

LLMs are Expert in Classifying Images: Hierarchically Prompting Vision-Language Models with Structured Linguistic Knowledge

Yubin Wang¹, Xinyang Jiang², De Cheng³, Wenli Sun¹, Dongsheng Li²,
Cairong Zhao^{1*}

^{1*}Department of Computer Science and Technology, Tongji University, Shanghai, China.

²Microsoft Research Asia, Shanghai, China.

³School of Telecommunications Engineering, Xidian University, Xi'an, China.

*Corresponding author(s). E-mail(s): zhaocairong@tongji.edu.cn;

Contributing authors: wangyubin2018@tongji.edu.cn; xinyangjiang@microsoft.com;
dcheng@xidian.edu.cn; 2233055@tongji.edu.cn; dongsheng.li@microsoft.com;

Abstract

Prompt learning has become a prevalent strategy for adapting vision-language foundation models (VLMs) such as CLIP to downstream tasks. With the emergence of large language models (LLMs), recent studies have explored the potential of using category-related descriptions to enhance prompt effectiveness. However, conventional descriptions lack explicit structured information necessary to represent the interconnections among key elements like entities or attributes with relation to a particular category. Since existing prompt tuning methods give little consideration to managing structured knowledge, this paper advocates leveraging LLMs to construct a graph for each description to prioritize such structured knowledge. Consequently, we propose a novel approach called Hierarchical Prompt Tuning (HPT), enabling 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. Finally, by enhancing multi-granularity corpus generation, redesigning the relationship-driven attention re-weighting module, and incorporating consistent constraints on the hierarchical text encoder, we propose HPT++, which further improves the performance of HPT. Our experiments are conducted across a wide range of evaluation settings, including base-to-novel generalization, cross-dataset evaluation, and domain generalization. Extensive results and ablation studies demonstrate the effectiveness of our methods, which consistently outperform existing SOTA methods.

Keywords: prompt learning, vision-language models, few-shot learning, domain generalization

1 Introduction

Vision-language foundation models (VLMs) (Radford et al., 2021; Jia et al., 2021) have significantly advanced in learning transferable representations. To effectively explore the potential of

1

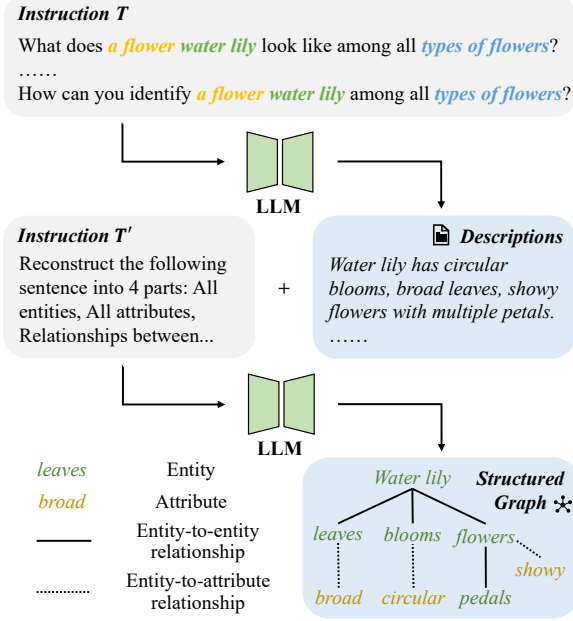


Fig. 1 We input a few hand-written instructions into LLMs to generate human-like category-related descriptions along with structured graphs based on each description.

them, prompt tuning methods (Zhou et al., 2022; Zhou et al., 2022; Khattak et al., 2023) learn continuous vectors, known as prompt vectors, and incorporate them into the input space, thereby enhancing the representation capability of the pre-trained network. However, when faced with ambiguous category names, models often struggle to accurately interpret the corresponding visual concepts, resulting in sub-optimal performance. Thus, using category names as text input without the assistance of more linguistic knowledge is an unsatisfactory choice. Recent methods (Zhang et al., 2023; Pratt et al., 2022; Menon and Vondrick, 2022) address this issue by employing large language models (LLMs) (Brown et al., 2020; AI@Meta, 2024). These methods use hand-written templates to generate human-like texts enriched with linguistic knowledge, thereby facilitating few-shot visual recognition.

In this paper, we propose a novel approach that enhances natural linguistic descriptions with structured knowledge representations. We assert that structured knowledge is essential for prompt tuning. Specifically, the descriptions of a category with unstructured knowledge consist of some key elements, such as entities and attributes, which

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

However, existing prompt tuning methods are inadequate for explicitly modeling such structured knowledge represented as a graph. To address this issue, we propose **Hierarchical Prompt Tuning (HPT)** to incorporate both structured and conventional linguistic knowledge from LLMs, enhancing prompt effectiveness hierarchically. To model complex structured information, HPT learns hierarchical prompts at different semantic levels. Specifically, HPT contains low-level prompts representing relationships among key elements of the category, high-level prompts with implicit category-related semantics derived from descriptions, and global-level prompts with task- or domain-specific knowledge shared across categories.

We introduce a relationship-guided attention module to leverage and model the LLM-generated pair-wise correspondences among entities and attributes, where learnable attention-based matrices are integrated into the text encoder. Furthermore, cross-level self-attention is adopted to model relationships between prompts from different levels to handle more complex and long-term relationships not fully exploited by LLMs. It effectively overcomes the limitations caused by relying solely on modeling 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., 2024), where the prompted image encoder and text encoder

are learned separately by aligning their output with the frozen encoder from the other modality. By replacing the vanilla-prompted text encoder, which learns only category-agnostic prompts, with a novel hierarchical prompted text encoder, text representations align better with corresponding visual concepts, resulting in superior recognition performance.

The contributions of our work are summarized as follows. 1) We raise the consideration that it is crucial to extract and leverage structured knowledge from descriptions to assist learning prompts. Thus, we firstly leverage large language models to generate category-related descriptions along with corresponding structured relationships. 2) We propose Hierarchical Prompt Tuning 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, including base-to-new generalization, cross-dataset evaluation and domain generalization, demonstrate remarkable improvements with our method, better than existing state-of-the-art methods.

However, some issues with HPT should be addressed. First, HPT utilizes LLMs to generate category-related descriptions via handcrafted prompt templates. However, this method may be ineffective as it does not guarantee that descriptions will be sufficiently discriminative among categories. Additionally, HPT models relationships as matrices and integrates them additively into attention computation, which is suboptimal since it treats all relationships of the same type equally. Furthermore, despite its effectiveness in modeling linguistic knowledge, hierarchical prompt learning is prone to overfitting and could perform poorly when conducting generalization.

To further enhance our model’s performance beyond that presented in our conference paper version (Wang et al., 2024), we propose HPT++ in Section 4. Specifically, we refine the corpus generation process, producing and merging coarse-grained and fine-grained descriptions into multi-granularity descriptions for generating structured graphs with more semantics. Additionally, we experiment with various methods to model structured information and re-design the relationship-driven attention re-weighting module, enabling

re-weighting of attention maps according to relationships between key elements with a predefined ratio. Finally, to avoid over-fitting in downstream generalization tasks, we incorporate a consistency constraint between prompt-tuned and pre-trained text encoders, where an adaptor is appended to the hierarchical text encoder to learn robust representations. These improvements and a comparison to HPT are validated in Section 5. The code are both available at <https://github.com/Vill-Lab/2024-AAAI-HPT>.

2 Related Work

2.1 Large Language Models

Large Language Models (LLMs) (Brown et al., 2020; Zhang et al., 2022; Chowdhery et al., 2022; AI@Meta, 2024; OpenAI, 2023), trained on extensive web-scale datasets, has gained widespread popularity due to its ability to generate text resembling human writing and to discern intricate patterns across diverse domains. Leveraging 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., 2022). Additionally, other studies investigate prompting vision-language models with LLMs for image classification, continuous learning, image caption generation, and action understanding (Zhang et al., 2023; Li et al., 2022; Wang et al., 2022). In this study, we aim to leverage the capabilities of LLMs in the field of image classification. When prompted with a target category, LLMs can generate related descriptions and corresponding structured relationships.

2.2 Visual-Language Models

Visual-language models (VLMs) have played a crucial role in advancing open vocabulary image classification, with CLIP (Radford et al., 2021) pioneering this domain. Notable approaches include scaling up models with larger amounts of data, batch sizes, and model sizes, 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

using models such as Florence (Yuan et al., 2021), UniCL (Yang et al., 2022), K-LITE (Shen et al., 2022), and REACT (Liu et al., 2023). Our study is motivated by the goal of enhancing CLIP’s capabilities through improved multi-modal prompts.

2.3 Prompt Learning for V-L Models

Prompt learning originates in natural language processing (NLP) and aims to enhance interaction with large language models (Liu et al., 2023; Brown et al., 2020; Wei et al., 2022). Some efforts (Menon and Vondrick, 2022; Pratt et al., 2022) propose leveraging pre-trained linguistic knowledge from LLMs to generate prompts, thereby enhancing vision-language 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, optimizing them during training, a process known as prompt tuning. With the emergence of visual prompt tuning (VPT) (Jia et al., 2022), recent methods (Khattak et al., 2023; Zhao et al., 2024) take a multi-modal approach, applying prompting to both modalities to improve alignment between vision and language representations. In contrast to prior studies, we generate diverse forms of linguistic knowledge and perform hierarchical prompt tuning based on this knowledge to produce more robust representations.

3 HPT

3.1 Overall Pipeline

We present the overall pipeline of our framework. As a baseline network, we apply a dual-path asymmetric network (Zhao et al., 2024) for prompt tuning with visual-language models. This network experts in addressing over-fitting issues of the learned prompts, particularly in few-shot learning scenarios. To perform 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, training the

prompted image encoder and text encoder separately, using the frozen encoder from the counterpart modality as guidance. Specifically, representations of prompted and frozen encoders from different modalities are aligned asymmetrically, generating two probabilities from the two frozen-prompted pairs, which are averaged to derive an overall probability. All three probabilities are used to calculate the asymmetric loss \mathcal{L}_{asy} for training, whereas only the overall probability is utilized during inference, following the previous work (Zhao et al., 2024).

For a specific category, we initially input a set of handcrafted templates filled with the category name as instruction into LLMs to generate human-like descriptions. Additionally, we input the generated descriptions, along with another instruction, into LLMs to capture the well-organized structure of each description, which includes category-related elements such as entities, attributes, and their relationships. Detailed exposition is provided in Section **Linguistic Data Generation**. Instead of modifying visual prompts, we focus primarily on prompt tuning for the text modality. Unlike the vanilla-prompted text encoder in the previous dual-path asymmetric network, Section **Hierarchical Prompt Tuning** offers a novel and detailed exploration of the core structure of this encoder for tuning prompts across different semantic levels. 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 finetuned for modeling structured information. To effectively capture the LLM-generated element-wise correspondences, the hierarchical prompted text encoder integrates a relationship-guided attention module, whose detailed implementation will be elaborated in Section **Relationship-guided Attention Module**.

3.2 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

Table 1 [CLASS] token and [TYPE] token for 11 image classification datasets. [X] denotes the category name.

Dataset	[CLASS]	[TYPE]
ImageNet	[X]	objects
OxfordPets	a pet [X]	types of pets
Caltech101	[X]	objects
DescribableTextures	a [X] texture	types of texture
EuroSAT	[X]	types of land in a centered satellite photo
FGVCAircraft	a [X] aircraft	types of aircraft
Food101	[X]	types of food
OxfordFlowers	a flower [X]	types of flowers
StanfordCars	a [X] car	types of car
SUN397	a [X] scene	types of scenes
UCF101	a person doing [X]	types of action

[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). A full list of [CLASS] token and [TYPE] token for all datasets is illustrated in Table 1. We denote the generated descriptions from T as $D = \{d_i\}_{i=1}^{N_h}$, formulated as:

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

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 , expressed as:

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

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.

3.3 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

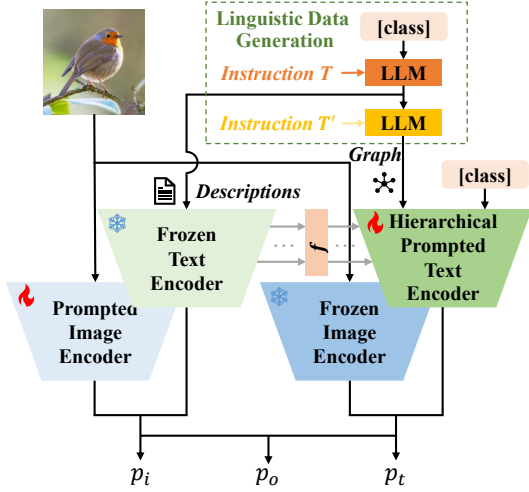
To model pair-wise relationships within a description, we select essential words from this description as the input of the text encoder. Specifically, for entities in the entity set E_i and attributes in the attribute set A_i , we simply concatenate them together as the low-level prompts p_l^0 for description 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

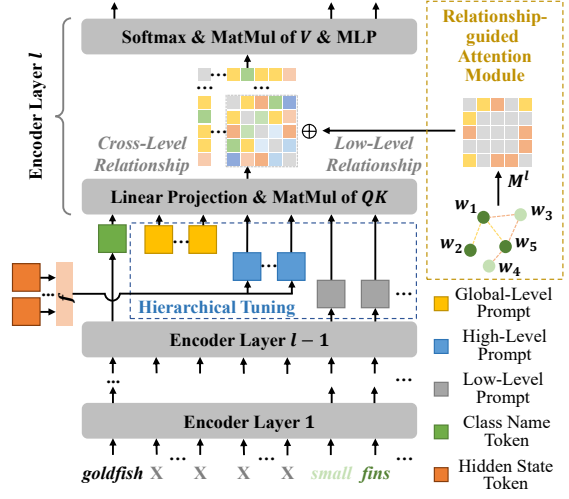
In order to capture more intricate associations between individual tokens and the complete description, we derive high-level prompts p_h that encapsulate the overall semantics of the category based on a series of descriptions. In detail, we feed descriptions D into the frozen text encoder. Instead of 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 , represented as:

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

where h_i^l represents the last token of description d_i at the l -th layer. These tokens are then



(a) Overall pipeline for hierarchical prompt tuning



(b) Structure of hierarchical prompted text encoder

Fig. 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 modeling structured knowledge.

concatenated together as the high-level prompts $p_h^l = [p_{h,1}^l; \dots; p_{h,N_h}^l]$ 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

$$\begin{aligned} [c^1, -, -, p_l^1] &= L_1([c, p_g^0, p_h^0, p_l^0]) \\ [c^i, -, -, p_l^i] &= L_i([c^{i-1}, p_g^{i-1}, p_h^{i-1}, p_l^{i-1}]), \quad (4) \\ i &= 2, 3, \dots, N \end{aligned}$$

where c represents the token of the class name. Via the projection head of the text encoder 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:

$$z = \text{TextProj}(x^N). \quad (5)$$

3.4 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 λ_{e2e}^l and λ_{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:

$$M_{i,j}^l = \begin{cases} \lambda_{e2e}^l & (w_i, w_j) \in R_{e2e} \\ \lambda_{e2a}^l & (w_i, w_j) \in R_{e2a} \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

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:

$$\text{Attention}^l(Q, K, V) = \text{softmax}\left(\frac{QK^\top + M^l}{\sqrt{d_k}}\right)V. \quad (7)$$

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.

4 HPT++

4.1 Overall Improvements

HPT, as introduced, can simultaneously model both structured and conventional linguistic knowledge. This capability makes it effective for handling complex and long-term relationships. We next explore several improvements to the original framework while maintaining the hierarchical structure. Specifically, in Section **Multi-Granularity Corpus Generation**, we refine the corpus generation process by merging coarse-grained and fine-grained descriptions into multi-granularity descriptions to create structured graphs with richer semantics. Additionally, we experiment with various strategies to model structured information and redesign the relationship-driven attention re-weighting module, allowing

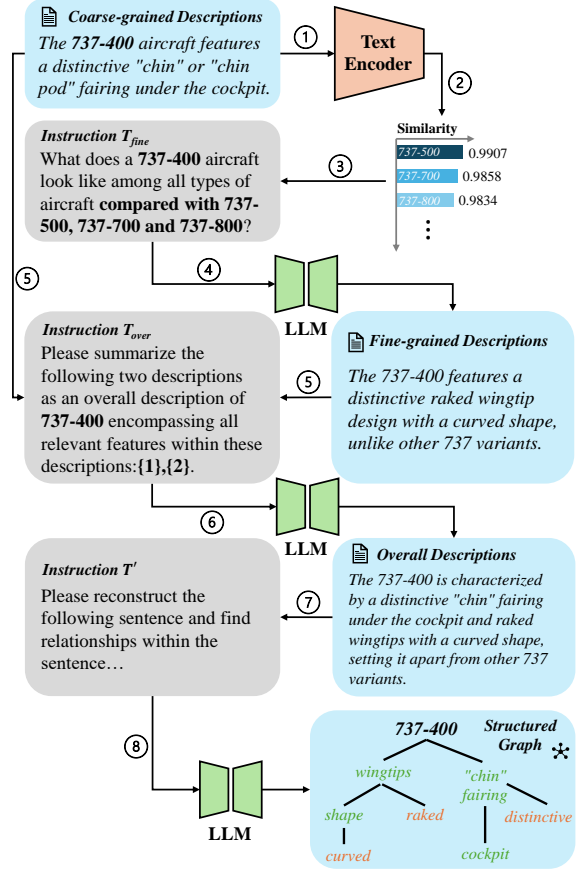


Fig. 3 Illustration of multi-granularity corpus generation. We firstly compute the similarity between descriptions of different categories, and then generate fine-grained descriptions for each category based on its closest categories. We integrate descriptions of both granularities to produce an overall description with multi-granularity semantics, which is subsequently used for generating structured graphs.

attention intensity to be scaled based on the generated relationships at a preset ratio. Detailed exposition is provided in Section **Relationship-Driven Attention Re-Weighting Module**. Finally, to avoid overfitting in downstream generalization tasks, we incorporate a consistency constraint between hierarchical prompted and pre-trained text encoders. An adapter is appended to the top of the hierarchical prompted text encoder to learn robust and adaptive representations, detailed in Section **Consistency Constraint on Hierarchical Prompted Text Encoder**.

4.2 Multi-Granularity Corpus Generation

HPT leverages the LLM to generate category-related descriptions based on handcraft prompt templates. This approach seems suboptimal since it does not ensure that the generated descriptions will be sufficiently discriminative to distinguish between different categories. For instance, in FGVC Aircraft, the descriptions of categories such as “737-400”, “737-500”, “737-700”, and “737-800” all share the characteristic of “a distinctive ‘chin’ fairing under the cockpit,” as these categories represent different variants of Boeing 737 series. This similarity makes it challenging for the VLM to correctly identify the category of an image. To prevent highly similar descriptions among certain categories, we propose multi-granularity corpus generation, as shown in Figure 3.

We firstly compute the similarity between descriptions of different categories using the pre-generated descriptions introduced in HPT (referred to as coarse-grained descriptions) and then generate fine-grained descriptions for each category based on its closest categories. Given $D_{coar}^c = \{d_{coar,i}^c\}_{i=1}^{N_h}$ as the coarse-grained descriptions of the c -th category, we input $d_{coar,i}^c$ into the frozen text encoder to obtain the corresponding text representation h_i^c , which is then normalized to \tilde{h}_i^c . We average the normalized representations for the c -th category as $\bar{h}^c = \frac{1}{N_h} \sum_{i=1}^{N_h} \tilde{h}_i^c$. We use cosine distance to find the C most relevant classes for each category, denoted as $[\text{CLASS}]_0, \dots, [\text{CLASS}]_{C-1}$. We reuse N_h question templates as T_{fine} , and for each handcrafted question template, we append “compared with $[\text{CLASS}]_0, \dots, [\text{CLASS}]_{C-1}$ ” as $T_{fine}(c)$ to prompt the LLM to generate fine-grained descriptions $D_{fine}^c = \{d_{fine,i}^c\}_{i=1}^{N_h}$ for the c -th category distinctive to other similar categories, expressed as:

$$D_{fine}^c = \text{LLM}(T_{fine}(c)). \quad (8)$$

Furthermore, we integrate descriptions of both granularities to produce a comprehensive description with multi-granularity semantics, which is subsequently used for generating structured knowledge. Specifically, we design an instruction like “Please summarize the following two descriptions as an overall description of $[\text{CLASS}]$

encompassing all relevant features within these descriptions: $\{d_1\}, \{d_2\}$.” as $T_{over}(d_1, d_2)$ for the LLM to output a summarized description. The process of obtaining the overall description corpus for the c -th category is represented as:

$$D_{over}^c = \text{LLM}(T_{over}(D_{coar}^c, D_{fine}^c)). \quad (9)$$

For descriptions in D_{over}^c , we revise the instruction T' in HPT to leverage the LLM for producing structured knowledge in a simpler form, only including relationships. We denote the relationships generated from D_{over}^c as R_{over}^c , formulated as:

$$R_{over}^c = \text{LLM}([T', D_{over}^c]). \quad (10)$$

Here $R_{over}^c = \{r_i^c\}_{i=1}^{N_h}$, where r_i^c represent the set of relationships based on the description $d_{over,i}^c$, including the subject, the verb and the direct object (or attribute) for each relationship. HPT++ utilizes both descriptions D_{over} and structured knowledge R_{over} as the final corpus to provide multi-granularity textual information for each category, leading to effective prompt tuning. It should be noted that unlike HPT, HPT++ does not generate entities and attributes for each description. Instead, it directly takes the description itself as the input to the hierarchical prompted text encoder, which is identical to the frozen text encoder. The tokens in the description serve as the low-level prompt, as they can be further processed by our proposed relationship-driven attention re-weighting module using the generated structured graph as connections.

4.3 Relationship-Driven Attention Re-Weighting Module

In HPT, relationships are modeled as matrices and applied additively to attention computation. We believe this approach is suboptimal, as the intensity of relationships within the same type is added equally (i.e., by adding the same scalar to the attention map). To address this issue, we investigate different approaches to modeling the relationship and propose a relationship-driven attention re-weighting module. Our findings indicate that re-weighting with element-wise multiplication enhances recognition performance more

effectively than simple addition, thereby better integrating structured knowledge into the model.

During the construction of the re-weighting matrix M^l for the l -th layer of the encoder, HPT++ uses a preset hyperparameter β to indicate the re-weighting intensity, instead of learning two scalar values λ_{e2e}^l and λ_{e2a}^l in HPT. When $\beta = 0$, no re-weighting operations are performed. As β increases, the attention intensity between tokens within a relationship is amplified, while that between tokens without a relationship diminishes. Based on the given relationship set R , we assign the value to the respective element in the matrix, expressed as:

$$M_{i,j}^l = \begin{cases} 1 + \beta & (w_i, w_j) \in R \\ \frac{1}{1+\beta} & \text{otherwise,} \end{cases} \quad (11)$$

where w_i indicates the i -th token in the sequence of the corresponding description. This re-weighting method proportionally scales the attention intensity between tokens, thereby preserving the relative intensity in the original attention map. We re-formulate Equation 7 as follows:

$$\text{Attention}^l(Q, K, V) = \text{softmax} \left(\frac{QK^\top \odot M^l}{\sqrt{d_k}} \right) V. \quad (12)$$

We also investigate and experiment with alternative re-weighting schemes for comparison, such as enhancing only the elements related to a relationship while keeping the other elements unchanged. Detailed studies can be found in the experimental section.

4.4 Consistent Constraint on Hierarchical Prompted Text Encoder

Although hierarchical prompt learning effectively models linguistic knowledge, it still faces potential overfitting issues and exhibits room for improvement in generalizing to new categories or domains. Inspired by PromptSRC (Khattak et al., 2023) and CoPrompt (Roy and Etemad, 2023), we impose a consistency constraint on the text branch, using cosine similarity between representations from the pre-trained and hierarchical prompted text encoders to regularize our hierarchical prompts. The asymmetric network mentioned earlier aligns the hierarchical text branch

with the pre-trained visual branch, while the consistency constraint aligns it with the pre-trained text branch. This dual alignment enhances the robustness and generalization of the learned representations by leveraging pre-trained knowledge from both modalities. Furthermore, to enhance learning capacity and improve adaptation, we introduce an adapter ϕ at the top of the hierarchical prompted text encoder. This adapter consists of trainable parameters designed to transform the embedding vector (Gao et al., 2024). The consistency loss is represented as:

$$\mathcal{L}_c = 1 - \frac{\phi(z) \cdot \Theta(t)}{\|\phi(z)\| \|\Theta(t)\|}. \quad (13)$$

Here Θ denotes the pre-trained text encoder and t stands for the input description. The final training loss \mathcal{L} of HPT++ is obtained by summing the asymmetric loss and the consistency loss with a balancing ratio λ , which is formulated as:

$$\mathcal{L} = \mathcal{L}_{asy} + \lambda \mathcal{L}_c. \quad (14)$$

5 Experimental Setup

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

5.1 Evaluation Protocols

5.1.1 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.

5.1.2 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.

5.1.3 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.

5.2 Datasets

For base-to-new generalization and cross-dataset evaluation, we 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 et al., 2004) for classification on generic objects; Oxford-Pets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback and Zisserman, 2008), Food101 (Bossard et al., 2014) and FGVC Aircraft (Maji et al., 2013) for fine-grained classification; SUN397 (Xiao et al., 2010) for scene recognition; UCF101 (Soomro et al., 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., 2021) and ImageNet-R (Hendrycks et al., 2021).

5.3 Implementation Details

We apply prompt tuning to the pre-trained CLIP model (Radford et al., 2021), using ViT-B/16 as the visual backbone. We employ SGD optimization with an initial learning rate of 0.0025 for base-to-novel generalization and 0.001 for other tasks, using a batch size of 8. The maximum number of epochs is set to 10 for base-to-novel generalization. For other tasks, we train our model for 3 epochs for HPT and 5 epochs for HPT++.

The length of global-level prompts N_g is set to 2, and the number of descriptions per category N_h , which also corresponds to the length of high-level prompts, is set to 5. We randomly select one description per category to conduct relationship modeling at each step during training to optimize memory usage, while leveraging all N_h descriptions per category for inference. We use GPT-3.5-turbo (OpenAI, 2023) and Llama-3-8B (AI@Meta, 2024) as LLMs in our study, both of which have comparable performance. Our research on performance related to different LLMs will be presented in the experimental section.

For HPT++, assuming N_c is the number of categories in a dataset, we determine the number of closest categories C for fine-grained description generation using the following function:

$$C = \lfloor \lg(N_c) \rfloor + 1. \quad (15)$$

The re-weighting intensity ratio β is set to 0.2, and the balancing ratio λ for the training loss is set to 1. Following prior works, we select 16 shots for training and use the entire test set for evaluation. For domain generalization and cross-dataset evaluation, we use the same hyperparameters across datasets, avoiding a separate search in CoPrompt (Roy and Etemad, 2023).

6 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 (Zhou et al., 2022; Zhou et al., 2022; Zhu et al., 2023; Yao et al., 2023; Khattak et al., 2023; Zhao et al., 2024; Roy and Etemad, 2023). In the case of CLIP, we use handcrafted 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.

6.1 Base-to-New Generalization

Table 2 presents the performance of HPT in base-to-new generalization setting on 11 recognition datasets. Compared to the state-of-the-art

Table 2 Comparison with existing methods on base-to-new generalization. B: Base Classes, N: New Classes, HM: Harmonic mean. HPT and HPT++ demonstrate strong generalization performance on 11 image recognition datasets.

Method		ImageNet	Caltech	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP (Radford et al., 2021)	B	72.43	96.84	91.17	63.37	72.08	90.10	27.19	69.36	53.24	56.48	70.53	69.34
	N	68.14	94.00	97.26	74.89	77.80	91.22	36.29	75.35	59.90	64.05	77.50	74.22
	H	70.22	95.40	94.12	68.65	74.83	90.66	31.09	72.23	56.37	60.03	73.85	71.70
CoOp (Zhou et al., 2022)	B	76.47	98.00	93.67	78.12	97.60	88.33	40.44	80.60	79.44	92.19	84.69	82.69
	N	67.88	89.81	95.29	60.40	59.67	82.26	22.30	65.89	41.18	54.74	56.05	63.22
	H	71.92	93.73	94.47	68.13	74.06	85.19	28.75	72.51	54.24	68.69	67.46	71.66
CoCoOp (Zhou et al., 2022)	B	75.98	97.96	95.20	70.49	94.87	90.70	33.41	79.74	77.01	87.49	82.33	80.47
	N	70.43	93.81	97.69	73.59	71.75	91.29	23.71	76.86	56.00	60.04	73.45	71.69
	H	73.10	95.84	96.43	72.01	81.71	90.99	27.74	78.27	64.85	71.21	77.64	75.83
ProGrad (Zhu et al., 2023)	B	77.02	98.02	95.07	77.68	95.54	90.37	40.54	81.26	77.35	90.11	84.33	82.48
	N	66.66	93.89	97.63	68.63	71.87	89.59	27.57	74.17	52.35	60.89	74.94	70.74
	H	71.46	95.91	96.33	72.88	82.03	89.98	32.82	77.55	62.45	72.67	79.35	76.16
KgCoOp (Yao et al., 2023)	B	75.83	97.72	94.65	71.76	95.00	90.50	36.21	80.29	77.55	85.64	82.89	80.73
	N	69.96	94.39	97.76	75.04	74.73	91.70	33.55	76.53	54.99	64.34	76.67	73.61
	H	72.78	96.03	96.18	73.36	83.65	91.09	34.83	78.36	64.35	73.48	79.65	77.00
MaPLe (Khattak et al., 2023)	B	76.66	97.74	95.43	72.94	95.92	90.71	37.44	80.82	80.36	94.07	83.00	82.28
	N	70.54	94.36	97.76	74.00	72.46	92.05	35.61	78.70	59.18	73.23	78.66	75.14
	H	73.47	96.02	96.58	73.47	82.56	91.38	36.50	79.75	68.16	82.35	80.77	78.55
PromptSRC (Khattak et al., 2023)	B	77.39	98.28	95.71	75.43	97.53	90.76	39.38	82.10	82.52	93.37	84.70	83.38
	N	71.06	94.58	96.98	74.43	74.54	91.77	37.59	79.01	60.10	78.34	78.56	76.09
	H	74.09	96.39	96.34	74.93	84.50	91.26	38.46	80.53	69.55	85.20	81.51	79.57
MetaPrompt (Zhao et al., 2024)	B	77.60	98.10	95.33	78.27	98.07	90.67	42.73	82.67	83.37	92.90	87.10	84.26
	N	70.73	94.03	97.30	74.97	76.50	91.53	37.87	78.47	62.97	73.90	78.80	76.10
	H	74.01	96.02	96.30	76.58	85.95	91.10	40.15	80.52	71.75	82.32	82.74	79.97
CoPrompt (Roy and Etemad, 2023)	B	77.67	98.27	95.67	76.97	97.27	90.73	40.20	82.63	83.13	94.60	86.90	84.00
	N	71.27	94.90	98.10	74.40	76.60	92.07	39.33	80.03	64.73	78.57	79.57	77.23
	H	74.33	96.55	96.87	75.66	85.71	91.40	39.76	81.31	72.79	85.84	83.07	80.48
HPT (Wang et al., 2024)	B	77.95	98.37	95.78	76.95	98.17	90.46	42.68	82.57	83.84	94.24	86.52	84.32
	N	70.74	94.98	97.65	74.23	78.37	91.57	38.13	79.26	63.33	77.12	80.06	76.86
	H	74.17	96.65	96.71	75.57	87.16	91.01	40.28	80.88	72.16	84.82	83.16	80.42
HPT++	B	77.66	98.17	95.94	76.99	97.50	90.56	40.50	82.40	84.18	95.31	86.26	84.13
	N	71.11	95.78	97.89	74.24	76.69	91.62	42.19	79.86	66.39	80.64	81.50	77.99
	H	74.24	96.96	96.91	75.59	85.85	91.09	41.33	81.11	74.23	87.36	83.81	80.95

Table 3 Comparison with existing methods on cross-dataset evaluation. The best results are highlighted in bold while the second best results are marked with an underline. HPT and HPT++ achieve competitive performance providing the highest average accuracy, indicating superior generalization abilities on other datasets.

	Source		Target									
	ImageNet	Caltech	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	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
PromptSRC	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
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
CoPrompt	70.80	94.50	90.73	<u>65.67</u>	72.30	86.43	24.00	67.57	47.07	51.90	69.73	67.00
HPT	<u>71.72</u>	94.20	92.63	66.33	74.84	86.21	<u>25.68</u>	<u>68.75</u>	50.87	47.36	70.50	<u>67.74</u>
HPT++	71.81	94.02	<u>92.16</u>	65.55	<u>72.43</u>	<u>86.34</u>	28.6	68.78	51.02	<u>50.76</u>	70.53	68.02

Table 4 Comparison with existing methods on domain generalization. The best results are highlighted in bold while the second best results are marked with an underline. Overall, HPT and HPT++ show consistent improvements on target variant datasets while achieving high accuracy on the source ImageNet dataset.

	Source	Target				
	ImNet	V2	S	A	R	Avg.
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
CoPrompt	70.80	64.25	<u>49.43</u>	50.50	77.51	60.42
PromptSRC	71.27	64.35	49.55	<u>50.90</u>	77.80	60.65
HPT	<u>71.72</u>	<u>65.25</u>	49.36	50.85	77.38	<u>60.71</u>
HPT++	71.81	65.31	49.28	51.18	<u>77.52</u>	60.82

prompt tuning method CoPrompt, HPT demonstrates comparable performance on base classes, while HPT++ achieves an improvement across all metrics on average. Specifically, HPT++ exhibits a 0.76% increase in average accuracy for new classes compared to CoPrompt, and a 1.13% increase over HPT, while maintaining competitive accuracy on seen classes. When considering both base and new classes, HPT++ exhibits an absolute average gain of approximately 0.5% in the harmonic mean over CoPrompt and HPT, achieving a favorable balance between in-domain and out-of-domain data. The most significant improvement over other baselines in the harmonic mean is observed for DTD and EuroSAT. When more linguistic knowledge beyond just category names is available, our methods demonstrate a significant improvement.

6.2 Cross-Dataset Evaluation

Table 3 shows the performance comparison between HPT, HPT++, and existing methods on cross-dataset evaluation. HPT and HPT++ demonstrate performance comparable to competing approaches on the ImageNet source dataset while exhibiting significantly superior generalization across most target datasets. Overall, HPT++ achieves the highest average accuracy of 68.02%, with an average gain of 1.02% over CoPrompt. Unlike other methods that merely transfer learned prompt vectors to new tasks, our approach provides a rich set of category-related knowledge, coupled with a novel hierarchical prompt learning strategy for modeling this knowledge. Compared

Table 5 Ablation study on different prompts in HPT.

Method	Global	High	Low	Base	New	HM
HPT	✓			84.02	75.20	79.37
	✓	✓		84.23	75.53	79.64
	✓		✓	84.05	76.11	79.88
	✓	✓	✓	84.32	76.86	80.42

to HPT, HPT++ applies a consistent constraint on the hierarchical text encoder, resulting in superior cross-domain performance. However, this module has adverse effects on some datasets, such as StanfordCars and Flowers102, where the hierarchical text encoder alone performs better than when complemented with pre-trained knowledge. This observation suggests that adaptive leveraging of pre-trained knowledge is crucial.

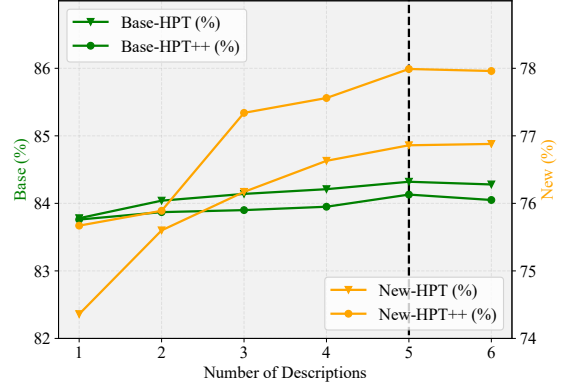
6.3 Domain Generalization

We evaluate the direct transferability of HPT and HPT++ trained on ImageNet to various out-of-domain datasets and observe consistent improvements over all existing approaches. As shown in Table 4, HPT and HPT++ outperform PromptSRC on the ImageNet source dataset as well as out-of-domain datasets in terms of average accuracy. Compared to HPT, HPT++ performs better on three OOD datasets, except for ImageNet-Sketch, where the lack of color information makes alignment with descriptions more difficult. Since variant datasets share identical or overlapping categories with ImageNet, relevant linguistic knowledge from the source domain can be easily transferred, aiding in the recognition of out-of-domain data.

6.4 Ablation Study

Prompts in Hierarchical Prompt Tuning

We conduct an ablation analysis on base-to-new generalization using various prompt combinations within the proposed hierarchical prompt tuning framework, as shown in Table 5. The baseline method is trained using only global-level prompts. Experimental results demonstrate that both low-level and high-level prompts positively impact recognition performance. Notably,

**Fig. 4** Performance of HPT and HPT++ using different numbers of descriptions.

low-level prompts significantly improve recognition of new classes, highlighting the effectiveness of explicitly modeling structured relationships within descriptions, thereby providing additional context for unfamiliar categories. High-level prompts also play a crucial role in enhancing performance by incorporating holistic semantics to handle more complex relationships. When all prompts are tuned simultaneously with cross-level self-attention, our model achieves optimal performance.

Number of Descriptions

We conduct experiments by varying the number of descriptions N_h for each category. As shown in Figure 4, increasing N_h enriches the knowledge related to a category, which consistently improves recognition accuracy. Notably, the impact on accuracy is significantly more pronounced for new classes compared to base classes. This is because, for unseen classes where training images are unavailable, performance relies primarily on the diversity of linguistic knowledge. We set $N_h = 5$ for implementation since the accuracy shows negligible improvement when more knowledge is provided.

HPT++ Improvements

Table 6 shows the contribution of each new component in our HPT++ model. Since our proposed multi-granularity corpus improves the quality of linguistic knowledge, enhancing the recognition of corresponding visual semantics, accuracy on both base and novel classes increases. However, a decrease of 0.15% in base accuracy is

Table 6 Ablation study on HPT++ components. We incrementally add each module to HPT to assess its contribution to the performance on base-to-new generalization.

Method	Base	New
HPT	84.32	76.86
+Multi-Granularity Corpus	84.36	77.23
+Attention Re-Weighting Module	84.21	77.51
+Consistent Constraint	84.13	77.99

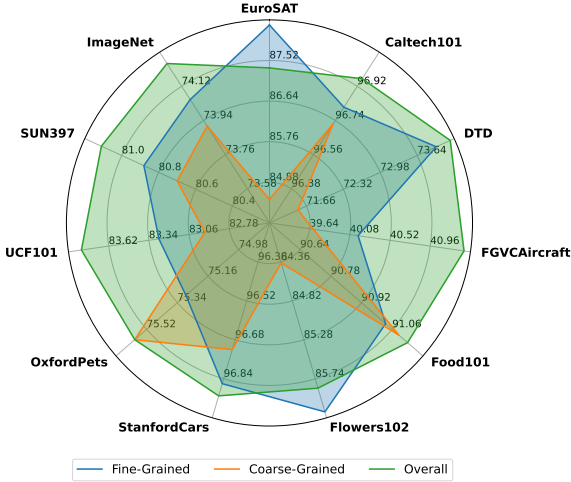


Fig. 5 Comprehensive comparison of the harmonic mean of HPT++ leveraging fine-grained corpus, coarse-grained corpus, and overall corpus with multiple granularities on 11 image recognition datasets for base-to-new generalization.

observed when the attention re-weighting module is applied. This improvement is attributed to replacing the learnable scalar in HPT with a preset hyperparameter that controls re-weighting intensity, helping to prevent overfitting on base classes and enhancing generalization to new classes. Additionally, by leveraging pre-trained knowledge, the consistency constraint further enhances generalization to new classes with an absolute gain of about 0.5%.

Corpus from Different Granularities

Given that different corpus granularities may capture distinct semantic aspects of a category, we examine the performance of leveraging various types of corpus as sources of linguistic knowledge for HPT++, as shown in Figure 5. The performance using the coarse-grained corpus is significantly worse than that of the fine-grained

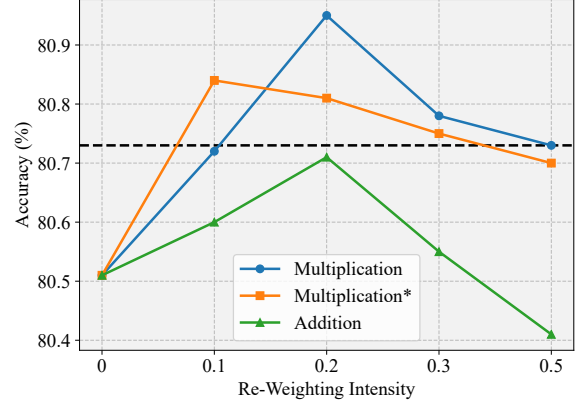


Fig. 6 Performance of different re-weighting strategies under various intensities in HPT++. Here “Multiplication*” indicates only conducting multiplication on interrelation elements in the attention map while keeping others unchanged. We also compare our method with the strategy employed in HPT, which uses learnable scalars for intensity indication. It is represented by a black dotted line.

one, whereas an overall corpus that combines multiple granularities achieves optimal performance on 9 out of 11 datasets, with the exceptions of EuroSAT and Flower102, where incorporating coarse-grained knowledge into the fine-grained corpus may lead to performance degradation. This finding also demonstrates that generating corpus with varying granularities enhances the quality of linguistic knowledge, thereby improving the ability to distinguish images across categories.

Choice of Attention Re-Weighting Strategy

We focus primarily on two types of attention-based re-weighting strategies for relationship modeling. One strategy adds the relationship-guided matrix, with identical values to indicate the relationship, to the attention map, while the other one uses the matrix as a weight for element-wise multiplication with the attention map in Transformer. Figure 6 presents the results under different re-weighting intensities, where we also compare our method with the strategy employed in HPT, which uses learnable scalars for intensity indication. The results indicate that the element-wise multiplication strategy for relationship modeling significantly outperforms the additive method, with optimal performance observed when the intensity is set to 0.2. Instead of merely enhancing interrelation elements while keeping others unchanged, we find that reducing the attention intensity between unrelated tokens in the

Table 7 Comparison with different LLMs on base-to-new generalization. Here “Avg.” refers to directly averaging the harmonic mean of all datasets, which differs from the approach used in base-to-novel generalization, where the average harmonic mean is computed using the average accuracy on base classes and new classes.

	LLM	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	Euro	UCF	ImNet	Avg.
HPT	Qwen2	96.63	96.65	75.34	86.73	91.00	39.86	80.47	71.85	83.75	83.23	73.87	79.94
HPT	GPT-3.5	96.65	96.71	75.57	87.16	91.01	40.28	80.88	72.16	84.82	83.16	74.17	80.23
HPT	Llama3	96.56	96.78	75.67	86.54	91.03	40.56	80.68	71.78	84.43	82.76	73.93	80.07
HPT++	Qwen2	96.85	96.85	75.44	85.76	91.10	40.89	81.09	74.35	86.35	83.45	73.97	80.55
HPT++	GPT-3.5	96.83	96.61	75.62	86.28	91.17	41.10	81.23	74.10	87.10	83.25	74.14	80.68
HPT++	Llama3	96.96	96.91	75.59	85.85	91.13	41.33	81.11	74.23	87.36	83.81	74.24	80.77

attention map is also crucial. Furthermore, compared to the learnable matrix in HPT, our method avoids overfitting to base classes, thereby ensuring better generalization.

Large Language Models for Corpus Generation

Since the performance of LLMs affects the quality of the generated corpus, thereby influencing the experimental results, we conduct ablation experiments on different LLMs, including the closed-source model GPT-3.5-turbo (OpenAI, 2023) and the open-source models Llama3-8B (AI@Meta, 2024) and Qwen2-7B (Yang et al., 2024). As shown in Table 7, each LLM may demonstrate superior performance on specific datasets compared to other LLMs, which can be attributed to the unique characteristics of this LLM. For example, GPT-3.5 shows a significant performance advantage over its competitors on Flowers102 dataset. However, the performance differences among various LLMs are generally not substantial, even though the superior performance of GPT-3.5 compared to other models has been demonstrated on many language understanding benchmarks. This suggests a weak correlation between recognition performance and the performance of LLMs, indicating that the latter does not play a decisive role. This also demonstrates that our corpus generation algorithm maintains a lower bound on recognition performance, regardless of the quality of linguistic knowledge.

Sample Analysis

To demonstrate the capability of HPT to capture category-related semantics, we provide sample analysis on three randomly selected categories from Caltech101. Figure 7 presents a comparison between our method and the baseline trained

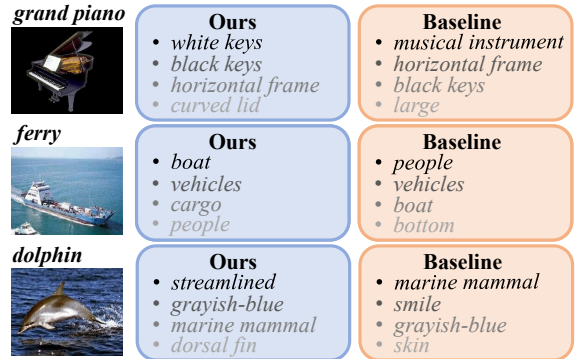


Fig. 7 Visualization of the top features with the highest attention scores according to the selected categories.

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.

7 Conclusion

In this paper, we assert that utilizing structured relationships from descriptions to enhance learning prompts is crucial. Consequently, we generate human-like descriptions along with their corresponding structured relationships and introduce hierarchical prompt tuning, a method that concurrently models both structured and conventional linguistic knowledge to significantly enhance prompt effectiveness. Our methods, including HPT and HPT++, demonstrate superior performance across three generalization tasks.

We hope that this work will attract greater attention to the role of structured knowledge in natural language prompt tuning, facilitating its application to a range of tasks beyond classification.

References

- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., *et al.*: Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* **35**, 23716–23736 (2022)
- AI@Meta: Llama 3 model card (2024)
- Bossard, L., Guillaumin, M., Gool, L.V.: Food-101—mining discriminative components with random forests. In: *European Conference on Computer Vision*, pp. 446–461 (2014). Springer
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020)
- Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M.: Visualgpt: Data-efficient adaptation of pre-trained language models for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18030–18040 (2022)
- Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., Pham, H., Dong, X., Luong, T., Hsieh, C.-J., *et al.*: Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675* (2023)
- Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613 (2014)
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., *et al.*: Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022)
- Cheng, D., Wang, G., Wang, B., Zhang, Q., Han, J., Zhang, D.: Hybrid routing transformer for zero-shot learning. *Pattern Recognition* **137**, 109270 (2023)
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255 (2009). Ieee
- Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 178–178 (2004). IEEE
- Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* **132**(2), 581–595 (2024)
- Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **12**(7), 2217–2226 (2019)
- Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., *et al.*: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349 (2021)
- Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271 (2021)
- Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.-N.: Visual prompt tuning. In: *European Conference on Computer Vision*, pp. 709–727 (2022). Springer
- Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig,

- T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, pp. 4904–4916 (2021). PMLR
- Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19113–19122 (2023)
- Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 554–561 (2013)
- Khattak, M.U., Wasim, S.T., Naseer, M., Khan, S., Yang, M.-H., Khan, F.S.: Self-regulating prompts: Foundational model adaptation without forgetting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15190–15200 (2023)
- Li, M., Chen, L., Duan, Y., Hu, Z., Feng, J., Zhou, J., Lu, J.: Bridge-prompt: Towards ordinal action understanding in instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19880–19889 (2022)
- Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5206–5215 (2022)
- Liu, H., Son, K., Yang, J., Liu, C., Gao, J., Lee, Y.J., Li, C.: Learning customized visual models with retrieval-augmented knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15148–15158 (2023)
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* **55**(9), 1–35 (2023)
- Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. In: European Conference on Computer Vision, pp. 529–544 (2022). Springer
- Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
- Menon, S., Vondrick, C.: Visual classification via description from large language models. arXiv preprint arXiv:2210.07183 (2022)
- Nilsback, M.-E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp. 722–729 (2008). IEEE
- OpenAI: GPT-4 Technical Report (2023)
- 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.: Combined scaling for zero-shot transfer learning. arXiv preprint arXiv:2111.10050 (2021)
- Pratt, S., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. arXiv preprint arXiv:2209.03320 (2022)
- Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3498–3505 (2012). IEEE
- Roy, S., Etemad, A.: Consistency-guided prompt learning for vision-language models. arXiv preprint arXiv:2306.01195 (2023)
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR
- Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International Conference on Machine Learning, pp. 5389–5400 (2019). PMLR
- Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu,

- Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18082–18091 (2022)
- Shen, S., Li, C., Hu, X., Xie, Y., Yang, J., Zhang, P., Gan, Z., Wang, L., Yuan, L., Liu, C., *et al.*: K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems* **35**, 15558–15573 (2022)
- Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012)
- Tay, Y., Tuan, L.A., Phan, M.C., Hui, S.C.: 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, pp. 1029–1038 (2017)
- Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816–16825 (2022)
- Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems* **32** (2019)
- Wang, Y., Jiang, X., Cheng, D., Li, D., Zhao, C.: Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 5749–5757 (2024)
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., *et al.*: Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* **35**, 24824–24837 (2022)
- Wang, Z., Zhang, Z., Lee, C.-Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 139–149 (2022)
- Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3485–3492 (2010). IEEE
- Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., *et al.*: Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021)
- Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783* (2021)
- Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., Gao, J.: Unified contrastive learning in image-text-label space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19163–19173 (2022)
- Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems* **35**, 124–141 (2022)
- Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Fan, Z.: Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024)

- Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6757–6767 (2023)
- Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Qiao, Y., Gao, P., Li, H.: Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15211–15222 (2023)
- Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15659–15669 (2023)
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
- Zhao, C., Wang, Y., Jiang, X., Shen, Y., Song, K., Li, D., Miao, D.: Learning domain invariant prompt for vision-language models. IEEE Transactions on Image Processing (2024)
- Zhang, Y., Wang, J., Yu, L.-C., Zhang, X.: Ma-bert: learning representation by incorporating multi-attribute knowledge in transformers. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 2338–2343 (2021)
- Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)