

Multi-modal Attribute Prompting for Vision-Language Models

Xin Liu , Jiamin Wu, Wenfei Yang[†], Xu Zhou, Tianzhu Zhang[†] 

Abstract—Pre-trained Vision-Language Models (VLMs), like CLIP, exhibit strong generalization ability to downstream tasks but struggle in few-shot scenarios. Existing prompting techniques primarily focus on global text and image representations, yet overlooking multi-modal attribute characteristics. This limitation hinders the model’s ability to perceive fine-grained visual details and restricts its generalization ability to a broader range of unseen classes. To address this issue, we propose a Multi-modal Attribute Prompting method (MAP) by jointly exploring textual attribute prompting, visual attribute prompting, and attribute-level alignment. The proposed MAP enjoys several merits. First, we introduce learnable visual attribute prompts enhanced by textual attribute semantics to adaptively capture visual attributes for images from unknown categories, boosting fine-grained visual perception capabilities for CLIP. Second, the proposed attribute-level alignment complements the global alignment to enhance the robustness of cross-modal alignment for open-vocabulary objects. To our knowledge, this is the first work to establish cross-modal attribute-level alignment for CLIP-based few-shot adaptation. Extensive experimental results on 11 datasets demonstrate that our method performs favorably against state-of-the-art approaches.

Index Terms—Few-shot classification, Prompt learning, Vision-language model, Attribute.

I. INTRODUCTION

PRE-TRAINED Vision-Language Models (VLMs), such as CLIP [1] and ALIGN [2], have demonstrated promising generalization power and transferability on a wide range of downstream tasks [3]–[9], including image classification [1], object detection [10], [11] and 3D understanding [12]–[14]. Through contrastive training on a large-scale dataset of image-text pairs, CLIP achieves a global alignment between images and textual descriptions by learning a joint embedding space. The robust cross-modal alignment empowers the CLIP model with the open-vocabulary visual recognition capability. In CLIP, class-specific weights for open vocabulary classification can be constructed by plugging the **class name** in a predefined prompt template like ‘A photo of a [CLASS].’ Despite its impressive generalization capability, it remains challenging to adapt CLIP to downstream tasks in few-shot scenarios. Due

to the large number of parameters in CLIP and the limited number of samples in few-shot task settings, naive fine-tuning of the entire model would likely lead to overfitting, resulting in performance degradation [15], [16].

To enhance the few-shot adaptation capability of CLIP, prompting techniques [17]–[23], such as CoOp [16] and Co-CoOp [18] have been proposed. These techniques replace hard template context with learnable context in combination with the class name to construct the text prompt. The classification result can be obtained by calculating the similarity between the global image feature and the encoded text prompt. However, as shown in Figure 1 (a), these prompting methods rely solely on class names and may struggle to fully encapsulate categorical semantics when new unseen classes emerge, causing an issue of ‘lexical weak tie’ where the class name has a tenuous link with its literal semantics. Consider ‘Rocky Road’ as an example, which textually resembles ‘rock’ and ‘road’ but refers to a dessert in reality. When introduced as a new class, the classification weight generated by the model may diverge from its true semantics, potentially causing misclassification. To address this issue, recent works [24]–[26], as shown in Figure 1 (b), introduce **textual attribute** descriptions obtained from Large Language Models [27]–[29]. These textual attribute descriptions are appended to the class name to construct text attribute prompts enriched with more semantics. The final classification result is determined by matching scores between the global image feature and the outputs of text attribute prompts across categories.

Despite the performance improvements demonstrated by prior methods, two crucial aspects have been overlooked. **(1) Visual Attribute Modeling.** Previous methods rely on a single global image feature for classification (see Figure 1 (a) and (b)). However, global image features may fall short in capturing fine-grained visual attribute information crucial for distinguishing visually similar classes in few-shot scenarios. As shown in Figure 2, the Moon Orchid and Japanese Anemone exhibit quite similar overall appearances, making it challenging to differentiate between them relying solely on global features. However, distinguishing them becomes much easier by relying on their distinct leaf shapes and reproductive structures. **(2) Attribute-Level Alignment.** The open-vocabulary visual recognition ability of the CLIP model stems from its global alignment between global image features and textual descriptions. However, when adapted to unseen tasks, the global alignment may lack robustness against disruptions from complex image backgrounds and irrelevant image details, hampering the image recognition ability. While previous methods have attempted to model class-specific tex-

[†]Corresponding author.

Xin Liu, Jiamin Wu, Wenfei Yang, and Tianzhu Zhang are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: xinliu99@mail.ustc.edu.cn; jiaminwu@mail.ustc.edu.cn; yangwf@ustc.edu.cn; tzzhang@ustc.edu.cn).

Xu Zhou is with the Sangfor Technologies Inc., Shenzhen 518000, China (e-mail: zhouxu@sangfor.com.cn).

Copyright © 2024 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

The definitive version of this paper can be found at: [10.1109/TCSVT.2024.3424566](https://doi.org/10.1109/TCSVT.2024.3424566)

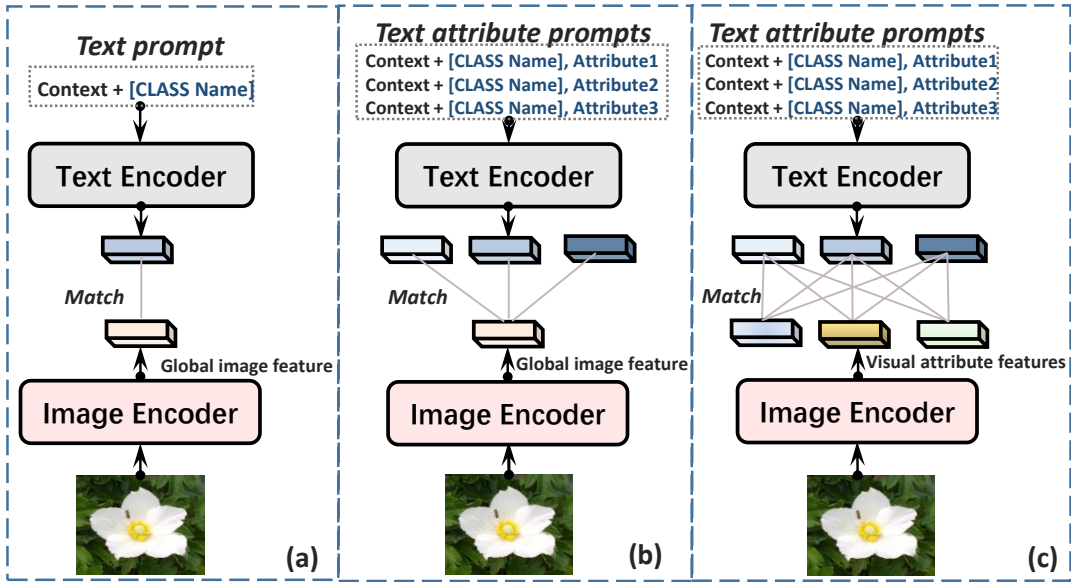


Fig. 1: (a) Conventional prompting methods use hand-crafted or learnable context in combination with the class name to construct the text prompt. (b) Recent methods introduce attribute descriptions to create text attribute prompts containing more semantic content. (c) Our method jointly explores multi-modal attributes and attribute-level alignment, enhancing fine-grained visual perception and achieving attribute-level alignment between images and text categories.

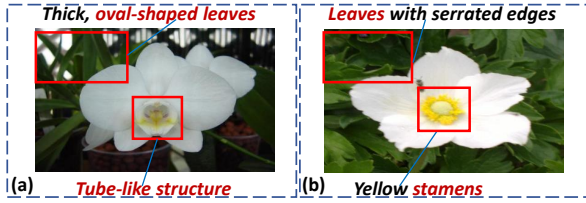


Fig. 2: (a) Moon Orchid and (b) Japanese Anemone exhibit strikingly similar overall appearances. Visual attributes play a crucial role in distinguishing between them, such as the central yellow stamens of Japanese Anemone.

tual attributes, as depicted in Figure 1 (b), they still focus on alignment with the global image features and fall short in addressing disruptions present in images. To address this issue, in addition to the global alignment, establishing **attribute-level alignment** is imperative, *i.e.*, alignment between fine-grained visual and textual attribute features (see Figure 1 (c)). This alignment empowers the model to selectively emphasize the distinctive visual attribute features described in the textual attributes, thereby enhancing the ability to handle disruptions in images.

Inspired by the above insights, we propose **Multi-modal Attribute Prompting** (MAP) by jointly exploring textual attribute prompting, visual attribute prompting, and attribute-level alignment to enhance the adaptability of CLIP in downstream few-shot tasks. For **textual attribute prompting**, we generate class-specific textual descriptions using a pre-trained large language model. Subsequently, these textual descriptions are utilized to create multiple textual attribute prompts, each encompassing context words, the class name, and an attribute description. It's challenging to directly capture appropriate dis-

criminative visual attributes in an unknown test image without prior information. Hence, for **visual attribute prompting**, first, we use learnable initial visual attribute prompts to aggregate regional features by interacting with image tokens. Then, we utilize the specially designed **Adaptive Visual Attribute Enhancement** (AVAE) module, in which the initial visual attribute prompts are enhanced by adaptively selected textual attribute prompts. Through interaction with both image tokens and textual attribute prompts, visual attribute prompts can adaptively capture visual attribute features in an unseen image. Finally, we reformulate the **attribute-level alignment** between visual attribute prompts and textual attribute prompts as an Optimal Transport problem [30] and use the Sinkhorn algorithm [31] to solve it. The ultimate classification result is determined by both the global matching score and the attribute-level matching score. This integration of additional attribute alignment, alongside global alignment, achieves multi-level robust alignment between images and text categories.

Our main contributions can be summarized as follows:

- We propose **Multi-modal Attribute Prompting**, which jointly explores textual attribute prompting, visual attribute prompting, and attribute-level alignment between images and text categories. To our knowledge, this is the first work to model visual attributes and establish attribute-level alignment between images and text categories for adapting the pre-trained CLIP model to downstream few-shot tasks.
- Extensive experimental results on 11 benchmark datasets demonstrate that our method performs favorably against state-of-the-art approaches.

II. RELATED WORKS

In this section, we introduce several lines of research in pre-trained vision-language models and prompt learning.

A. Vision-Language Models.

In recent years, pre-trained vision-language models [3], [4], [32]–[36] have shown exceptional performance in diverse downstream tasks. Among them, CLIP [1] stands out as a representative approach. By training its vision and text encoders to map both modalities closely in a shared embedding space, CLIP establishes a comprehensive global alignment between images and their corresponding textual descriptions, enabling open-vocabulary classification tasks. The classification result can be obtained by computing the similarity scores of the global image feature with class names encoded by the text encoder. However, as classification relies solely on the global matching score, the accuracy may be affected by disruptions in images, such as complex backgrounds, especially in few-shot settings [37]–[43], where only a few training samples are available. To improve the robustness of cross-modal alignment, we achieve multi-level alignment for CLIP by introducing additional attribute-level alignment between dynamically learned textual and visual attribute features. In this manner, our method enhances the fine-grained perception capability with the pre-trained global knowledge preserved.

B. Prompt Learning.

Prompt learning is initially introduced in the field of natural language processing (NLP) [44]–[48]. With language models frozen, prompt learning methods effectively facilitate the adaptation of pre-trained language models to downstream few-shot tasks by involving additional hand-crafted or learnable prompt tokens. Prompt learning has recently been employed to enhance the adaptation of the CLIP model to downstream few-shot tasks, where limited training samples are available. CoOp [16] constructs prompts by concatenating learnable continuous vectors and class name tokens. CoCoOp [18] extends CoOp by further learning a lightweight neural network to generate an input-conditional vector for each image, tackling the poor generalizability to broader unseen classes in CoOp [16]. ProDA [21] optimizes a set of prompts by learning the distribution of prompts. Instead of focusing on text-modal prompts, VPT [49] introduces learnable vectors to the Vision Transformer [50] to refine image features within the frozen vision encoder. DAPT [19], RPO [22], and MaPLe [23] improve the generalization ability of VLMs via multimodal prompting. PromptSRC [20] introduces regularization loss to prompt learning. These methods rely solely on class names for text prompt construction and may struggle to fully encapsulate categorical semantics.

C. Textual Attribute Prompts.

To enrich the semantic description for different classes, recent works [24]–[26], instead of relying solely on class names, have shifted towards the utilization of attribute descriptions to construct textual attribute prompts for each class. This shift is

facilitated by the development of pre-trained large language models (LLMs) like the GPT family [27], [28]. Attribute descriptions can be easily obtained by querying the LLM with suitable question templates. However, these methods focus on attributes in text space only, neglecting the modeling of visual attributes, leading to limited visual perception capabilities of the model and misalignment between global visual and local textual features. In contrast, we jointly model visual and textual attribute features and establish attribute-level alignment between images and text categories.

D. Visual Attributes.

Visual attributes refer to intuitive properties of objects, encompassing low-level semantics (e.g., color, texture, and shape) and high-level semantics (e.g., head, body, and tail of objects) [51]. Utilizing visual attributes has led to significant progress in various vision tasks, including image search [52], image recognition [53], and scene understanding [54]. Some previous works on learning attributes [52], [55], [56] usually require extensive manual attribute annotations, which are labor-intensive. Dealing with this issue, a recent work [57] developed an encoder-decoder network to unsupervisedly distill high-level attribute-specific vectors without requiring attribute annotations. VAPNet [58] achieves semantic details by utilizing local image patches to distill visual attributes from these discovered semantics. Different from these methods, our approach uniquely leverages visual prompts to model visual attributes. By incorporating visual attribute prompts as learnable tokens within Vision Transformers, our method captures and aggregates relevant image features effectively.

III. METHODOLOGY

In this section, we first provide a concise overview of CLIP [1]. Then, we present a comprehensive introduction to our proposed multi-modal attribute prompting, as illustrated in Figure 3, including textual attribute prompting, visual attribute prompting, and attribute-level alignment. The main symbols and instructions are shown in Table I.

A. Review of CLIP

The Contrastive Language-Image Pre-training (CLIP) model [1] is a well-known vision-language model trained on large-scale image-text pairs. CLIP consists of two primary components: an image encoder $\phi(\cdot)$ for converting input images into visual embeddings and a text encoder $\theta(\cdot)$ for encoding textual information. During pre-training, CLIP trains encoders using a contrastive loss objective [59], with the purpose of achieving a global alignment between images and textual descriptions. The CLIP model can be easily applied to downstream tasks.

Given a set \mathcal{V} of \mathcal{C} class names, the text prompts $\{t_i\}_{i=1}^{\mathcal{C}}$ are formulated as manually designed templates, such as ‘A photo of a [CLASS].’ The classification vectors $\{w_i\}_{i=1}^{\mathcal{C}}$ are derived by passing text prompts $\{t_i\}_{i=1}^{\mathcal{C}}$ to the text encoder: $w_i = \theta(t_i)$. Given an image x and its label y , the global image

TABLE I
MAIN SYMBOLS AND INSTRUCTIONS

Symbol	Instruction
$\phi(\cdot)$	the image encoder
$\theta(\cdot)$	the text encoder
\mathcal{V}	set of class names
\mathcal{C}	the number of class names
x	the input image
y	the ground-truth label
f	the global image feature
p_k^n	the n -th textual attribute prompt of k -th class
g_k^n	encoded n -th textual attribute prompt of k -th class
\mathbf{G}_k	encoded textual attribute prompts of the k -th class
l_j	the j -th ViT layer
E_j	image tokens output from j -th ViT layer
s_j	[CLS] token output from j -th ViT layer
U_j	visual attribute prompts output from j -th ViT layer
\mathbf{F}	visual attribute prompts output from ViT
T^*	the optimal transportation plan
Γ	adaptive visual attribute enhancement module
$\psi(\cdot, \cdot)$	similarity function
M	the number of visual attribute prompts
N	the number of textual attribute prompts
L	the number of transformer layers in ViT
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	queries, keys, and values in the attention layer
W_Q, W_K, W_V	linear projections of the attention layer
$\mathbf{1}_N$	N -dimensional all-one vector
p, q	discrete distributions
μ, ν	discrete probability vectors

feature f is extracted by the image encoder: $f = \phi(x)$. The classification probability is formulated as

$$P(y = i|x) = \frac{\exp(\cos(w_i, f)/\tau)}{\sum_{j=1}^{\mathcal{C}} \exp(\cos(w_j, f)/\tau)}, \quad (1)$$

where τ is a temperature parameter and $\cos(\cdot, \cdot)$ denotes the cosine similarity.

B. Textual Attribute Prompting

To address the potential ‘lexical weak tie’ issue of relying solely on class names for text prompt construction, we create multiple textual attribute prompts for each class, which helps enrich the semantic content in text prompts.

Attribute Descriptions. Consistent with previous methods [24]–[26], we obtain category attribute descriptions by querying a Large Language Model (LLM) using a predefined question template: ‘What are useful visual features for distinguishing a [CLASS] in an image?’ In response, the LLM provides discriminative attribute descriptions for the queried class. We select N descriptions for each class from the query results.

Textual Attribute Prompt Construction. We formulate N textual attribute prompts for each class by combining attribute description sentences with a standardized prompt template. For instance, for the k -th class, with the template ‘A photo of a [CLASS]’ we construct a textual attribute prompt: $p_k^n = \{A \text{ photo of a class } (k), t_k^n\}$, where class (k) denotes the class name corresponding to the k -th class, and t_k^n denotes the n -th attribute description for the k -th class. To enhance the adaptability of textual attribute prompts, we replace the hand-crafted context, *i.e.*, ‘A photo of a’ with several learnable context vectors. Following CoOp [16], we use four learnable class-agnostic context vectors, concatenated with the class name and attribute description to construct the textual attribute prompt. These vectors are optimized during training to better adapt to downstream tasks, providing a more flexible context.

By feeding the textual attribute prompts into the text encoder θ , we can obtain encoded textual attribute prompts:

$$\mathbf{G}_k = \{g_k^n |_{n=1}^N\}, g_k^n = \theta(p_k^n), \quad (2)$$

where \mathbf{G}_k is the textual attribute prompt set for the k -class.

C. Visual Attribute Prompting

To improve fine-grained visual perception, we model visual attributes with visual attribute prompts. However, it is challenging to directly learn discriminative visual attributes for an unknown image without prior information. Therefore, we design an adaptive visual attribute enhancement module to adaptively establish visual attribute prompts under the guidance of textual attribute information.

Learnable Visual Attribute Prompts. We model visual attributes by introducing M visual attribute prompts $U = \{u_i\}_{i=1}^M$, where each attribute prompt u_i is a randomly initialized learnable vector with the dimension of d_v . $\{u_i\}_{i=1}^M$ are inserted into the first Vision Transformer (ViT) layer and are then propagated into deeper layers. For the j -th ViT layer l_j , visual attribute prompts U_{j-1} output from the $(j-1)$ -th ViT layer are concatenated with image tokens E_{j-1} and the learnable classification token s_{j-1} ([CLS]), forming the input sequence of the current layer. Formally,

$$[s_j, U_j, E_j] = l_j([s_{j-1}, U_{j-1}, E_{j-1}]), j = 1, 2, \dots, L, \quad (3)$$

where $[\cdot, \cdot]$ indicates the concatenation along the sequence length dimension. In early layers of ViT, the visual attribute prompts progressively aggregate image regional features through interaction with image tokens facilitated by the attention mechanism. Learnable visual attribute prompts compute similarity with image tokens and aggregate information accordingly. Similar to the [CLS] token in models like BERT [60] and ViT [50], visual prompts can read and aggregate visual information from image tokens [22]. Previous research [61], [62] indicates that ViTs will attend to local information in early layers. This property, together with the attention mechanism, helps aggregate image regional features.

Adaptive Visual Attribute Enhancement Module. AVAE, represented as Γ , is designed to dynamically refine visual attribute prompts with textual attribute guidance for arbitrary images from unseen classes. As the category of the test

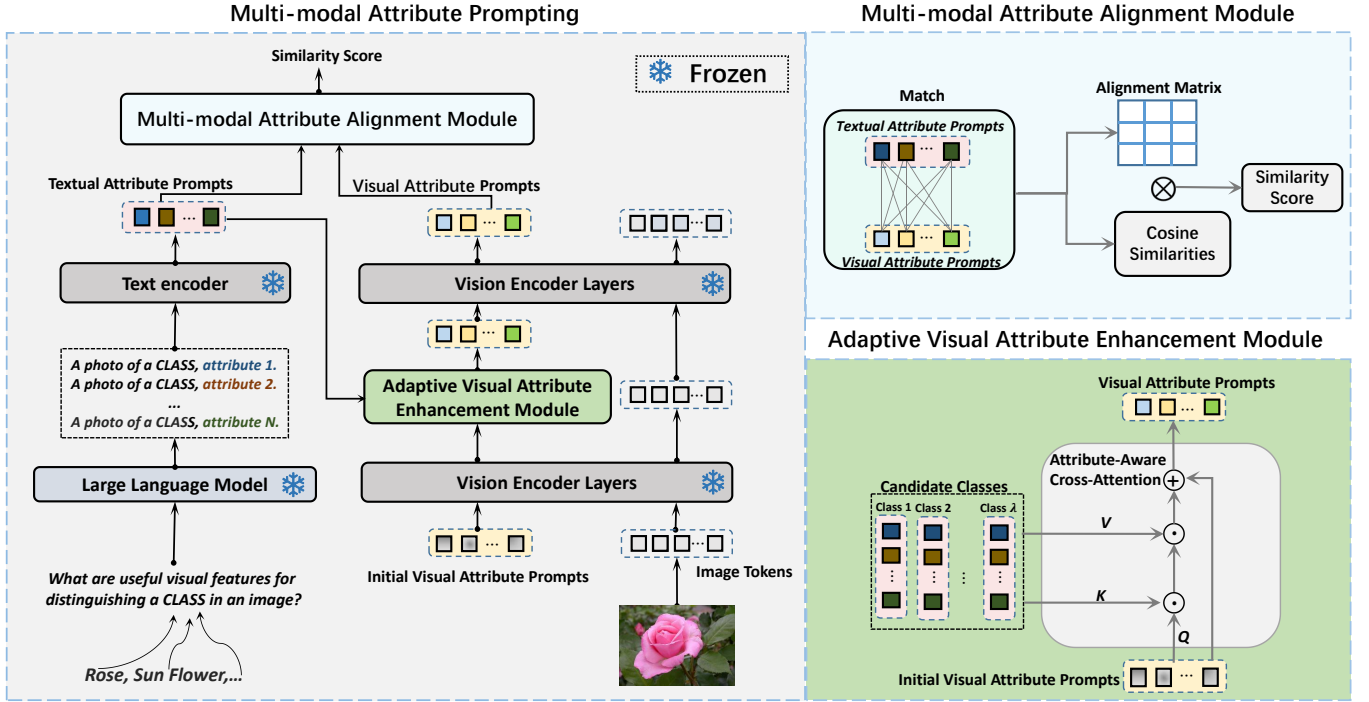


Fig. 3: The architecture of our method: **MAP** leverages textual attribute descriptions to construct textual attribute prompts and incorporates learnable visual attribute prompts for capturing visual attributes. In the **Adaptive Visual Attribute Enhancement** module, initial visual attribute prompts are enhanced by textual attribute prompts via the attribute-aware cross-attention layer. The **Multi-modal Attribute Alignment** module calculates the similarity score between visual attributes and textual attributes with the optimal transport.

image is unknown, we select possibly related textual attribute prompts from the most similar classes. Specifically, we first compute the similarities between the global image feature, *i.e.*, the classification token s , and textual category embeddings represented by the mean of textual attribute prompts. Based on these similarities, we select the most similar λ categories as the candidate classes and gather their textual attribute prompts as $\mathbf{G}' = \{g_j\}_{j=1}^{\lambda N}$. Subsequently, the textual attribute prompts \mathbf{G}' are employed as the semantic guidance to enhance visual attribute prompts at the l -th ViT layer:

$$\{\tilde{u}_i^{(l)}\}_{i=1}^M = \Gamma(\{u_i^{(l)}\}_{i=1}^M, \mathbf{G}'), \quad (4)$$

where Γ takes the initial visual attribute prompts $\{u_i^{(l)}\}_{i=1}^M$ generated from l -th layer as the input, and refine them conditioned on textual attribute prompts \mathbf{G}' . Then the enhanced visual attribute prompt $\tilde{u}_i^{(l)}$ is inserted into the $(l+1)$ -th layer for progressive attribute learning.

To better inject the semantic clues of selected textual prompts into visual attribute prompts, we design an attribute-aware cross-attention layer in Γ . Here, the visual attribute prompt tokens $\{u_i^{(l)}\}_{i=1}^M$ function as queries \mathbf{Q} . Simultaneously, the textual attribute prompt features \mathbf{G}' of candidate classes are utilized as keys \mathbf{K} and values \mathbf{V} . The enhanced visual attribute prompt $\tilde{u}_i^{(l)}$ is formulated as

$$\tilde{\alpha}_{ij} = \frac{\exp(\alpha_{ij})}{\sum_{j'=1}^{\lambda N} \exp(\alpha_{ij'})}, \alpha_{ij} = \frac{u_i^{(l)} W_Q \cdot (g_j W_K)^T}{\sqrt{d_K}}, \quad (5)$$

$$\tilde{u}_i^{(l)} = u_i^{(l)} + \sum_{j=1}^{\lambda N} \tilde{\alpha}_{ij} (g_j W_V), i = 1, 2, \dots, \lambda N, \quad (6)$$

where W_Q, W_K and W_V are linear projections of the attention layer. Attention scores $\tilde{\alpha}_{ij}$ indicate the correspondence between visual and textual attribute prompts, emphasizing relevant image-specific semantic attribute patterns for enhancing the visual attribute prompts. After the text-guided enhancement, the refined visual attribute prompts $\{\tilde{u}_i^{(l)}\}_{i=1}^M$ are propagated into the remaining vision encoder layers and continue to capture visual attributes through interaction with image tokens.

D. Attribute-Level Alignment

To achieve precise alignment between visual attribute prompts $\{u_i^{(L)}\}_{i=1}^M$ and textual attribute prompts $\mathbf{G}_k = \{g_k^n\}_{n=1}^N$, we formulate the attribute-level matching task as an Optimal Transport (OT) problem [30]. For simplicity, we refer to $\{u_i^{(L)}\}_{i=1}^M$ as $\mathbf{F} = \{f_m\}_{m=1}^M$ hereafter. Optimal Transport (OT) [30] is a powerful tool to measure the distance between two distributions. Given two sets of feature points $\mathbf{F} = \{f_m\}_{m=1}^M$ and $\mathbf{G}_k = \{g_k^n\}_{n=1}^N$, their distributions can be formulated as $p = \sum_{m=1}^M \mu_m \delta_{f_m}$, $q = \sum_{n=1}^N \nu_n \delta_{g_k^n}$, δ_{f_m} is a Dirac delta function centered at a specific point f_m in

the embedding space. Here, $\mu \in \mathbb{R}^M$, $\nu \in \mathbb{R}^N$ are two discrete distribution vectors. We define the cost matrix between $\mathbf{F} = \{f_m\}_{m=1}^M$ and $\mathbf{G}_k = \{g_k^n\}_{n=1}^N$ as $\mathbf{C} \in \mathbb{R}^{M \times N}$, where $C_{m,n} = 1 - \langle f_m, g_k^n \rangle$ is the transport cost from f_m to g_k^n . The transport cost between p and q is $\langle \mathbf{T}, \mathbf{C} \rangle$, where \mathbf{T} is the transport plan, and $\mathbf{T}_{m,n}$ is the probability or ‘‘flow’’ of transporting from f_m to g_k^n . The goal of OT is to transport p to q at the smallest cost with the optimal transport plan \mathbf{T}^* :

$$\begin{aligned} \mathbf{T}^* &= \arg \min_{\mathbf{T} \in \Pi(p,q)} \langle \mathbf{T}, \mathbf{C} \rangle, \\ \text{s.t. } \quad \mathbf{T} \mathbf{1}_N &= \mu, \mathbf{T}^T \mathbf{1}_M = \nu, \end{aligned} \quad (7)$$

where $\Pi(p, q)$ is the joint distribution with marginals μ and ν , and $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. To accelerate the solving process, we use the Sinkhorn algorithm, which introduces the entropic regularization term to the transport cost to encourage smoother solutions: $\min_{\mathbf{T}} \langle \mathbf{T}, \mathbf{C} \rangle - \gamma h(\mathbf{T})$, γ is a constant hyperparameter controlling the intensity of regularization term. Instead of solving the constrained optimization directly, the Sinkhorn algorithm [31] employs an iterative procedure:

$$\begin{aligned} \mathbf{T}^* &= \text{diag}(U(t)) \mathbf{A} \text{diag}(V(t)), \\ \mathbf{A} &= \exp(-\mathbf{C}/\gamma) \end{aligned} \quad (8)$$

where in the t -th iteration, $U(t) = \mu / (\mathbf{A} V(t-1))$, $V(t) = \nu / \mathbf{A}^T U(t)$, with the initiation $V(0) = \mathbf{1}$. With Equation (8), we can obtain \mathbf{T}^* to serve as the alignment matrix, and then define the final similarity score between the visual attribute prompts \mathbf{F} and textual attribute prompts \mathbf{G}_k as:

$$\psi(\mathbf{F}, \mathbf{G}_k) = \sum_{m=1}^M \sum_{n=1}^N \langle f_m, g_k^n \rangle \mathbf{T}_{m,n}^*, \quad (9)$$

where $\psi(\cdot, \cdot)$ denotes the similarity function.

E. Training Objectives

Based on the attribute-level alignment, we can classify the image x with fine-grained visual attributes:

$$P_a(y = i|x) = \frac{\exp(\psi(\mathbf{F}, \mathbf{G}_i)/\tau)}{\sum_{j=1}^C \exp(\psi(\mathbf{F}, \mathbf{G}_j)/\tau)}. \quad (10)$$

Furthermore, relying on the global alignment in CLIP, the prediction probability is computed as

$$P_g(y = i|x) = \frac{\exp(\cos(\langle \mathbf{f}, \bar{\mathbf{g}}_i \rangle / \tau))}{\sum_{j=1}^C \exp(\cos(\langle \mathbf{f}, \bar{\mathbf{g}}_j \rangle / \tau))}, \quad (11)$$

where \mathbf{f} is the global feature of the image x , *i.e.*, the class token s_L , and $\bar{\mathbf{g}}_i$ is the textual categorical embedding of the i -th class, *i.e.*, the mean value of textual prompts in \mathbf{G}_i . The final prediction probability is

$$P(y = i|x) = P_g(y = i|x) + \beta P_a(y = i|x), \quad (12)$$

which incorporates both global-level prediction scores and additional attribute-level matching scores, achieving multi-level robust alignment between images and categorical texts. Naturally, the classification loss is formulated as:

$$L_{cls} = -\frac{1}{B} \sum_{i=1}^B \log(P(y = y_i|x_i)), \quad (13)$$

where B is the batch size of image-text pairs, and y_i denotes the ground-truth label of the input image x_i .

IV. EXPERIMENTS

In this section, we begin by introducing the benchmark settings and implementation details, followed by a comprehensive presentation of the experimental results.

All the models used are based on the open-source CLIP [1] model. We evaluate the adaptation and generalization capability of MAP in four distinct settings following previous methods [16], [18].

Base-to-novel generalization. Datasets are split into base and novel classes. The model is trained on the training dataset, which is constructed by randomly selecting 16 images per class from base classes. Then the model is evaluated on both base and novel classes. The evaluation encompasses 11 image recognition datasets, including Food101 (Foo) [64], DTD [65], ImageNet (Img) [66], Caltech101 (Cal) [67], EuroSAT (Eur) [68], StanfordCars (Car) [69], FGVCaircraft (FGV) [70], Flowers102 (Flo) [71], OxfordPets (Pet) [72], UCF101 (UCF) [72], and SUN397 (SUN) [73].

Few-shot image classification. To evaluate the learning capacity under extremely limited supervision, we assess the model’s performance across varying shot scenarios, namely, 1, 2, 4, 8, and 16 shots. Similar to the base-to-novel generalization setting, we employ the same 11 datasets.

Domain generalization. To assess the robustness under domain shifts, we train the model using the source dataset ImageNet and subsequently evaluate its performance on out-of-distribution target datasets, namely ImageNet-R (-R) [74], ImageNet-A (-A) [75], ImageNetV2 (V2) [76], and ImageNet-Sketch (-S) [77].

Cross-dataset evaluation. In the cross-dataset transfer setting, we train the models on the source dataset ImageNet and directly evaluate them on target datasets. Specifically, the target datasets include Food101, DTD, Caltech101, EuroSAT, StanfordCars, FGVCaircraft, Flowers102, OxfordPets, UCF101, and SUN397.

Implementation Details. In all the experiments, we use the pre-trained CLIP [1] with ViT-B/16 image encoder backbone as the base model. We use the GPT-3.5 as the large language model. For MAP, we set the number of textual attribute prompts N to 4, and the number of visual attribute prompts M to 4. The AVAE module is inserted into the 7th transformer layer in the Vision Transformer (ViT). The default value of λ is set as 10. β is set as 1. We train the model using the SGD optimizer with a learning rate of 0.002. For the base-to-novel generalization setting, the model is trained for 20 epochs with a batch size of 16. For few-shot image classification, the maximum epoch is set to 200 for 16/8 shots, 100 for 4/2 shots, and 50 for 1 shot (except for ImageNet, where the maximum epoch is fixed to 50).

A. Base-to-Novel Generalization

To demonstrate generalization to label-shift, where labels are divided into base and novel classes for each dataset, we train the model on training datasets constructed by randomly

TABLE II

COMPARISON WITH CLIP, CoOp AND CoCoOp IN THE BASE-TO-NOVEL GENERALIZATION SETTING. THE RESULTS DEMONSTRATE THE STRONG GENERALIZABILITY TO NOVEL CLASSES OF OUR MAP. HM: HARMONIC MEAN TO HIGHLIGHT THE GENERALIZATION TRADE-OFF [63]. THE BEST RESULTS IN EACH COLUMN ARE SHOWN IN **BOLD FONT**.

(A) AVERAGE RESULTS				(B) IMAGENET				(C) CALTECH101			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	69.34	74.22	71.70	CLIP	72.43	68.14	70.22	CLIP	96.84	94.00	95.40
CoOp	82.69	63.22	71.66	CoOp	76.47	67.88	71.92	CoOp	98.00	89.81	93.73
CoCoOp	80.47	71.69	75.83	CoCoOp	75.98	70.43	73.10	CoCoOp	97.96	93.81	95.84
Ours	83.66	75.76	79.36	Ours	76.60	70.60	73.48	Ours	98.30	93.80	96.00

(D) DTD				(E) EUROSAT				(F) UCF101			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	53.24	59.90	56.37	CLIP	56.48	64.05	60.03	CLIP	70.53	77.50	73.85
CoOp	79.44	41.18	54.24	CoOp	92.19	54.74	68.69	CoOp	84.69	56.05	67.46
CoCoOp	77.01	56.00	64.85	CoCoOp	87.49	60.04	71.21	CoCoOp	82.33	73.45	77.64
Ours	82.63	66.23	73.53	Ours	92.13	76.10	83.33	Ours	86.67	78.77	82.52

(G) OXFORDPETS				(H) STANFORDCARS				(I) FLOWERS102			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	91.17	97.26	94.12	CLIP	63.37	74.89	68.65	CLIP	72.08	77.80	74.83
CoOp	93.67	95.29	94.47	CoOp	78.12	60.40	68.13	CoOp	97.60	59.67	74.06
CoCoOp	95.20	97.69	96.43	CoCoOp	70.49	73.59	72.01	CoCoOp	94.87	71.75	81.71
Ours	95.43	96.90	96.16	Ours	76.70	73.73	75.18	Ours	97.57	75.23	84.95

(J) FOOD101				(K) FGVC AIRCRAFT				(L) SUN397			
Method	Base	Novel	HM	Method	Base	Novel	HM	Method	Base	Novel	HM
CLIP	90.10	91.22	90.66	CLIP	27.19	36.29	31.09	CLIP	69.36	75.35	72.23
CoOp	88.33	82.26	85.19	CoOp	40.44	22.30	28.75	CoOp	80.60	65.89	72.51
CoCoOp	90.70	91.29	90.99	CoCoOp	33.41	23.71	27.74	CoCoOp	79.74	76.86	78.27
Ours	90.30	89.30	89.80	Ours	41.63	36.43	38.84	Ours	82.33	76.30	79.20

TABLE III

COMPARING MAP AGAINST MORE METHODS ON THE AVERAGE ACCURACY OVER 11 DATASETS.

Method	Base	Novel	HM
CLIP [1]	69.34	74.22	71.70
CoOp [16]	82.69	63.22	71.66
CoCoOp [18]	80.47	71.69	75.83
ProDA [21]	81.56	72.30	76.65
RPO [22]	81.13	75.00	77.78
VDT-Adapter [26]	82.48	74.51	78.09
MaPLe [23]	82.28	75.14	78.55
MAP	83.66	75.76	79.36

selecting 16 images per class from base classes. The model is trained using this few-shot sampled data for 3 random seeds, and the results are averaged. We evaluate accuracy on test data corresponding to both the base and novel classes and use their

harmonic mean [63] as the final evaluation metric.

Compared to CoOp, MAP exhibits higher harmonic mean accuracy across all datasets. As shown in Table II, MAP, on average, increases novel accuracy by 12.54% and base accuracy by 0.97%. This demonstrates that MAP not only enhances the model’s generalization to novel classes but also achieves better alignment between visual and textual modalities within base classes.

Compared to CoCoOp, MAP demonstrates superior generalization to novel classes, achieving an impressive average gain of up to 4.07%. When considering both base and novel classes, MAP outperforms CoCoOp with an absolute average gain of 3.53%. Among the 11 datasets, MAP exhibits higher accuracy than CoCoOp in 10 base datasets and 7 novel datasets.

We present the average accuracy results across 11 datasets for MAP compared with several other methods in Table III. MAP outperforms other methods by a significant margin, demonstrating our superior performance over other methods. It’s worth noting that VDT-Adapter [26], which leverages

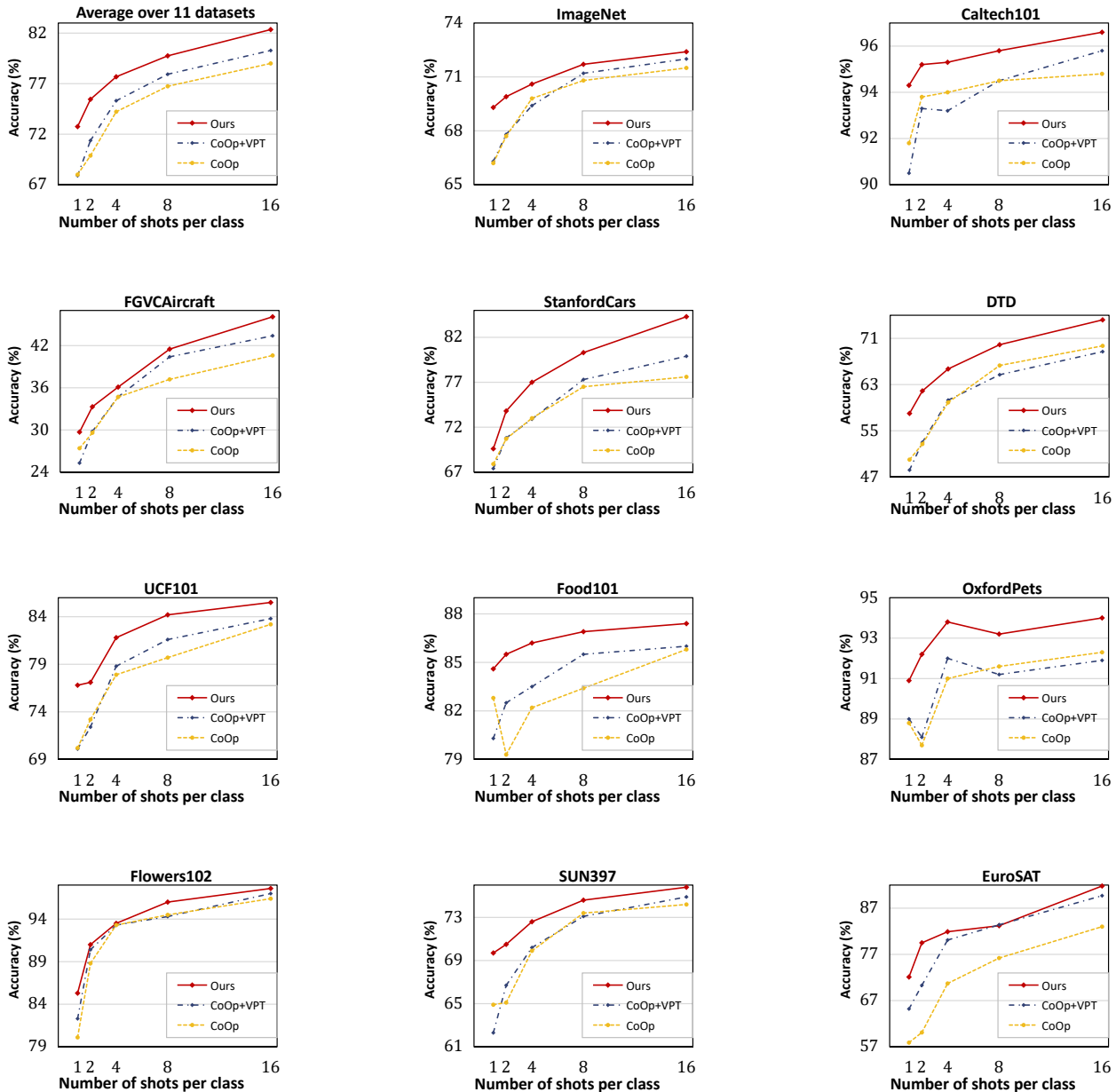


Fig. 4: Main results of few-shot image classification on 11 datasets. MAP consistently outperforms other CLIP adaptation methods across all datasets, demonstrating the strong few-shot adaptability of MAP.

textual attributes obtained from GPT-4 to formulate prompts, improves the novel accuracy compared to CoOp. However, it neglects modeling visual attributes and fails to leverage the role of attributes fully. MAP outperforms VDT-Adapter 1.18% in base classes and 1.25% in novel classes.

B. Few-Shot Image Classification

To evaluate few-shot learning ability, we adopt the few-shot evaluation protocol from CLIP [1], utilizing 1, 2, 4, 8, and 16 shots per class for training and deploying models in full test sets. Figure 4 summarizes the performance of MAP in few-shot learning on 11 datasets. Each plot compares MAP with CoOp and CoOp+VPT. CoOp+VPT refers to the

combination of CoOp and VPT, *i.e.*, the integration of both learnable text prompts and learnable visual prompts [49] into the CLIP model simultaneously. In terms of the overall performance (Figure 4, top-left), compared to CoOp, the combination of CoOp and VPT shows some improvement, though not significant. However, in the 1-shot setting, the performance of the combination is even worse than CoOp alone. This suggests that simply introducing more learnable parameters in the vision encoder brings limited performance improvement in the extreme few-shot setting. However, MAP consistently delivers significant performance improvements, even in scenarios with very few training samples (*e.g.*, 1-shot), showcasing the effectiveness of our visual attribute prompts

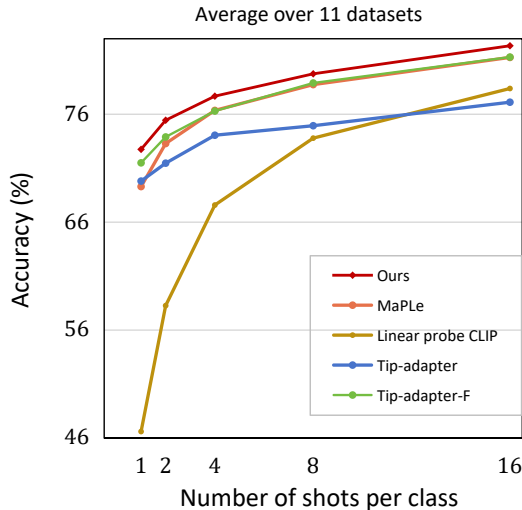


Fig. 5: The average few-shot image classification results of more methods across 11 datasets.

enhanced by textual guidance. Furthermore, on certain datasets (Caltech101, Flowers102, DTD, SUN397, and OxfordPets), CoOp+VPT does not outperform CoOp alone, whereas MAP consistently achieves superior performance across all benchmark datasets, demonstrating the generalizability of MAP across diverse datasets.

In Figure 5, we present the performance results of additional methods for few-shot image classification. Tip-adapter-F [78], the fine-tuned version of Tip-adapter, requires fine-tuning on the few-shot training data to update the adapter. The results show that Tip-adapter-F consistently achieves better performance than Tip-adapter and Linear probe CLIP. MaPLE [23] achieves performance comparable to Tip-adapter-F overall. Notably, MAP consistently outperforms both MaPLE [23] and Tip-adapter-F [78] in few-shot image classification across various shot settings, highlighting the effectiveness of our proposed approach.

C. Domain Generalization

To evaluate the model’s robustness under domain shifts, we initially train the model using the source dataset, ImageNet [66]. Subsequently, we evaluate its performance on target out-of-distribution datasets, namely ImageNetV2 [76], ImageNet-Sketch [77], ImageNet-A [75] and ImageNet-R [74]. The overall results are summarized in Table IV. From the experimental results, the fully fine-tuned CLIP model shows poorer performance compared to the zero-shot CLIP on the ImageNet dataset and variants of ImageNet. This demonstrates that naive fine-tuning of the entire CLIP model may cause overfitting on the training set, leading to performance degradation. MAP achieves remarkable performance on unseen data compared to zero-shot CLIP [1], linear probe CLIP, CoOp [16] and CoCoOp [18]. Compared to MaPLE, MAP shows slightly lower performance on ImageNet-Sketch but outperforms MaPLE [23] on other target datasets (Ima-

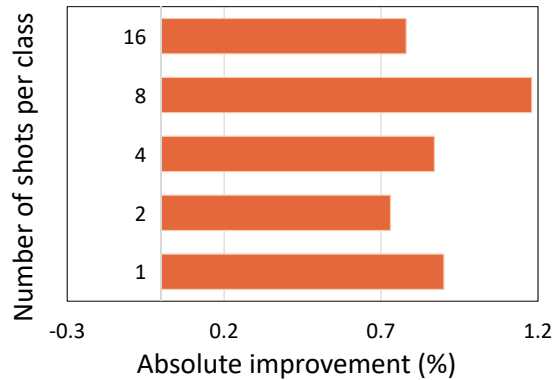


Fig. 6: The absolute accuracy improvements provided by using AVAE compared to scenarios without AVAE.

geNetV2, ImageNet-A, and ImageNet-R). This underscores the robustness of MAP to domain shifts.

D. Cross-Dataset Evaluation

To demonstrate the model’s capacity for generalization beyond a single dataset, we conduct training on ImageNet [66] and subsequently evaluate its performance on the other 10 datasets. When transferring to other datasets, textual attribute prompts are constructed using class attribute descriptions of the target dataset classes, which are also collected from the LLM. The learned parameters can be directly transferred, allowing effective inference despite category differences between the source and target datasets. Table V presents a comprehensive overview of the performance comparison between MAP and previous methodologies on the cross-dataset evaluation benchmark. On the source dataset, MAP achieves the highest score, underscoring its effectiveness in the source domain. When compared with CoOp [16], CoCoOp [18], and MaPLE [23], MAP demonstrates a superior capacity for generalization across diverse datasets. Specifically, it outperforms these methodologies in 7 out of 10, 6 out of 10, and 6 out of 10 datasets, respectively. This suggests that MAP exhibits robustness to varied data distributions.

E. Ablation Study

In this section, we perform ablation studies to demonstrate the effectiveness of each design of the proposed method.

Effectiveness of Attribute Prompts. We denote Textual Attribute Prompts as TAP and Visual Attribute Prompts as VAP. We remove TAP and VAP from MAP as our baseline. The results in Table VI are analyzed as follows: (1) Compared to the baseline, utilizing TAP powered by the LLM effectively improves the novel accuracy, achieving an accuracy gain of 1.43%, which demonstrates textual attributes enrich the semantics for novel classes. (2) The incorporation of VAP shows a distinct performance boost on both base (+1.6%) and novel classes (+2.11%). This proves that VAP contributes to enhancing fine-grained visual perception ability by capturing visual attributes.

TABLE IV
DOMAIN GENERALIZATION EVALUATION. METHODS ARE TRAINED ON THE SOURCE DATASET IMAGENET AND EVALUATED ON DATASETS WITH DOMAIN SHIFTS, INCLUDING IMAGENETV2, IMAGENET-S, IMAGENET-A, AND IMAGENET-R.

	Source		Target			Avg.
	ImageNet	ImageNetV2	ImageNet-S	ImageNet-A	ImageNet-R	
CLIP [1]	66.73	60.83	46.15	47.77	73.96	57.18
Fully Fine-Tuned CLIP	61.65	52.70	26.10	17.55	50.15	36.63
Linear probe CLIP [1]	67.42	57.19	35.97	36.19	60.10	47.36
CoOp [16]	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp [18]	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe [23]	70.72	64.07	49.15	50.90	76.98	60.27
MAP	71.60	64.47	49.07	51.07	77.37	60.49

TABLE V
CROSS-DATASET EVALUATION. MODELS ARE TRAINED ON IMAGENET AND EVALUATED ON TARGET DATASETS. MAP ACHIEVES OVERALL FAVORABLE PERFORMANCE.

	Source		Target								
	ImageNet	Cal	Pet	Car	Flo	Foo	Air	SUN	DTD	Eur	UCF
CoOp [16]	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55
CoCoOp [18]	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21
MaPLe [23]	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69
MAP	71.60	93.93	90.80	63.00	68.40	86.07	24.87	68.10	51.87	42.63	68.73

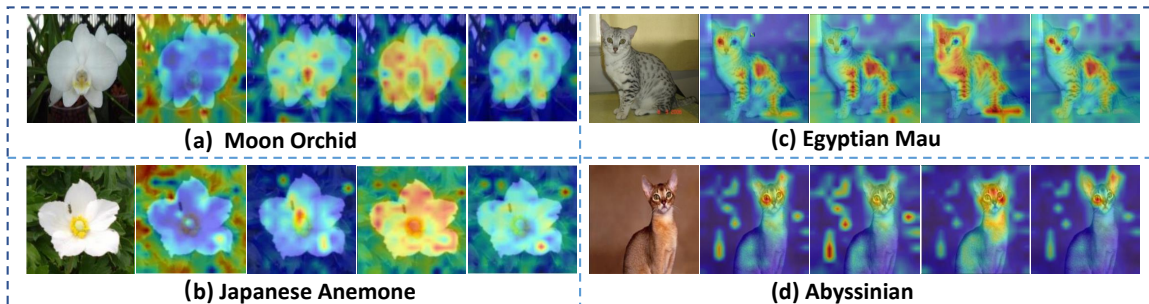


Fig. 7: The visualization of visual attribute prompts. Guided by textual attribute semantics, visual attribute prompts focus on distinctive visual details, such as the different leaf shapes of the Moon Orchid and Japanese Anemone, the spotted coat of the Egyptian Mau, and the large ears of the Abyssinian.

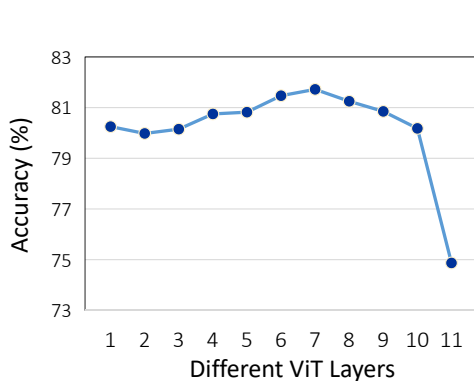


Fig. 8: The impact of inserting AVAE into different layers of ViT with 1 shot per class.

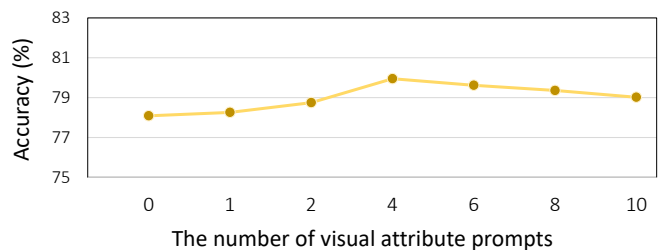


Fig. 9: The impact of the number of visual attribute prompts in the base-to-novel generalization setting.

Effectiveness of Adaptive Visual Attribute Enhancement. To verify the accuracy improvement when using AVAE, we conduct few-shot image classification experiments on 6 datasets (Flowers102, DTD, UCF101, OxfordPets, Cal-

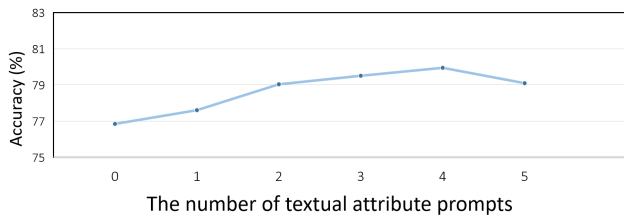


Fig. 10: The impact of the number of textual attribute prompts per class in the base-to-novel generalization setting.

TABLE VI
ABLATION RESULTS.

Method	Base	Novel	HM
Baseline	82.20	72.22	76.41
+TAP(LLM)	82.06	73.65	77.36
+TAP+VAP (MAP)	83.66	75.76	79.36

TABLE VII
COMPLEXITY RESULTS.

	CoCoOp	MaPLe	MAP
parameters	0.04M	3.56M	0.74M
GFLOPs	83.83	55.23	84.80
test time	56.70s	9.58s	9.79s

TABLE VIII
THE IMPACT OF USING DIFFERENT LLMs.

Method	Base	Novel	HM
Qwen-1.8B-Chat	97.47	73.23	83.63
GPT-3.5	97.57	75.23	84.95
Qwen1.5-72B-Chat	97.77	75.30	85.08

tech101, Food101). As shown in Figure 6, the employment of AVAE brings remarkable performance gains. Furthermore, we investigate the impact of placing AVAE into different ViT layers. As observed from Figure 8, placing AVAE in the middle layers (Layer 6-8) attains superior performance. When applying AVAE in the shallow or deep layers, the performance deteriorates obviously compared to the middle layers. Therefore, the AVAE module should be placed in the middle layers. Initial visual attribute prompts can aggregate visual regional features in shallow layers and continue to capture visual attributes in the remaining layers after enhancement by AVAE.

Analysis of Number of Visual Attribute Prompts. Figure 9 illustrates the averaged harmonic mean accuracy of using varying numbers of visual prompts over 10 datasets in the base-to-novel generalization setting. When the number is as small as 1, the performance gain is quite limited. The accuracy increases with more visual attribute prompts, as more visual attribute characteristics can be captured. However, the accuracy decreases slightly when the number is beyond 4, as an excessive amount of visual attribute prompts may contain redundancy and noises.

Analysis of Number of Textual Attribute Prompts. Fig-

ure 10 illustrates the averaged harmonic accuracy of using different numbers of textual attribute prompts. According to the experimental results, the introduction of textual attribute prompts indeed improves the performance, demonstrating the effectiveness of textual attribute prompts. The accuracy improves with the incorporation of more textual attribute prompts, as this introduces more descriptive information. However, when the number of textual attribute prompts exceeds four, the performance decreases. This may be attributed to the fact that additional prompts introduce more redundancy. The initial prompts are usually the most relevant and effective, while later ones may include less useful or intuitive descriptions. Increased complexity and less discriminative attributes like size or height can also burden the model, resulting in reduced performance. Overall, the accuracy changes relatively smoothly with different prompt numbers.

Impact of Different LLMs. We conduct experiments using other large language models (LLMs), specifically Qwen-1.8B-Chat and Qwen-1.5-72B-Chat [79], and examine performance variations on the Flowers102 dataset. The results in Table VIII show that Qwen-1.5-72B-Chat achieves performance comparable to GPT-3.5. However, when using Qwen-1.8B-Chat, there is a significant performance drop compared to using GPT-3.5 and Qwen-1.5-72B-Chat. This decline may be attributed to the fact that the outputs from Qwen-1.8B-Chat are sometimes inconsistent, noisy, and occasionally lack meaningful information. These findings suggest that selecting a large language model capable of generating consistent and clear outputs is crucial for maintaining performance.

Analysis of Complexity. We compare different prompting methods about the number of parameters, the GFLOPs, and the test time in Table VII. MaPLe [23] and MAP enjoy faster inference speeds than CoCoOp [18]. Compared with MaPLe, MAP is more parameter-efficient (0.74M vs 3.56M). The computation cost (GFLOPs) of MAP is higher, but considering the performance improvement, it is acceptable.

Visualization of Visual Attribute Prompts. We visualize visual attribute prompts output by the Vision Transformer in Figure 7. It can be observed that different visual attribute prompts focus on various aspects of the image and highlight distinctive visual details. This visualization demonstrates the capacity of visual attribute prompts to augment the model’s fine-grained visual perception ability.

V. LIMITATION AND FUTURE WORK

We use text attributes directly from GPT without manual filtering. Text attributes may contain noise that may hinder accurate classification, such as attributes with high uncertainty, like colors of toad lilies (white, purple, pink, or yellow). On Flowers102 [71], we manually filter improper attributes, resulting in an improvement of 0.37% in HM. Filtering improper ones has the potential to improve results. We’ll design an automatic filter plan in the future.

VI. CONCLUSION

In this paper, we propose a Multi-modal Attribute Prompting method to adapt pre-trained Vision-Language models for

downstream few-shot tasks. Our method involves modeling visual attributes to enhance the visual fine-grained perception ability. We establish attribute-level alignment, complementing the global alignment to achieve multi-level robust alignment between images and text categories. Extensive experimental results demonstrate the effectiveness.

ACKNOWLEDGMENTS

This work was supported by National Defense Basic Scientific Research Program of China (JCKY2020903B002), National Natural Science Foundation of China (62306294), Anhui Provincial Natural Science Foundation (2308085QF222), China Postdoctoral Science Foundation (2023M743385) and Youth Innovation Promotion Association CAS.

REFERENCES

- [1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
- [2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.
- [3] T. Mei, J. J. Corso, G. Kim, J. Luo, C. Shen, and H. Zhang, "Guest editorial introduction to the special section on video and language," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 1–4, 2022.
- [4] W. Zhang, C. Ma, Q. Wu, and X. Yang, "Language-guided navigation via cross-modal grounding and alternate adversarial learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3469–3481, 2020.
- [5] Z. Wei, Z. Zhang, P. Wu, J. Wang, P. Wang, and Y. Zhang, "Fine-granularity alignment for text-based person retrieval via semantics-centric visual division," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [6] H. Zhu, C. Zhang, Y. Wei, S. Huang, and Y. Zhao, "Esa: External space attention aggregation for image-text retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [7] W. Zhou and Z. Zhou, "Unsupervised domain adaptation harnessing vision-language pre-training," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [8] X. Lin, M. Zhu, R. Dang, G. Zhou, S. Shu, F. Lin, C. Liu, and Q. Chen, "Clipose: Category-level object pose estimation with pre-trained vision-language knowledge," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [9] L. Wang, H. Qiu, B. Qiu, F. Meng, Q. Wu, and H. Li, "Tridentcap: Image-fact-style trident semantic framework for stylized image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [10] R. Arandjelović, A. Andonian, A. Mensch, O. J. Hénaff, J.-B. Alayrac, and A. Zisserman, "Three ways to improve feature alignment for open vocabulary detection," *arXiv preprint arXiv:2303.13518*, 2023.
- [11] P. Kaul, W. Xie, and A. Zisserman, "Multi-modal classifiers for open-vocabulary object detection," in *International Conference on Machine Learning*. PMLR, 2023, pp. 15946–15969.
- [12] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 815–824.
- [13] C. Zhu, W. Zhang, T. Wang, X. Liu, and K. Chen, "Object2scene: Putting objects in context for open-vocabulary 3d detection," *arXiv preprint arXiv:2309.09456*, 2023.
- [14] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "Openmask3d: Open-vocabulary 3d instance segmentation," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems*, 2023.
- [15] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.
- [16] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [17] C. Ma, Y. Liu, J. Deng, L. Xie, W. Dong, and C. Xu, "Understanding and mitigating overfitting in prompt tuning for vision-language models," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [18] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16816–16825.
- [19] E. Cho, J. Kim, and H. J. Kim, "Distribution-aware prompt tuning for vision-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22004–22013.
- [20] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan, "Self-regulating prompts: Foundational model adaptation without forgetting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15190–15200.
- [21] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt distribution learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5206–5215.
- [22] D. Lee, S. Song, J. Suh, J. Choi, S. Lee, and H. J. Kim, "Read-only prompt optimization for vision-language few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1401–1411.
- [23] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19113–19122.
- [24] Z. Feng, A. Bair, and J. Z. Kolter, "Leveraging multiple descriptive features for robust few-shot image learning," *arXiv preprint arXiv:2307.04317*, 2023.
- [25] S. Menon and C. Vondrick, "Visual classification via description from large language models," in *International Conference on Learning Representations*, 2023.
- [26] M. Maniparambil, C. Vorster, D. Molloy, N. Murphy, K. McGuinness, and N. E. O'Connor, "Enhancing clip with gpt-4: Harnessing visual descriptions as prompts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 262–271.
- [27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.
- [28] R. OpenAI, "Gpt-4 technical report. arxiv 2303.08774," *View in Article*, 2023.
- [29] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.
- [30] C. Villani, *Optimal transport: old and new*. Springer, 2009, vol. 338.
- [31] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.
- [32] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4467–4480, 2019.
- [33] Z. Yang, T. Kumar, T. Chen, J. Su, and J. Luo, "Grounding-tracking-integration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3433–3443, 2020.
- [34] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "Flava: A foundational language and vision alignment model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15638–15650.
- [35] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18123–18133.
- [36] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.
- [37] W. Jiang, K. Huang, J. Geng, and X. Deng, "Multi-scale metric learning for few-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1091–1102, 2020.
- [38] M. Cheng, H. Wang, and Y. Long, "Meta-learning-based incremental few-shot object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2158–2169, 2021.

- [39] X. Wang, X. Wang, B. Jiang, and B. Luo, "Few-shot learning meets transformer: Unified query-support transformers for few-shot classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [40] R. Xu, L. Xing, S. Shao, L. Zhao, B. Liu, W. Liu, and Y. Zhou, "Gct: Graph co-training for semi-supervised few-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8674–8687, 2022.
- [41] M. Zhang, M. Shi, and L. Li, "Mfnet: Multiclass few-shot segmentation network with pixel-wise metric learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8586–8598, 2022.
- [42] C. Zhang, C. Li, and J. Cheng, "Few-shot visual classification using image pairs with binary transformation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2867–2871, 2019.
- [43] Z. Dang, M. Luo, C. Jia, C. Yan, X. Chang, and Q. Zheng, "Counterfactual generation framework for few-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [44] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020.
- [45] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 4582–4597.
- [46] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059.
- [47] Y. Gu, X. Han, Z. Liu, and M. Huang, "PPT: pre-trained prompt tuning for few-shot learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 8410–8423.
- [48] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," *AI Open*, 2023.
- [49] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.
- [50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations, ICLR 2021*.
- [51] V. Ferrari and A. Zisserman, "Learning visual attributes," *Advances in neural information processing systems*, vol. 20, 2007.
- [52] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, 2011.
- [53] S. Wang, Z. Wang, H. Li, J. Chang, W. Ouyang, and Q. Tian, "Accurate fine-grained object recognition with structure-driven relation graph networks," *International Journal of Computer Vision*, vol. 132, no. 1, pp. 137–160, 2024.
- [54] G. Patterson, C. Xu, H. Su, and J. Hays, "The sun attribute database: Beyond categories for deeper scene understanding," *International Journal of Computer Vision*, vol. 108, pp. 59–81, 2014.
- [55] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1062–1070.
- [56] H. Zhang, X. Cao, and R. Wang, "Audio visual attribute discovery for fine-grained object recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [57] X.-S. Wei, Y. Shen, X. Sun, H.-J. Ye, and J. Yang, "Learning attribute-aware hash codes for large-scale fine-grained image retrieval," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5720–5730, 2021.
- [58] S. Wang, J. Chang, H. Li, Z. Wang, W. Ouyang, and Q. Tian, "Learning to parameterize visual attributes for open-set fine-grained retrieval," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [59] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.
- [60] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [61] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in neural information processing systems*, vol. 34, pp. 12 116–12 128, 2021.
- [62] D. Jiang, Y. Liu, S. Liu, X. Zhang, J. Li, H. Xiong, and Q. Tian, "From clip to dino: Visual encoders shout in multi-modal large language models," 2023.
- [63] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning—the good, the bad and the ugly," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4582–4591.
- [64] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101—mining discriminative components with random forests," in *European Conference on Computer Vision*. Springer, 2014, pp. 446–461.
- [65] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3606–3613.
- [66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.
- [67] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2004, pp. 178–178.
- [68] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.
- [69] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.
- [70] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.
- [71] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conference on Computer Vision, Graphics & Image processing*. IEEE, 2008, pp. 722–729.
- [72] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3498–3505.
- [73] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3485–3492.
- [74] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.
- [75] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 262–15 271.
- [76] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 5389–5400.
- [77] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [78] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adaptor: Training-free adaption of clip for few-shot classification," in *European conference on computer vision*. Springer, 2022, pp. 493–510.
- [79] J. B. *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.



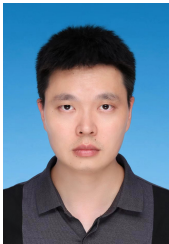
Xin Liu received a bachelor's degree in Information Security from the University of Science and Technology of China in 2022. She is now pursuing a master degree in Control Science and Engineering at University of Science and Technology of China. Her research interests include computer vision and deep learning, especially few-shot learning and multi-modal learning.



Jiamin Wu received the bachelor's degree in the School of Electronic Engineering, Xidian University, Xian, Shaanxi, China. She is studying for her doctorate in the Department of Automation, University of Science and Technology of China, Hefei, Anhui, China. Her research interests include pattern recognition, computer vision and deep learning. She is currently focusing on zero-shot and few-shot learning.



Wenfei Yang received the bachelor's degree in Electronic Engineering and Information Science in 2017, and the Ph.D. degree in pattern recognition and intelligent systems from the department of Automation, University of Science and Technology of China, Hefei, China, in 2022. Currently, he is a post-doctor in Control Science and Engineering, University of Science and Technology of China. His current research interests include computer vision and machine learning, especially action detection and object detection.



Xu Zhou received the PhD degree in computer science and technology from Huazhong University of Science and Technology in 2016. His research interests span the areas of large language model, NLP system design and reinforcement learning.



Tianzhu Zhang received the bachelor's degree in communications and information technology from Beijing Institute of Technology, Beijing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. Currently, he is a Professor at the Department of Automation, University of Science and Technology of China, Hefei, Anhui, China. His current research interests include computer vision and multimedia, especially action recognition, object classification, object tracking, and social event analysis.