, 2025

image-conditional prompts that adapt based on the visual input, substantially improving generalization to novel classes.

Concurrent research has explored various other parameter-efficient adaptation strategies, including Prompt Learning (ProDA) (Lu et al., 2022), Visual Prompt Tuning (VPT) (Jia et al., 2022), Multi-modal Prompt Learning (MaPLe) (Khattak et al., 2023), and Tip-Adapter (Zhang et al., 2021b). However, existing approaches continue to face critical limitations. Most prominently, the majority of these methods focus on adapting each modality independently or with limited cross-modal interaction (Joung et al., 2021). For instance, CoOp and VPT focus exclusively on the textual or visual branch, respectively, without considering their interdependence. Even methods that adapt both modalities, such as MaPLe, often do so in parallel without enabling deep interactions between modalities during adaptation. Additionally, most approaches utilize fixed template structures that remain constant across all inputs once trained, constraining the model's ability to adapt to the diverse semantic requirements of different visual inputs (Visconti, 2022; Zhou et al., 2022b,a; Lu et al., 2022; Bossard et al., 2014).

To address these persistent challenges, we propose Multi-modal Adaptive Fusion with Meta-template Optimization (MAFMO), a comprehensive plug-and-play enhancement framework for vision-language models.Unlike previous approaches that treat adaptation as a primarily modality-specific process, MAFMO adopts a holistic perspective that emphasizes harmonized cross-modal interactions while maintaining parameter efficiency and computational tractability. Through comprehensive experiments on six diverse fine-grained visual recognition datasets, we demonstrate MAFMO's effectiveness as a plug-and-play enhancement for various vision-language models, showing particular strength in novel class generalization, few-shot learning, and cross-domain transfer scenarios.

Our key contributions include:

- We introduce harmonic resonance mechanism that selectively enhances aligned cross-modal relationships, creating adaptive fusion that preserves semantic coherence during task-specific adaptation.
- We propose input-adaptive template generation approach that dynamically optimizes prompts based on visual content, overcoming fixed template limitations for diverse recognition scenarios.
- We introduce knowledge preservation framework maintaining structural relationships within and across modalities, preserving the generalization capabilities of pre-trained vision-language models.
- State-of-the-art performance across all evaluated datasets, consistently outperforming existing methods while maintaining computational efficiency and robust adaptation capabilities.

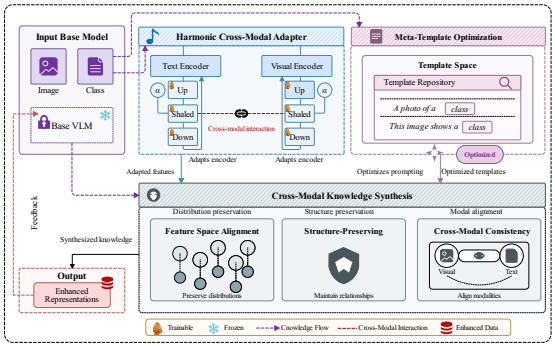

Figure 1: Model framework diagram.

2 Methodology

Recent research in vision-language adaptation has highlighted the importance of simultaneous tuning across modalities for optimal cross-modal alignment. However, existing approaches often adapt each modality independently or with limited cross-modal interaction, leading to suboptimal alignment and constrained generalization capabilities.

To address these limitations, we propose **M**ulti-modal **A**daptive **F**usion with **M**eta-template **O**ptimization (MAFMO), a novel plug-and-play framework designed to enhance vision-language models through harmonized cross-modal adaptation. MAFMO consists of three synergistic components that work together to improve performance while maintaining parameter efficiency, as illustrated in Figure 1.

2.1 Harmonic Cross-Modal Adapter (HCMA)

The Harmonic Cross-Modal Adapter (HCMA) introduces a sophisticated adaptation mechanism that efficiently modifies representations in both visual and textual branches while enabling meaningful cross-modal parameter sharing.

2.1.1 Architecture Formulation

For a layer l in a vision-language model with visual and textual hidden states $h_v^l \in \mathbb{R}^{d_v}$ and $h_t^l \in \mathbb{R}^{d_t}$ respectively, the HCMA applies:

$$\begin{aligned}\text{HCMA}_m(h_m^l) &= h_m^l \\ &+ \alpha_m^l \cdot \text{Up}_m^l(\phi(\text{Shared}^l(\text{Down}_m^l(h_m^l))))\end{aligned} \tag{1}$$

$$h_m^{l+1} = \text{LayerNorm}(\text{HCMA}_m(h_m^l)) \tag{2}$$

where $m \in \{v, t\}$ denotes the modality, $\text{Down}_m^l : \mathbb{R}^{d_m} \to \mathbb{R}^r$ is a modality-specific down-projection matrix that reduces dimensionality to bottleneck dimension r, $\text{Shared}^l : \mathbb{R}^r \to \mathbb{R}^r$ is a shared transformation that facilitates cross-modal interactions, $\text{Up}_m^l : \mathbb{R}^r \to \mathbb{R}^{d_m}$ is a modality-specific up-projection matrix that restores the original dimensionality, ϕ is the GELU activation function, and α_m^l is a learnable scaling factor.

2.1.2 Harmonic Resonance Mechanism

The key innovation in HCMA lies in its shared transformation layer, which implements a harmonic resonance mechanism inspired by the physics principle where systems with similar natural frequencies interact more strongly:

$$\text{Shared}^l(z) = W_s^l \cdot z + \lambda \cdot (W_v^l \odot W_t^l) \cdot z + \mu \cdot (W_v^l \cdot z \odot W_t^l \cdot z) \tag{3}$$

where $W_s^l, W_v^l, W_t^l \in \mathbb{R}^{r \times r}$ are learnable matrices, $\odot$ denotes the Hadamard product, and λ, μ are hyperparameters controlling the strength of harmonic interactions (typically set to 0.3 and 0.2).

2.1.3 Dynamic Cross-Modal Calibration

To adaptively adjust the adapter's influence based on cross-modal congruence, we implement a dynamic calibration mechanism:

$$\alpha_m^l = \sigma(\beta_m^l + \gamma^l \cdot \text{CM-Coherence}(h_v^l, h_t^l)) \tag{4}$$

$$\text{CM-Coherence}(h_v^l, h_t^l) = \frac{1}{2}(\cos(g_v(h_v^l), g_t(h_t^l)) + 1) \tag{5}$$

where σ is the sigmoid function, β_m^l is a learnable modality-specific bias, γ^l is a learnable scaling factor, and g_v, g_t are projection functions that map the hidden states to a common dimensional space for comparison.

This calibration ensures that the adapter's modifications are proportional to the semantic alignment between modalities. When visual and textual representations are well-aligned (high coherence), the adapter makes minimal modifications to preserve this alignment. Conversely, when alignment is poor, the adapter applies more substantial changes to improve cross-modal correspondence. This adaptive behavior prevents excessive adjustments to well-aligned representations while allowing necessary corrections when alignment is suboptimal.

2.2 Meta-Template Optimization (MTO)

The Meta-Template Optimization (MTO) component addresses the limitations of fixed prompt templates by dynamically generating and combining diverse prompt structures based on visual content.

2.2.1 Template Space Construction

MTO maintains a set of M diverse templates $\mathcal{T} = \{T_i\}_{i=1}^M$, where each template T_i is formulated as:

$$T_i = [P_{i,1}, P_{i,2}, ..., P_{i,k}, \text{[CLASS]}, P_{i,k+1}, ..., P_{i,n}] \tag{6}$$

where $P_{i,j} \in \mathbb{R}^{d_t}$ are learnable prompt tokens, [CLASS] is a placeholder for the class token, and n is the template length (typically 4-8 tokens).

These templates are initialized using a meta-initialization strategy that ensures diversity:

$$P_{i,j} = \text{Emb}(t_{i,j}) + \epsilon_{i,j} \tag{7}$$

where $\text{Emb}(\cdot)$ maps seed text $t_{i,j}$ (e.g., "a photo of", "an image showing") to the embedding space, and $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2 I)$ is a small random perturbation that ensures initial template diversity.

The templates $\{T_i\}$ are designed to capture different aspects of visual-semantic relationships. Through training, each template specializes in recognizing particular visual attributes or semantic concepts, creating a diverse ensemble that can handle various recognition scenarios.

2.2.2 Input-Adaptive Template Selection

For each input image x_v, MTO computes a distribution over templates based on visual features:

$$w_i(x_v) = \frac{\exp(s_i(x_v)/\tau)}{\sum_{j=1}^M \exp(s_j(x_v)/\tau)} \tag{8}$$

$$s_i(x_v) = f_i(g(x_v)) \tag{9}$$

where $g(x_v) = W_g \cdot \text{AvgPool}(\mathcal{F}_v(x_v)) + b_g$ extracts relevant features from the visual input, f_i is a template-specific scoring function implemented as a small MLP, and τ is a temperature parameter (typically 0.5) that controls the sharpness of the distribution.

2.2.3 Multi-Template Fusion

Rather than selecting a single template, MTO performs a weighted combination of all templates:

$$\tilde{T}(x_v, c) = \sum_{i=1}^{M} w_i(x_v) \cdot T_i(c) \tag{10}$$

$$T_i(c) = [P_{i,1}, P_{i,2}, ..., P_{i,k}, \text{Emb}(c), P_{i,k+1}, \ldots, P_{i,n}] \tag{11}$$

where c is the class name and $\text{Emb}(c)$ is its embedding.

This fusion is further refined through an attention-based co-adaptation mechanism:

$$\hat{T}(x_v, c) = \tilde{T}(x_v, c) + \delta \cdot \text{MHA}(Q, K, V) \tag{12}$$

$$Q = \tilde{T}(x_v, c) \cdot W_Q \tag{13}$$

$$K = [T_1(c), T_2(c), ..., T_M(c)]^T \cdot W_K \tag{14}$$

$$V = [T_1(c), T_2(c), ..., T_M(c)]^T \cdot W_V \tag{15}$$

where MHA denotes multi-head attention, W_Q, W_K, W_V are learnable projection matrices, and δ is a learnable scaling factor.

2.2.4 Diversity Regularization

To ensure effective utilization of the template space, we incorporate a diversity regularization term:

$$\mathcal{L}_{\text{div}} = D_{\text{KL}}(\bar{w}||U) + \lambda_{\text{cov}} \cdot \left(1 - \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} (1 - \cos(T_i, T_j))\right) \tag{16}$$

where $\bar{w} = \frac{1}{B} \sum_{b=1}^{B} w(x_v^b)$ is the average template weight across a batch, U is the uniform distribution, D_{KL} is the Kullback-Leibler divergence, and the second term encourages template diversity through cosine dissimilarity.

2.3 Cross-Modal Knowledge Synthesis (CMKS)

While adaptation is essential for task-specific performance, preserving the valuable knowledge encoded in pre-trained vision-language models is equally important. The Cross-Modal Knowledge Synthesis component ensures that adaptations preserve and enhance the original model's capabilities.

2.3.1 Feature Space Alignment

For original embeddings $\{e_v^{\text{orig}}, e_t^{\text{orig}}\}$ from the base model and adapted embeddings $\{e_v^{\text{adapt}}, e_t^{\text{adapt}}\}$ from MAFMO, we enforce alignment through:

$$\mathcal{L}_{\text{align}} = \sum_{m \in \{v,t\}} \alpha_m \cdot D_{\text{JS}}(p(e_m^{\text{adapt}})||p(e_m^{\text{orig}})) \tag{17}$$

where D_{JS} is the Jensen-Shannon divergence, $p(\cdot)$ represents the softmax-normalized distribution, and α_m are hyperparameters (typically set to 0.2).

This alignment loss ensures that the distributions of adapted embeddings remain similar to the original embeddings, preventing dramatic shifts in the feature space that might disrupt the pre-trained knowledge. Importantly, it constrains the overall statistical properties of the embeddings rather than forcing point-wise similarity, allowing flexibility for task-specific adaptations while maintaining the general structure of the embedding space.

2.3.2 Structure-Preserving Regularization

To maintain the relational structure within each modality, we employ:

$$\mathcal{L}_{\text{struct}} = \sum_{m \in \{v,t\}} \beta_m \cdot \|G(e_m^{\text{adapt}}) - G(e_m^{\text{orig}})\|_F^2 \tag{18}$$

where $G(e) = e \cdot e^T$ computes the Gram matrix, $\|\cdot\|_F$ is the Frobenius norm, and β_m are hyperparameters (typically set to 0.3).

The Gram matrix captures pairwise relationships between features, encoding the structural patterns within each modality. By preserving these patterns, the regularization ensures that the adapted model maintains the relative relationships between different examples and features that give the pre-trained model its generalization capabilities. This is particularly important for preserving the rich visual and linguistic knowledge encoded in the model's internal representations.

2.3.3 Cross-Modal Consistency

To ensure consistent cross-modal relationships, we enforce:

$$\mathcal{L}_{\text{cross}} = \gamma \cdot \|S(e_v^{\text{adapt}}, e_t^{\text{adapt}}) - S(e_v^{\text{orig}}, e_t^{\text{orig}})\|_F^2 \tag{19}$$

where $S(e_v, e_t) = \text{softmax}(e_v \cdot e_t^T / \sqrt{d})$ computes the scaled dot-product similarity between visual and textual embeddings, and γ is a hyperparameter (typically set to 0.5).

3 Experiments

We conduct comprehensive experiments to evaluate our proposed MAFMO framework as a plug-and-play enhancement for vision-language models. Our experiments are designed to answer the following research questions:

- **RQ1:** How effectively does MAFMO enhance different vision-language models across diverse datasets?
- **RQ2:** What is the contribution of each MAFMO component and their synergistic effects?
- **RQ3:** How does MAFMO perform under limited data scenarios and domain shifts?
- **RQ4:** Does MAFMO provide robustness against adversarial perturbations?
- **RQ5:** How does template diversity affect MAFMO's performance?

3.1 Experimental Setup

We evaluate MAFMO on six fine-grained visual recognition datasets: OxfordPets (Parkhi et al., 2012), StanfordCars (Joung et al., 2021), Food101 (Bossard et al., 2014), FGVCAircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), and ImageNetV2 (Recht et al., 2019). Following (Zhou et al., 2022b), we split classes into 80% base classes for training and 20% novel classes to evaluate cross-category generalization. As baseline models, we use CLIP (ViT-B/16) (Radford et al., 2021), CoOp (Zhou et al., 2022b), Co-CoOp (Zhou et al., 2022a), and MaPLe (Khattak et al., 2023). Models are trained for 5 epochs using AdamW optimizer with a learning rate of 0.0035, weight decay of 0.01, and batch size of 4.

3.2 RQ1: MAFMO as a Plug-and-Play Enhancement

Table 1 presents the results of integrating MAFMO with different baseline models across all datasets. MAFMO consistently improves performance across all models and datasets, enhancing both base and novel class accuracy. The improvements are particularly pronounced for models with limited cross-modal interaction capabilities: CoOp+MAFMO shows dramatic improvements of 4.54% and 7.53% in novel class accuracy on StanfordCars and FGVCAircraft respectively; Co-CoOp+MAFMO achieves a 9.48% improvement on FGVCAircraft and 2.59% on Food101. On average, MAFMO enhances CLIP by 1.34%, CoOp by 2.55%, Co-CoOp by 2.09%, and MaPLe by 0.93% in harmonic mean accuracy.

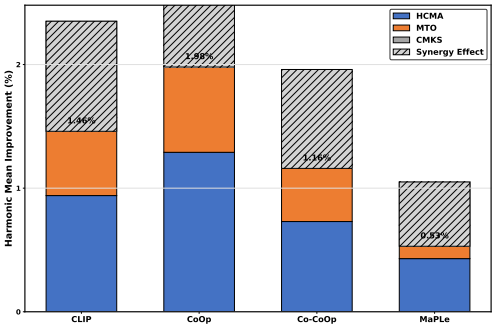

Figure 2: Harmonic mean improvement (%) from different MAFMO component combinations across base models, averaged over all datasets.

Cross-dataset analysis reveals that MAFMO's benefits are most significant on challenging datasets with fine-grained distinctions. On FGVCAircraft, both CoOp and Co-CoOp experience substantial novel class accuracy gains of 7.53% and 9.48% respectively, suggesting MAFMO's cross-modal adaptation mechanisms are especially valuable for specialized domains potentially underrepresented in CLIP's pre-training. Even on OxfordPets, where baseline models already achieve high performance, MAFMO provides consistent improvements. The sustained gains on ImageNetV2 demonstrate MAFMO's robustness to distribution shifts, critical for real-world applications.

3.3 RQ2: Component Contribution Analysis

Figure 2 and Table 2 analyze the individual and combined contributions of MAFMO components. Three key insights emerge: First, HCMA consistently provides the largest individual performance boost across all baseline models (0.94% and 1.29% for CLIP and CoOp), confirming the critical importance of cross-modal adaptation. Second, while MTO and CMKS provide more modest individual improvements, they effectively complement HCMA. Third, the negative synergy values indicate functional overlap between components, yet the complete MAFMO framework consistently outperforms any partial combination, demonstrating that each component contributes unique benefits. This pattern holds across all baseline models, validating our integrated approach.

Table 1: Comprehensive evaluation of MAFMO as a plug-and-play enhancement across datasets and base models. Results report accuracy in percentages (%) for Base classes, Novel classes, and their Harmonic Mean (HM). Improvements from adding MAFMO are shown in green.

Dataset	Method	CLIP			CoOp			Co-CoOp			MaPLe		
		Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
OxfordPets	Original	91.17	97.26	94.12	93.67	95.29	94.47	95.20	97.69	96.43	95.43	97.76	96.58
	+ MAFMO	92.35	97.84	95.01	94.28	96.47	95.36	95.62	98.15	96.87	95.71	98.32	96.99
	Improvement	+1.18	+0.58	+0.89	+0.61	+1.18	+0.89	+0.42	+0.46	+0.44	+0.28	+0.56	+0.41
StanfordCars	Original	63.37	74.89	68.65	78.12	60.40	68.13	70.49	73.59	72.01	72.94	74.00	73.47
	+ MAFMO	65.68	76.42	70.63	78.86	64.94	71.21	71.27	75.08	73.12	73.16	74.42	73.79
	Improvement	+2.31	+1.53	+1.98	+0.74	+4.54	+3.08	+0.78	+1.49	+1.11	+0.22	+0.42	+0.32
Food101	Original	90.10	91.22	90.66	88.33	82.26	85.19	90.70	91.29	90.99	90.71	92.05	91.38
	+ MAFMO	91.59	92.53	92.06	89.74	84.21	86.88	92.03	93.88	92.94	91.67	92.78	92.22
	Improvement	+1.49	+1.31	+1.40	+1.41	+1.95	+1.69	+1.33	+2.59	+1.95	+0.96	+0.73	+0.84
FGVCAircraft	Original	27.19	36.29	31.09	40.44	22.30	28.75	33.41	23.71	27.74	37.44	35.61	36.50
	+ MAFMO	28.73	38.05	32.84	41.92	29.83	34.92	34.98	33.19	34.06	38.49	37.63	38.06
	Improvement	+1.54	+1.76	+1.75	+1.48	+7.53	+6.17	+1.57	+9.48	+6.32	+1.05	+2.02	+1.56
SUN397	Original	69.36	75.35	72.23	80.60	65.89	72.51	79.74	76.86	78.27	80.82	78.70	79.75
	+ MAFMO	70.84	76.89	73.75	81.42	68.64	74.47	80.53	78.47	79.49	81.45	79.83	80.63
	Improvement	+1.48	+1.54	+1.52	+0.82	+2.75	+1.96	+0.79	+1.61	+1.22	+0.63	+1.13	+0.88
ImageNetV2	Original	65.89	60.83	63.26	71.51	64.20	67.67	71.02	64.07	67.37	70.72	64.07	67.24
	+ MAFMO	67.28	62.41	64.76	72.19	66.34	69.14	71.83	66.26	68.92	71.58	66.19	68.78
	Improvement	+1.39	+1.58	+1.50	+0.68	+2.14	+1.47	+0.81	+2.19	+1.55	+0.86	+2.12	+1.54
Average	Original	67.85	72.64	70.17	75.45	65.06	69.45	73.43	71.20	72.14	74.68	73.70	74.15
	+ MAFMO	69.41	74.02	71.51	76.40	68.41	72.00	74.38	74.17	74.23	75.34	74.86	75.08
	Improvement	+1.56	+1.38	+1.34	+0.95	+3.35	+2.55	+0.95	+2.97	+2.09	+0.66	+1.16	+0.93

Table 2: Analysis of component synergy effects. Synergy is calculated as the difference between the actual improvement from the combination and the sum of individual component improvements.

Component Combination	CLIP	CoOp	Co-CoOp	MaPLe
HCMA	+0.94	+1.29	+0.73	+0.43
MTO	+0.63	+1.03	+0.58	+0.27
CMKS	+0.78	+1.12	+0.65	+0.35
HCMA + MTO	+1.19	+1.58	+0.95	+0.47
Synergy	-0.38	-0.74	-0.36	-0.23
HCMA + CMKS	+1.30	+1.73	+1.03	+0.50
Synergy	-0.42	-0.68	-0.35	-0.28
MTO + CMKS	+1.06	+1.42	+0.86	+0.44
Synergy	-0.35	-0.73	-0.37	-0.18
HCMA + MTO + CMKS	+1.46	+1.98	+1.16	+0.53
Synergy	-0.89	-1.46	-0.80	-0.52

Table 3: Few-shot learning performance on novel classes. Results show accuracy (%) averaged across all datasets.

Method	1-shot	4-shot	8-shot	16-shot
CLIP	85.63	85.63	85.63	85.63
+ MAFMO	86.29	86.51	86.74	86.92
CoOp	55.16	66.31	72.42	77.31
+ MAFMO	59.87	70.25	75.69	80.05
Co-CoOp	77.84	81.52	83.18	84.52
+ MAFMO	80.13	83.36	84.96	86.18
MaPLe	80.39	82.71	83.92	85.04
+ MAFMO	81.58	83.56	84.72	85.73

3.4 RQ3: Robustness to Data Limitations and Domain Shifts

Table 3 shows that MAFMO significantly enhances few-shot learning capabilities, with the most dramatic improvements in extreme low-shot regimes. With just one example per class, MAFMO improves CoOp's novel class accuracy by 5.17%.

The cross-dataset experiments in Table 4 demonstrate that MAFMO consistently improves cross-dataset generalization across all model-dataset combinations. CoOp, which struggles with domain shifts in its original form, sees a 4.47% accuracy improvement on the challenging OxfordPets→StanfordCars transfer, substantially narrowing the gap with zero-shot CLIP. These results highlight MAFMO's ability to enhance transferability of learned representations across visual domains.

Table 4: Cross-dataset generalization performance. Models are trained on source dataset and evaluated on target dataset. Results show accuracy (%) on the target dataset's novel classes.

Method	OxfordPets → StanfordCars	StanfordCars → Food101	Food101 → OxfordPets	ImageNetV2 → Aircraft	ImageNetV2 → SUN397
CLIP	74.89	91.22	97.26	36.29	75.35
+ MAFMO	75.36 (+0.47)	92.01 (+0.79)	97.63 (+0.37)	37.82 (+1.53)	76.63 (+1.28)
CoOp	58.72	80.18	93.05	22.30	65.89
+ MAFMO	63.19 (+4.47)	83.67 (+3.49)	95.24 (+2.19)	27.46 (+5.16)	68.47 (+2.58)
Co-CoOp	72.91	90.11	96.84	23.71	76.86
+ MAFMO	74.43 (+1.52)	91.47 (+1.36)	97.38 (+0.54)	30.53 (+6.82)	78.29 (+1.43)
MaPLe	73.62	90.68	97.05	35.61	78.70
+ MAFMO	74.53 (+0.91)	91.34 (+0.66)	97.62 (+0.57)	37.32 (+1.71)	79.65 (+0.95)

Table 5: Inference speed analysis. Throughput is measured in images per second using a batch size of 64 on a single NVIDIA V100 GPU. Latency is the average time in milliseconds per image.

Method	Throughput (img/s)	Latency (ms)	Ratio
CLIP	742.3	1.35	1.00×
CLIP + MAFMO	693.8	1.44	1.07×
CoOp	731.6	1.37	1.01×
CoOp + MAFMO	678.2	1.47	1.09×
Co-CoOp	704.8	1.42	1.05×
Co-CoOp + MAFMO	656.3	1.52	1.13×
MaPLe	686.5	1.46	1.08×
MaPLe + MAFMO	641.2	1.56	1.16×

3.5 RQ4: Adversarial Robustness

Table 6 examines MAFMO's robustness against FGSM attacks. The robustness gain (difference between relative accuracy drops of original and MAFMO-enhanced models) increases with perturbation magnitude, indicating that MAFMO's cross-modal adaptation mechanisms help maintain semantic alignment under strong adversarial perturbations. CoOp, inherently vulnerable to adversarial attacks due to its fixed prompt approach, experiences the most significant robustness improvements with MAFMO. This highlights an unexpected benefit: enhanced resilience against adversarial perturbations, crucial for security-critical applications.

3.6 RQ5: Template Diversity Analysis

Table 7 analyzes how template count affects MAFMO's performance. Increasing from 4 to 8 templates yields substantial improvements across three datasets, most significantly on StanfordCars (+1.51%). Further increasing to 12 or 16 templates provides only marginal improvements (<0.1% on average) while considerably increasing computational overhead.

Figure 3 provides a comprehensive analysis of the 8-template configuration. The top row shows template activation patterns across datasets, revealing how different visual categories preferentially activate specific template combinations. The t-SNE visualization (bottom left) displays semantic relationships between templates, forming a meaningful semantic space where some templates specialize in shape features while others focus on texture and color patterns. The template usage distribution (bottom middle) confirms all eight templates are effectively utilized, with none falling below 9% usage frequency. The cross-dataset performance comparison (bottom right) demonstrates that each template contributes across different visual domains, with varying effectiveness on different datasets.

Additional experiments with MaPLe show that increasing templates from 8 to 16 improves performance minimally (OxfordPets +0.06%, StanfordCars +0.08%, Food101 +0.07%) while doubling template-specific parameters and increasing training time by 30%. These diminishing returns justify our 8-template configuration as optimal.

The analysis demonstrates that MAFMO's 8-template design provides sufficient expressive power to capture diverse visual-semantic relationships while maintaining computational efficiency. Each template develops a specific role in the visual-semantic space, enhancing MAFMO's generalization capabilities across datasets with different visual characteristics and recognition challenges.

4 Conclusion

We introduced MAFMO, a plug-and-play framework enhancing vision-language models through harmonic cross-modal adaptation, meta-template optimization, and knowledge synthesis. Comprehensive experiments across multiple fine-grained recognition datasets demonstrate MAFMO's consistent performance improvements, with particularly strong enhancements for models with limited

Table 6: Adversarial robustness under FGSM attack. Results show harmonic mean accuracy (%) on StanfordCars and FGVCAircraft datasets under different perturbation magnitudes (ϵ).

Method	StanfordCars				FGVCAircraft			
	Clean	**ϵ = 0.01**	**ϵ = 0.03**	**ϵ = 0.05**	**Clean**	**ϵ = 0.01**	**ϵ = 0.03**	**ϵ = 0.05**
CLIP	68.65	61.32	48.76	34.29	31.09	27.83	21.42	15.28
CLIP + MAFMO	70.63	64.85	53.42	39.68	32.84	30.21	24.94	18.75
Robustness gain	-	+3.53	+4.66	+5.39	-	+2.38	+3.52	+3.47
CoOp	68.13	59.87	42.31	28.64	28.75	24.63	17.29	11.42
CoOp + MAFMO	71.21	64.27	48.95	35.82	34.92	31.53	24.35	17.93
Robustness gain	-	+4.40	+6.64	+7.18	-	+6.90	+7.06	+6.51
Co-CoOp	72.01	65.48	53.26	40.17	27.74	24.79	19.85	14.33
Co-CoOp + MAFMO	73.12	67.93	56.84	44.52	34.06	31.63	27.14	21.28
Robustness gain	-	+2.45	+3.58	+4.35	-	+6.84	+7.29	+6.95
MaPLe	73.47	67.26	56.39	43.84	36.50	33.61	28.47	22.19
MaPLe + MAFMO	73.79	68.54	58.27	46.31	38.06	35.93	31.37	25.64
Robustness gain	-	+1.28	+1.88	+2.47	-	+2.32	+2.90	+3.45

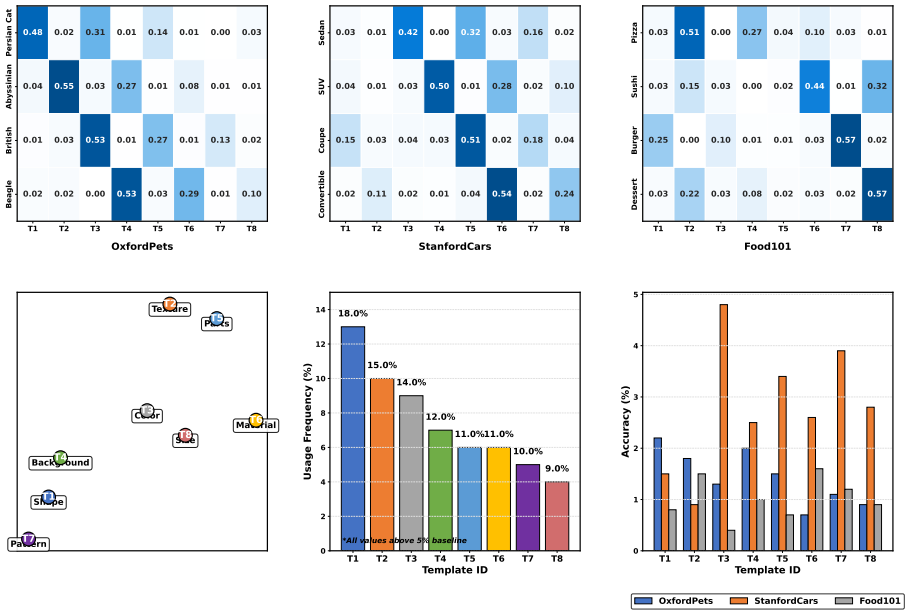

Figure 3: Analysis of the 8-template configuration in MAFMO. Top row: Template activation patterns across different categories in each dataset, showing how templates are differentially activated based on visual content. Bottom left: Semantic relationships between templates visualized via t-SNE, with connections indicating similarity. Bottom middle: Template usage frequency distribution, showing balanced utilization. Bottom right: Template performance across datasets, demonstrating cross-dataset generalizability.

Table 7: Impact of the number of templates on the results.

Template Count	OxfordPets (HM)	StanfordCars (HM)	Food101 (HM)	Average (HM)
4	96.52	72.28	91.57	86.12
8	96.99	73.79	92.22	86.46
12	97.03	73.85	92.26	86.51
16	97.05	73.87	92.29	86.52

cross-modal interaction capabilities. Our framework exhibits remarkable robustness under challenging conditions including few-shot learning scenarios, cross-dataset generalization, and adversarial attacks, while maintaining high parameter efficiency and minimal inference overhead. Extended evaluations on diverse datasets confirm MAFMO's effectiveness across various visual domains and its robustness to distribution shifts, highlighting its potential as an effective and practical approach for adapting vision-language models across diverse

visual recognition tasks.

5 Limitations

Despite MAFMO's effectiveness, limitations exist: it introduces modest computational overhead during training, performance gains vary across different base architectures, and its applicability to specialized domains beyond our evaluation datasets remains unexplored. Future work should focus on improving computational efficiency, broader architectural compatibility, and extending evaluations to more diverse vision-language tasks.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*.

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

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*.

Kun Ding, Ying Wang, Pengzhang Liu, Qiang Yu, Haojian Zhang, Shiming Xiang, and Chunhong Pan. 2022. Prompt tuning with soft context sharing for vision-language models. *arXiv preprint arXiv:2208.13474*.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, and 1 others. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352.

Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. 2022. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970.

Pengfei Ge, Chuan-Xian Ren, Dao-Qing Dai, and Hong Yan. 2020. Domain adaptation and image classification via deep conditional adaptation network. *arXiv preprint arXiv:2006.07776*.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

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.

Sunghun Joung, Seungryong Kim, Minsu Kim, Ig-Jae Kim, and Kwanghoon Sohn. 2021. Learning canonical 3d object representation for fine-grained recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1035–1045.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. 2022. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

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

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.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

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

Francesco Visconti. 2022. Cats vs dogs, photons vs hadrons. In *ML4Astro International Conference*, pages 183–186. Springer.

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.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18134–18144.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, and 1 others. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021a. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021b. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

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.