MAD-SP: MULTI-AGENT DYNAMIC SEMANTIC PROMPTING FRAMEWORK FOR FEW-SHOT IMAGE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot learning (FSL) recognizes novel classes from few labeled examples, but scarce support data yield unstable visual representations. Multimodal methods alleviate this limitation by incorporating textual semantics as prompts, but LLM-generated captions may be overly verbose and redundant. To address these issues, we propose Multi-Agent Dynamic Semantic Prompting (MAD-SP), which improves both the construction and utilization of semantic prompts. First, Multi-Agent Caption Refinement (MACR) coordinates generator, critic, and refiner agents under CLIP-based alignment and discriminative scoring to produce concise, visually grounded captions, thereby providing higher-quality semantic prompts for visual recognition. Building on these refined prompts, Dynamic Prompt Injection (DPI) enables their instance-adaptive utilization by employing a lightweight learnable decision network to estimate the compatibility between each semantic prompt and visual features at different blocks, thereby selecting an appropriate injection layer for each image. To ensure reliable alignment between semantic prompts and layer-specific visual representations during varying-depth injection, Contrastive Modality Alignment (CMA) further supports DPI by introducing a bottleneck alignment adapter after the text encoder and optimizing it with the InfoNCE loss. Together, MACR improves the quality of semantic prompts, while DPI and CMA enable their effective and adaptive integration into the visual encoder. Extensive experiments show that MAD-SP achieves state-of-the-art results in all eight settings on four standard benchmarks and in seven of eight settings across four cross-domain benchmarks.

1 INTRODUCTION

Few-shot learning (FSL) aims to recognize novel classes from only a few labeled examples (Snell et al., 2017). Prototype-based metric methods (Snell et al., 2017; Vinyals et al., 2016; Sung et al., 2018; Zhang et al., 2020; Bateni et al., 2020; Chen et al., 2021a) address this problem by representing each class with a prototype constructed from its support images. However, scarce support samples often provide incomplete visual evidence, resulting in inaccurate and unstable prototypes (Chen et al., 2024; Oreshkin et al., 2018; Li et al., 2019; Tian et al., 2020; Ye et al., 2020). To complement limited visual observations, recent approaches introduce language semantics into FSL (Chen et al., 2023). They encode class names or textual descriptions as semantic prompts and inject them into the visual encoder to guide discriminative feature extraction (Xu & Le, 2022; Zhang et al., 2023). Class names alone, however, are often ambiguous and provide little information about fine-grained visual characteristics (Xing et al., 2019; Li et al., 2020). Consequently, several methods use Large Language Models (LLMs) (Achiam et al., 2023; Bai et al., 2025) to expand class labels or images into more detailed descriptions (Zhang et al., 2021; Liu et al., 2025; Zhang et al., 2024a; Li et al., 2026; Wu et al., 2026).

However, does providing more elaborate textual descriptions necessarily lead to better semantic prompts? Without adequate supervision, one-pass LLM captions may contain irrelevant or visually unsupported attributes, introducing noisy guidance and mismatching textual and visual representations. A reliable prompt must instead be visually faithful, discriminative across classes, and concise. Because a one-pass generator cannot diagnose or correct its own output, caption quality must be

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

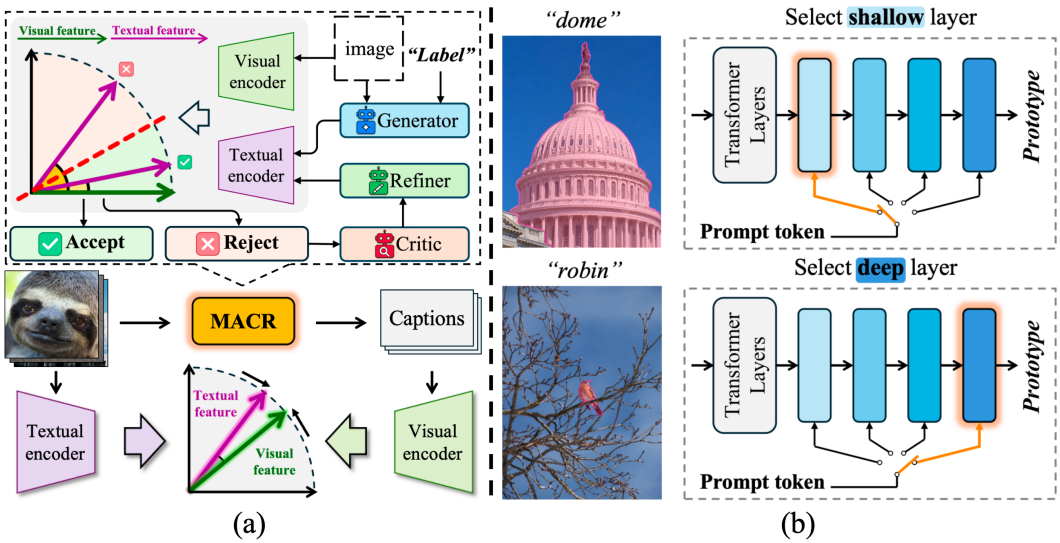


Figure 1: Motivation of MAD-SP. (a) MACR uses a generator to produce an initial caption, evaluates it with alignment and discriminative scores, and activates the critic–refiner loop when the caption fails to meet the quality threshold, yielding a visually grounded semantic prompt. (b) DPI selects an instance-specific injection layer according to the texture and detail patterns of each image: images dominated by local visual details can interact with prompts at shallower layers, whereas those requiring higher-level semantic abstraction can use deeper layers.

explicitly assessed before the text is used for recognition. Objective scores identify deficient captions, while specialized agents with complementary roles diagnose their weaknesses and perform targeted revisions (Li et al., 2023; Madaan et al., 2023; Gou et al., 2024; Chan et al., 2024). Moreover, existing semantic prompt injection methods (Chen et al., 2023; Zhang et al., 2024a) generally insert prompt tokens at a fixed layer, which may be suboptimal because different objects and visual patterns can require different levels of semantic abstraction (Wu et al., 2026). Since shallow and deep layers capture different visual information, the interaction between visual features and semantic prompts should ideally be adapted to the characteristics of each input. When semantic prompts are injected at different layers, both cross-modal alignment and layer selection therefore require careful consideration.

With these considerations in mind, we seek to develop a dynamic network framework equipped with a caption-generation supervisor, ensuring image-faithful descriptions without oversimplification or unnecessary verbosity, with dynamic selection of the most suitable prompt-injection layer for each image. From another perspective, the former improves the semantic quality of the prompts, whereas the latter enables more effective utilization of the semantic information they contain.

To this end, we propose MAD-SP, a novel Multi-Agent Dynamic Semantic Prompting framework for few-shot learning. First, the Multi-Agent Caption Refinement (MACR) module employs a generator agent, a critic agent, and a refiner agent (Li et al., 2023; Madaan et al., 2023; Gou et al., 2024; Chan et al., 2024) based on LLMs to collaboratively produce and improve semantic prompts. As shown in Fig. 1(a), CLIP-based (Radford et al., 2021) alignment and similarity scores provide objective feedback during the iterative refinement process. Compared with directly using captions generated by an LLM, captions supervised by the scoring mechanism and iteratively refined by multiple agents are more closely aligned with the image content, making them suitable for subsequent encoding by the text encoder and enabling the generation of high-quality semantic prompts.

Next, inspired by the concept of dynamic networks, we propose a Dynamic Prompt Injection (DPI) mechanism, which uses a lightweight dynamic decision network to select the most suitable semantic prompt injection layer for each image according to its instance-specific visual characteristics, as illustrated in Fig. 1(b). This design enables more effective utilization of the semantic information contained in the prompts. Furthermore, we introduce the Contrastive Modality Alignment (CMA) module to adapt prompt tokens to different injection layers and further improve cross-modal alignment. Specifically, a bottleneck alignment adapter is added after the text encoder and trained using the InfoNCE loss (van den Oord et al., 2018). Overall, MAD-SP generates higher-quality seman-

tic prompts while efficiently exploiting their semantic information through instance-level dynamic prompt injection.

Our contributions are summarized as follows:

- We introduce MAD-SP, a unified framework designed to improve both the construction and utilization of semantic prompts. It leverages multi-agent collaboration to generate high-quality, visually grounded descriptions and dynamically selects an appropriate visual layer for instance-specific prompt injection.
- To improve semantic prompt quality, we introduce MACR, a multi-agent collaborative framework that performs iterative caption generation with explicit alignment and discriminativeness supervision, producing visually grounded and discriminative prompts while mitigating class-label ambiguity and redundant LLM-generated details.
- To utilize semantic information more effectively, DPI introduces dynamic routing into prompt injection for FSL, automatically selecting the most suitable layer per image according to its visual characteristics. CMA supports DPI by aligning semantic prompts with visual features across layers, using a bottleneck adapter and InfoNCE.
- Extensive experiments demonstrate that MAD-SP achieves state-of-the-art performance in both the 1-shot and 5-shot settings on all four standard benchmarks. Across the four standard and four cross-domain benchmarks, MAD-SP achieves average margins of 2.72 and 1.97 percentage points over the strongest competing methods, respectively.

2 Related Work

Few-Shot Learning. Few-shot learning (Snell et al., 2017) aims to extract class-discriminative features from limited support samples. Existing methods mainly follow the metric-based paradigm. Metric-based methods (Song et al., 2024; Chen et al., 2023; Xing et al., 2019; Li et al., 2020; Xu & Le, 2022; Zhang et al., 2023) learn an embedding space that clusters same-class samples and separates different classes. At inference, queries are classified by their distance or similarity to support-derived prototypes. To extract more discriminative information from limited support images, many methods (Zhang et al., 2021; Liu et al., 2025; Zhang et al., 2024a; Li et al., 2026; Wu et al., 2026) introduce textual semantics from the language modality as auxiliary guidance. Many of these approaches (Chen et al., 2023; Xing et al., 2019; Zhang et al., 2021) rely solely on short class names or manually crafted brief descriptions, whose vague and limited content can introduce ambiguity. Recently, methods such as SemFew (Zhang et al., 2024a), ECER (Liu et al., 2025), MPA (Wu et al., 2026), and DVLA-RL (Li et al., 2026) have leveraged LLMs to generate more fine-grained textual descriptions. However, their prompt-generation processes lack explicit supervision, potentially resulting in excessive or redundant semantic details. MAD-SP addresses this limitation by combining multi-agent caption refinement with objective alignment feedback.

Dynamic Network. Dynamic networks (Zhao et al., 2024; Cao et al., 2024; Ye et al., 2025) in Vision Transformers have been widely studied for efficient computation. These methods (Xu et al., 2025; Tang et al., 2023; Zhu et al., 2025) generally train a lightweight decision network to predict routing decisions, allowing the model to skip unnecessary Transformer layers or attention heads and discard less informative tokens (Luo et al., 2025; Huang et al., 2025; Guo et al., 2025). MAD-SP transfers this idea to prompt injection. Although previous semantic prompting methods leverage semantic prompts as auxiliary guidance, their fixed-layer injection schemes may fail to fully exploit the semantic information carried by these prompts. To address this issue, DPI constrains each sample to select only one Transformer layer for prompt injection and uses the current visual representation together with the semantic prompt to determine the most appropriate layer.

3 Preliminaries

Few-shot learning (FSL) aims to classify novel categories using only a few labeled samples. The dataset is divided into a base set $\mathcal{D}_{base} = \{(x, y) \mid x \in \mathcal{X}_{base}, y \in \mathcal{C}_{base}\}$ for model training and a novel set $\mathcal{D}_{novel} = \{(x, y) \mid x \in \mathcal{X}_{novel}, y \in \mathcal{C}_{novel}\}$ for evaluation, where x and y denote an image and its label, respectively. The two label spaces are disjoint, i.e., $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$.

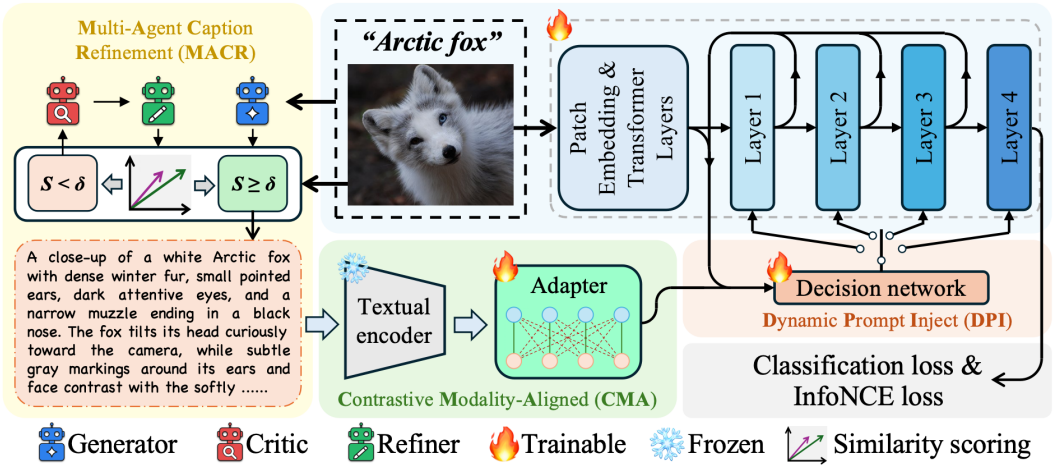


Figure 2: Overview of MAD-SP. MACR generates and refines semantic prompts through generator–critic–refiner collaboration, while CMA aligns semantic prompts with visual representations. DPI then matches each semantic prompt with intermediate visual features from different Transformer layers and selects the most compatible layer for prompt injection. The framework is jointly optimized using classification and InfoNCE losses.

During evaluation, an N-way K-shot task consists of a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N\times K}$ and a query set $\mathcal{Q} = \{(x_i, y_i)\}_{i=1}^{N\times M}$, where N is the number of classes, K is the number of support samples per class, and M is the number of query samples per class. The support set defines episode-specific prototypes, while the query set measures generalization. Given a feature extractor $f(\cdot)$, the prototype of class t is computed as

$$p_t = \frac{1}{|\mathcal{S}_t|} \sum_{x_i \in \mathcal{S}_t} f(x_i), \tag{1}$$

where $\mathcal{S}_t$ denotes the support samples belonging to class t. The probability that a query sample q belongs to class t is defined as

$$P(y = t \mid q) = \frac{\exp(d(f(q), p_t))}{\sum_{i=1}^{N} \exp(d(f(q), p_i))}, \tag{2}$$

where $d(\cdot, \cdot)$ denotes a similarity metric, such as cosine similarity. The query sample is assigned to the class with the highest prototype similarity.

4 Methodology

This section introduces the three components of the proposed framework: (1) MACR generates high-quality semantic prompts through collaborative caption refinement; (2) under DPI, each semantic prompt automatically selects the most suitable visual interaction layer for insertion; and (3) CMA establishes modal alignment between semantic prompts and visual features by jointly considering the adapter architecture and the contrastive learning objective.

4.1 Multi-Agent Caption Refinement

MACR employs multiple LLM agents to generate instance-aware, fine-grained descriptions as semantic prompts for few-shot image recognition, covering attributes such as color, shape, texture, pose, and class-specific details. For each labeled training or support image x with class name c, the Generator Agent produces the initial caption $p_0 = \mathcal{G}(x, c)$, where $\mathcal{G}(\cdot)$ denotes the generator. Conditioning on both inputs preserves class semantics while grounding the caption in the current instance, yielding a more informative candidate than a generic class-level description. Since LLM-generated captions may contain hallucinations, omit visual details, or lack class discriminability, MACR objectively evaluates each prompt before semantic feature extraction.

The scoring mechanism combines text-image alignment and class discriminativeness. The alignment score measures the caption's consistency with the image:

$$S_{\text{align}}(x, p) = \cos\big(E_I(x), E_T(p)\big), \tag{3}$$

where $E_I(\cdot)$ and $E_T(\cdot)$ are the CLIP image and text encoders, respectively. A higher S_{align} indicates stronger image-caption alignment.

Alignment alone may favor generic captions, such as "a photo of an animal", which provide little discriminative information. We therefore define:

$$S_{\text{dis}}(p, c) = \cos\big(E_T(p), E_T(c)\big) - \max_{c' \neq c} \cos\big(E_T(p), E_T(c')\big). \tag{4}$$

where c is the ground-truth class and c' denotes other candidate classes. This score encourages similarity to the target class while separating the caption from semantically similar classes.

The two terms form the final quality score:

$$S(x, c, p) = (1 - \omega) S_{\text{align}}(x, p) + \omega S_{\text{dis}}(p, c), \qquad \omega \in [0, 1], \tag{5}$$

where ω balances visual alignment and class discriminativeness. The caption is accepted when $S(x, c, p) \geq \delta$, where δ is the acceptance threshold; otherwise, feedback-based refinement begins.

For a rejected caption, the Critic Agent assesses its faithfulness, specificity, and discriminativeness, providing structured feedback on missing attributes, visual inconsistencies, or insufficient distinction from similar classes. The Refiner Agent rewrites the caption accordingly, after which it is rescored. This process continues until the caption meets the threshold or reaches the maximum number of iterations.

This generator-critic-refiner collaboration forms a closed-loop prompt evolution process that produces visually faithful, detailed, and discriminative prompts for reliable semantic guidance in few-shot recognition.

4.2 Dynamic Prompt Injection

Semantic prompts interact differently with hierarchical visual features: early layers capture low-level patterns, whereas deeper layers encode class-specific semantics. Dynamic Prompt Injection (DPI) therefore selects an injection layer for each support image-prompt pair instead of using a fixed layer.

DPI considers the L blocks in the third stage of the Visformer backbone. Before block l, it pools the current feature x_l into $v_l = \text{GAP}(x_l)$ and combines it with semantic prompt t to predict the injection probability:

$$p_l = \sigma\big(w_l^\top \text{ReLU}\big(W_v v_l \,\|\, W_t t\big) + b_l\big), \tag{6}$$

where W_v and W_t project the two modalities into a shared space, $\|$ denotes concatenation, and the sigmoid σ yields the suitability $p_l \in (0, 1)$ of block l.

A hard gate $d_l = \mathbb{I}(p_l > \xi)$ activates the first block whose probability exceeds threshold ξ; later decisions are disabled, and the last block serves as a fallback. Since this gate is non-differentiable, the straight-through estimator $g_l = d_l + p_l - \text{sg}(p_l)$, where sg denotes stop-gradient, uses d_l in the forward pass while preserving the gradient of p_l during backpropagation.

At the selected block, g_l controls spatial and channel-wise interaction. Spatial interaction appends the projected prompt p^{sp} as an additional spatial row before stage 3 and activates it as $x_l^{\text{sp}} = [x_l, g_l p^{\text{sp}}]$, where $[\cdot, \cdot]$ denotes spatial concatenation. Channel-wise interaction concatenates the global descriptor $\text{GAP}(x_l)$ with the projected prompt p^{ch} and feeds them to a squeeze-and-excitation block, yielding $c_l = \text{SEBlock}([\text{GAP}(x_l) \,\|\, p^{\text{ch}}])$ and $x_l^{\text{ch}} = x_l + g_l c_l$. Here, GAP denotes global average pooling, SEBlock generates the channel modulation vector c_l, and x_l^{sp} and x_l^{ch} are the resulting features. The two modes are used jointly.

4.3 Contrastive Modality Alignment

Contrastive Modality Alignment (CMA) projects semantic prompts into the visual space and aligns them with prompt-enhanced visual features at the class level. For the k-th support image of class i,

its prompt embedding $s_{i,k}$ is transformed by a lightweight bottleneck adapter:

$$\hat{s}_{i,k} = W_2\,\mathrm{GELU}(W_1 s_{i,k}), \tag{7}$$

where $W_1 \in \mathbb{R}^{r\times d_t}$ and $W_2 \in \mathbb{R}^{d_v\times r}$ are learnable projections, GELU is the activation function, and r, d_t, and d_v denote the bottleneck, textual, and visual dimensions, respectively. This design uses fewer parameters and provides stronger regularization than a full-width projector, which is beneficial under limited supervision.

Given an N-way K-shot episode, CMA averages the support representations into a visual prototype $v_i = K^{-1}\sum_{k=1}^{K} f(x_{i,k})$ and a prompt prototype $p_i = K^{-1}\sum_{k=1}^{K} \hat{s}_{i,k}$, where $f(\cdot)$ is the visual encoder with semantic prompt injection.

CMA treats (p_i, v_i) as a positive pair and (p_i, v_j) for $j \neq i$ as negatives. With cosine similarity and temperature τ, the class-level InfoNCE loss is

$$\mathcal{L}_{\mathrm{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(p_i, v_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(p_i, v_j)/\tau)}, \tag{8}$$

which aligns corresponding prompt and visual prototypes while increasing inter-class separation. Finally, the alignment loss is combined with the cross-entropy loss to form the overall training objective:

$$\mathcal{L}_{total} = -\frac{1}{|\mathcal{Q}|}\sum_{(q_i,y_i)\in\mathcal{Q}}\log P(y_i \mid q_i) + \lambda\mathcal{L}_{\mathrm{InfoNCE}}, \tag{9}$$

where y_i is the ground-truth label of query sample q_i, $|\mathcal{Q}| = NM$, and λ controls the contribution of the InfoNCE loss. During episodic inference, semantic prompts enhance only labeled support samples to form class prototypes, using language priors from known support labels to shift data-scarce, unstable prototypes toward more reliable positions. The same backbone encodes queries without prompt injection and classifies them by prototype similarity without accessing query labels.

5 Experiments

5.1 Experimental Details

Datasets. We evaluate MAD-SP on eight benchmarks spanning standard and cross-domain few-shot classification. For standard FSL, we use miniImageNet (Vinyals et al., 2016) and tieredImageNet (Ren et al., 2018), two ImageNet subsets, together with CIFAR-FS (Bertinetto et al., 2019) and FC100 (Oreshkin et al., 2018) from CIFAR-100. For cross-domain FSL, we train on miniImageNet base classes and evaluate on CUB (Wah et al., 2011), Places (Zhou et al., 2018), Plantae (Van Horn et al., 2018), and ChestX (Wang et al., 2017), covering fine-grained recognition, scene understanding, and medical imaging. Dataset splits and statistics are provided in Appendix A.3.

Implementation Details. We use Visformer-T (Chen et al., 2021b) as the visual backbone and Qwen3-VL-8B (Bai et al., 2025) as the shared foundation model for the generator, critic, and refiner agents. Semantic prompts are encoded by the CLIP ViT-B/32 text encoder (Radford et al., 2021), which produces 512-dimensional representations. Training consists of a pre-training stage followed by episodic meta-training (Dong et al., 2022). Pre-training lasts 300 epochs for tieredImageNet and 800 epochs for the other source datasets, with a batch size of 512. The model is then meta-trained on the corresponding source dataset for 100 epochs. We use AdamW (Loshchilov & Hutter, 2019) with an initial learning rate of 5×10^{-4} and a weight decay of 5×10^{-2}; the learning rate is warmed up at the beginning of training and subsequently decayed using a cosine schedule (Loshchilov & Hutter, 2017). We set ω, δ, ξ, τ, and λ to 0.3, 0.3, 0.5, 0.2, and 0.2, respectively. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU.

Evaluation Protocol. We follow the standard episodic evaluation paradigm (Chen et al., 2023; Xing et al., 2019; Li et al., 2020; Xu & Le, 2022; Zhang et al., 2023) and evaluate the proposed method under the 5-way 1-shot and 5-way 5-shot settings. Each episode contains 15 query images per class. We randomly generate 2,000 test episodes and report the accuracy averaged over all episodes, together with the 95% confidence interval.

Table 1: Few-shot classification results (%) on miniImageNet and tieredImageNet. We report the mean accuracy with the 95% confidence interval. The best and second-best results are shown in bold and underlined, respectively.

Model	Backbone	miniImageNet		tieredImageNet	
		1-shot	5-shot	1-shot	5-shot
SUN (Dong et al., 2022)	Visformer-S	67.80 ± 0.45	83.25 ± 0.30	72.99 ± 0.50	86.74 ± 0.33
SVAE (Xu & Le, 2022)	ResNet-12	74.84 ± 0.23	83.28 ± 0.40	76.98 ± 0.65	85.77 ± 0.50
Meta-AdaM (Sun & Gao, 2023)	ResNet-12	59.89 ± 0.49	77.92 ± 0.43	65.31 ± 0.48	85.24 ± 0.35
NIW-Meta (Kim & Hospedales, 2024)	WRN28-10	68.54 ± 0.26	84.81 ± 0.28	74.59 ± 0.33	89.76 ± 0.23
FeatWalk (Chen et al., 2024)	ResNet-12	70.21 ± 0.44	87.38 ± 0.27	75.25 ± 0.48	89.92 ± 0.29
MatchingNet (Vinyals et al., 2016)	ResNet-12	65.64 ± 0.20	78.72 ± 0.15	68.50 ± 0.92	80.60 ± 0.71
ProtoNet (Snell et al., 2017)	ResNet-12	62.39 ± 0.21	80.53 ± 0.14	68.23 ± 0.23	84.03 ± 0.16
AM3 (Xing et al., 2019)	ResNet-12	65.30 ± 0.49	78.10 ± 0.36	69.08 ± 0.47	82.58 ± 0.31
DeepEMD (Zhang et al., 2020)	ResNet-12	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
PCPK (Zhang et al., 2021)	ResNet-12	73.13 ± 0.85	82.06 ± 0.54	81.04 ± 0.89	87.42 ± 0.57
FewTURE (Hiller et al., 2022)	ViT-S/16	68.02 ± 0.88	84.51 ± 0.53	72.96 ± 0.92	86.43 ± 0.67
ProtoDiff (Du et al., 2023)	ResNet-12	66.63 ± 0.21	83.48 ± 0.15	72.95 ± 0.24	85.15 ± 0.18
CPEA (Hao et al., 2023)	ViT-S/16	71.97 ± 0.65	87.06 ± 0.38	76.93 ± 0.70	90.12 ± 0.45
SP (Chen et al., 2023)	Visformer-T	72.31 ± 0.40	83.42 ± 0.30	78.03 ± 0.46	88.55 ± 0.32
SIFT (Pan et al., 2024)	WRN-28-10	77.31 ± 0.67	86.95 ± 0.53	77.86 ± 0.77	89.89 ± 0.52
SemFew (Zhang et al., 2024a)	Swin-T	78.94 ± 0.66	86.49 ± 0.50	82.37 ± 0.77	89.89 ± 0.52
ECER (Liu et al., 2025)	Visformer-T	81.14 ± 0.15	–	81.81 ± 0.51	–
DVLA-RL (Li et al., 2026)	Visformer-T	81.69 ± 0.36	88.25 ± 0.28	83.02 ± 0.43	91.71 ± 0.29
MAD-SP	Visformer-T	$\mathbf{85.98 \pm 0.37}$	$\mathbf{89.39 \pm 0.28}$	$\mathbf{89.51 \pm 0.40}$	$\mathbf{92.27 \pm 0.34}$

Table 2: Cross-domain few-shot classification results (%) when training on miniImageNet and evaluating on CUB, Places, Plantae, and ChestX. The best and second-best results are shown in bold and underlined, respectively.

Method	CUB		Places		Plantae		ChestX	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
FWT (Tseng et al., 2020)	45.50	64.97	53.44	70.70	32.56	49.66	22.04	25.18
AFA (Hu & Ma, 2022)	46.86	68.25	54.04	76.21	36.76	54.26	22.92	25.02
UCD (Oh et al., 2022)	40.65	58.54	51.84	72.19	37.28	54.15	22.64	26.26
ATA (Wang & Deng, 2021)	45.00	66.22	53.57	75.48	34.42	52.69	22.10	24.32
StyleAdv (Fu et al., 2023)	48.49	68.72	58.58	77.73	41.13	<u>61.52</u>	22.64	26.07
FAP (Zhang et al., 2024b)	50.56	64.17	57.34	72.05	37.44	53.58	21.56	24.15
FLoR (Zou et al., 2024)	49.99	70.39	53.18	72.31	40.10	55.80	23.11	26.70
MEFP (Zhou et al., 2024)	51.55	73.61	52.06	73.78	<u>41.55</u>	61.39	22.82	26.53
SVasP (Li et al., 2025)	49.49	68.95	59.07	77.78	41.22	60.63	23.23	26.87
GNN (Garcia & Bruna, 2018)	44.40	62.87	52.42	70.91	33.60	48.51	22.00	25.27
LDP-net (Zhou et al., 2023)	49.82	70.39	53.82	72.90	39.84	58.49	23.01	26.67
DVLA-RL (Li et al., 2026)	<u>67.46</u>	<u>78.99</u>	<u>69.26</u>	<u>80.70</u>	–	–	**23.47**	<u>26.94</u>
MAD-SP	**67.85**	**81.44**	**75.31**	**82.14**	**46.51**	**61.88**	<u>23.39</u>	**27.11**

5.2 Main Results

Tables 1–3 summarize the standard and cross-domain few-shot classification results. Methods above and below the internal horizontal rule correspond to *Optimization-/Adaptation-based Methods* and *Metric-/Prototype-based Methods*, respectively.

Few-Shot Classification. Tables 1 and 3 compare MAD-SP with representative methods under the 5-way 1-shot and 5-shot settings. MAD-SP establishes new state-of-the-art results on all four benchmarks. Compared with the previous best methods, it improves the 1-shot/5-shot accuracy by 4.29/1.14 points on miniImageNet, 6.49/0.56 on tieredImageNet, 2.18/1.44 on CIFAR-FS, and 2.25/3.39 on FC100, yielding an average gain of 2.72 points across the eight settings. The strong 1-shot improvements on miniImageNet, tieredImageNet, and CIFAR-FS show that reliable semantic prompts effectively compensate for extremely limited visual evidence. The consistent 5-shot gains indicate that the advantage remains as more support samples

Table 3: Few-shot classification results (%) on CIFAR-FS and FC100. The best and second-best results are shown in bold and underlined, respectively.

Model	CIFAR-FS		FC100	
	1-shot	5-shot	1-shot	5-shot
MABAS (Kim et al., 2020)	73.51	85.65	42.31	58.16
Meta-AdaM (Sun & Gao, 2023)	–	–	41.12	56.14
ALFA (Baik et al., 2024)	76.32	86.73	44.54	58.44
LastShot (Ye et al., 2024)	76.76	87.49	44.08	59.14
ProtoNet (Snell et al., 2017)	72.20	83.50	41.54	57.08
FewTURE (Hiller et al., 2022)	77.76	88.90	47.68	63.81
SP (Chen et al., 2023)	82.18	88.24	48.53	61.55
SemFew (Zhang et al., 2024a)	84.34	89.11	54.27	<u>65.02</u>
ECER (Liu et al., 2025)	86.01	–	<u>57.34</u>	–
DVLA-RL (Li et al., 2026)	<u>87.18</u>	<u>90.59</u>	–	–
MAD-SP	**89.36**	**92.03**	**59.59**	**68.41**

Table 4: Ablation study of MACR, DPI, and CMA on two standard few-shot benchmarks and the cross-domain CUB benchmark. We report the mean accuracy with the 95% confidence interval, with the best results shown in bold.

MACR	DPI	CMA	miniImageNet		CIFAR-FS		CUB	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
			73.43 ± 0.45	82.61 ± 0.30	77.18 ± 0.36	86.13 ± 0.28	54.52 ± 0.55	75.91 ± 0.35
✓			80.11 ± 0.41	83.12 ± 0.29	83.91 ± 0.34	87.04 ± 0.26	62.15 ± 0.49	77.10 ± 0.34
	✓		80.08 ± 0.41	86.34 ± 0.29	84.22 ± 0.35	90.07 ± 0.26	60.70 ± 0.48	79.46 ± 0.35
✓	✓		84.72 ± 0.38	88.92 ± 0.28	88.02 ± 0.33	91.77 ± 0.26	66.95 ± 0.46	81.08 ± 0.34
	✓	✓	81.17 ± 0.41	87.89 ± 0.30	84.74 ± 0.35	90.36 ± 0.25	61.97 ± 0.48	80.15 ± 0.35
✓	✓	✓	$\mathbf{85.98 \pm 0.37}$	$\mathbf{89.39 \pm 0.28}$	$\mathbf{89.36 \pm 0.33}$	$\mathbf{92.03 \pm 0.24}$	$\mathbf{67.85 \pm 0.46}$	$\mathbf{81.44 \pm 0.33}$

Table 5: Comparison of different semantic prompts (%) on miniImageNet, CIFAR-FS, and CUB under the 1-shot and 5-shot settings. We report the mean accuracy with the 95% confidence interval, with the best results shown in bold.

Prompt	miniImageNet		CIFAR-FS		CUB	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Name	81.17 ± 0.41	87.89 ± 0.30	84.74 ± 0.35	90.36 ± 0.25	61.97 ± 0.48	80.15 ± 0.35
Definition	83.01 ± 0.39	88.35 ± 0.31	85.53 ± 0.37	90.84 ± 0.26	63.72 ± 0.48	80.78 ± 0.35
SemEvo	84.06 ± 0.41	88.69 ± 0.32	86.58 ± 0.37	91.52 ± 0.25	65.95 ± 0.49	80.99 ± 0.35
MACR	$\mathbf{85.98 \pm 0.37}$	$\mathbf{89.39 \pm 0.28}$	$\mathbf{89.36 \pm 0.33}$	$\mathbf{92.03 \pm 0.24}$	$\mathbf{67.85 \pm 0.46}$	$\mathbf{81.44 \pm 0.33}$

become available. Moreover, although MAD-SP employs the lightweight Visformer-T backbone with only 10.0M parameters, it outperforms several methods using larger backbones, confirming that the gains do not merely result from increased model capacity. These results demonstrate the benefit of combining visually grounded caption refinement with instance-adaptive prompt injection.

Cross-Domain Classification. Following the standard cross-domain protocol, we train on miniImageNet and evaluate on CUB, Places, Plantae, and ChestX. As shown in Table 2, MAD-SP achieves state-of-the-art performance in seven of the eight settings, improving the previous best results by 0.17–6.05 percentage points, and ranks second on ChestX 1-shot by only 0.08 points. Across all eight settings, MAD-SP achieves an average margin of 1.97 percentage points over the strongest competing results. This consistent performance across diverse target domains demonstrates the strong cross-domain generalization of MAD-SP. MACR generates visually grounded captions through multi-agent refinement, while CMA and DPI align and adaptively integrate these transferable semantics under domain shifts.

5.3 Ablation Studies

Table 4 evaluates the three components of MAD-SP. When MACR is disabled, class names are used as semantic prompts; when DPI is disabled, prompts are injected at the stage3-2 block (the third Transformer layer of stage 3); when CMA is disabled, the dimensionality-matching projector is retained but the InfoNCE loss is removed. (1) MACR improves the baseline by an average of 7.01 and 0.87 percentage points across the three benchmarks under the 1-shot and 5-shot settings, respectively, showing the effectiveness of refined, visually grounded captions. On cross-domain CUB, the gains reach 7.63 and 1.19 points, indicating that such captions provide reliable semantic cues when visual representations are affected by domain shift. (2) DPI yields average gains of 6.62 and 3.74 points in the 1-shot and 5-shot settings, respectively. Adding CMA further improves DPI across all settings, validating adaptive prompt injection and the supporting role of cross-modal alignment. (3) Combining MACR, DPI, and CMA achieves the best performance throughout, demonstrating their complementarity in constructing, aligning, and utilizing semantic prompts.

5.4 Model Analysis

Comparison with Different Semantic Prompts. To evaluate the effect of textual prompt quality, we compare MACR with class names (Name), short definitions (Definition), and Semantic Evolution (SemEvo) (Zhang et al., 2024a) while keeping the remaining framework unchanged. As shown in Table 5, accuracy consistently increases from Name to MACR across all datasets and shot settings. MACR achieves the best results throughout, demonstrating that score-guided multi-agent refinement produces more visually grounded and discriminative semantics than class-level or directly generated

Table 6: Accuracy (%) for fixed, random, and dynamic prompt injection on miniImageNet, CIFAR-FS, and CUB in 1-shot and 5-shot settings. Values are means with 95% confidence intervals. S2-L1 denotes the first layer of stage 2; the best results are bold.

Injection	miniImageNet		CIFAR-FS		CUB	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
S2-L1	55.76 ± 0.66	67.81 ± 0.54	53.74 ± 0.58	71.85 ± 0.47	34.93 ± 0.81	66.33 ± 0.64
S2-L2	54.15 ± 0.66	68.17 ± 0.55	55.04 ± 0.58	72.89 ± 0.48	34.77 ± 0.81	66.59 ± 0.65
S2-L3	61.28 ± 0.65	71.72 ± 0.53	56.87 ± 0.57	74.67 ± 0.46	39.25 ± 0.80	69.07 ± 0.62
S2-L4	65.76 ± 0.62	72.51 ± 0.51	71.39 ± 0.55	77.86 ± 0.45	48.75 ± 0.76	69.62 ± 0.60
S3-L1	76.62 ± 0.44	82.37 ± 0.40	78.94 ± 0.39	85.64 ± 0.35	57.96 ± 0.54	76.53 ± 0.47
S3-L2	77.31 ± 0.45	82.66 ± 0.40	80.10 ± 0.40	86.01 ± 0.35	58.88 ± 0.56	76.73 ± 0.47
S3-L3	80.75 ± 0.41	83.92 ± 0.39	84.63 ± 0.36	87.38 ± 0.34	62.87 ± 0.51	77.61 ± 0.46
S3-L4	78.13 ± 0.44	83.58 ± 0.39	83.02 ± 0.39	87.34 ± 0.34	60.75 ± 0.54	77.37 ± 0.46
Random (S2+S3)	67.30 ± 0.53	80.01 ± 0.40	67.51 ± 0.47	83.05 ± 0.35	47.58 ± 0.65	74.87 ± 0.47
Random (S3)	72.96 ± 0.47	81.62 ± 0.37	81.02 ± 0.42	86.42 ± 0.32	57.17 ± 0.58	76.00 ± 0.44
DPI (S2+S3)	82.34 ± 0.40	88.13 ± 0.30	86.40 ± 0.36	91.10 ± 0.26	64.55 ± 0.50	80.56 ± 0.35
DPI (S3)	**85.98 ± 0.37**	**89.39 ± 0.28**	**89.36 ± 0.33**	**92.03 ± 0.24**	**67.85 ± 0.46**	**81.44 ± 0.33**

descriptions. The improvement is more pronounced in the 1-shot setting, confirming that reliable semantic guidance is particularly valuable when visual evidence is extremely scarce.

Comparison of Different Prompt Injection Layers. Visformer-T stages 2 and 3 each contain four Transformer layers. Table 6 compares fixed, random, and dynamic injection over stages 2–3 or stage 3 on three datasets in both shot settings. Stage-3 layers generally outperform stage-2 layers, although the last is slightly weaker because it leaves less depth for prompt interaction. DPI surpasses random selection over both candidate ranges in all six settings, and stage-3 DPI performs best throughout, supporting the exclusion of poorly matched early features.

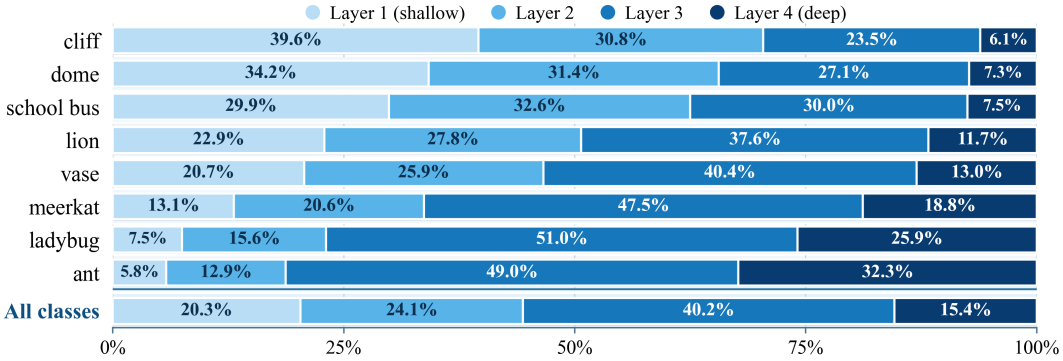


Figure 3: Distribution of instance-level prompt injection decisions aggregated over images from eight representative miniImageNet categories and the full dataset. Layers 1–4 denote the four Transformer layers in stage 3 of Visformer-T, ordered from shallow to deep.

Distribution of Instance-Level Injection Decisions. Figure 3 aggregates per-image layer selections within the displayed categories. The results reveal a scale-related tendency, not a class-conditioned routing rule: 70.4% of cliff images and 65.6% of dome images are routed to shallow Layers 1–2, whereas 76.9% of ladybug images and 81.3% of ant images select deeper Layers 3–4. Shallow interaction preserves structural cues already evident for large targets, while deeper representations provide stronger semantic abstraction and contextual aggregation for small foreground objects. Although Layer 3 is selected most often overall (40.2%), the different aggregate distributions show that DPI responds to instance-specific visual scale instead of collapsing to a fixed layer.

6 Conclusion

We presented MAD-SP, a unified framework for constructing and utilizing semantic prompts in few-shot image classification. MACR performs score-guided multi-agent caption refinement, while DPI and CMA enable instance-adaptive injection and cross-modal alignment. Across four standard and four cross-domain benchmarks, MAD-SP achieves state-of-the-art results in all eight standard settings and seven of eight cross-domain settings, with average gains of 2.72 and 1.97 points. Ablation studies and model analyses validate the components' effectiveness and complementarity.

AI USE STATEMENT

In this work, we used generative AI tools for translation and to assist with the qualitative analysis of experimental results. We did not use generative AI tools for any other tasks with required disclosure; the remaining required-disclosure tasks were either completed solely by the authors or were not applicable to this work. Additionally, we used generative AI tools to improve the clarity and readability of the manuscript, to write or revise supporting software code, and to format the LaTeX source. AI-assisted code use did not include implementation of the proposed method. We have reviewed all AI-assisted work: translated and revised text was checked by the authors, qualitative analyses were verified against the underlying experimental results, and code and LaTeX changes were reviewed before inclusion. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Sungyong Baik, Myungsub Choi, Janghoon Choi, Heewon Kim, and Kyoung Mu Lee. Learning to learn task-adaptive hyperparameters for few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1441–1454, 2024. doi: 10.1109/TPAMI.2023.3261387.

Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14493–14502, 2020.

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxnZh0ct7.

Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. MADTP: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15710–15719, 2024.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations*, 2024.

Dalong Chen, Jianjia Zhang, Wei-Shi Zheng, and Ruixuan Wang. FeatWalk: Enhancing few-shot classification through local view leveraging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1019–1027, 2024. doi: 10.1609/aaai.v38i2.27862.

Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan. Semantic prompt for few-shot image recognition. *arXiv preprint arXiv:2303.14123*, 2023. doi: 10.48550/arXiv.2303.14123.

Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9062–9071, 2021a.

Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 589–598, 2021b.

Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Self-promoted supervision for few-shot transformer. In *European Conference on Computer Vision*, pp. 329–347, 2022.

Yingjun Du, Zehao Xiao, Shengcai Liao, and Cees G. M. Snoek. ProtoDiff: Learning to learn prototypical networks by task-guided diffusion. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46304–46322, 2023.

Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. StyleAdv: Meta style adversarial training for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24575–24584, 2023.

Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024.

Yongxin Guo, Zhenglin Cheng, Xiaoying Tang, Zhaopeng Tu, and Tao Lin. Dynamic mixture of experts: An auto-tuning approach for efficient transformer models. In *International Conference on Learning Representations*, 2025.

Fusheng Hao, Fengxiang He, Liu Liu, Fuxiang Wu, Dacheng Tao, and Jun Cheng. Class-aware patch embedding adaptation for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18905–18915, 2023.

Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond. Rethinking generalization in few-shot classification. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3582–3595, 2022.

Yanxu Hu and Andy J. Ma. Adversarial feature augmentation for cross-domain few-shot classification. In *European Conference on Computer Vision*, pp. 20–37, 2022.

Wenxuan Huang, Zijie Zhai, Yunhang Shen, Shaosheng Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, and Shaohui Lin. Dynamic-LLaVA: Efficient multimodal large language models via dynamic vision-language context sparsification. In *International Conference on Learning Representations*, 2025.

Jaekyeom Kim, Hyoungseok Kim, and Gunhee Kim. Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning. In *European Conference on Computer Vision*, pp. 599–617, 2020.

Minyoung Kim and Timothy M. Hospedales. A hierarchical bayesian model for few-shot meta learning. In *International Conference on Learning Representations*, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12576–12584, 2020.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008, 2023.

Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1–10, 2019.

Wenhao Li, Xianjing Meng, Qiangchang Wang, Zhongyi Han, Zhibin Wu, and Yilong Yin. DVLA-RL: Dual-level vision-language alignment with reinforcement learning gating for few-shot learning. In *International Conference on Learning Representations*, 2026.

Wenqian Li, Pengfei Fang, and Hui Xue. SVasP: Self-versatility adversarial style perturbation for cross-domain few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15275–15283, 2025. doi: 10.1609/aaai.v39i15.33676.

Mushui Liu, Fangtai Wu, Bozheng Li, Ziqian Lu, Yunlong Yu, and Xi Li. Envisioning class entity reasoning by large language models for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18906–18914, 2025. doi: 10.1609/aaai.v39i18.34081.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yaxin Luo, Gen Luo, Jiayi Ji, Yiyi Zhou, Xiaoshuai Sun, Zhiqiang Shen, and Rongrong Ji. γ-MoD: Exploring mixture-of-depth adaptation for multimodal large language models. In *International Conference on Learning Representations*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023.

Jaehoon Oh, Sungnyun Kim, Namgyu Ho, Jin-Hwa Kim, Hwanjun Song, and Se-Young Yun. Understanding cross-domain few-shot learning based on domain similarity and few-shot difficulty. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2622–2636, 2022.

Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Mei-Hong Pan, Hong-Yi Xin, and Hong-Bin Shen. Semantic-based implicit feature transform for few-shot classification. *International Journal of Computer Vision*, 132(11):5014–5029, 2024. doi: 10.1007/s11263-024-02113-8.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations*, 2018.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Mingchen Song, Huiqiang Wang, and Guoqiang Zhong. Self-prompt mechanism for few-shot image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4934–4942, 2024. doi: 10.1609/aaai.v38i5.28297.

Siyuan Sun and Hongyang Gao. Meta-AdaM: A meta-learned adaptive optimizer with momentum for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 65441–65455, 2023.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Quan Tang, Bowen Zhang, Jiajun Liu, Fagui Liu, and Yifan Liu. Dynamic token pruning in plain vision transformers for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 777–786, 2023.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *Proceedings of the European Conference on Computer Vision*, pp. 266–282, 2020.

Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *International Conference on Learning Representations*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pp. 3630–3638, 2016.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Haoqing Wang and Zhi-Hong Deng. Cross-domain few-shot classification via adversarial task augmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 1075–1081, 2021. doi: 10.24963/ijcai.2021/149.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.

Liwen Wu, Wei Wang, Lei Zhao, Zhan Gao, Qika Lin, Shaowen Yao, Zuozhu Liu, and Bin Pu. MPA: Multimodal prototype augmentation for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26965–26973, 2026. doi: 10.1609/aaai.v40i32.39909.

Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O. O. Pinheiro. Adaptive cross-modal few-shot learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jingyi Xu and Hieu Le. Generating representative samples for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9003–9013, 2022.

Yalong Xu, Lin Zhao, Chen Gong, Guangyu Li, Di Wang, and Nannan Wang. DynPose: Largely improving the efficiency of human pose estimation by a simple dynamic framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1160–1169, 2025.

Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8808–8817, 2020.

Han-Jia Ye, Lu Ming, De-Chuan Zhan, and Wei-Lun Chao. Few-shot learning with a strong teacher. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1425–1440, 2024. doi: 10.1109/TPAMI.2022.3160362.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. ATP-LLaVA: Adaptive token pruning for large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24972–24982, 2025.

Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. Prototype completion with primitive knowledge for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3754–3762, 2021.

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12203–12213, 2020.

Hai Zhang, Junzhe Xu, Shanlin Jiang, and Zhenan He. Simple semantic-aided few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28588–28597, 2024a.

Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15211–15222, 2023.

Tiange Zhang, Qing Cai, Feng Gao, Lin Qi, and Junyu Dong. Exploring cross-domain few-shot classification via frequency-aware prompting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5490–5498, 2024b. doi: 10.24963/ijcai.2024/607.

Wangbo Zhao, Jiasheng Tang, Yizeng Han, Yibing Song, Kai Wang, Gao Huang, Fan Wang, and Yang You. Dynamic tuning towards parameter and inference efficiency for ViT adaptation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 114765–114796, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, and Yanning Zhang. Revisiting prototypical network for cross domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20061–20070, 2023.

Fei Zhou, Peng Wang, Lei Zhang, Zhenghua Chen, Wei Wei, Chen Ding, Guosheng Lin, and Yanning Zhang. Meta-exploiting frequency prior for cross-domain few-shot learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 116783–116814, 2024.

Chen Zhu, Wangbo Zhao, Huiwen Zhang, Yuhao Zhou, Weidong Tang, Shuo Wang, Zhihang Yuan, Yuzhang Shang, Xiaojiang Peng, Kai Wang, and Dawei Yang. EA-ViT: Efficient adaptation for elastic vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1038–1047, 2025.

Yixiong Zou, Yicong Liu, Yiman Hu, Yuhua Li, and Ruixuan Li. Flatten long-range loss landscapes for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23575–23584, 2024.

A Appendix

A.1 Training Procedure

Algorithm 1: Training procedure of the proposed MAD-SP.

Input: Base training set $\mathcal{D}_{base}$; visual backbone f_θ; frozen CLIP encoders E_I, E_T; frozen Generator, Critic, and Refiner agents $\mathcal{G}, \mathcal{C}, \mathcal{R}$; DPI selector D_ϕ; CMA adapter A_η
Output: Trained f_θ, D_ϕ, and A_η

Pre-train f_θ on $\mathcal{D}_{base}$ using supervised classification.
for all labeled training images $(x, c) \in \mathcal{D}_{base}$ **do**
  Generate $p_0 = \mathcal{G}(x, c)$, compute its scores using Eqs. 3–5, and set $r \leftarrow 0$.
  while $S(x, c, p_r) < \delta$ and $r < R_{\max}$ **do**
    Obtain feedback from $\mathcal{C}$, refine $p_{r+1} = \mathcal{R}(x, c, p_r)$, rescore it, and set $r \leftarrow r + 1$.
  end while
  Retain the accepted caption, or the highest-scoring candidate at $R_{\max}$, and cache $t = E_T(p^\star)$.
end for
while not converged **do**
  Sample an N-way K-shot episode $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$ from $\mathcal{D}_{base}$.
  for all support images $x_{i,k} \in \mathcal{S}$ **do**
    Retrieve $t_{i,k}$ and obtain the aligned prompt $\hat{s}_{i,k} = A_\eta(t_{i,k})$ using Eq. 7.
    Select $l^\star = D_\phi(x_{i,k}, t_{i,k})$ using Eq. 6 and extract $z^s_{i,k} = f_\theta(x_{i,k}; t_{i,k}, l^\star)$.
  end for
  Construct $v_i = K^{-1} \sum_k z^s_{i,k}$ as in Eq. 1 and $p_i = K^{-1} \sum_k \hat{s}_{i,k}$.
  Compute $\mathcal{L}_{\text{InfoNCE}}$ from $\{(p_i, v_i)\}_{i=1}^N$ using Eq. 8.
  for all query images $q_j \in \mathcal{Q}$ **do**
    Encode $z^q_j = f_\theta(q_j)$ without prompt injection and classify it using Eq. 2.
  end for
  Compute $\mathcal{L}_{total}$ using Eq. 9 and update θ, ϕ, and η by backpropagation.
end while

Algorithm 1 summarizes the two-stage optimization of MAD-SP. The visual backbone is first pre-trained on base classes, while MACR generates and caches a refined semantic prompt for each labeled image independently of gradient optimization. During episodic meta-training, DPI selects an injection layer for every support image and CMA aligns the resulting visual and prompt prototypes. Query images are encoded without semantic prompts and classified against the support-derived visual prototypes. The classification and InfoNCE losses jointly optimize the backbone, DPI selector, and alignment adapter, while the LLM agents and CLIP encoders remain frozen.

A.2 MACR Text Generation Scheme

Class names provide a semantic anchor for recognition, but expanding them with an LLM may introduce visually unsupported attributes. MACR conditions caption generation on both the image and its class name, and separates generation, diagnosis, and revision into three agent roles. A frozen CLIP ViT-B/32 model evaluates each candidate to guide refinement, yielding an instance-level semantic prompt grounded in the input image.

Agent roles and outputs. All three agents share one frozen Qwen3-VL-8B-Instruct model (Bai et al., 2025) with distinct system prompts. Each call includes the same image and the relevant caption, scores, or feedback; no persistent conversation history is used. The roles are defined as follows:

- **Generator Agent:** Takes the image x and class name c and produces an initial caption p_0: one English sentence describing visible attributes such as color, shape, texture, parts, and pose.
- **Critic Agent:** Inspects the image, current caption, target class, CLIP scores, and hardest negative class. It returns feedback f_r in a requested JSON format covering unsupported claims, missing visible attributes, redundancy, class confusion, and revision instructions, without rewriting the caption.
- **Refiner Agent:** Uses the image, current caption, target class, alignment and discriminative scores, hardest negative class, and feedback f_r to produce a revised sentence p_r. It

preserves correct details, removes unsupported statements, and adds visible discriminative attributes.

Score-guided coordination. MACR scores p_0 using the frozen CLIP encoders. The candidate set $\mathcal{C}$ defaults to all distinct class names in the labeled data manifest and can be restricted to an episode through an explicit list. Class names are encoded directly, without a text template. For caption p_{r-1}, the hardest negative is

$$c_r^- = \arg\max_{c' \in \mathcal{C} \setminus \{c\}} \cos\big(E_T(p_{r-1}), E_T(c')\big), \tag{10}$$

where $\mathcal{C}$ contains c and at least one negative class. The critic and refiner receive c_r^-, not the full candidate list. A caption is accepted immediately when its combined score reaches δ; otherwise, the critic diagnoses it, the refiner revises it, and CLIP scores the revision. After at most three refinement rounds, the highest-scoring candidate is retained if none meets the threshold; ties favor the earlier caption. This fallback remains marked as below threshold. Each image therefore requires one to seven LLM calls. Only the selected caption enters semantic encoding; scores and feedback are retained for analysis.

Generation settings. Table 7 lists the implementation defaults. All agents use greedy decoding (temperature 0; nucleus sampling inactive). Generator and refiner outputs are cleaned of formatting wrappers and truncated to 45 whitespace-delimited words before scoring. Word counts differ from token counts: CLIP separately truncates text at its configured context limit. The critic has no additional word limit. Its JSON schema is prompt-specified, without constrained decoding; nonempty malformed responses remain textual feedback. The caption-scoring weight and acceptance threshold are set to $\omega = 0.3$ and $\delta = 0.3$, respectively.

Table 7: Default generation settings for the three MACR agents. Token budgets count newly generated Qwen tokens and exclude the input image and prompt.

Agent	Maximum new tokens	Caption word limit	Output
Generator	96	45	Initial caption p_0
Critic	320	Not applicable	Structured feedback f_r
Refiner	96	45	Revised caption p_r

Prompt templates. The templates below reproduce the implemented instructions with runtime placeholders and the default word limit. Each user message also contains the image; scores use six decimal places. In round r, {`caption`} and {`feedback`} denote p_{r-1} and f_r.

Prompt Template for the Generator Agent

System message
You are the Generator Agent in a scientific image-captioning pipeline. Describe only visible evidence. Your caption will be used as a semantic prompt for fine-grained image classification, so it must be concise, image-faithful, and class-discriminative. Do not invent hidden properties, context, identity, or uncertain attributes. Return only one English caption without a prefix, analysis, markdown, or quotation marks.

User message
Target class: {`class_name`}
Write a single natural sentence that names the target class and describes the most useful visible color, shape, texture, parts, pose, and context when they are actually visible. Use at most 45 words. Avoid generic filler and comparisons not supported by the image.

Prompt Template for the Critic Agent

System message
You are the Critic Agent in a scientific image-captioning pipeline. Inspect the image, proposed caption, target class, competing classes, and objective CLIP scores. Diagnose faithfulness, specificity, and discriminativeness. Never rewrite the caption. Return exactly one JSON object with string-array fields: `unsupported_claims`, `missing_visible_attributes`, `generic_or_redundant_phrases`, `confusing_with`, and `revision_instructions`. Use empty arrays when there is no issue; do not add markdown fences.

User message
Refinement round: {iteration}
Target class: {class_name}
Caption: {caption}
Objective scores (higher is better):
- image-text alignment: {alignment_score}
- class discriminative margin: {discriminative_score}
- combined quality: {total_score}
- hardest negative class: {hardest_negative_class}
Identify concrete changes that can improve visual grounding and separation from the hardest negative while avoiding fabricated details.

Prompt Template for the Refiner Agent

System message
You are the Refiner Agent in a scientific image-captioning pipeline. Rewrite the proposed caption using the critic's feedback and the visible image. Preserve correct visual details, remove unsupported or redundant statements, add only clearly visible discriminative attributes, and remain consistent with the target class. Return only one concise English caption without a prefix, analysis, markdown, or quotation marks.

User message
Refinement round: {iteration}
Target class: {class_name}
Current caption: {caption}
Current alignment score: {alignment_score}
Current discriminative margin: {discriminative_score}
Hardest negative class: {hardest_negative_class}
Critic feedback: {feedback}
Return one revised sentence of at most 45 words.

Use in few-shot recognition. This label-conditioned scheme applies to training images and labeled support images, without accessing hidden query labels. Caption generation is independent of downstream gradient optimization; the selected text can be cached for subsequent semantic encoding in MAD-SP.

A.3 Dataset Details

Table 8 summarizes the class partitions and image counts of the eight benchmarks used in our experiments. Counts refer to the selected few-shot subsets, not necessarily the complete source collections. The datasets cover standard and cross-domain few-shot classification.

Table 8: Dataset statistics for standard (upper group) and cross-domain (lower group) evaluation. "Test" denotes the image pool for sampling both support and query sets.

Dataset	# Classes			# Images	
	$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{val}}$	$\mathcal{D}_{\text{test}}$	Test	Total
miniImageNet (Vinyals et al., 2016)	64	16	20	12,000	60,000
tieredImageNet (Ren et al., 2018)	351	97	160	206,209	779,165
CIFAR-FS (Bertinetto et al., 2019)	64	16	20	12,000	60,000
FC100 (Oreshkin et al., 2018)	60	20	20	12,000	60,000
CUB-200-2011 (Wah et al., 2011)	100	50	50	2,958	11,788
Places (Zhou et al., 2018)	183	91	91	18,200	73,000
Plantae (Van Horn et al., 2018)	100	50	50	17,253	68,461
ChestX (Wang et al., 2017)	–	–	7	25,847	25,847

Class columns describe benchmark partitions. Cross-domain source-model training uses only miniImageNet base classes, not the target-domain training/validation partitions. Dashes indicate that no training/validation partition is used for ChestX in this protocol.

Standard few-shot benchmarks. Training, validation, and test classes are disjoint.

miniImageNet (Vinyals et al., 2016) is an ImageNet subset containing 100 object classes and 600 images per class, for a total of 60,000 images. Its 64/16/20 class split yields 38,400 training, 9,600 validation, and 12,000 test images.

tieredImageNet (Ren et al., 2018) contains 779,165 images from 608 ImageNet classes, partitioned into 351 training, 97 validation, and 160 test classes. The split follows higher-level semantic groups to reduce semantic overlap between partitions. The test partition contains 206,209 images.

CIFAR-FS (Bertinetto et al., 2019) reorganizes CIFAR-100 (Krizhevsky, 2009) into a few-shot benchmark with 64/16/20 training/validation/test classes. Each of the 100 classes contains 600 images, giving 60,000 images overall and 12,000 test images. The original images have a resolution of 32×32 pixels.

FC100 (Oreshkin et al., 2018) also uses the 60,000 CIFAR-100 images, but partitions classes by semantic superclass. Training comprises 60 classes from 12 superclasses; validation and testing each comprise 20 classes from four separate superclasses. This separation limits semantic overlap across partitions. The test set contains 12,000 images.

Cross-domain benchmarks. The model trained on the 64 miniImageNet base classes is evaluated on the following target-domain test sets, spanning birds, scenes, plants, and radiographs.

CUB-200-2011 (CUB) (Wah et al., 2011) contains 11,788 images of 200 bird species. The adopted few-shot partition contains 100/50/50 training/validation/test classes, with 2,958 images in the test partition. Distinguishing these species requires fine-grained visual cues, such as plumage patterns and beak shape.

Places (Zhou et al., 2018) uses a scene-recognition subset with 73,000 images across 365 categories, divided into 183 training, 91 validation, and 91 test classes. Its 18,200 test images assess transfer from object-centered source images to categories defined by scene layout and environmental context.

Plantae (Van Horn et al., 2018) is a plant-species subset of iNaturalist containing 68,461 images from 200 classes. The 100/50/50 class partition provides 17,253 test images. Similar appearances across species make leaf shape, surface texture, and other botanical details relevant to recognition.

ChestX (Wang et al., 2017) contains 25,847 chest radiographs from seven categories for target-domain evaluation. Its grayscale medical images and pathology-related categories produce a substantial shift from the natural-image source domain.

A.4 More Discussion on MACR

Table 9: Internal ablation of MACR on miniImageNet and CIFAR-FS. "MACR w/o Scoring" performs one fixed Critic–Refiner round without CLIP-based scoring. Mean accuracy (%) is reported, with the best results shown in bold.

Variant	Generator	Critic	Refiner	Scoring	miniImageNet		CIFAR-FS	
					1-shot	5-shot	1-shot	5-shot
Generator Only	✓				82.35	88.27	85.76	90.44
MACR w/o Scoring	✓	✓	✓		84.62	89.02	88.41	91.87
Full MACR	✓	✓	✓	✓	**85.98**	**89.39**	**89.36**	**92.03**

Internal ablation. Table 9 decomposes the caption refinement pipeline. Applying one Critic–Refiner round without scoring consistently improves over direct generation, showing the benefit of multi-agent feedback. Full MACR further improves all settings and exceeds Generator Only by an average of 3.62 and 1.36 percentage points in the 1-shot and 5-shot settings, respectively. These gains indicate that score-guided collaboration corrects unreliable descriptions and favors more image-faithful, discriminative captions, thereby providing higher-quality semantic prompts.

Comparison of Caption Scoring Strategies. Table 10 shows that either score alone reduces accuracy. S_{align} favors visual consistency but weakens class separation, whereas S_{dis} favors class distinction without ensuring image faithfulness. Their combination performs best, confirming their complementarity.

Table 10: Comparison of caption scoring strategies (%) on miniImageNet and CIFAR-FS.

Scoring Strategy	miniImageNet		CIFAR-FS	
	1-shot	5-shot	1-shot	5-shot
S_{align}	85.22	89.01	88.68	91.84
S_{dis}	83.54	88.32	86.87	91.21
$S_{\text{align}} + S_{\text{dis}}$	**85.98**	**89.39**	**89.36**	**92.03**

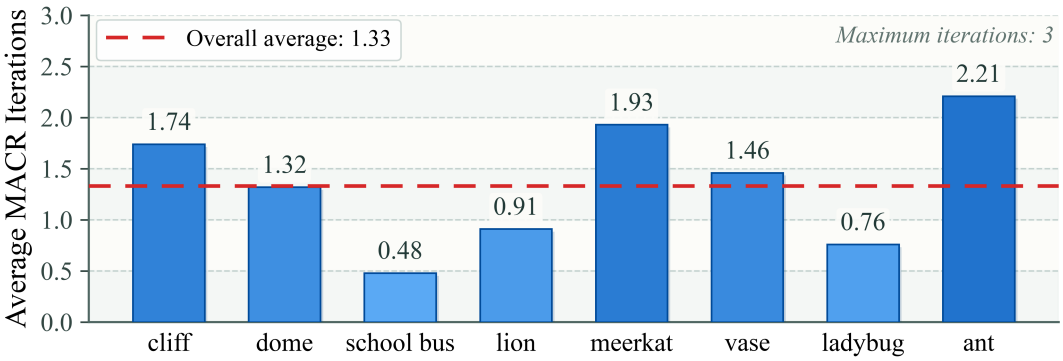


Figure 4: Average numbers of MACR refinement iterations for eight representative miniImageNet classes. The dashed line denotes the overall mean across all miniImageNet classes (1.33), and the maximum number of refinement rounds is three.



Figure 5: Representative images assigned by DPI to the four prompt-injection layers in stage 3 of Visformer-T, ordered from shallow to deep. Magenta overlays delineate the target objects for visual clarity.

Refinement statistics. Figure 4 shows that refinement demand varies with visual complexity. Images with stable, salient cues generally yield reliable initial captions and require fewer revisions. More refinement is needed under large intra-class variation, small or occluded objects, cluttered backgrounds, or fine-grained ambiguity, where unsupported context must be removed and discriminative details supplemented. Across all miniImageNet classes, MACR averages only 1.33 refinement rounds under a maximum of three, indicating that score-based early stopping concentrates additional LLM calls on difficult samples instead of refining every caption uniformly.

A.5 More Discussion on DPI

Qualitative Examples of Layer Selection. Figure 5 illustrates instance-level layer selection. Layer 1 receives large objects or global structures, including aircraft carriers, cliffs, domes, and school buses, whose structure is evident in shallow features. Layer 2 handles medium-scale animals and vehicles whose salient silhouettes require contextual aggregation. Layer 3 handles appearance-dependent objects requiring texture, part, and category cues. Layer 4 handles small or fine-grained targets, including ladybugs, spiders, mushrooms, and birds, whose limited extent demands semantic abstraction. Consistent with Fig. 3, these examples show DPI selects input-specific depths according to visual scale and recognition demands rather than using a fixed layer.

A.6 More Discussion on CMA

Table 11 compares three adapters for mapping textual prompts into the visual space. The bottleneck MLP performs best in all settings. Unlike the linear projector, it captures nonlinear cross-modal relations, while its low-dimensional hidden layer uses fewer parameters than the full-width MLP and provides stronger regularization. The linear projector lacks sufficient capacity to bridge the nonlinear modality gap, whereas the full-width MLP introduces more task-specific parameters and is therefore more susceptible to overfitting. The bottleneck design provides a favorable balance between expressive alignment and parameter efficiency, making it more suitable for few-shot learning.

A.7 Hyperparameter Analysis

Table 11: Comparison of alignment adapters for mapping textual prompts into the visual space (%) on miniImageNet and CIFAR-FS.

Adapter	miniImageNet		CIFAR-FS	
	1-shot	5-shot	1-shot	5-shot
Linear Projector	84.98	88.94	88.32	91.71
Full-width MLP	85.51	89.16	88.73	91.85
Bottleneck MLP	**85.98**	**89.39**	**89.36**	**92.03**

Table 12: Effect of the caption-scoring balance weight ω on miniImageNet and CIFAR-FS. Mean accuracy (%) is reported.

Value (ω)	miniImageNet		CIFAR-FS	
	1-shot	5-shot	1-shot	5-shot
0.1	85.36	89.14	88.70	91.77
0.3	**85.98**	**89.39**	**89.36**	**92.03**
0.5	85.71	88.98	89.21	91.89
0.7	84.60	88.75	88.03	91.33
0.9	83.38	87.64	86.25	90.42

Table 13: Effect of the DPI selection threshold ξ on miniImageNet and CIFAR-FS. Mean accuracy (%) is reported.

Value (ξ)	miniImageNet		CIFAR-FS	
	1-shot	5-shot	1-shot	5-shot
0.1	78.42	83.73	80.69	85.10
0.3	83.18	87.90	87.61	90.41
0.5	**85.98**	**89.39**	**89.36**	**92.03**
0.7	84.02	88.49	88.03	91.15
0.9	82.71	87.53	86.91	90.18

Table 14: Effect of the CMA loss weight λ on miniImageNet and CIFAR-FS. Mean accuracy (%) is reported.

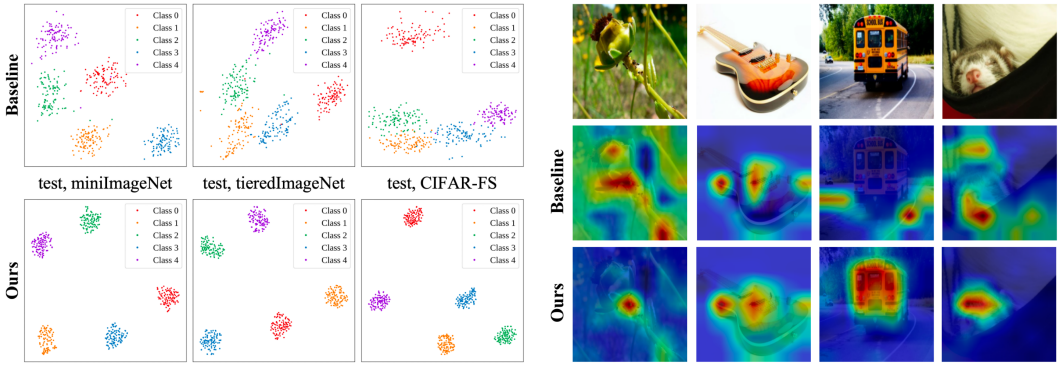
Value (λ)	miniImageNet		CIFAR-FS	
	1-shot	5-shot	1-shot	5-shot
0.1	83.48	88.12	86.92	90.33
0.2	**85.98**	**89.39**	**89.36**	**92.03**
0.3	85.53	89.21	89.12	91.90
0.4	85.41	88.87	88.31	91.69
0.5	84.65	88.50	87.64	91.02

Caption-scoring balance ω. As shown in Table 12, ω in Eq. 5 balances visual alignment and class discriminativeness. A small value underweights class separation and may retain visually plausible but generic captions, whereas a large value can favor label-discriminative yet visually unsupported details. The best overall performance at $\omega = 0.3$ indicates that visual grounding should remain the dominant criterion.

DPI selection threshold ξ. As shown in Table 13, the threshold in Eq. 6 determines when an instance activates the first eligible injection layer. A low threshold triggers prompt insertion too readily at shallow layers, while an overly high threshold delays selection and routes more samples to deep or fallback layers. The optimum at $\xi = 0.5$ provides a balanced instance-adaptive depth criterion.

CMA loss weight λ. As shown in Table 14, λ in Eq. 9 controls the contribution of prototype-level cross-modal alignment. Too small a value provides insufficient alignment between semantic and visual prototypes; too large a value overemphasizes alignment relative to episodic classification and can constrain discriminative visual learning. The best trade-off is obtained at $\lambda = 0.2$.

A.8 VISUALIZATION



(a)T-SNE visualization on novel classes (b)Visualization of attention maps

Figure 6: Visualization results. (a) t-SNE embeddings on three benchmarks. (b) Attention maps for representative images.

Figure 6 compares the baseline and MAD-SP. Across all three datasets, MAD-SP forms more compact intra-class clusters and clearer inter-class separation in the t-SNE plots. Its attention also focuses more consistently on foreground objects, whereas the baseline is often distracted by background regions. These results show that semantic prompting improves feature discriminability and the localization of class-relevant visual cues.