Prompt Tuning In a Compact Attribute Space

Shiyu Hou[1,2], **Tianfei Zhou**[1], **Shuai Zhang**[2], **Ye Yuan**[1,*], **Guoren Wang**[1]

[1] Beijing Institute of Technology
[2] Beijing Zhongguancun Laboratory

Abstract

Prompt tuning (PT) has emerged as a key to unlocking the power of visual-language models like CLIP for various downstream tasks. Predominant approaches learn a small set of task-relevant soft prompts by solving an image-class matching problem. Nevertheless, by optimizing merely with respect to class names, they face challenges in learning high performant prompts capable of capturing fine-grained, diverse characteristics of each class, and tends to overfit potentially biased distribution of base classes. In this work, we propose **PTinCAS** to tackle prompt tuning in a compact attribute space, driven by the premise that attributes offer *detailed* class interpretations and can facilitate *transfer* across related categories. Particularly, **PTinCAS** is grounded in two innovative designs. First, we create a compact attribute space by properly prompting large language models to generate factual descriptions about categories, which are subsequently clustered to form a concise attribute vocabulary. Second, we leverage attributes as a source of supervision in PT to transfer the inherent common sense knowledge in attributes to soft prompts. An object-aware visual prompting mechanism is developed to effortlessly highlight intended regions in the original image, which guides the model towards learning visual attributes associated with object regions rather than the background. We show that **PTinCAS** not only improves few-shot *generalizability* compared to existing PT methods, but also provides some level of inherent *explainability* that helps us understand why a class name is determined based on the attributes activated in an image.

Code — https://github.com/hhhoushiyu/PTinCAS

Introduction

Foundation vision-language models (VLMs), represented by CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), have established their preeminence in a wide range of computer vision tasks like image classification (Zhang et al. 2022; Abdelfattah et al. 2023), object detection (Gu et al. 2021; Shi et al. 2022) and semantic segmentation (Zhou, Loy, and Dai 2022; Jiao et al. 2023; Zhou et al. 2024). These models are constratively trained on extensive corpora of paired image-text data sourced from the internet, and offer prosperous representations that can be effectively guided through well-designed input prompts (*e.g.*, `A photo of a [CLASS]`) to execute tasks in a zero-shot fashion. Nevertheless, hand-crafting prompts is a specialized alchemy, demanding intensive human efforts and potentially introducing artificial bias.

An alternative approach popularized by CoOp (Zhou et al. 2022b) is *Prompt Tuning*, which replaces hand-crafted discrete prompts with a sequence of continuous-valued context vectors, known as soft prompts. The method achieves optimal prompts by minimally updating context vectors using task-specific training examples, keeping the pre-trained weights of VLMs frozen. This data-driven paradigm has increasingly garnered significant interests within the community (Zhou et al. 2022a; Khattak et al. 2023a; Yao, Zhang, and Xu 2023; Lu et al. 2022; Zhang et al. 2024b; Hantao Yao 2024; Cho, Kim, and Kim 2023; Chen et al. 2023a). These approaches tune prompts based on the most common *naming* criteria, for example, the image in Fig. 1 is named as `Cat`. However, class names entail limited semantic information and exhibit poor transferability. The original contrastive learning strategy mainly focuses on high-level semantic information while overlooking visual details (Wang et al. 2024). These pose challenges to learn high-performing prompts capable of perceiving fine-grained visual patterns and generalizing to unseen scenarios.

Upon re-examining the image in Fig. 1, we can perceive additional detailed information beyond mere recognition of a cat, such as "`a cat with pointed ears`", "`shiny coat`", or "`large, rounded eyes`". These descriptions are commonly referred to as visual attributes. In general, a visual attribute refers to any visual property that has a semantic connotation, delineating aspects such as parts ("`ears`", "`eyes`"), size ("`shiny coat`"), and shape ("`large and rounded`"). With this regard, attributes serve to characterize objects in far greater details than the standard phraseology of a class name (Murphy 2004). Furthermore, attributes can well characterize the connections among categories, and thus play important roles in recognizing unknown categories (Akata et al. 2015; Xu et al. 2020; Huynh and Elhamifar 2020). Last, attributes are associated with real-world meanings and human-understandable, making it more suitable to learn interpretable classifiers com-

*Corresponding author

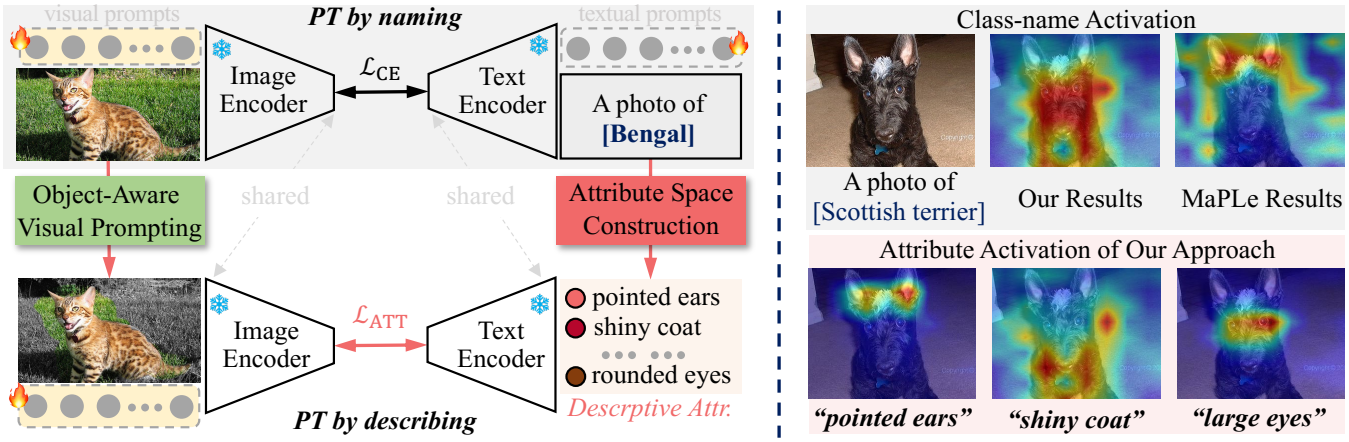

Figure 1: (Left) Popular methods learn soft prompts as a *PT by naming* task, *i.e.*, align image with descriptive name (*e.g.*, `a photo of Bengal`) via a cross-entropy loss $\mathcal{L}_{\text{CE}}$. In contrast, **PTinCAS** additionally introduces *PT by describing* as a core task. It builds a compact attribute space and learns via $\mathcal{L}_{\text{ATT}}$ (Eq. 3) to describe images with more fine-grained, generalizable, and interpretable primitive attributes. (Right) The new objective instructs **PTinCAS** to look at the right region related to the unknown category (`Scottish terrier`), and surprisingly, to localize the attributes without any localization annotations.

pared to other features (Liu et al. 2019). Despite this, there remains a notable gap in understanding how visual attributes can inform the creation of better prompts and ultimately enhance the transfer of VLMs. While exceptions (Tian et al. 2024; Kan et al. 2023) exist, they primarily leverage linguistic attribute descriptions to enrich the input context of the language encoder, and the attributes employed are uninformative, *e.g.*, (Tian et al. 2024) only extracts three attributes for each image. Thus, they remain constrained in fully exploring the potential of attribute-centric learning to enhance generalizability and interpretability of VLMs.

In light of the foregoing discussions, we are motivated to address the gap by tackling PT from an attribute-based learning perspective. Moving beyond the straightforward *naming* criterion, we make *describing*, *i.e.*, inferring visual attributes, a core problem in prompt optimization. In particular, we investigate how detailed supervision of attributes can be harnessed to develop more effective prompts. Facing this issue involves two main challenges. • The first challenge is to find an expressive set of attributes that is amendable to describe various visual objects. The process must be scalable and automated to facilitate seamless transfer across downstream tasks. • The second challenge is leveraging these attributes to optimally guide the learning of soft prompts.

We propose a new approach **PTinCAS** that performs PT in a compact attribute space. • First, **PTinCAS** employs Large Language Models (LLMs) like GPT (Brown et al. 2020) as a commonsense knowledge engine to generate a plethora of factual attribute descriptions for each category. These raw LLM-generated attributes often include noisy information and fail to convey the inherent structure of common attributes shared by multiple classes (Yan et al. 2023). To tackle this, **PTinCAS** clusters raw attribute descriptions into a concise set of representative attributes, thereby creating a compact yet transferable attribute space, in which each cluster center is related to a primitive attribute. This addresses the first challenge. • To tackle the second challenge, **PTinCAS** takes the class-attribute association derived from clustering and formulates attribute-based learning as a multi-label classification problem to guide the model towards inferring potential attributes present within each image. For more reliable perception of object attributes, we incorporate an object-aware visual prompting scheme which highlights intended objects and reduces background interference. Notably, **PTinCAS** represents a multi-task learning framework (Fig. 1), which learns to predict visual attributes in concert with class names. This differs from many studies (Yu et al. 2013; Farhadi et al. 2009; Yang et al. 2023; Menon and Vondrick 2023) that treat attributes as intermediate representations, and allows our method to learn shared prompts amenable to both naming and describing tasks.

PTinCAS shows a few attractive qualities. First, it effectively enhances recognition capability. The advantage is particularly strong for unseen classes, likely because attributes facilitate better transfer across categories. Second, it gains fine-grained localization ability *w.r.t.* attributes without explicit supervision (Fig. 1(right)). Last, in addition to simply naming the images, **PTinCAS** enables models to classify images as interpretable attributes like "`large eyes`". This helps us understand why the model makes a particular naming decision, enhancing the explainability of VLMs.

Related Work

Prompt Tuning in VLMs. PT, which originates from the language domain, has emerged as an efficient adaptation technique to transfer pre-trained large VLMs like CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) to downstream tasks. It follows the data-driven paradigm to learn a small set of prompt tokens, while keeping the pre-trained weights unchanged. CoOp (Zhou et al. 2022b) is a pioneering effort that learns a continuous set of textual prompt vectors alongside the class name. As CoOp suffers

in severe overfitting, numerous efforts have been dedicated to solve the generalization issue by, *e.g.*, explicitly conditioning prompts on image instances (Zhou et al. 2022a; Derakhshani et al. 2023), regularizing soft prompts with general knowledge embedded in hand-crafted prompts (Yao, Zhang, and Xu 2023; Zhu et al. 2023a,b), jointly learning visual-textual prompts (Khattak et al. 2023a), investigating multiple prompt ensemble (Khattak et al. 2023b; Chen et al. 2023a; Lee et al. 2023; Chen et al. 2023b), modeling the probabilistic distribution of prompts (Derakhshani et al. 2023; Lu et al. 2022), decoupling knowledge in features (Zhang et al. 2023), or incorporating external knowledge (Chen et al. 2024; Zhang et al. 2024a; Kan et al. 2023).

Though impressive, these solutions address PT by simply naming objects at a basic level. In this study, we push further the boundaries by advancing PT to describe objects with attributes. Our approach establishes a compact attribute space by extracting an expressive set of descriptive attributes commonly shared across categories, thereby facilitating the learning of more generalizable prompts. This differs from (Tian et al. 2024), which conditions the model on a limited number (*i.e.*, 3) of image-specific attributes and thus requires an attribute generator even at inference time. Moreover, our approach uniquely unifies the PT-by-naming and PT-by-describing tasks within a multi-task learning framework, which allows the creation of shared prompts for both tasks. While DoubleRight (Mao et al. 2023) also adopts a dual-focus approach, it necessitates a costly process to manually collect ground-truth attributes (called rationals) for images. Finally, an appealing advantage of our solution is that, it enhances the interpretability of original VLMs, a facet that remains largely unexplored in current PT research.

Attribute-based Recognition. Attributes exhibits intriguing characteristics (Murphy 2004), such as inherently generalizable across classes, independently useful for describing known or unknown entities, and serving as mid-level features for classification. Consequently, they have been extensively explored in building classifiers with improved zero-shot performance (Akata et al. 2015; Xu et al. 2020; Farhadi et al. 2009; Yu et al. 2013; Xie et al. 2021a; Huynh and Elhamifar 2020; Xie et al. 2021b) or model interpretability (Menon and Vondrick 2023; Yang et al. 2023; Zhou and Wang 2024; Chiquier, Mall, and Vondrick 2024; Mao et al. 2023; Wang et al. 2023). Drawing inspiration from these past efforts, we propose an attribute-centric approach to enable more effective transfer of VLMs. Following recent advances (Yan et al. 2023; Chiquier, Mall, and Vondrick 2024; Saha, Horn, and Maji 2024; Liu et al. 2024) that treat LLMs as a common sense knowledge base, our method builds a nuanced and human-interpretable attribute space, within which soft prompts are optimized to accurately perceive attributes.

Fine-Grained Visual Prompting. Our work is also related to recent advances in fine-grained visual prompting. CPT (Yao et al. 2021) and RedCircle (Shtedritski, Rupprecht, and Vedaldi 2023) demonstrate that incorporating visual markers, such as colorful boxes or red ciricles, can enhance VLMs' ability to differentiate objects. FGVP (Yang et al. 2024) takes this further by developing detailed markings based on high-quality segmentation masks derived from Segment Anything (SAM) (Kirillov et al. 2023). Much like FGVP, our approach incorporates an object-aware visual prompting scheme to emphasize object regions, facilitating more precise perception of object-aware attributes. However, our approach distinguishes from FGVP by deriving segmentation masks from CLIP activations, which is much more cost-effective and eliminates the need for external, computationally intensive segmentors.

Our Approach

Preliminary

Contrastive Language-Image Pre-training (CLIP). Given an image x with its associated label c, CLIP extracts the image feature from an image encoder as $\boldsymbol{x} = \Phi_{image}(\boldsymbol{v}(x)) \in \mathbb{R}^d$, and the textual feature from a text encoder $\boldsymbol{y}_c = \Phi_{text}(\boldsymbol{t}(y_c)) \in \mathbb{R}^d$, where y_c refers to the sentence "`a photo of {class name of c}`". Here, $\boldsymbol{v}(\cdot)$ and $\boldsymbol{t}(\cdot)$ compute image and language embeddings, respectively, and d is the feature dimensionality. The probability of classifying image x as class c is computed as:

$$P_{\texttt{CLIP}}(c|x) = \frac{\exp(\text{sim}(\boldsymbol{x}, \boldsymbol{y}_c)/\gamma)}{\sum_{j=1}^{N} \exp(\text{sim}(\boldsymbol{x}, \boldsymbol{y}_{c_j})/\gamma)}, \tag{1}$$

where N is the total number of classes, $\text{sim}(\cdot,\cdot)$ represents a metric function such as the cosine similarity, and γ denotes the temperature in Softmax.

Prompt Tuning of CLIP. For zero-shot recognition, CLIP requires engineering the template y_c before computing the above probability. Instead, PT methods like CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) learn a set of vectors $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m\}$, each of which is interpreted as a pseudo word embedding and m represents the length of these tokens. As such, for a given class c, these contextual vectors are concatenated with the embedding of the class name of c to form a textual prompt, denoted as $\tilde{\boldsymbol{t}}(y_c) = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m, \boldsymbol{t}(y_c)\}$. Moreover, methods like MaPLe (Khattak et al. 2023a) introduce n learnable vectors to the visual branch and construct a prompt $\tilde{\boldsymbol{v}}(x) = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n, \boldsymbol{v}(x)\}$. Following CLIP, the visual and textual features after prompting can be separately computed as: $\tilde{\boldsymbol{x}} = \Phi_{image}(\tilde{\boldsymbol{v}}(x))$ and $\tilde{\boldsymbol{y}}_c = \Phi_{text}(\tilde{\boldsymbol{t}}(y_c))$. In this manner, the prediction probability can be derived as follows:

$$P_{\texttt{PT}}(c|x) = \frac{\exp(\text{sim}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_c)/\gamma)}{\sum_{j=1}^{N} \exp(\text{sim}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_{c_j})/\gamma)}. \tag{2}$$

Then, the prompt parameters $\{\boldsymbol{t}_i\}_{i=1}^m$ and $\{\boldsymbol{v}_i\}_{i=1}^n$ can be optimized by minimizing the cross-entropy loss function $\mathcal{L}_{\texttt{CE}}(c, x) = -\log P_{\texttt{PT}}(c|x)$.

Prompt Tuning in a Compact Attribute Space

Attributes can inherently characterize the nature of object categories and their connections. Therefore, rather than merely addressing a naming task, our approach introduces an attribute-based criterion to enhance the soft prompts' awareness to fine-grained attributes. As illustrated in Fig. 1, this leads to a multi-task learning system in **`PTinCAS`**, which is solved by optimizing the following objective:

$$\mathcal{L} = \mathcal{L}_{\texttt{CE}}(c, x) + \lambda \mathcal{L}_{\texttt{ATT}}(\boldsymbol{A}^*, x). \tag{3}$$

(a) LLM Prompt Template (b) Descriptive Attributes from LLMs (c) T-SNE Illustration of Attribute Space

Figure 2: **Compact Attribute Space Construction.** **PTinCAS** internally leverages the world knowledge in GPT-3.5 via (a) linguistically prompting it to reason (b) descriptive attributes about each class name (*e.g.*, `british_shorthair`). Analogous attributes are further consolidated to create (c) a compact attribute space.

Here $\mathcal{L}_{\text{ATT}}$ measures the compliance of image x *w.r.t.* a set of attributes $\boldsymbol{A}^*$. The balancing weight λ captures the trade-off between the two terms. To solve Eq. 3, **PTinCAS** first constructs the attribute space $\boldsymbol{A}^*$ that describes categories, and then learns the prompts in the attribute space.

Compact Attribute Space Construction

(1) Acquiring Descriptive Attributes from LLMs. The initial step involves identifying relevant attributes that define each class. To ensure our method general-purpose and minimize human intervention, we tap into the expert knowledge of LLMs to collect visually-related attributes by asking the right language prompts. Concretely, we use GPT-3.5 to acquire descriptive attributes for each class c by providing its class name as input and receive a list of useful attributes as output: $\mathcal{Z}_c = \text{GPT-3.5}(\chi, c)$. Here χ refers to LLM-prompt, and we craft a template as in Fig. 2(a) to guide GPT generating descriptive attributes with consistent detail and uniform information types. $\mathcal{Z}_c = \{z_c^1, z_c^2, \ldots, z_n^{M_c}\}$ denote a set of M_c attributes related to class c, wherein each z_c^i is a linguistic description, *e.g.*, $z_c^i =$"*ears are feathered and hang down*" (see Fig. 2(b)).

(2) Attribute Space Construction with Descriptions. LLMs generate attributes for each class individually, and as a result, the descriptive attributes are not completely generalizable. Additionally, since current LLMs condense world knowledge noisily, the raw attributes they produce are not optimized for recognition tasks. To address these challenges, we propose streamlining the attribute space by consolidating analogous attributes, so as to uncover primitive attributes that are both generalizable across categories and accessible for computational analysis and human interpretation. Particularly, we begin by transforming the descriptive attributes, *e.g.*, z_c^i of class c, into a representation space by employing the pre-trained text encoder Φ_{text} of CLIP: $\boldsymbol{z}_c^i = \Phi_{text}(\boldsymbol{t}(z_c^i)) \in \mathbb{R}^d$. Next, we cluster attribute embeddings of *all* classes into K clusters by solving the problem:

$$\min_{\boldsymbol{A},\boldsymbol{S}} \sum\nolimits_{c,i} \|\boldsymbol{z}_c^i - \boldsymbol{A}\boldsymbol{s}_c^i\|, \quad s.t.\ \ \boldsymbol{s}_c^i \in \{0,1\}^K, \quad \mathbf{1}^\top \boldsymbol{s}_c^i = 1. \quad (4)$$

Here $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K] \in \mathbb{R}^{d\times K}$ represents the cluster centroid matrix, where $\boldsymbol{a}_k \in \mathbb{R}^d$ refers to the centroid (or prototype) of the k-*th* cluster. $\boldsymbol{S} = [\boldsymbol{s}_c^i]$ stores the cluster assignments of all attributes, and $\mathbf{1}$ is a K-dimensional all-one vector. While various clustering methods have been designed to solve Eq. 4, for simplicity, we use the most classic one – k-means, which determines the optimal $\boldsymbol{A}^*$ and $\boldsymbol{S}^*$ in an EM fashion. Here the K attribute prototypes in $\boldsymbol{A}^*$ span the attribute space (see Fig. 2(c)).

In addition, based on the assignment matrix $\boldsymbol{S}^*$, it is easy to embed a class c into the K-dim attribute space, denoted as $\boldsymbol{\rho}_c = [\rho_c^1, \ldots, \rho_c^K] \in \{0,1\}^K$. Here $\boldsymbol{\rho}_c$ indicates the desired label of class c in the attribute space, and serves as attribute supervision in subsequent learning stage. Its element $\rho_c^k = 1$ if at least one descriptive attribute related to class c is assigned to the cluster k; otherwise, $\rho_c^k = 0$.

Object-Aware Prompt Tuning under Attribute Supervision

In this section, we delve into PT within the attribute space. The task poses two challenges. First, while it is straightforward to solve $\mathcal{L}_{\text{ATT}}(\boldsymbol{A}^*, x)$ by deriving attribute predictions from x and imposing supervision from $\boldsymbol{\rho}_c$, the presence of task-unrelated information, such as background clutter, can complicate attribute inference and cause spurious correlations. Second, the attribute space is exclusively constructed within the linguistic domain, lacking integration of task-specific visual cues. Although this design benefits generalization, it may not sufficiently capture the specific distribution of attributes in downstream tasks. To address the first issue, we propose an object-aware visual prompting scheme aimed at transforming x into an object-sensitive counterpart $\hat{x}$, from which attributes can be more easily inferred. To address the second issue, we advocate for continuous updating $\boldsymbol{A}^*$ using task-specific cues during training to align with downstream task distribution.

(1) Object-aware Visual Prompting (OVP). OVP is a fine-grained visual prompting technique, sharing a similar spirit to (Yang et al. 2024). However, unlike (Yang et al. 2024) depending on SAM (Kirillov et al. 2023) to obtain segmentation masks, OVP is self-contained and does not rely on any external models. Our core idea is that class activation mappings (CAMs) (Zhou et al. 2016) in a classifier inherently accentuate object regions and serve as a valuable resource for extracting object masks. More concretely, for each input

(a) Original Image x (b) GradCAM (c) OVP Image $\hat{x}$

Figure 3: For image x labelling as, *e.g.*, `Maine Coon`, OVP computes its class activation and creates an object-aware prompt $\hat{x}$ via background grayscaling.

image x (Fig. 3(a)), we utilize GradCAM (Selvaraju et al. 2017) to derive the activations of the desired class name (Fig. 3(b)) from the pre-trained image encoder Φ_{image} in CLIP, *i.e.*, $\boldsymbol{m} = \texttt{GradCAM}_{\Phi_{image}}(x)$. The mask is then normalized within the range of $[0, 1]$, and converted to a binary mask via thresholding. Leveraging the binary mask, we generate an object-aware visual prompt (Fig. 3(c)) by modulating background pixels to grayscale, while preserving the foreground pixel values. Notably, we found that grayscaling is superior than alternative methods, *e.g.*, masking out or blurring the background area, because it reduces the impact of background meanwhile retains the global image context. This aspect is markedly crucial for high-quality attribute prediction, as certain attributes may be intricately linked to the background (*e.g.*, environment accompanying objects). In summary, OVP transforms an original image x to $\hat{x}$ by:

$$\hat{x} = \texttt{BackgroundGrayscaling}(\texttt{GradCAM}_{\Phi_{image}}(x)). \quad (5)$$

(2) Prompt Tuning in the Attribute Space. Upon receiving the prompted image $\hat{x}$, we compute its feature as $\hat{\boldsymbol{x}} = \Phi_{image}(\boldsymbol{v}(\hat{x})) \in \mathbb{R}^{hw \times d}$, where we retain its spatial dimension (*i.e.*, hw). Then, attribute predictions are obtained as:

$$\boldsymbol{H} = \hat{\boldsymbol{x}}\boldsymbol{A}^* \in \mathbb{R}^{hw \times K}, \quad \boldsymbol{\varphi} = \texttt{avgpool}(\boldsymbol{H}) \in \mathbb{R}^{K}. \quad (6)$$

Here $\boldsymbol{H}$ is an attention map, with each element $\boldsymbol{H}_{i,k}$ measuring the feature similarity between the *i-th* pixel and the *k-th* attribute prototype. Based on $\boldsymbol{\varphi}$, we are able to compute the probability of $\hat{x}$ belonging to each of the K attributes as: $P_{\texttt{ATT}}(\boldsymbol{A}^*|\hat{x}) = \texttt{sigmoid}(\boldsymbol{\varphi})$. Last, we employ the binary cross entropy loss as the attribute-aware criterion, *i.e.*, $\mathcal{L}_{\texttt{ATT}} = -\boldsymbol{\rho_c} \log P_{\texttt{ATT}}(\boldsymbol{A}^*|\hat{x})$.

(3) Task-Specific Attribute Updating. To get more task-aware attribute prototypes, our approach continuously refines prototypes during training by integrating pertinent visual cues from the specialized domain studied. Particularly, the updating is executed in a momentum manner as follows:

$$\boldsymbol{A}^* \leftarrow \alpha \boldsymbol{A}^* + (1-\alpha)\hat{\boldsymbol{x}}^\top \boldsymbol{H}, \quad (7)$$

where α is a momentum coefficient controlling the updating.

Experiment

Experimental Setting

As conventions (Zhou et al. 2022b; Khattak et al. 2023a,b; Zhang et al. 2024b), we evaluate **PTinCAS** on three types of tasks: 1) base-to-novel generalization; 2) cross-domain generalization; 3) few-shot classification.

Dataset and Metric. The experiments are conducted on 11 visual recognition datasets, including ImageNet (Deng et al. 2009), Caltech101 (Fei-Fei, Fergus, and Perona 2004), OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Gool 2014), FGVCAircraft (Maji et al. 2013), SUN397 (Xiao et al. 2010), DTD (Cimpoi et al. 2014), EuroSAT (Helber et al. 2019) and UCF101 (Soomro, Zamir, and Shah 2012). For cross-domain generalization, we additionally evaluate on four ImageNet variants including ImageNetV2 (Recht et al. 2019), ImageNet-Sketch (Wang et al. 2019), ImageNet-A (Hendrycks et al. 2021b) and ImageNet-R (Hendrycks et al. 2021a). We report recognition accuracy (%) and harmonic mean (HM) averaged over 3 seeds as final scores.

Training Detail. Following (Khattak et al. 2023a,b; Zhou et al. 2022a), we adopt a pre-trained ViT-B/16 CLIP model and predominantly employ a few-shot training approach. Specifically, experiments excluding few-shot classification are conducted utilizing a selection of 16 randomly sampled shots per class. **PTinCAS** can function as a plug-and-play module to be integrated with existing methods. Hence, we comprehensively evaluate with four PT baselines, *i.e.*, CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), MaPLe (Khattak et al. 2023a) and PromptSRC (Khattak et al. 2023b). We strictly adhere to the training setting specified in the official implementation of each baseline. All models are trained on a NVIDIA RTX 3090 GPU equipped with a 24GB memory. By default, the coefficient λ for $L_{\texttt{ATT}}$ is set as 2, and the clustering number K is set as 64.

Inference Detail. In testing phase, the attribute branch in **PTinCAS** can be entirely discarded, and the model executes in exactly the same manner to methods like (Khattak et al. 2023a,b; Zhou et al. 2022a), without introducing any computation overhead. This is superior over (Tian et al. 2024; Kan et al. 2023) that necessitate LLMs during inference.

Base-to-Novel Class Generalization

Setup. To assess **PTinCAS**'s generalizability to unseen classes, we follow (Khattak et al. 2023a) to divide each dataset into base and novel classes. The model is trained on base classes and tested on base and novel classes.

Result. Table 1 reports the performance of all methods averaged over 11 datasets. **PTinCAS** obtains consistent improvements on both base and novel classes, yielding gains of **2.66%** over CoOp, **0.99%** over CoCoOp, **0.83%** over MaPLe, and **0.75%** over PromptSRC in terms of HM. Particularly, **PTinCAS** achieves notable improvements to novel classes (**4.10%/1.55%/1.19%/1.04%** over CoOp/CoCoOp/MaPLe/PromptSRC), verifying the efficacy of attribute-based learning in engendering better generalizability.

Cross-Domain Generalization

Setup. This experiment is to evaluate the robustness of **PTinCAS** on out-of-distribution datasets. Models are trained on ImageNet (source) and evaluated on downstream

Method	Base	Novel	HM
CoOp	82.69	63.22	71.66
+ **PTinCAS**	**82.94** ↑ 0.25	**67.32** ↑ 4.10	**74.32** ↑ 2.66
CoCoOp	80.47	71.69	75.83
+ **PTinCAS**	**80.76** ↑ 0.29	**73.24** ↑ 1.55	**76.82** ↑ 0.99
MaPLe	82.28	75.14	78.55
+ **PTinCAS**	**82.69** ↑ 0.41	**76.33** ↑ 1.19	**79.38** ↑ 0.83
PromptSRC	84.26	76.10	79.97
+ **PTinCAS**	**84.64** ↑ 0.38	**77.14** ↑ 1.04	**80.72** ↑ 0.75

Table 1: Base-to-novel generalization performance average.

Method	Source	Target				
	ImageNet	-V2	-Sketch	-A	-R	**Avg.**
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
+PTinCAS	**71.87**	**64.46**	**48.33**	**50.30**	**75.77**	**59.71**
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91
+PTinCAS	**71.23**	**64.53**	**48.90**	**51.27**	**76.23**	**60.23**
MaPLe	70.72	64.07	49.15	50.90	76.98	60.27
+PTinCAS	**70.93**	**64.13**	**49.77**	**51.87**	**77.57**	**60.84**
PromptSRC	71.27	64.35	**49.55**	50.90	**77.80**	60.65
+PTinCAS	**71.93**	**65.23**	49.47	**51.13**	77.67	**60.88**

Table 2: Cross-domain generalization performance.

Method	1 shot	2 shots	4 shots	8 shots	16 shots
CoOp	67.56	70.65	74.02	76.98	79.89
+PTinCAS	**68.15**	**71.18**	**74.57**	**77.08**	**80.03**
CoCoOp	66.79	67.65	71.21	72.96	74.90
+PTinCAS	**68.92**	**69.95**	**71.81**	**73.61**	**75.45**
MaPLe	69.27	72.58	75.37	78.89	81.79
+PTinCAS	**71.42**	**74.69**	**76.79**	**79.51**	**82.88**
PromptSRC	72.32	75.29	78.35	80.69	82.87
+PTinCAS	**72.93**	**76.14**	**78.94**	**81.12**	**83.33**

Table 3: Few-shot classification performance.

datasets (target) in a zero-shot manner. For each of the 1K classes in ImageNet, we sample 16 examples for training.

Result. As shown in Table 2, **PTinCAS** sustains a competitive edge in source domain (71.87% *vs* 71.51% of CoOp, 71.23% *vs* 71.02% of CoCoOp, 70.93% *vs* 70.72% of MaPLe, 71.93% *vs* 71.27% of PromptSRC) and exhibits compelling overall performance within a diverse array of target downstream domains (**59.71%** *vs* 59.28% of CoOp, **60.23%** *vs* 59.91% of CoCoOp, **60.84%** *vs* 60.27% of MaPLe, **60.88%** *vs* 60.65% of PromptSRC). These results provide further corroboration that **PTinCAS** successfully learns more generalized knowledge, thereby enhancing the robustness to domain shifts.

Few-shot Classification

Setup. This experiment evaluates the model performance in an extremely limited data regime. We train models with $\{1, 2, 4, 8, 16\}$ examples per class for each dataset.

Result. We report the average performance over 11 datasets in Table 3. Benefiting from the role of attributes, **PTinCAS** demonstrates superior performance across different shots of training samples on four baselines and relatively larger gains in minimal data cases such as $\{1,2\}$ shots. These results further confirm the effectiveness of our approach.

λ (Eq. 3)	Base	Novel	HM
0	82.28	75.14	78.55
1	82.61	75.64	78.97
2	**82.69**	**76.33**	**79.38**
3	82.67	75.94	79.16

Table 4: Efficacy of $\mathcal{L}_{\text{ATT}}$.

Variants	Base	Novel	HM
MaPLe Baseline	82.28	75.14	78.55
Original image	82.29	75.50	78.75
Foreground-only	81.81	75.18	78.35
Background-blurring	82.62	75.92	79.13
Our OVP	**82.69**	**76.33**	**79.38**

Table 5: Efficacy of Object-aware Visual Prompting.

K (Eq. 4)	Base	Novel	HM
No Clustering	71.04	59.73	64.90
16	82.23	75.84	78.91
32	82.53	76.07	79.17
64	**82.69**	**76.33**	**79.38**
96	82.59	75.79	79.04

Table 6: Clustering number K.

Ablation Study

We conduct ablative experiments to investigate the effect of core designs in our approach. We use MaPLe as the baseline method and report average results across 11 datasets under the base-to-novel generalization setting.

Attribute Objective $\mathcal{L}_{\text{ATT}}$. We start by investigating the attribute objective $\mathcal{L}_{\text{ATT}}$ in Eq. 3. We analyze the impact of the loss coefficient λ and list the results in Table 4. Here $\lambda = 0$ refers to the baseline without $\mathcal{L}_{\text{ATT}}$. We observe non-trivial performance improvements by setting λ to 1. The gains become larger when increasing to 2. However, the model tends to degrade at $\lambda = 3$. Hence, we set $\lambda = 2$ by default.

Object-aware Visual Prompting. Next, we examine the effect of OVP by comparing it with three variants: 1) *original image* – employing original image as $\hat{x}$, 2) *foreground-only* – setting background pixels to zero, and 3) *background-blurring* – another fine-grained prompting scheme that applies Gaussian blurring with a 21×21 kernel to background pixels. The results are summarized in Table 5. We see that directly using *original image* leads to a slight improvement. The *foreground-only* variant leads to performance degradation. In contrast, by retaining background information, *background-blurring* and OVP lead to much better performance, and verifies our motivation that highlighting the foreground and weakening the background facilitates the learning of object-related attributes. Moreover, OVP achieves better performance compared to *background-blurring*.

Clustering Number K. Further, we study the impact of the clustering number K to **PTinCAS** in Table 6. Clustering emerges as a crucial component in **PTinCAS**, as evidenced

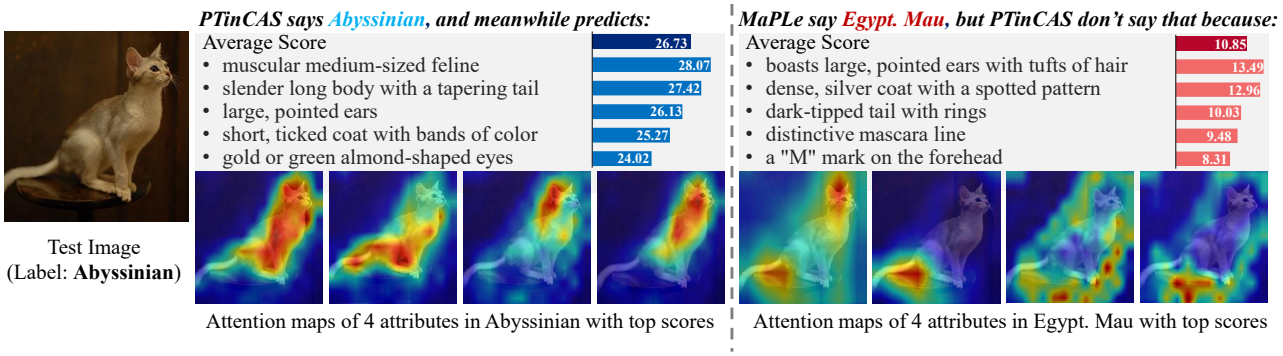

Figure 4: **PTinCAS provides explanations to its class-name decisions.** For an image labeled as `Abyssinian`, we show class-name decisions of **PTinCAS** and MaPLe. **PTinCAS** predicts the correct label `Abyssinian`, while its predictions to attributes associated with `Abyssinian` help justify the decision. In contrast, MaPLe makes the wrong decision `Egypt. Mau`, and **PTinCAS**'s predictions on attributes related to `Egypt. Mau` justify why **PTinCAS** did not select the answer. Meanwhile, **PTinCAS** is properly activated *w.r.t* fine-grained attributes associated with `Abyssinian`, but not to `Egypt. Mau`.

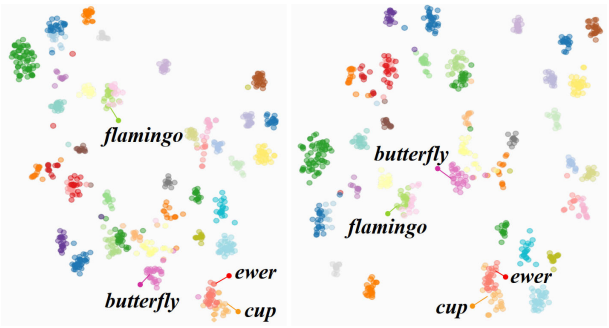

Figure 5: T-SNE visualization of visual embedding of MaPLe (left) and Our method (right).

by the significantly poor performance of the "no clustering" variant. Among the various cluster numbers we examined, the best performance is reached at $K = 64$. Nonetheless, **PTinCAS** maintains relatively stable performance within the range of $[16, 96]$, indicating that our model's performance is not sensitive to the selection of K. By default, we set $K = 64$ for all downstream tasks without specific tuning.

Study of Model Explainability

Furthermore, we show that **PTinCAS** can gain enhanced model explainability by incorporating attribute decisions. Notably, all attribute cluster assignments are stored in $\boldsymbol{S}^*$, which allows for the straightforward anchoring of each attribute prototype (cluster center) in $\boldsymbol{A}^*$ to a set of actual attribute descriptions generated by LLMs. **PTinCAS** enables its class-name decision to be human-understandable. More specifically, users can view and verify the model's confidence to specific attributes associated with the determined class-name. In Fig. 4, an example is correctly classified by **PTinCAS**, because it more confidently perceives the attributes such as `muscular medium-sized feline`, which are related to the correct label `Abyssinian`. However, MaPLe makes an incorrect class-name decision of `Egypt. Mau`, but **PTinCAS** disregards this error because it does not identify relevant attributes. Markedly, we extract attention maps from $\boldsymbol{H}$ (Eq. 6) for four attributes with the highest scores. And based on $\boldsymbol{S}^*$, we locate corresponding attribute descriptions. **PTinCAS** provides accurate pixel responses to descriptive attributes in the correct category, while remaining properly unactivated *w.r.t.* attributes in the wrong category. This fine-grained localization capability is impressive, especially considering that no any localization annotations are provided to the model.

Analysis of Visual Embedding

Fig. 5 presents the T-SNE visualization of visual embedding of MaPLe and **PTinCAS** on the Caltech101 dataset. As a PT-by-naming method, while MaPLe can distinguish between categories, sometimes encodes visually different images into similar embeddings. **PTinCAS** can learn the relationships between classes through attributes, encoding images with similar attributes into similar embeddings, thereby producing a more meaningful feature space (*e.g.*, `butterfly` is more closer to `flamingo`). The benefit of PT-by-describing to image encoder is further proved.

Conclusion

The vast majority of recent effort in PT seek to learn meaningful soft prompts by solving a naming task. In contrast, this paper explores an attribute-based learning regime to tackle PT as a describing task. This yields **PTinCAS**, which is comprised of two key components: the creation of a concise attribute space using the common sense knowledge from LLMs, and object-aware prompt tuning under the supervision of fine-grained attributes. **PTinCAS** operates as a multi-task learning framework that integrates our proposed PT-by-describing task together with the traditional PT-by-naming task. This allows for creating shared prompts that are useful for each of the two tasks. Through extensive experiments across various real-world datasets, we demonstrate the superior recognition and generalization performance of **PTinCAS** compared to existing PT algorithms. Beyond this, **PTinCAS** enjoys strong explainability in providing transparent model decisions.

Acknowledgments

This work was sponsored by the National Key R&D Program of China (Grant No.2022YFB2702100), the NSFC (Grant Nos. 61932004, 62225203, U21A20516), and CAAI-Lenovo Blue Sky Research Fund.

References

Abdelfattah, R.; Guo, Q.; Li, X.; Wang, X.; and Wang, S. 2023. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *ICCV*.

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2015. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(7): 1425–1438.

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101–mining discriminative components with random forests. In *ECCV*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Chen, G.; Tao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2023a. PLOT: Prompt Learning with Optimal Transport for Vision-Language Models. In *ICLR*.

Chen, H.; Li, Y.; Huang, Z.; Hong, Y.; Xu, Z.; Gu, Z.; Lan, J.; Zhu, H.; and Wang, W. 2024. Conditional Prototype Rectification Prompt Learning. In *CVPR*.

Chen, Z.; Huang, X.; Guan, Q.; Lin, L.; and Luo, W. 2023b. A Retrospect to Multi-prompt Learning across Vision and Language. In *ICCV*.

Chiquier, M.; Mall, U.; and Vondrick, C. 2024. Evolving Interpretable Visual Classifiers with Large Language Models. *arXiv preprint arXiv:2404.09941*.

Cho, E.; Kim, J.; and Kim, H. J. 2023. Distribution-Aware Prompt Tuning for Vision-Language Models. In *ICCV*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Derakhshani, M. M.; Sanchez, E.; Bulat, A.; da Costa, V. G. T.; Snoek, C. G.; Tzimiropoulos, G.; and Martinez, B. 2023. Bayesian prompt learning for image-language model generalization. In *ICCV*.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *CVPR*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Hantao Yao, C. X., Rui Zhang. 2024. TCP: Textual-based Class-aware Prompt tuning for Visual-Language Model. In *CVPR*.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *CVPR*.

Huynh, D.; and Elhamifar, E. 2020. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Jiao, S.; Wei, Y.; Wang, Y.; Zhao, Y.; and Shi, H. 2023. Learning mask-aware clip representations for zero-shot segmentation. *NeurIPS*.

Kan, B.; Wang, T.; Lu, W.; Zhen, X.; Guan, W.; and Zheng, F. 2023. Knowledge-aware prompt tuning for generalizable vision-language models. In *ICCV*.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023a. Maple: Multi-modal prompt learning. In *CVPR*.

Khattak, M. U.; Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2023b. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *ICCV*.

Lee, D.; Song, S.; Suh, J.; Choi, J.; Lee, S.; and Kim, H. J. 2023. Read-only Prompt Optimization for Vision-Language Few-shot Learning. In *ICCV*.

Liu, H.; Wang, R.; Shan, S.; and Chen, X. 2019. What is a tabby? Interpretable model decisions by learning attribute-based classification criteria. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5): 1791–1807.

Liu, M.; Roy, S.; Li, W.; Zhong, Z.; Sebe, N.; and Ricci, E. 2024. Democratizing Fine-grained Visual Recognition with Large Language Models. In *ICLR*.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt Distribution Learning. In *CVPR*.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Mao, C.; Teotia, R.; Sundar, A.; Menon, S.; Yang, J.; Wang, X.; and Vondrick, C. 2023. Doubly right object recognition: A why prompt for visual rationales. In *CVPR*.

Menon, S.; and Vondrick, C. 2023. Visual Classification via Description from Large Language Models. In *ICLR*.

Murphy, G. 2004. *The big book of concepts*. MIT press.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *ICVGIP*.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *ICML*.

Saha, O.; Horn, G. V.; and Maji, S. 2024. Improved Zero-Shot Classification by Adapting VLMs with Text Descriptions. *arXiv preprint arXiv:2401.02460*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Shi, H.; Hayat, M.; Wu, Y.; and Cai, J. 2022. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *CVPR*.

Shtedritski, A.; Rupprecht, C.; and Vedaldi, A. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *ICCV*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tian, X.; Zou, S.; Yang, Z.; and Zhang, J. 2024. ArGue: Attribute-Guided Prompt Tuning for Vision-Language Models. In *CVPR*.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. In *NeurIPS*.

Wang, W.; Han, C.; Zhou, T.; and Liu, D. 2023. Visual Recognition with Deep Nearest Centroids. In *ICLR*.

Wang, W.; Sun, Q.; Zhang, F.; Tang, Y.; Liu, J.; and Wang, X. 2024. Diffusion Feedback Helps CLIP See Better. *arXiv preprint arXiv:2407.20171*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*.

Xie, G.-S.; Liu, J.; Xiong, H.; and Shao, L. 2021a. Scale-aware graph neural network for few-shot semantic segmentation. In *CVPR*.

Xie, G.-S.; Xiong, H.; Liu, J.; Yao, Y.; and Shao, L. 2021b. Few-shot semantic segmentation with cyclic memory network. In *ICCV*.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2020. Attribute prototype network for zero-shot learning. In *NeurIPS*.

Yan, A.; Wang, Y.; Zhong, Y.; Dong, C.; He, Z.; Lu, Y.; Wang, W. Y.; Shang, J.; and McAuley, J. 2023. Learning concise and descriptive attributes for visual recognition. In *ICCV*.

Yang, L.; Wang, Y.; Li, X.; Wang, X.; and Yang, J. 2024. Fine-grained visual prompting. *NeurIPS*.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*.

Yao, H.; Zhang, R.; and Xu, C. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*.

Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.-S.; and Sun, M. 2021. CPT: Colorful Prompt Tuning for Pre-trained Vision-Language Models. *arXiv preprint arXiv:2109.11797*.

Yu, F. X.; Cao, L.; Feris, R. S.; Smith, J. R.; and Chang, S.-F. 2013. Designing Category-Level Attributes for Discriminative Visual Recognition. In *CVPR*.

Zhang, E.; Chen, Y.; Miao, Q.; Tang, M.; Wang, J.; et al. 2024a. Optimization of Prompt Learning via Multi-Knowledge Representation for Vision-Language Models. *arXiv preprint arXiv:2404.10357*.

Zhang, J.; Wu, S.; Gao, L.; Shen, H.; and Song, J. 2023. Dept: Decoupled prompt tuning. *arXiv preprint arXiv:2309.07439*.

Zhang, J.; Wu, S.; Gao, L.; Shen, H. T.; and Song, J. 2024b. DePT: Decoupled Prompt Tuning. In *CVPR*.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *ECCV*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *CVPR*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 1–12.

Zhou, T.; and Wang, W. 2024. Prototype-based semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*

Zhou, T.; Xia, W.; Zhang, F.; Chang, B.; Wang, W.; Yuan, Y.; Konukoglu, E.; and Cremers, D. 2024. Image segmentation in foundation model era: A survey. *arXiv preprint arXiv:2408.12957*.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023a. Prompt-aligned Gradient for Prompt Tuning. In *ICCV*.

Zhu, B.; Niu, Y.; Lee, S.; Hur, M.; and Zhang, H. 2023b. Debiased Fine-Tuning for Vision-Language Models by Prompt Regularization. In *AAAI*.