

Multi-modal Attribute Prompting for Vision-Language Models

Xin Liu, Jiamin Wu, and Tianzhu Zhang

University of Science and Technology of China
`{attcb63442, jiaminwu}@mail.ustc.edu.cn, tzzhang@ustc.edu.cn`

Abstract. Large 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.

Keywords: Vision-Language Model · Few-Shot · Adaptation

1 Introduction

Large-scale pre-trained Vision-Language Models (VLMs), such as CLIP [38] and ALIGN [17], have demonstrated promising generalization power and transferability on a wide range of downstream tasks, including image classification [38], object detection [2, 20, 24] and 3D understanding [36, 42, 51]. 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. The extensive size of the CLIP model and the limited

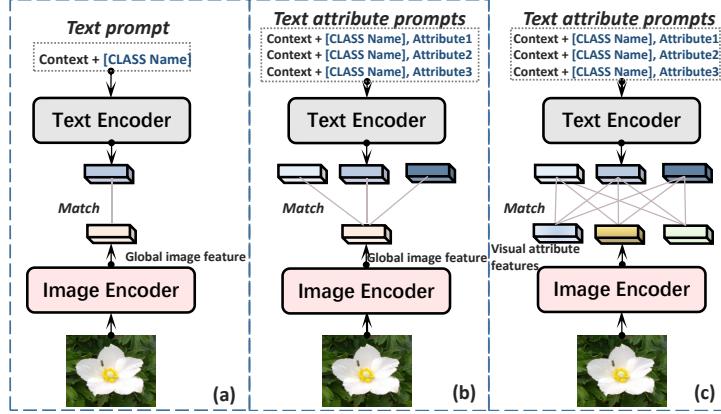


Fig. 1: (a) Conventional prompting methods use hand-crafted or learnable context in combination with the class name to construct the text prompt. (b) Recent methods introduce attribute descriptions to create text attribute prompts containing more semantic content. (c) Our method jointly explores multi-modal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

training data available make it unfeasible to fine-tune the complete model for use in downstream few-shot tasks.

To enhance the few-shot adaptation capability of CLIP, prompting techniques [1, 5, 6, 21, 25, 29, 50], such as CoOp [5] and CoCoOp [50] 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 [12, 31, 32], as shown in Figure 1 (b), introduce **textual attribute** descriptions obtained from Large Language Models [4, 34, 49]. 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 sim-

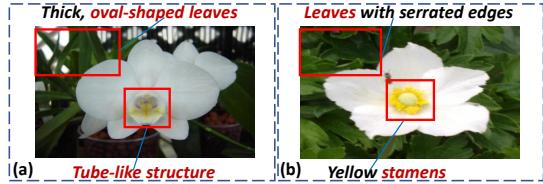


Fig. 2: (a) Moon Orchid and (b) Japanese Anemone exhibit strikingly similar overall appearances. Visual attributes play a crucial role in distinguishing between them, such as the central yellow stamens of Japanese Anemone.

ilar 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 textual 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 discriminative 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 [43] and use

the Sinkhorn algorithm [8] 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.

The contributions of our method could be summarized in three-fold: (1) We propose **Multi-modal Attribute Prompting**, which jointly explores textual attribute prompting, visual attribute prompting, and attribute-level alignment between images and text categories. (2) We enhance fine-grained visual perception ability by modeling visual attribute features with visual attribute prompts. Moreover, we introduce attribute-level alignment, complementing global alignment, to achieve multi-level robust 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. (3) Extensive experimental results on 11 benchmark datasets demonstrate that our method performs favorably against state-of-the-art approaches.

2 Related work

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

Pre-trained Vision-Language Models. In recent years, pre-trained vision-language models [17, 38, 41, 47, 48], have shown exceptional performance in diverse downstream tasks. Among them, CLIP 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. 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.

Prompt Learning. Prompt learning is initially introduced in the field of natural language processing (NLP) [13, 19, 26–28, 37, 40]. 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 [5] constructs prompts by concatenating learnable continuous vectors and class name tokens. CoCoOp [50] 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 [5]. ProDA [29] optimizes a set of prompts by learning the distribution of prompts. Instead of focusing on text-modal prompts, VPT [18] introduces learnable vectors to the Vision Transformer [10] to refine image features within the frozen vision encoder. DAPT [6], RPO [25], and MaPLe [21] improve the generalization ability of VLMs via multimodal prompting. These methods rely solely on class names for text prompt construction and may struggle to fully encapsulate categorical semantics.

Textual Attribute Prompts. To enrich the semantic description for different classes, recent works [12, 31, 32], 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 [4, 34]. 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.

3 Method

In this section, we first provide a concise overview of CLIP [38]. 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.

3.1 Review of CLIP

The Contrastive Language-Image Pre-training (CLIP) model [38] 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 [22], 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 C class names, the text prompts $\{t_i\}_{i=1}^C$ are formulated as manually designed templates, such as ‘A photo of a [CLASS].’ The classification vectors $\{w_i\}_{i=1}^C$ are derived by passing text prompts $\{t_i\}_{i=1}^C$ to the text encoder: $w_i = \theta(t_i)$. Given an image x and its label y , the global image feature f is extracted by the image encoder: $f = \phi(x)$. The classification probability is formulated as

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

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

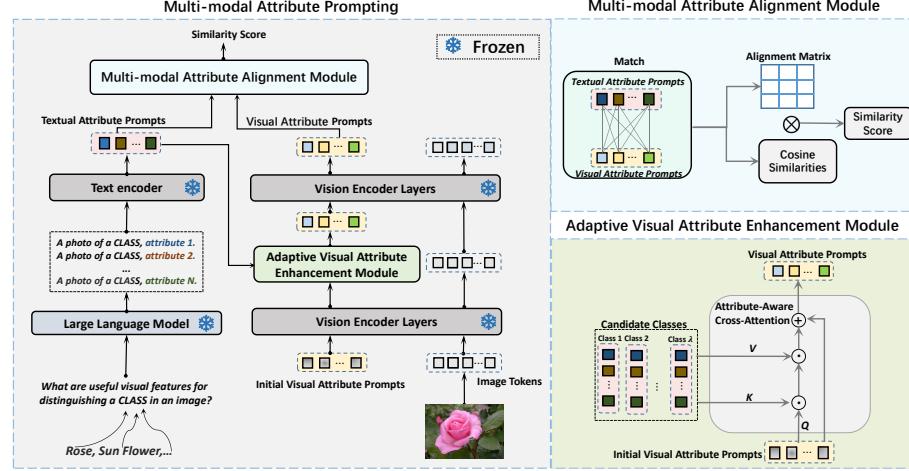


Fig. 3: The architecture of our method: **MAP** leverages textual attribute descriptions to construct textual attribute prompts and incorporates learnable visual attribute prompts for capturing visual attributes. In the **Adaptive Visual Attribute Enhancement** module, initial visual attribute prompts 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.

3.2 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 [12, 31, 32], 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: $p_k^n = \{A \text{ photo of a class } (k), t_{n,k}\}$, where class (k) denotes the class name corresponding to the k -th class, and $t_{n,k}$ 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. 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.

3.3 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 $[., .]$ 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.

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 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.

3.4 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 [43]. For simplicity, we refer to $\{u_i^{(L)}\}_{i=1}^M$ as $\mathbf{F} = \{f_m\}_{m=1}^M$ hereafter. OT aims to find the optimal transportation plan with the minimal cost between two discrete distributions $\mu \in \mathbb{R}^M$, $\nu \in \mathbb{R}^N$. The optimal transportation plan \mathbf{T}^* is obtained by minimizing the transportation cost:

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

where $\Pi(\mu, \nu)$ is the joint distribution with marginals μ and ν , $\langle \cdot, \cdot \rangle$ denotes the cosine similarity, and $\mathbf{C} \in \mathbb{R}^{M \times N}$ represents the cost matrix of transporting μ to ν . The Problem in Equation (7) can be efficiently solved by the Sinkhorn algorithm [8]. To apply OT to the multi-modal attribute matching task, we formulate the attribute prompt sets as discrete uniform distributions. The cost matrix \mathbf{C} is defined as the distances between attribute prompts. By solving Equation (7), 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:

$$\psi(\mathbf{F}, \mathbf{G}_k) = \sum_{m=1}^M \sum_{n=1}^N \langle f_m, g_k^n \rangle \mathbf{T}_{mn}^*, \quad (8)$$

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

3.5 Training Objectives

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

$$P_a(y = i|x) = \frac{\exp(\psi((\mathbf{F}, \mathbf{G}_i)/\tau))}{\sum_{j=1}^C \exp(\psi(\mathbf{F}, \mathbf{G}_j/\tau))}. \quad (9)$$

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

$$P_g(y = i|x) = \frac{\exp(\cos((\mathbf{f}, \bar{\mathbf{g}}_i)/\tau))}{\sum_{j=1}^C \exp(\cos(\mathbf{f}, \bar{\mathbf{g}}_j/\tau))}, \quad (10)$$

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

$$P(y = i|x) = P_g(y = i|x) + \beta P_a(y = i|x), \quad (11)$$

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:

$$L_{cls} = -\frac{1}{B} \sum_{i=1}^B \log P(y = y_i|x_i), \quad (12)$$

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

4 Experiments

We evaluate MAP in four settings: (1) generalization from base to novel classes within a dataset; (2) few-shot image classification; (3) domain generalization; (4) cross-dataset evaluation. All models used are based on the open-source CLIP [38].

Datasets. We use the 11 image recognition datasets following the setup in CoOp [5] for the first two settings. The benchmark includes Food101 (Foo) [3], DTD [7], Imagenet (Img) [9], Caltech101 (Cal) [11], EuroSAT (Eur) [14], StanfordCars (Car) [23], FGVC Aircraft (FGV) [30], Flowers102 (Flo) [33], Oxford-Pets (Pet) [35], UCF101 (UCF) [35], and SUN397 (SUN) [46]. In the domain generalization setting, we utilize ImageNet as the source dataset, and its distinct domain variants, namely ImageNet-R (-R) [15], ImageNet-A (-A) [16], ImageNetV2 (V2) [39], and ImageNet-Sketch (-S) [44] as target datasets.

Implementation Details. In all the experiments, we use the pre-trained CLIP [38] with ViT-B/16 image encoder backbone as the base model. We use GPT-3.5 as the LLM. 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

Table 1: Accuracy comparison on Base-to-novel generalization of MAP with previous methods. HM: Harmonic mean to highlight the generalization trade-off [45].

| Dataset | | CLIP [38] | CoOp [5] | CoCoOp [50] | ProDA [29] | VDT-Adapter [31] | MaPLe [21] | MAP (Ours) |
|---------------------------|-------|--------------|-------------|----------------|---------------|---------------------|---------------|---------------|
| Average on 11 datasets | Base | 69.34 | 82.69 | 80.47 | 81.56 | 82.48 | 82.28 | 83.66 |
| | Novel | 74.22 | 63.22 | 71.69 | 72.30 | 74.51 | 75.14 | 75.76 |
| | HM | 71.70 | 71.66 | 75.83 | 76.65 | 78.09 | 78.55 | 79.36 |
| ImageNet | Base | 72.43 | 76.47 | 75.98 | 75.40 | 76.4 | 76.66 | 76.60 |
| | Novel | 68.14 | 67.88 | 70.43 | 70.23 | 68.3 | 70.54 | 70.60 |
| | HM | 70.22 | 71.92 | 73.10 | 72.72 | 72.12 | 73.47 | 73.48 |
| Caltech101 | Base | 96.84 | 98.00 | 97.96 | 98.27 | 98.3 | 97.74 | 98.30 |
| | Novel | 94.00 | 89.81 | 93.81 | 93.23 | 95.9 | 94.36 | 93.80 |
| | HM | 95.40 | 93.73 | 95.84 | 95.68 | 97.09 | 96.02 | 96.00 |
| OxfordPets | Base | 91.17 | 93.67 | 95.20 | 95.43 | 94.4 | 95.43 | 95.43 |
| | Novel | 97.26 | 95.29 | 97.69 | 97.83 | 97.0 | 97.76 | 96.90 |
| | HM | 94.12 | 94.47 | 96.43 | 96.62 | 95.68 | 96.58 | 96.16 |
| Stanford Cars | Base | 63.37 | 78.12 | 70.49 | 74.70 | 76.8 | 72.94 | 76.70 |
| | Novel | 74.89 | 60.40 | 73.59 | 71.20 | 72.9 | 74.00 | 73.73 |
| | HM | 68.65 | 68.13 | 72.01 | 72.91 | 74.80 | 73.47 | 75.18 |
| Flowers102 | Base | 72.08 | 97.60 | 94.87 | 97.70 | 97.4 | 95.92 | 97.57 |
| | Novel | 77.80 | 59.67 | 71.75 | 68.68 | 75.3 | 72.46 | 75.23 |
| | HM | 74.83 | 74.06 | 81.71 | 80.66 | 84.94 | 82.56 | 84.95 |
| Food101 | Base | 90.10 | 88.33 | 90.70 | 90.30 | 90.4 | 90.71 | 90.30 |
| | Novel | 91.22 | 82.26 | 91.29 | 88.57 | 91.2 | 92.05 | 89.30 |
| | HM | 90.66 | 85.19 | 90.99 | 89.43 | 90.80 | 91.38 | 89.80 |
| FGVC Aircraft | Base | 27.19 | 40.44 | 33.41 | 36.90 | 37.8 | 37.44 | 41.63 |
| | Novel | 36.29 | 22.30 | 23.71 | 34.13 | 33.0 | 35.61 | 36.43 |
| | HM | 31.09 | 28.75 | 27.74 | 35.46 | 35.24 | 36.50 | 38.84 |
| SUN397 | Base | 69.36 | 80.60 | 79.74 | 78.67 | 81.4 | 80.82 | 82.33 |
| | Novel | 75.35 | 65.89 | 76.86 | 76.93 | 76.8 | 78.70 | 76.30 |
| | HM | 72.23 | 72.51 | 78.27 | 77.79 | 79.03 | 79.75 | 79.20 |
| DTD | Base | 53.24 | 79.44 | 77.01 | 80.67 | 81.8 | 80.36 | 82.63 |
| | Novel | 59.90 | 41.18 | 56.00 | 56.48 | 62.3 | 59.18 | 66.23 |
| | HM | 56.37 | 54.24 | 64.85 | 66.44 | 70.73 | 68.16 | 73.53 |
| EuroSAT | Base | 56.48 | 92.19 | 87.49 | 83.90 | 88.5 | 94.07 | 92.13 |
| | Novel | 64.05 | 54.74 | 60.04 | 66.00 | 70.5 | 73.23 | 76.10 |
| | HM | 60.03 | 68.69 | 71.21 | 73.88 | 78.48 | 82.35 | 83.33 |
| UCF101 | Base | 70.53 | 84.69 | 82.33 | 85.23 | 84.1 | 83.00 | 86.67 |
| | Novel | 77.50 | 56.05 | 73.45 | 71.97 | 76.4 | 78.66 | 78.77 |
| | HM | 73.85 | 67.46 | 77.64 | 78.04 | 80.07 | 80.77 | 82.52 |

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).

4.1 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 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

Table 2: Domain generalization. Prompting methods are trained on ImageNet and evaluated on datasets with domain shifts.

| Source | Target | | | | | |
|--------|--------------|--------------|--------------|--------------|--------------|--------------|
| | ImageNet | ImageNet-V2 | ImageNet-S | ImageNet-A | ImageNet-R | Avg. |
| CoOp | 71.51 | 64.20 | 47.99 | 49.71 | 75.21 | 59.28 |
| CoCoOp | 71.02 | 64.07 | 48.75 | 50.63 | 76.18 | 59.91 |
| MaPLe | 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 3: Cross-dataset evaluation. Prompting methods 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 | 71.51 | 93.70 | 89.14 | 64.51 | 68.71 | 85.30 | 18.47 | 64.15 | 41.92 | 46.39 | 66.55 |
| CoCoOp | 71.02 | 94.43 | 90.14 | 65.32 | 71.88 | 86.06 | 22.94 | 67.36 | 45.73 | 45.37 | 68.21 |
| MaPLe | 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 |

are averaged. We evaluate accuracy on test data corresponding to both the base and novel classes and use their harmonic mean [45] as the final evaluation metric.

Compared to CoOp, MAP exhibits higher harmonic mean accuracy across all datasets. As shown in Table 1, 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 also provide results of several recent methods in Table 1, MAP outperforms other methods in both base classes and novel classes. It’s worth noting that VDT-Adapter [31], which utilizes textual attributes obtained from GPT-4 to construct prompts, improves novel accuracy compared to CoOp. However, it neglects modeling visual attributes and fails to fully leverage the role of attributes. MAP outperforms VDT-Adapter 1.18% in base classes and 1.25% in novel classes.

4.2 Few-Shot Image Classification

To evaluate few-shot learning ability, we adopt the few-shot evaluation protocol from CLIP [38], 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

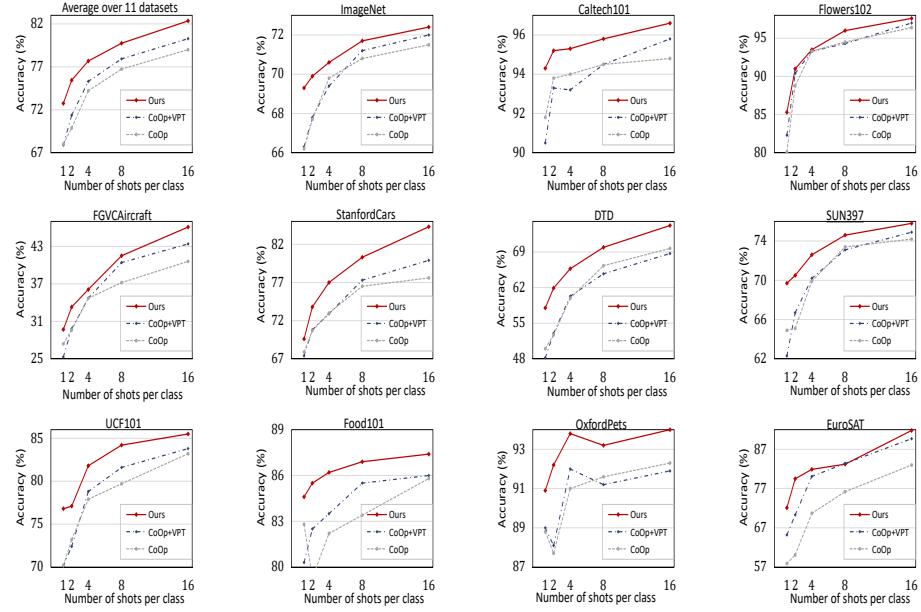


Fig. 4: Main results of few-shot image classification on 11 datasets. MAP consistently outperforms other CLIP adaptation methods across all datasets, demonstrating the strong few-shot adaptability of MAP.

prompts [18] 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 enhanced by textual guidance. Furthermore, on certain datasets (Caltech101, Flowers102, DTD, SUN397, and Oxford-Pets), 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.

4.3 Domain Generalization

We evaluate the generalizability of MAP on out-of-distribution data through a comparison with CoOp, CoCoOp, and MaPLe. The overall results are summarized in Table 2. MAP not only attains the highest accuracy on ImageNet but also exhibits superior performance on ImageNetV2, ImageNet-A, and ImageNet-R, demonstrating the robustness of MAP.

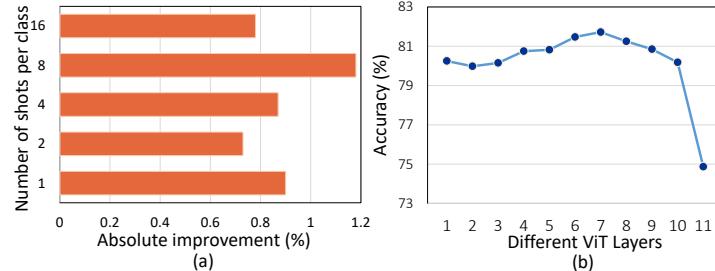


Fig. 5: Average few-shot image classification results over 6 datasets. (a) Absolute accuracy improvements provided by using **AVAE** compared to scenarios without **AVAE**. (b) The impact of inserting **AVAE** into different layers of ViT with 1 shot per class.

4.4 Cross-Dataset Evaluation

Table 3 summarizes the results of MAP and previous methods on cross-dataset evaluation benchmark. On the source dataset, MAP achieves the highest score. Compared with CoOp, CoCoOp, and MaPLE, MAP shows favorable performance and achieves better generalization in 7/10, 6/10, and 6/10 datasets respectively.

4.5 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 4 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 4: Ablation results averaged over 11 datasets in the base-to-novel setting.

| 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 |

Effectiveness of Adaptive Visual Attribute Enhancement. To verify the accuracy improvement when using AVAE, we conduct few-shot image classification experiments on 6 datasets (Flo, DTD, UCF, Pet, Cal, Foo). As shown in Figure 5 (a), the employment of AVAE brings remarkable performance gains. Furthermore, we investigate the impact of placing AVAE into different ViT layers. As observed from Figure 5 (b), 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 6 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.

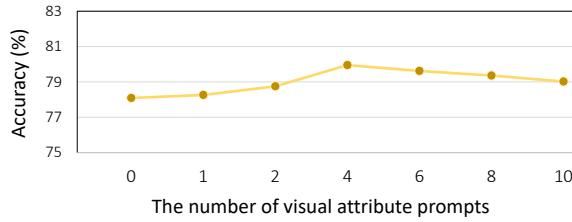


Fig. 6: The impact of the number of visual attribute prompts in the base-to-novel generalization setting.

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.

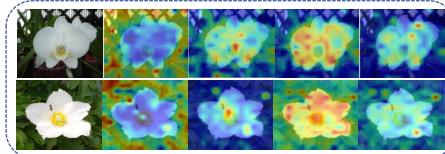


Fig. 7: The visualization of visual attribute prompts for the category “Moon Orchid” and “Japanese Anemone”. Guided by textual attribute semantics, visual attribute prompts focus on distinctive visual details, such as different leaf shapes.

5 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.

References

1. et al., K.: Self-regulating prompts: Foundational model adaptation without forgetting. In: ICCV (2023) [2](#)
2. Arandjelović, R., Andonian, A., Mensch, A., Hénaff, O.J., Alayrac, J.B., Zisserman, A.: Three ways to improve feature alignment for open vocabulary detection. arXiv preprint arXiv:2303.13518 (2023) [1](#)
3. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101-mining discriminative components with random forests. In: European Conference on Computer Vision. pp. 446–461. Springer (2014) [9](#)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020) [2, 5](#)
5. Cho, E., Kim, J., Kim, H.J.: Distribution-aware prompt tuning for vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22004–22013 (2023) [2, 4, 5, 9, 10](#)
6. Cho, E., Kim, J., Kim, H.J.: Distribution-aware prompt tuning for vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22004–22013 (2023) [2, 5](#)
7. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3606–3613 (2014) [9](#)
8. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013) [4, 8](#)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. Ieee (2009) [9](#)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) [5](#)
11. Fei-Fei, L., Fergus, R., Perona, P.: 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. pp. 178–178. IEEE (2004) [9](#)
12. Feng, Z., Bair, A., Kolter, J.Z.: Leveraging multiple descriptive features for robust few-shot image learning. arXiv preprint arXiv:2307.04317 (2023) [2, 5, 6](#)
13. Gu, Y., Han, X., Liu, Z., Huang, M.: Ppt: Pre-trained prompt tuning for few-shot learning. arXiv preprint arXiv:2109.04332 (2021) [4](#)
14. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019) [9](#)
15. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., 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. pp. 8340–8349 (2021) [9](#)
16. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021) [9](#)

17. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021) [1](#), [4](#)
18. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022) [5](#), [12](#)
19. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? Transactions of the Association for Computational Linguistics **8**, 423–438 (2020) [4](#)
20. Kaul, P., Xie, W., Zisserman, A.: Multi-modal classifiers for open-vocabulary object detection. arXiv preprint arXiv:2306.05493 (2023) [1](#)
21. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023) [2](#), [5](#), [10](#)
22. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020) [5](#)
23. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 554–561 (2013) [9](#)
24. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint arXiv:2209.15639 (2022) [1](#)
25. Lee, D., Song, S., Suh, J., Choi, J., Lee, S., Kim, H.J.: Read-only prompt optimization for vision-language few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1401–1411 (2023) [2](#), [5](#)
26. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021) [4](#)
27. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021) [4](#)
28. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: Gpt understands, too. AI Open (2023) [4](#)
29. Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5206–5215 (2022) [2](#), [5](#), [10](#)
30. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013) [9](#)
31. Maniparambil, M., Vorster, C., Molloy, D., Murphy, N., McGuinness, K., O'Connor, N.E.: Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 262–271 (2023) [2](#), [5](#), [6](#), [10](#), [11](#)
32. Menon, S., Vondrick, C.: Visual classification via description from large language models. arXiv preprint arXiv:2210.07183 (2022) [2](#), [5](#), [6](#)
33. Nilshback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics & Image processing. pp. 722–729. IEEE (2008) [9](#)
34. OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article (2023) [2](#), [5](#)
35. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3498–3505. IEEE (2012) [9](#)

36. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 815–824 (2023) [1](#)
37. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H., Riedel, S.: Language models as knowledge bases? arXiv preprint arXiv:1909.01066 (2019) [4](#)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) [1](#), [4](#), [5](#), [9](#), [10](#), [11](#)
39. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International Conference on Machine Learning. pp. 5389–5400. PMLR (2019) [9](#)
40. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980 (2020) [4](#)
41. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022) [4](#)
42. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. arXiv preprint arXiv:2306.13631 (2023) [1](#)
43. Villani, C.: Optimal transport: old and new, vol. 338. Springer (2009) [3](#), [8](#)
44. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in Neural Information Processing Systems **32** (2019) [9](#)
45. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning—the good, the bad and the ugly. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4582–4591 (2017) [10](#), [11](#)
46. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3485–3492. IEEE (2010) [9](#)
47. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021) [4](#)
48. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18123–18133 (2022) [4](#)
49. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023) [2](#)
50. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022) [2](#), [4](#), [10](#)
51. Zhu, C., Zhang, W., Wang, T., Liu, X., Chen, K.: Object2scene: Putting objects in context for open-vocabulary 3d detection. arXiv preprint arXiv:2309.09456 (2023) [1](#)