# Learning Hierarchical Prompt with Structured Linguistic Knowledge for Vision-Language Models

**Yubin Wang**[1], **Xinyang Jiang**[2], **De Cheng**[3], **Dongsheng Li**[2], **Cairong Zhao**[1]*

[1]Tongji University
[2]Microsoft Research Asia
[3]Xidian University

## Abstract

Prompt learning has become a prevalent strategy for adapting vision-language foundation models to downstream tasks. As large language models (LLMs) have emerged, recent studies have explored the use of category-related descriptions as input to enhance prompt effectiveness. Nevertheless, conventional descriptions fall short of structured information that effectively represents the interconnections among entities or attributes linked to a particular category. To address this limitation and prioritize harnessing structured knowledge, this paper advocates for leveraging LLMs to build a graph for each description to model the entities and attributes describing the category, as well as their correlations. Preexisting prompt tuning methods exhibit inadequacies in managing this structured knowledge. Consequently, we propose a novel approach called Hierarchical Prompt Tuning (HPT), which enables simultaneous modeling of both structured and conventional linguistic knowledge. Specifically, we introduce a relationship-guided attention module to capture pair-wise associations among entities and attributes for low-level prompt learning. In addition, by incorporating high-level and global-level prompts modeling overall semantics, the proposed hierarchical structure forges cross-level interlinks and empowers the model to handle more complex and long-term relationships. Extensive experiments demonstrate that our HPT shows strong effectiveness and generalizes much better than existing SOTA methods. Our code is available at https://github.com/Vill-Lab/2024-AAAI-HPT.

## Introduction

Vision-Language foundation models (VLMs) (Radford et al. 2021; Jia et al. 2021), trained on large-scale datasets of image-text pairs, have made remarkable advancements in learning transferable representations. To effectively explore the potential of these powerful foundation models, prompt tuning methods (Zhou et al. 2022; Zhou et al. 2022; Khattak et al. 2023) aim to learn a set of continuous vectors known as prompt vectors and incorporate them in the input space, endowing the pre-trained network with a powerful representation capability. However, when confronted with ambiguous category names, models frequently struggle to make accurate judgments regarding the corresponding visual concepts, leading to underwhelming performance. Therefore, utilizing category names as text input without the assistance of linguistic knowledge seems to be a suboptimal choice. Recent methods (Zhang et al. 2023; Pratt, Liu, and Farhadi 2022; Menon and Vondrick 2022) have addressed this issue by using large language models (LLMs), such as GPT-3 (Brown et al. 2020). They take hand-written templates as input and generate human-like texts, containing rich linguistic knowledge that complements few-shot visual recognition.

In this paper, we propose a novel approach to complement natural linguistic descriptions with a structured representation of knowledge. We assert that this structured knowledge is essential for prompt tuning. Specifically, the descriptions of a category with unstructured knowledge consist of key entities and attributes that define the category.

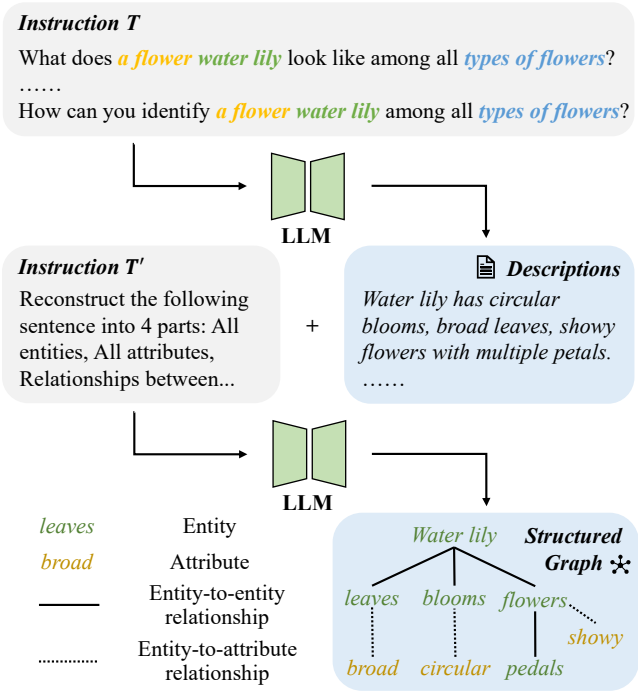



Figure 1: We input a few hand-written instructions into LLM to generate human-like category-related descriptions along with structured graphs based on each description.

*Corresponding Author (zhaocairong@tongji.edu.cn)

For example, the category 'water lily' is defined by entities like 'leaves', 'blooms', 'flowers', each linked to category-specific attributes. Following related works on knowledge graphs (Tay et al. 2017; Zhang et al. 2021), we represent these entities, attributes, and their correlations as a graph for semantic understanding. This graph-based representation offers a more organized way to present information, leading to improved data comprehension. It facilitates the discovery of implicit connections that may not be evident in original descriptions. In this work, we leverage existing large language models to obtain the structured information from vanilla descriptions, as shown in Figure 1. Given a specific category, we feed hand-crafted instructions into LLMs, intending to generate human-like descriptions, as well as structured relationships within each description, including entities, attributes, and relationships among them.

However, existing prompt tuning methods are inadequate to explicitly model the structured knowledge represented in a graph. To this end, we propose **H**ierarchical **P**rompt **T**uning (HPT) to incorporate both structured and conventional linguistic knowledge from LLMs for enhancing prompt effectiveness in a hierarchical manner. To model the complex structured information, HPT learns hierarchical prompts with different semantic levels. Specifically, HPT contains low-level prompts representing the entities and attributes, high-level prompts with category-related information derived from descriptions, and global-level prompts with category-agnostic knowledge shared across categories.

To capture the LLM-generated pair-wise correspondences among entities and attributes, we introduce a relationship-guided attention module, where learnable attention-based matrices are integrated into the text encoder. Furthermore, to handle more complex and long-term relationships not fully exploited by LLMs, cross-level self-attention is adopted to model relationships between prompts from different levels. It effectively overcomes the limitations caused by relying solely on the modeling of low-level tokens and allowing for a more comprehensive understanding of the category. Our prompts are trained under a dual-path asymmetric framework (Zhao et al. 2022), where the prompted image encoder and text encoder are learned separately by aligning the output with the frozen encoder from the other modality respectively. By replacing the vanilla-prompted text encoder, which learns only category-agnostic prompts, with a novel hierarchical prompted text encoder, text representations can be better aligned with corresponding visual concepts, leading to excellent recognition performance.

The contributions of our work are summarized as follows. 1) We raise the consideration that it is crucial to use structured knowledge from descriptions to assist learning prompts. Thus, we leverage large language models to generate category-related descriptions along with corresponding structured relationships; 2) We propose Hierarchical Prompt Tuning (HPT) for simultaneously modeling both structured and conventional linguistic knowledge. By incorporating both forms of knowledge, we can enhance prompt effectiveness with more category-related information; 3) Extensive experiments on three commonly used evaluation settings demonstrate remarkable improvements with our method.

# Related Work

## Large Language Models

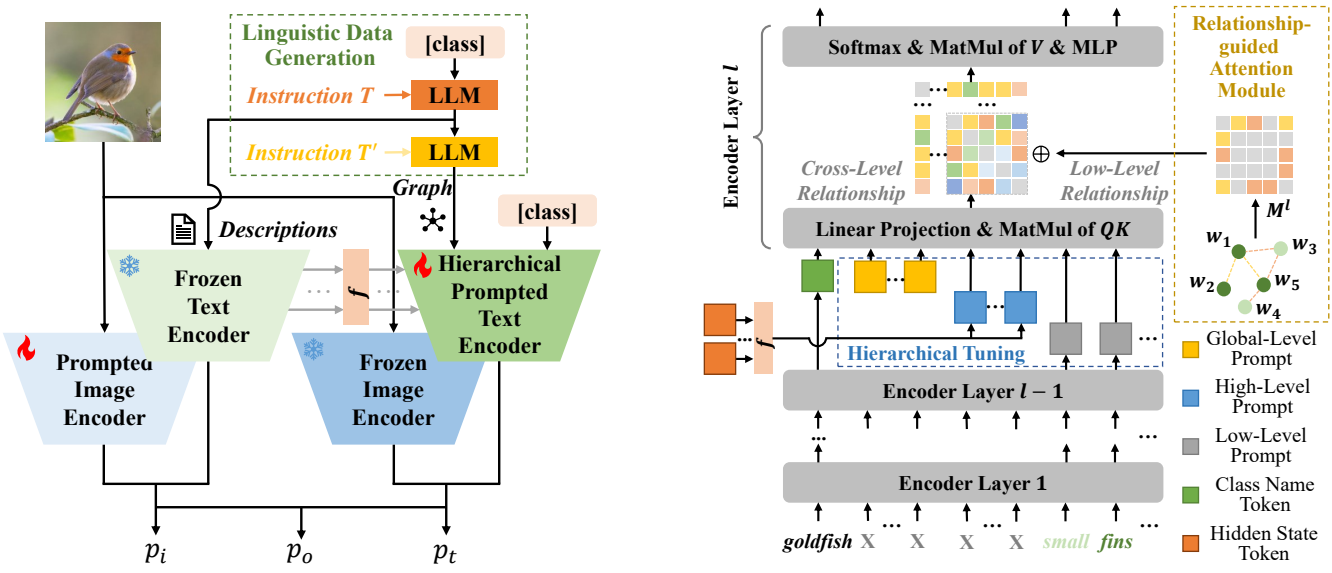
Large Language Models (LLMs), such as GPT-3 (Brown et al. 2020), OPT (Zhang et al. 2022), and PaLM (Chowdhery et al. 2022), are trained on extensive web-scale datasets. Recently, ChatGPT (OpenAI 2023) has gained widespread popularity due to its strong ability to generate text resembling human-like writing and discern intricate patterns across diverse domains. Taking advantage of the vast potential of LLMs, recent studies have demonstrated their effectiveness in addressing various vision-language tasks (Chen et al. 2022; Alayrac et al. 2022; Cheng et al. 2023; Yang et al. 2022a). Additionally, other studies investigate prompting vision-language models (Zhang et al. 2023; Li et al. 2022; Wang et al. 2022) with LLMs for image classification, continuous learning, image caption generation, and action understanding. In this study, we aim to leverage the capabilities of LLMs in the field of the image classification task. When prompted with the target category, LLMs are able to generate related descriptions as well as corresponding structured relationships.

## Visual-Language Models

Large visual-language models (VLMs) have been instrumental in driving open vocabulary image classification, with CLIP (Radford et al. 2021) being the pioneering work in this domain. Notable approaches include scaling up the models by using larger amounts of data, larger batch sizes, and bigger models, such as Align (Jia et al. 2021) and Basic (Pham et al. 2021), refining objective functions with models like SLIP (Mu et al. 2022), FILIP (Yao et al. 2021), and Lion (Chen et al. 2023), and incorporating supplementary information during training through models such as Florence (Yuan et al. 2021), UniCL (Yang et al. 2022b), K-LITE (Shen et al. 2022), and REACT (Liu et al. 2023a). Our study is motivated by the desire to enhance the capabilities of CLIP with improved multi-modal prompts.

## Prompt Learning for V-L Models

Prompt learning has its roots in natural language processing (NLP) and aims to enhance interaction with large language models (Liu et al. 2023b; Brown et al. 2020; Wei et al. 2022). Certain endeavors (Menon and Vondrick 2022; Pratt, Liu, and Farhadi 2022) propose leveraging pre-trained linguistic knowledge from LLMs to generate prompts, thereby enhancing V-L models without requiring additional training or labeling. To automate prompt engineering and explore optimal prompts, other studies (Rao et al. 2022; Zhou et al. 2022; Zhou et al. 2022; Lu et al. 2022) employ learnable text inputs and optimize them during training, known as prompt tuning. With the emergency of visual prompt tuning (VPT) (Jia et al. 2022), recent methods (Khattak et al. 2023; Zhao et al. 2022) take a multi-modal approach applying prompting on both modalities to improve alignment between vision and language representations. In contrast to prior studies, we generate diverse forms of linguistic knowledge and conduct hierarchical prompt tuning based on them to generate more robust representations.




(a) Overall pipeline for hierarchical prompt tuning

(b) Structure of hierarchical prompted text encoder

Figure 2: Our HPT applies a dual-path asymmetric network as the framework. Descriptions and relationship-guided graphs with class names are used as input for the frozen text encoder and the hierarchical prompted text encoder respectively. In the hierarchical prompted text encoder, we apply three types of prompts, low-level prompts, high-level prompts, and global-level prompts for hierarchical tuning, and design a relationship-guided attention module for better modeling structure knowledge.

# Methodology

## Overall Pipeline

In this subsection, we will present the overall pipeline of our proposed method, as shown in Figure 2(a). In the context of a specific category, we initially input it with a set of hand-crafted templates as instruction into LLMs to generate human-like descriptions. Moreover, we further feed generated descriptions with another instruction into LLMs, aiming to capture the well-organized structure within each description, encompassing entities, attributes, and their relationships. We will provide a more detailed exposition in Section **Linguistic Data Generation**.

Given generated data, we apply a dual-path asymmetric network (Zhao et al. 2022) for prompt tuning with visual-language models. This network experts in addressing overfitting issues associated with learned prompts, particularly in a few-shot learning scenario. To conduct prompt tuning for transformer-like encoders, learnable vectors are introduced at each transformer layer's input space as prompts. The framework incorporates a novel asymmetric contrastive loss, which trains the prompted image encoder and text encoder separately with the frozen encoder from the opposite modality as guidance. Specifically, representations of prompted and frozen encoders from different modalities are aligned in an asymmetric way, leading to generating two probabilities $p_i$ and $p_t$ from the two frozen-prompted pairs. They are then averaged to derive an overall prediction $p_o$.

Rather than making any modifications to visual prompts, we will mainly focus on prompt tuning for the text modality. In contrast to the prior dual-path asymmetric network, wherein two text encoders process identical text inputs, our approach adopts a distinct strategy that the frozen and prompted text encoders take entirely different inputs. In particular, unstructured descriptions are fed into the frozen encoder, while relationship-guided graphs along with the corresponding category name are fed into the novel hierarchical prompted encoder, which is specifically designed and fine-tuned for modeling structured information. In Section **Hierarchical Prompt Tuning**, we will dive into the core structure of this encoder for more details of tuning prompts from different semantic levels. To effectively capture the LLM-generated pair-wise correspondences among entities and attributes, the hierarchical prompted text encoder integrates a relationship-guided attention module, whose detailed implementation will be elaborated in Section **Relationship-guided Attention Module**.

## Linguistic Data Generation

To acquire linguistic knowledge, we use one of the most powerful LLMs, ChatGPT (OpenAI 2023), to generate descriptions with corresponding structured relationships. As shown in Figure 1, we adopt $N_h$ question templates as the language instruction $T$ for LLMs, e.g., "What does a [CLASS] look like among all a [TYPE]?" or "What are the distinct features of [CLASS] for recognition among all [TYPE]?", etc. [CLASS] denotes a specific category name with a modifier, like "a pet Abyssinian". [TYPE] indicates the type of objects related to the dataset, like "types of pets" for OxfordPets (Parkhi et al. 2012). We denote the generated descriptions from $T$ as $D = \{d_i\}_{i=1}^{N_h}$, formulated as

$$D = \text{LLM}(T). \tag{1}$$

For descriptions in $D$, we design an extra instruction $T^{'}$ to leverage LLMs for producing structured knowledge, including entities, attributes, and relationships among them. We denote the structured knowledge generated from $D$ as $R$, formulated as

$$R = \text{LLM}([T^{'}, D]). \tag{2}$$

Here $R = \{r_i\}_{i=1}^{N_h}$, $r_i = \{E_i, A_i, R_{e2e,i}, R_{e2a,i}\}$, where $E_i$, $A_i$, $R_{e2e,i}$, $R_{e2a,i}$ represent the entity set, the attribute set, the set of entity-entity relationships, and the set of entity-attribute relationships based on description $d_i$.

Our method utilizes both descriptions $D$ and structured knowledge $R$ as the source of category-related textual information, leading to effective prompt tuning.

## Hierarchical Prompt Tuning

Given descriptions $D$ and structured knowledge $R$, we aspire to simultaneously model both structured and conventional linguistic knowledge. Therefore, we propose a novel approach called Hierarchical Prompt Tuning (HPT), which leverages both forms of knowledge for learning prompts in a hierarchical manner, as shown in Figure 2(b). HPT contains low-level prompts, high-level prompts, and global-level prompts, respectively denoted as $p_l$, $p_h$, $p_g$.

**Low-Level Prompt** To model pair-wise relationships within a description, we select essential words from this description as the input of the text encoder. Specifically, for entities in the entity set $E_i$ and attributes in the attribute set $A_i$, we simply concatenate them together as the low-level prompts $p_l^0$ for description $d_i$ and feed them into the first layer of the encoder. These prompts are seen as nodes in a relationship-guided graph, whose relationships are further processed by a novel relationship-guided attention module.

**High-Level Prompt** In order to capture more intricate associations between individual tokens and the complete description, we derive high-level prompts $p_h$ that encapsulate the overall semantics of the category based on a series of descriptions. In detail, we feed descriptions $D$ into the frozen text encoder. Instead of simply utilizing representations from the last layer, we extract the last tokens from each layer containing rich semantics and feed them into a learnable prompt generator $f$, formulated as

$$p_{h,i}^l = f\left(h_i^l\right), \tag{3}$$

where $h_i^l$ represents the last token of description $d_i$ at the $l$-th layer. These tokens are then concatenated together as the high-level prompts $p_h^l = [p_{h,1}^l; ...; p_{h,N_h}^l]$ of this category, which are further integrated into the corresponding layer of the hierarchical prompted encoder.

**Global-Level Prompt** To represent category-shared knowledge pertinent to the task, we employ the standard approach for tuning the global-level prompts $p_g$. Instead of leveraging any form of knowledge, we automatically learn $N_g$ category-agnostic continuous vectors shared across categories as contexts and concatenate them with other prompts for each layer.

**Hierarchical Tuning** Based on the above prompts, we conduct the proposed hierarchical prompt tuning on the hierarchical prompted text encoder, formulated as

$$\left[c^1, \_, \_, p_l^1\right] = L_1\left(\left[c, p_g^0, p_h^0, p_l^0\right]\right) \tag{4}$$

$$\left[c^i, \_, \_, p_l^i\right] = L_i\left(\left[c^{i-1}, p_g^{i-1}, p_h^{i-1}, p_l^{i-1}\right]\right), \quad i = 2, 3, ..., N \tag{5}$$

where $c$ represents the token of the class name. Via the projection head of the text encoder $\text{TextProj}$, the final text representation $z$ is acquired by projecting the text embeddings $x^N$ corresponding to the last token of the last transformer block $L_N$ to a common V-L latent embedding space,

$$z = \text{TextProj}\left(x^N\right). \tag{6}$$

## Relationship-guided Attention Module

We introduce a relationship-guided attention module to model structured knowledge $R$ to capture pair-wise correspondences among entities and attributes in a layer-wise manner. For the $l$-th layer of a transformer-like encoder, an attention-based matrix $M^l$ is constructed based on generated relationships from each description. Two types of scalar values $\lambda_{e2e}^l$ and $\lambda_{e2a}^l$ are learned to indicate the strength of the relationship of entity-entity pairs and entity-attribute pairs separately. We assign the value to the respective element in the matrix, formulated as

$$M_{i,j}^l = \begin{cases} \lambda_{e2e}^l & (w_i, w_j) \in R_{e2e} \\ \lambda_{e2a}^l & (w_i, w_j) \in R_{e2a} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $w_i$ indicates the entity or attribute associated with the $i$-th token in the sequence of low-level prompts.

Guided by structured knowledge, the learned attention-based matrices are integrated into layers of the text encoder. In practice, we compute the attention function on a set of queries simultaneously, packed together into a matrix $Q$. The keys and values are also packed together into matrices $K$ and $V$. For the $l$-th layer, with the attention-based matrix $M^l$, the output of self-attention is computed as

$$\text{Attention}^l(Q, K, V) = \text{softmax}\left(\frac{QK^\top + M^l}{\sqrt{d_k}}\right)V. \tag{8}$$

By explicitly adding $M^l$ into the calculation of self-attention, our model explicitly represents rich structured relationships within each description, thus enhancing crucial information associated with the category.

To deal with more intricate relationships, we include high-level and global-level prompts for the construction of long-term relationships. Unlike modeling correspondences with matrices, we automatically leverage the implicit associations through cross-level self-attention itself without any manual intervention. This design, as a hierarchical knowledge modeling approach, blends holistic semantics from multiple levels with structured relationships, thereby helping us discover complex associations that LLMs have failed to identify.

## Experimental Setup

To evaluate our method, we follow the experiment setup established in previous works such as CoOp (Zhou et al. 2022), CoCoOp (Zhou et al. 2022), and MaPLe (Khattak et al. 2023). We first describe evaluation protocols and datasets, followed by a discussion on implementation details.

### Evaluation Protocols

**Base-to-New Generalization** Aiming to evaluate the generalizability across various classes, this process involves dividing the dataset into base (seen) and new (unseen) classes and then training the model using a small number of samples from the base classes. Finally, we evaluate the model's performance on both base (few-shot performance) and new (zero-shot performance) classes. Additionally, we calculate the harmonic mean over the accuracy on both base and new classes to highlight the generalization trade-off.

**Cross-Dataset Evaluation** This evaluation approach aims to assess the zero-shot ability of the model on a cross-dataset setup. To validate the potential of our approach in cross-dataset transfer, we train our model on all ImageNet classes in a few-shot manner and evaluate it directly on ten other unseen datasets with unknown categories in a zero-shot regime.

**Domain Generalization** To evaluate the robustness of our method on out-of-distribution datasets, we consider ImageNet as the source domain and its other variants as the target domain. We finetune our model on ImageNet in a few-shot setting and evaluate it on four variants of ImageNet with identical classes or subsets while manifesting diverse domain shifts.

### Datasets

For base-to-new generalization and cross-dataset evaluation, we follow the prior work (Zhou et al. 2022) and evaluate the performance of our method on 11 image recognition datasets, which cover a wide range of recognition tasks. Specifically, the benchmark includes ImageNet (Deng et al. 2009) and Caltech101 (Fei-Fei, Fergus, and Perona 2004) for classification on generic objects; OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Gool 2014) and FGVCAircraft (Maji et al. 2013) for fine-grained classification; SUN397 (Xiao et al. 2010) for scene recognition; UCF101 (Soomro, Zamir, and Shah 2012) for action recognition; DTD (Cimpoi et al. 2014) for texture classification; and finally EuroSAT (Helber et al. 2019) for satellite imagery recognition. For domain generalization, we utilize ImageNet as the source dataset and its four variants as target datasets 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).

### Implementation Details

We apply prompt tuning to the pre-trained CLIP (Radford et al. 2021) model, using ViT-B/16 as the visual backbone. We utilize SGD optimization with an initial learning rate of 0.0025 for base-to-new generalization and 0.001 for other tasks. Following the prior work (Zhao et al. 2022), the cross-entropy loss is adopted to equally minimize the discrepancy between the ground-truth label and the three aforementioned distributions $p_i$, $p_t$, $p_o$, while the overall distribution $p_o$ is used for inference. We randomly pick one description for each category to conduct relationship-guided attention learning during training for saving memory while leveraging all $N_h$ descriptions per category for inference.

For base-to-new generalization, the maximum epoch is set to 10, with a batch size of 8. The length of global-level prompts $N_g$ is set to 2, and the number of descriptions for each category $N_h$, which is also the length of high-level prompts is set to 5. In accordance with the prior work (Zhou

| Dataset | | CLIP | CoCoOp | MaPLe* | **HPT** | Δ |
|---|---|---|---|---|---|---|
| *Average* | B | 69.34 | 80.47 | 82.28 | **84.32** | +2.04 |
| | N | 74.22 | 71.69 | 75.14 | **76.86** | +1.72 |
| | H | 71.70 | 75.83 | 78.55 | **80.23** | +1.68 |
| ImageNet | B | 72.43 | 75.98 | 76.66 | **77.95** | +1.29 |
| | N | 68.14 | 70.43 | 70.54 | **70.74** | +0.20 |
| | H | 70.22 | 73.10 | 73.47 | **74.17** | +0.70 |
| Caltech101 | B | 96.84 | 97.96 | 97.74 | **98.37** | +0.41 |
| | N | 94.00 | 93.81 | 94.36 | **94.98** | +0.62 |
| | H | 95.40 | 95.84 | 96.02 | **96.65** | +0.63 |
| OxfordPets | B | 91.17 | 95.20 | 95.43 | **95.78** | +0.35 |
| | N | 97.26 | 97.69 | **97.76** | 97.65 | -0.11 |
| | H | 94.12 | 96.43 | 96.58 | **96.71** | +0.13 |
| StanfordCars | B | 63.37 | 70.49 | 72.94 | **76.95** | +4.01 |
| | N | 74.89 | 73.59 | 74.00 | **74.23** | +0.23 |
| | H | 68.65 | 72.01 | 73.47 | **75.57** | +2.10 |
| Flowers102 | B | 72.08 | 94.87 | 95.92 | **98.17** | +2.25 |
| | N | 77.80 | 71.75 | 72.46 | **78.37** | +0.57 |
| | H | 74.83 | 81.71 | 82.56 | **87.16** | +4.60 |
| Food101 | B | 90.10 | 90.70 | **90.71** | 90.46 | -0.25 |
| | N | 91.22 | 91.29 | **92.05** | 91.57 | -0.48 |
| | H | 90.66 | 90.99 | **91.38** | 91.01 | -0.37 |
| FGVCAircraft | B | 27.19 | 33.41 | 37.44 | **42.68** | +5.24 |
| | N | 36.29 | 23.71 | 35.61 | **38.13** | +1.84 |
| | H | 31.09 | 27.74 | 36.50 | **40.28** | +3.78 |
| SUN397 | B | 69.36 | 79.74 | 80.82 | **82.57** | +1.75 |
| | N | 75.35 | 76.86 | 78.70 | **79.26** | +0.56 |
| | H | 72.23 | 78.27 | 79.75 | **80.88** | +1.13 |
| DTD | B | 53.24 | 77.01 | 80.36 | **83.84** | +3.48 |
| | N | 59.90 | 56.00 | 59.18 | **63.33** | +3.43 |
| | H | 56.37 | 64.85 | 68.16 | **72.16** | +4.00 |
| EuroSAT | B | 56.48 | 87.49 | 94.07 | **94.24** | +0.17 |
| | N | 64.05 | 60.04 | 73.23 | **77.12** | +3.89 |
| | H | 60.03 | 71.21 | 82.35 | **84.82** | +2.48 |
| UCF101 | B | 70.53 | 82.33 | 83.00 | **86.52** | +3.52 |
| | N | 77.50 | 73.45 | 78.66 | **80.06** | +1.40 |
| | H | 73.85 | 77.64 | 80.77 | **83.16** | +2.39 |

* Previous SOTA method, the same for other generalization tasks.

Table 1: Comparison with existing methods on base-to-new generalization. B: Base Classes. N: New Classes. HM: Harmonic mean. Δ: absolute improvement of HPT over the previous best result. HPT demonstrates strong generalization performance on 11 image recognition datasets.

| | Source | Target | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ImNet | Caltech | Pets | Cars | Flowers | Food | Aircraft | SUN397 | DTD | EuroSAT | UCF | *Average* |
| CoOp | 71.51 | 93.70 | 89.14 | 64.51 | 68.71 | 85.30 | 18.47 | 64.15 | 41.92 | 46.39 | 66.55 | 63.88 |
| CoCoOp | 71.02 | **94.43** | 90.14 | 65.32 | 71.88 | 86.06 | 22.94 | 67.36 | 45.73 | 45.37 | 68.21 | 65.74 |
| MaPLe | 70.72 | 93.53 | 90.49 | 65.57 | 72.23 | 86.20 | 24.74 | 67.01 | 46.49 | **48.06** | 68.69 | 66.30 |
| **HPT** | **71.72** | 94.20 | **92.63** | **66.33** | **74.84** | **86.21** | **25.68** | **68.75** | **50.87** | 47.36 | **70.50** | **67.74** |

Table 2: Comparison with existing methods on cross-dataset evaluation. HPT achieves competitive performance providing the highest average accuracy, indicating superior generalization abilities on other datasets.

| | Source | Target | | | | |
|---|---|---|---|---|---|---|
| | ImageNet | ImageNetV2 | ImageNet-S | ImageNet-A | ImageNet-R | *Average* |
| CLIP | 66.73 | 60.83 | 46.15 | 47.77 | 73.96 | 57.17 |
| CoOp | 71.51 | 64.20 | 47.99 | 49.71 | 75.21 | 59.28 |
| CoCoOp | 71.02 | 64.07 | 48.75 | 50.63 | 76.18 | 59.90 |
| MaPLe | 70.72 | 64.07 | 49.15 | **50.90** | 76.98 | 60.26 |
| **HPT** | **71.72** | **65.25** | **49.36** | 50.85 | **77.38** | **60.71** |

Table 3: Comparison with existing methods on domain generalization. Overall, HPT shows consistent improvements on target variant datasets while achieving the highest accuracy on ImageNet.

et al. 2022), we select 16 shots for training and the entire test set for evaluation. For domain generalization and cross-dataset evaluation, the maximum epoch is set to 3, with a batch size of 8, where we use the same hyperparameters for each dataset instead of a separate search.

## Experiments

We evaluate our approach in three generalization settings, i.e. base-to-new generalization, cross-dataset evaluation, and domain generalization. We compare its performance with zero-shot CLIP (Radford et al. 2021) and recent prompt learning works as strong baselines including CoOp (Zhou et al. 2022) and CoCoOp (Zhou et al. 2022), as well as the state-of-the-art method MaPLe (Khattak et al. 2023). In the case of CLIP, we use hand-crafted prompts specifically designed for each dataset. We further conduct several ablation experiments and sample analyses to better demonstrate the effectiveness of the proposed hierarchical prompt tuning.

### Base-to-New Generalization

Table 1 presents the performance of HPT in base-to-new generalization setting on 11 recognition datasets. Compared to the state-of-the-art prompt tuning method MaPLe, our approach achieves an enhancement of 1.72% in terms of average accuracy for new classes, while simultaneously maintaining high accuracy on seen classes, even surpassing MaPLe by 2.04%. When considering both base and new classes, HPT shows an absolute average gain of 1.68% on the harmonic mean over MaPLe, achieving a good trade-off between in-domain and out-of-domain data. The highest improvement of 4.64% over the previous SOTA in the harmonic mean is observed for Flowers102. With more available linguistic knowledge instead of only category names, our model trained by hierarchical prompt tuning shows a significant improvement.

### Cross-Dataset Evaluation

Table 2 shows the performance comparison between our HPT and existing methods on cross-dataset evaluation. On the ImageNet source dataset, HPT demonstrates comparable performance to competing approaches, yet it exhibits significantly superior generalization across 8 out of 10 datasets. Overall, HPT shows competitive performance leading to the highest average accuracy of 67.74% with a gain of 1.44% compared to the previous SOTA. Unlike other methods that simply transfer the learned prompt vectors to new tasks, we provide a rich set of category-related knowledge as well as a novel hierarchical learning strategy for modeling the knowledge, leading to superior cross-domain performance.

### Domain Generalization

We evaluate the direct transferability of our HPT trained on ImageNet to various out-of-domain datasets and observe that HPT consistently improves against all the existing approaches, as indicated in Table 3. Compared to MaPLe, HPT performs slightly worse on ImageNet-A but better on the other three. As variant datasets share identical categories or subsets of categories with ImageNet, related linguistic knowledge from the source domain can be easily transferred, thereby assisting in recognizing out-of-domain data.

### Ablation Experiments

**Influence of Different Prompts in HPT** We perform an ablation analysis on base-to-new generalization with various prompt combinations in HPT, as illustrated in Table 4. The baseline method trains simply with global-level prompts. Experimental results show that both low-level and high-level

| Global | High | Low | Base | New | HM |
|---|---|---|---|---|---|
| ✓ | | | 84.02 | 75.20 | 78.99 |
| ✓ | ✓ | | 84.23 | 75.53 | 79.33 |
| ✓ | | ✓ | 84.05 | 76.11 | 79.59 |
| ✓ | ✓ | ✓ | **84.32** | **76.86** | **80.23** |

Table 4: Ablation on different prompts in HPT.

prompts positively affect recognition performance. Among them, low-level prompts demonstrate a significant improvement in new classes, which shows the effectiveness of explicitly modeling structured relationships within descriptions thereby providing additional information linked to unfamiliar categories. High-level prompts also play an inseparable role in boosting performance by incorporating holistic semantics to handle more complex relationships. When all prompts are tuned with cross-level self-attention simultaneously, our model achieves optimal performance.

**Influence of Components in Relationship-guided Attention Module** As shown in Table 5, we perform an ablation study on combinations of components in the relationship-guided attention module, including entities and attributes, along with their relationships. Entities and attributes contribute essential insights extracted from descriptions indicating pertinent information. Consequently, they play an important role in aligning category-related text with corresponding visual concepts. Furthermore, by incorporating relationships that capture pair-wise correspondences among entities and attributes, we comprehensively model structured knowledge with vital information linked to the category, thereby leading to additional performance enhancements.

**Influence of the Number of Descriptions** We conduct experiments by varying the value $N_h$, the number of descriptions for each category. In Figure 3, as $N_h$ increases, the knowledge related to a category becomes richer, thus leading to consistent improvement in recognition accuracy. Notably, the impact on accuracy is considerably more pronounced for new classes compared to base classes. This is because, in the case of unseen classes where training images are unavailable, performance mainly relies on the diversity of linguistic knowledge. We set $N_h = 5$ for implementation as the accuracy barely changes when more information is provided.

| Ent. | Attr. | Rel. | Base | New | HM |
|---|---|---|---|---|---|
| | | | 84.23 | 75.53 | 79.33 |
| ✓ | | | 84.21 | 75.76 | 79.49 |
| | ✓ | | 84.25 | 75.86 | 79.56 |
| ✓ | | ✓ | **84.34** | 76.00 | 79.71 |
| ✓ | ✓ | | 84.11 | 76.43 | 79.85 |
| ✓ | ✓ | ✓ | 84.32 | **76.86** | **80.23** |

Table 5: Ablation on entities, attributes and their relationships in relationship-guided attention module.

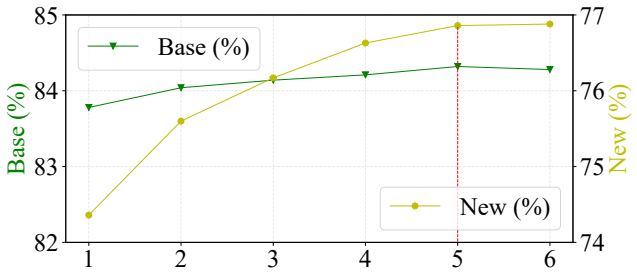



Figure 3: Performance of HPT using different values of $N_h$.

## Sample Analysis

In order to demonstrate the capability of HPT to capture category-related semantics, we provide sample analysis on three randomly selected categories from Caltech101. Figure 4 presents a comparison between our method and the baseline trained with the global-level prompts only. We observe the attention scores between tokens of entities and attributes from descriptions and the last token at the last layer of the prompted encoder. The top four features with the highest scores are displayed. It proves that HPT is capable of identifying discriminative visual concepts that significantly contribute to image recognition, leading to a substantial enhancement in the quality of text representations.

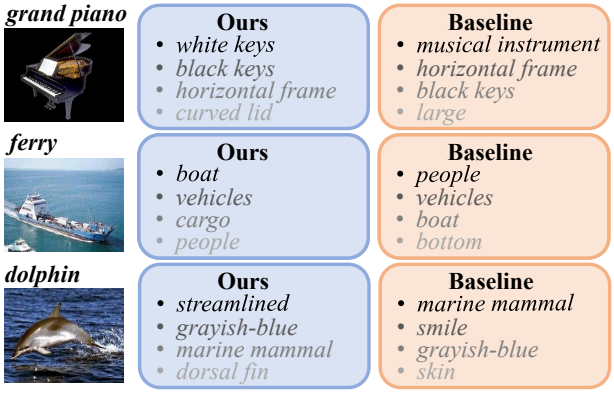



Figure 4: Visualization of the top features with the highest attention scores according to the selected categories.

## Conclusion

In this paper, we posit that utilizing structured relationships from descriptions to aid learning prompts is crucial. Consequently, we produce human-like descriptions accompanied by their corresponding structured relationships and present Hierarchical Prompt Tuning (HPT), a method that concurrently models both structured and conventional linguistic knowledge to strongly enhance prompt effectiveness. Our method demonstrates superior performance across three generalization tasks. We aspire that this work will garner increased attention toward the role of structured knowledge in natural language for prompt tuning, enabling its application to diverse tasks beyond classification.

## Ethical Statement

The integration of ChatGPT in studies carries ethical implications with broad social ramifications. It enables inclusive communication but raises concerns about misinformation and biases. Ethical considerations demand transparency, bias mitigation, and ongoing evaluation to harness its benefits responsibly.

## Acknowledgments

This work was supported by National Natural Science Fund of China (62076184, 61976158, 61976160, 62076182, 62276190, 62176198), in part by Fundamental Research Funds for the Central Universities and State Key Laboratory of Integrated Services Networks (Xidian University), in part by Shanghai Innovation Action Project of Science and Technology (20511100700) and Shanghai Natural Science Foundation (22ZR1466700), in part by Fundamental Research Funds for the Central Universities under grant NO. ZYTS23173.

## References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, 446–461. Springer.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, J.; Guo, H.; Yi, K.; Li, B.; and Elhoseiny, M. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18030–18040.

Chen, X.; Liang, C.; Huang, D.; Real, E.; Wang, K.; Liu, Y.; Pham, H.; Dong, X.; Luong, T.; Hsieh, C.-J.; et al. 2023. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*.

Cheng, D.; Wang, G.; Wang, B.; Zhang, Q.; Han, J.; and Zhang, D. 2023. Hybrid routing transformer for zero-shot learning. *Pattern Recognition*, 137: 109270.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

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 *2004 conference on computer vision and pattern recognition workshop*, 178–178. IEEE.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15262–15271.

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 *International conference on machine learning*, 4904–4916. PMLR.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Li, M.; Chen, L.; Duan, Y.; Hu, Z.; Feng, J.; Zhou, J.; and Lu, J. 2022. Bridge-prompt: Towards ordinal action understanding in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19880–19889.

Liu, H.; Son, K.; Yang, J.; Liu, C.; Gao, J.; Lee, Y. J.; and Li, C. 2023a. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15148–15158.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5206–5215.

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

Menon, S.; and Vondrick, C. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2022. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, 529–544. Springer.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729. IEEE.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Pham, H.; Dai, Z.; Ghiasi, G.; Kawaguchi, K.; Liu, H.; Yu, A. W.; Yu, J.; Chen, Y.-T.; Luong, M.-T.; Wu, Y.; et al. 2021. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*.

Pratt, S.; Liu, R.; and Farhadi, A. 2022. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*.

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 *International conference on machine learning*, 8748–8763. PMLR.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18082–18091.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, 5389–5400. PMLR.

Shen, S.; Li, C.; Hu, X.; Xie, Y.; Yang, J.; Zhang, P.; Gan, Z.; Wang, L.; Yuan, L.; Liu, C.; et al. 2022. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35: 15558–15573.

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*.

Tay, Y.; Tuan, L. A.; Phan, M. C.; and Hui, S. C. 2017. Multi-task neural network for non-discrete attribute prediction in knowledge graphs. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1029–1038.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022a. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35: 124–141.

Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; and Gao, J. 2022b. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19163–19173.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Zhang, R.; Hu, X.; Li, B.; Huang, S.; Deng, H.; Qiao, Y.; Gao, P.; and Li, H. 2023. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15211–15222.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, Y.; Wang, J.; Yu, L.-C.; and Zhang, X. 2021. MA-BERT: learning representation by incorporating multi-attribute knowledge in transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2338–2343.

Zhao, C.; Wang, Y.; Jiang, X.; Shen, Y.; Song, K.; Li, D.; and Miao, D. 2022. Learning domain invariant prompt for vision-language models. *arXiv preprint arXiv:2212.04196*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.