Pattern Recognition 174 (2026) 112986

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Image-free multi-label image recognition via LLM-powered hierarchical prompt tuning

Shuo Yang [a,*], Zirui Shang [b], Yongqi Wang [b], Derong Deng [a], Hongwei Chen [a], Xinxiao Wu [a,b], Qiyuan Cheng [a]

[a] Guangdong Provincial Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, 518172, Shenzhen, China
[b] Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, 100081, Beijing, China

ARTICLE INFO

Keywords:
Hierarchical prompt tuning
Multi-label image recognition
Image-free
LLM

ABSTRACT

This paper proposes a novel framework for multi-label image recognition without any training images, namely image-free framework, which uses knowledge of pre-trained Large Language Model (LLM) to learn prompts to adapt a pre-trained Vision-Language Model (VLM) like Contrastive Language–Image Pre-training (CLIP) to multi-label classification. Through asking LLM well-designed questions, we acquire comprehensive knowledge about the characteristics and contexts of objects, which provides valuable text descriptions for learning prompts. Then, we propose a hierarchical prompt learning method by taking the multi-label dependency into consideration, wherein a subset of category-specific prompt tokens is shared when the corresponding objects exhibit similar attributes or are more likely to co-occur. Benefiting from the remarkable alignment between visual and linguistic semantics of CLIP, the hierarchical prompts learned from text descriptions are applied to perform classification of images during inference. Our framework presents a new way to explore the synergies between multiple pre-trained models for novel category recognition. Extensive experiments on three public datasets, *i.e.*, Microsoft Common Objects in Context (MS-COCO), Visual Object Classes 2007 (VOC2007), and National University of Singapore Web Image Database (NUS-WIDE), demonstrate that our method achieves better results than the state-of-the-art methods.

1. Introduction

Image recognition, a core problem in computer vision, aims to automatically identify and categorize visual objects within images. Over the past decades, it has evolved into a cornerstone technology with wide-ranging applications in various domains, including medical imaging for disease diagnosis, autonomous driving for scene understanding, remote sensing for environmental monitoring, and industrial inspection for quality control. These diverse applications demonstrate the importance and versatility of image recognition systems in addressing real-world challenges.

Within this field, multi-label image recognition aims to recognize all objects present in an image. This task is challenging due to the emergence of novel objects and scenes [1] during inference in real-world scenarios, as shown in Fig. 1(a). Recent large-scale pre-trained Vision-Language Models (VLMs) like CLIP [2] spawn the training-free zero-shot methods [3], which can handle new categories by calculating similarities between images and texts in a well-aligned embedding space. To further effectively adapt VLMs to enhance the performance of novel categories, several methods have been proposed to learn adapters [4] or prompts [5] using sufficient annotated images, as shown in Fig. 1(b). However, the performance of these prompt learning methods may be limited when it is infeasible to collect sufficient fully annotated images.

To address this issue, Sun et al. [5] propose dual context optimization to quickly adapt CLIP to multi-label recognition using partially labeled images, where only a few categories for each training image are annotated, significantly reducing the annotation burden. Guo et al. [6] propose texts as images in prompt tuning to adapt CLIP, where the text descriptions are human-written image captions from existing datasets and serve as alternatives to images. This method presents a more practical and efficient way for prompting, as text descriptions are more easily accessible than images. And their following work [7] further learn a pseudo-visual prompt for each category and then adopts a co-learning strategy with a dual-adapter module to transfer visual knowledge from pseudo-visual prompt to text prompt.

In this paper, we propose an image-free framework for multi-label image recognition without any annotated images or image captions for training. It leverages knowledge of objects from a pre-trained Large

* Corresponding author.
E-mail address: yangshuo@smbu.edu.cn (S. Yang).

https://doi.org/10.1016/j.patcog.2025.112986
Received 28 May 2025; Received in revised form 22 October 2025; Accepted 22 December 2025
Available online 24 December 2025

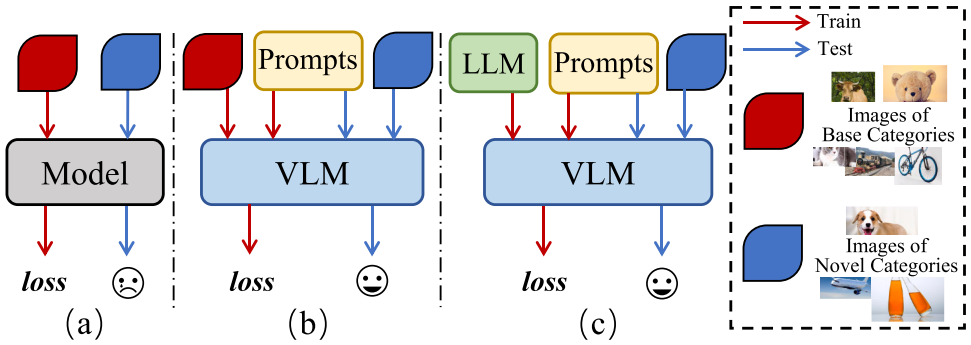

Fig. 1. Illustration of different ways to handle novel categories. (a) Traditional methods train on base categories but fail on novel categories. (b) Recent prompting methods successfully adapt VLM to novel categories but need annotated images for prompt tuning. (c) Our image-free framework only performs prompt tuning to adapt VLM to novel categories by LLM.

Language Model (LLM) to adapt CLIP to multi-label classification by textual prompt tuning, as shown in Fig. 1(c). Specifically, we propose to collect comprehensive information of objects by designing different types of questions posed to LLM. Starting with asking LLM category-agnostic questions like *[object lists], please summarize 90 attributes that may be common to the above 80 words* to acquire common attributes, such as shape, color, and material, shared by all categories, similarly we then acquire particular attributes for each category by category-specific questions like *please summarize 30 attributes of [object]*. Finally, we acquire text descriptions of the attributes by category-description questions like *please help me generate 100 different sentences about [category] from the angle of the [attribute]*. Moreover, we design scene-related questions like *generate ten sentences to describe different scenes involving [category1] and [category2]* to acquire text descriptions of contextual relationships between multiple object categories in real-world scenes, namely relationship knowledge.

Along with category labels, the acquired text descriptions of attribute and relationship knowledge from LLM are used as images for prompt tuning CLIP to multi-label recognition. To incorporate the relationship information between multiple objects into prompt learning to further improve the performance, we propose a hierarchical prompt learning method, which categorizes the prompt tokens into three types: (1) shared tokens shared by all object categories; (2) partial-shared tokens shared by the object categories of the same subgroups with a co-occurrence relationship or similar attributes; (3) category-specific tokens specific to each individual object category. Through designing these hierarchical tokens, we learn prompts that absorb both task-specific knowledge and object-specific knowledge, as well as the relationship knowledge between objects. Benefiting from the remarkable alignment between visual and linguistic semantics of CLIP, the hierarchical prompts learned from text description are applied to perform classification of images during inference.

In summary, the contributions of our work are as follows:

- We propose an image-free framework for multi-label image recognition without any annotated image or image captions, which leverages rich knowledge in LLM to prompt tune CLIP. Our framework introduces a promising avenue for handling new objects in visual recognition, relying solely on pre-trained models, and also paves an effective way to explore the synergies between multiple pre-trained models.
- We propose a hierarchical prompt learning method to adapt CLIP by using the acquired knowledge of objects from LLM. It incorporates relationships between different categories into learnable prompts, thus further improving the multi-label recognition performance.
- We propose to collect comprehensive information about object attributes and relationships from LLM by designing different types of questions.

Extensive experiments on three public datasets, such as the MS-COCO dataset, the VOC2007 dataset, and the NUS-WIDE dataset, demonstrate that our method achieves better results than the existing methods. This paper is an updated version of our previous arXiv paper [8], in which we refine the term "data-free" to the more precise "image-free" and incorporate additional experimental results for comparison with more recent works. The code is publicly available at https://github.com/shuoyang129/image-free-multi-label-image-recognition.

2. Related work

2.1. Multi-label image recognition

Early approaches to multi-label image recognition, such as HCP [9], regarded this task as training multiple independent binary classifiers (e.g., Binary Relevance [10] and Human Reporting Bias Modelling [11]). Although straightforward, these methods neglected label correlations. To address this, subsequent work explicitly modeled label dependencies—for example, NUS-WIDE [12] and ML-GCN [13] employed graph-based reasoning with category co-occurrence graphs, while RLS-DNet [14] and Orderless RNN [15] adopted sequence modeling strategies to capture label ordering. Attention-driven methods, including MCAR [16] and ADGNN [17], enhanced focus on discriminative image regions. Loss reweighting schemes, such as the Asymmetric Loss (ASL) [18], further improved performance by mitigating label imbalance and emphasizing hard positive predictions.

Despite significant progress, these approaches rely heavily on large-scale annotated images, limiting their scalability in data-scarce regimes. Recent research has therefore explored few-shot and zero-shot recognition, for example by Shared Multi-Attention [19], Semantic Diversity Learning (SDL) [1], and language-driven cross-modal models [20]. In parallel, methods for partial-label learning such as PL-MCL [21], SARB [22], SST [23], and G^2NetPL [24] alleviate annotation sparsity. By exploring the synergies between LLM and VLM, we take a significant step forward in multi-label image recognition by introducing an image-free framework where no annotated images or image captions are provided.

2.2. Adapting CLIP to visual tasks

Vision–Language Models (VLMs) have demonstrated impressive capabilities in learning generic representations, such as CLIP [2]. In order to adapt VLMs to specific downstream tasks, many prompt-tuning methods have been proposed to learn task-specific prompts while keeping most parameters frozen. Representative examples include Visual Prompt Tuning (VPT) [25], which prepends learnable visual tokens to the input sequence, ProGrad [26], which regularizes gradients to balance adapted and zero-shot performance, AD-CLIP [27], which adapts prompts across

domains, and VPT-Deep [28], which inserts prompts at deeper transformer layers for stronger adaptation. These approaches attract significant attention for their combination of high accuracy and parameter efficiency. To further bridge the domain gap between the pre-training data of VLMs and specific target tasks, dedicated adapter-based strategies have been designed and integrated into CLIP. Notable examples include Tip-Adapter [29], which stores training features in a key–value cache for fast retrieval, ProbVLM [30], which introduces probabilistic adapters to model uncertainty in predictions, and CLIP-Driven Unsupervised Learning(CDUL) [4], which exploits unsupervised contrastive learning to adjust CLIP to specialized domains, all of which avoid full model fine-tuning.

The works most relevant to ours are DualCoOp [5], DualCoOp++ [31], TaI-DPT [6], and TaI++ [7]. DualCoOp learns a pair of differentiable prompts to provide positive and negative contexts for the target class using partial-annotated images, and DualCoOp++ further learns evidential prompts to serve as guidance to aggregate positive and negative contexts from the spatial domain of the image. TaI-DPT uses image captioning collected from existing datasets as images to learn text prompts for each class, and TaI++ further learns pseudo-visual prompts and then adopts a co-learning strategy with a dual-adapter module to transfer visual knowledge from pseudo-visual prompts to text prompts, and it also adopts a simple strategy to query LLM to generate pseudo-texts for training. In contrast, our method inquires LLMs to acquire comprehensive knowledge of object categories and scene-related relationship knowledge as text descriptions for prompt learning. Moreover, our method learns relationship-aware hierarchical prompts that define a multi-level token architecture — shared tokens (global task priors), partial-shared tokens (attributes or scenes common to specific category subgroups), and category-specific tokens (fine-grained discriminative cues). This structure allows the prompts to capture both intra-class details and inter-class dependencies, which is especially beneficial for multi-label recognition scenarios.

2.3. LLM-enhanced visual understanding

Large Language Models (LLMs) have increasingly been used to augment visual understanding tasks owing to their *emergent abilities* to reason from in-context examples [32,33]. By converting visual content into textual descriptions, methods such as GPT-3 for KB-VQA [34] and Visual Programming [35] leverage LLMs to perform compositional visual reasoning without additional task-specific training. LLM-based grounding systems, including LLM-Grounder [36], achieve open-vocabulary 3D visual grounding by treating the LLM as an agent, while ChatGPT-powered Comparison Tree [37] enhances CLIP's zero-shot classification through structured semantic comparisons.

In contrast to these approaches, our method positions the LLM not as an auxiliary reasoner, but as a *textual knowledge generator*, querying it with carefully designed attribute and scene-relation questions. This produces structured, multi-granularity descriptions that empower the hierarchical prompt tuning of CLIP for multi-label recognition in an entirely image-free setting.

3. Our method

3.1. Overview

We propose an image-free framework for multi-label image recognition without any annotated images or image captions. A large language model, *i.e.*, ChatGLM, is employed as a repository of encyclopedic knowledge due to its open-source nature and strong bilingual (Chinese–English) reasoning capabilities. We leverage it to acquire comprehensive knowledge of object categories by formulating diverse question types specifically designed for knowledge extraction. This is motivated by the fact that we can effectively identify an object in a picture if provided with a linguistic description. Then, a pre-trained vision-language model, *i.e.*, CLIP, is prompt-tuned using the acquired knowledge to enhance multi-label classification, based on the robust cross-modal alignment between images and text. We propose a hierarchical prompt learning method to incorporate relationships between objects into the learnable prompts. Fig. 2 shows an overview of our framework.

Given an input image $\mathbf{x}$, multi-label image recognition aims to identify all object categories in it, formulated as $\mathbf{S} = f_{\Phi,\Psi}(\mathbf{x})$, where $f_{\Phi,\Psi}$ denotes the recognition model, Φ denotes ChatGLM [38], Ψ denotes CLIP, including a text encoding module Ψ_t and an image encoding module Ψ_i, and $\mathbf{S} \in \mathbb{R}^N$ is the predicted probability scores for all N categories $\mathbb{Y} = \{Y_1, \cdots, Y_N\}$.

3.2. Knowledge acquirement

To describe an object, it is crucial to have detailed information about its color, shape, texture, and other attributes. To obtain this information, we engage with ChatGLM, a highly knowledgeable language model that functions as a chatbot and responds to carefully crafted questions. Its extensive encyclopedic knowledge allows us to extract the necessary details for adequate object description.

Coarse Attribute Description. To capture diverse aspects of objects, we begin by extracting common attributes shared by all categories using category-agnostic questions and particular attributions for individual categories using category-specific questions, formulated as

$$(\mathbb{A}_c, \mathbb{A}_s) = \Phi(\Pi_1(\mathbb{Y})) \tag{1}$$

where $\Pi_1(\cdot)$ denotes the common and category-specific questions like *[object lists], please summarize 90 attributes that may be common to the above 80 words*. $\mathbb{A}_c = \{a_1, \cdots, a_{n_1}\}$ denotes n_1 common attributes of all categories. $\mathbb{A}_s = \{\mathbb{A}_{s,1}, \cdots, \mathbb{A}_{s,N}\}$ denotes the category-specific attribute sets, where $\mathbb{A}_{s,i} = \{a_{i,1}, \cdots, a_{i,n_2}\}$ denotes n_2 attributes of the i-th category.

Then we obtain the text descriptions of each category by asking additional questions. Note that these text descriptions of attributes may contain noise, and we call them coarse attribute descriptions. This process is formulated by

$$\mathbb{D}_i^c = \Phi(\Pi_2(\mathbb{A}_c \cup \mathbb{A}_{s,i}, Y_i)) \tag{2}$$

where $\Pi_2(\cdot)$ denotes the questions about describing the attributes of objects, like *please help me generate 100 different sentences about [category] from the angle of the [attribute]*, and $\mathbb{D}_i^c$ denotes the coarse attribution descriptions of the i-th category. Let $\mathbb{D}^c = \{\mathbb{D}_1^c, \cdots, \mathbb{D}_N^c\}$ denote the coarse attribute description sets. Here, the union operation "$\cup$" denotes a set-theoretic merge between the common attribute set $\mathbb{A}_c$ and the category-specific attribute set $\mathbb{A}_{s,i}$ for the i-th category. $\mathbb{A}_c$ contains general attributes shared by all categories, while $\mathbb{A}_{s,i}$ contains attributes unique to that category. The merged set provides a richer and more complete attribute pool, which helps generate textual descriptions that capture both inter-category commonalities and intra-category distinctions.

Fine-grained Attribute Description. We design several questions to remove the noisy attributes that are irrelevant to the specific category, resulting in a fine-grained attribute set for each category. We similar inquire ChatGLM to acquire the fine-grained attribute descriptions $\mathbb{D}^f = \{\mathbb{D}_1^f, \cdots, \mathbb{D}_N^f\}$ by asking questions. This process is formulated by

$$\mathbb{D}_i^f = \Phi(\Pi_3(\mathbb{D}_i^c, Y_i)) \tag{3}$$

where $\Pi_3(\cdot)$ denotes the questions about how to remove irrelevant attributes, like *[attribute list], please delete the above attribute words given that are not very relevant to [category]. Finally, 70 attribute words remain.* No explicit relevance rules are provided. The LLM autonomously selects attributes it considers most representative based on its knowledge. This rule-free approach enables an extensible pipeline capable of handling newly emerging object categories without additional human intervention.

Relationship Description. In multi-label image recognition, the co-occurrence relationships between different categories contribute significantly to the performance [13,39]. To simulate this scenario, we first

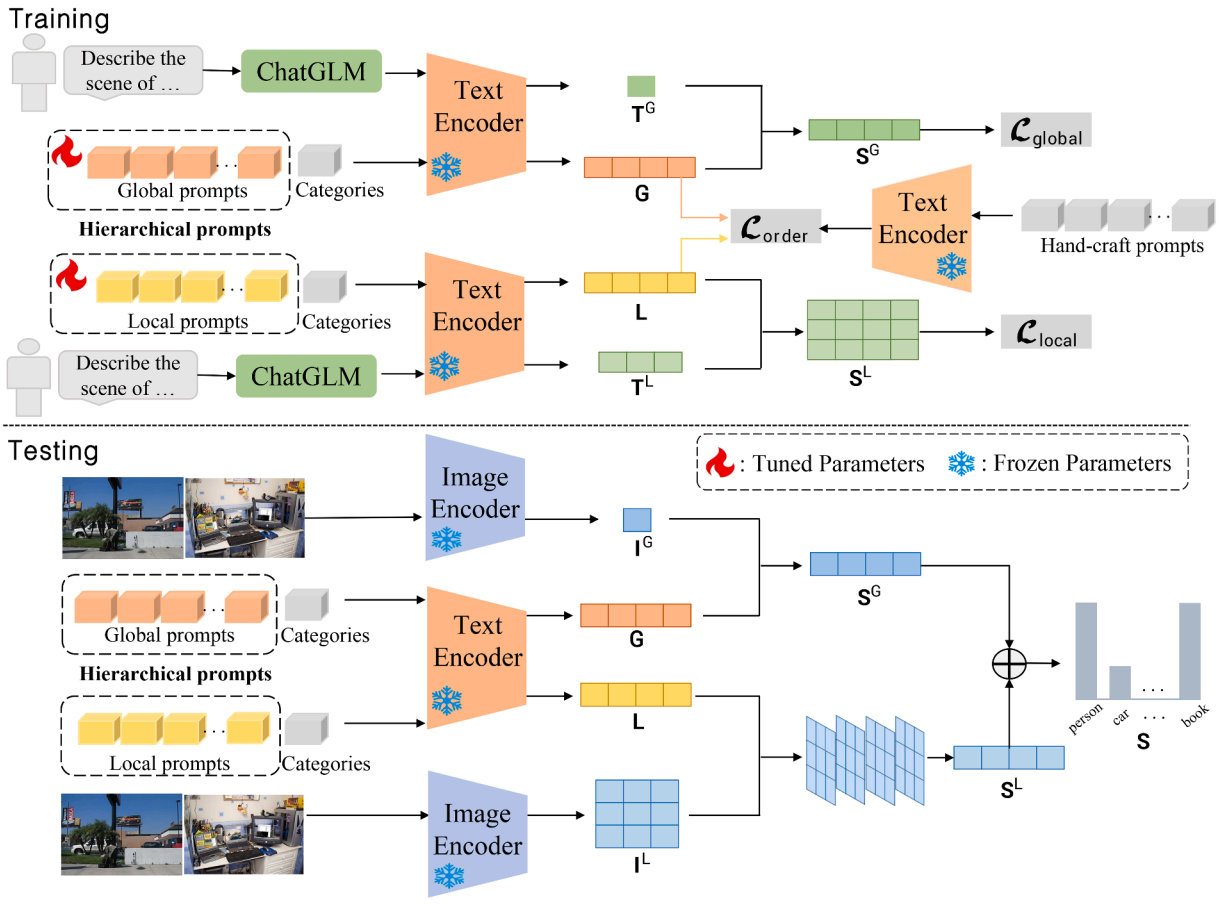

Fig. 2. Overview of our framework.

split all categories into multiple scene-related subgroups by ChatGLM:

$$\mathbb{G} = \{\mathbb{G}_i\}_{i=1}^{n_3} = \Phi(\Pi_4(\mathbb{Y})) \tag{4}$$

where $\Pi_4(\cdot)$ denotes questions about how to divide categories into subgroups based on their relationships, like *[category list], categorize the above words according to possible common occurrences in a scene*, and n_3 is the number of subgroups. For each subgroup, two categories are selected to formulate scene-related questions $\Pi_5(\cdot)$, like *generate 100 different descriptive sentences for a scene containing [category1] and [category2]*, and fed into ChatGLM to obtain relationship descriptions:

$$\mathbb{D}_i^r = \Phi(\Pi_5(\mathbb{G}_i)) \tag{5}$$

where $\mathbb{D}_i^r$ denotes the fine-grained attribute descriptions of the i-th categroy. Let $\mathbb{D}^r = \{\mathbb{D}_1^r, \cdots, \mathbb{D}_N^r\}$ denote the fine-grained attribute description sets.

Fig. 3 illustrates an example of the designed questions and their corresponding answers from ChatGLM.

Discussion. Our description generation relies on LLMs (e.g., ChatGLM), which are known to occasionally produce hallucinations and noisy outputs. This could be problematic in specialized domains where training data of LLMs is scarce (e.g., medical data), potentially leading to unreliable generations. However, in common scenarios, LLMs generally perform well, and while some noise exists, the majority of generated data remains useful, especially as training data for downstream tasks. Although a small fraction of the automatically generated descriptions may contain noise, this has minimal impact on the downstream network training. The large-scale dataset ensures that most samples are accurate, and deep models are generally robust to low-level description noise. Thus, our focus is on maintaining overall quality rather than eliminating every noisy instance.

3.3. Hierarchical prompt learning

Based on the previous generated text descriptions $\mathbb{D} = \mathbb{D}^k, k \in \{c, f, r\}$, we propose hierarchical prompt learning to adapt CLIP to multi-label recognition. This approach designs hierarchical prompts to model the complex relationships between categories, introducing multi-level prompt learning to grasp the discriminability of features.

Hierarchical Prompts. A learnable prompt usually consists of several learnable tokens and a placeholder to put the category label, denoted as $\mathbf{p}_i = [\mathbf{t}_1^i, \mathbf{t}_2^i, \cdots, \mathbf{t}_M^i, Y_i]^\top$, where M is the number of learnable tokens, and Y_i is the i-th category label. For different categories, there are two common prompt types: shared prompts, where tokens are shared across all categories, and category-specific prompts, where tokens are distinct for each category. Both types of prompts have been demonstrated to be effective in recent works [26,40], but they neglect the rich, structured relationships between categories, leading to sub-optimal performance.

To capture the complex inter-category relationships, we propose hierarchical prompts $\mathbf{P}^h$ that are structured into four distinct levels, moving from general to specific. The learnable tokens for each category are a composite of:

(1) **Shared Tokens ($\mathbf{t}_s$):** These tokens are shared across all categories, learning global, universal context (e.g., "a photo of a …").

(2) **Coarse-grained Partial-Shared Tokens ($\mathbf{t}_{cg}$):** These tokens are shared by categories belonging to the same pre-defined super-category. For instance, based on dataset annotations, categories like "bicycle," "car," and "bus" all belong to the "vehicle" super-category and would share a specific token $\mathbf{t}_{cg,\text{vehicle}}$ at this position, which is different from the token $\mathbf{t}_{cg,\text{furniture}}$ shared by "sofa" and "chair".

(3) **Fine-grained Partial-Shared Tokens ($\mathbf{t}_{fg}$):** This level captures functional or contextual co-occurrence patterns. We first construct a

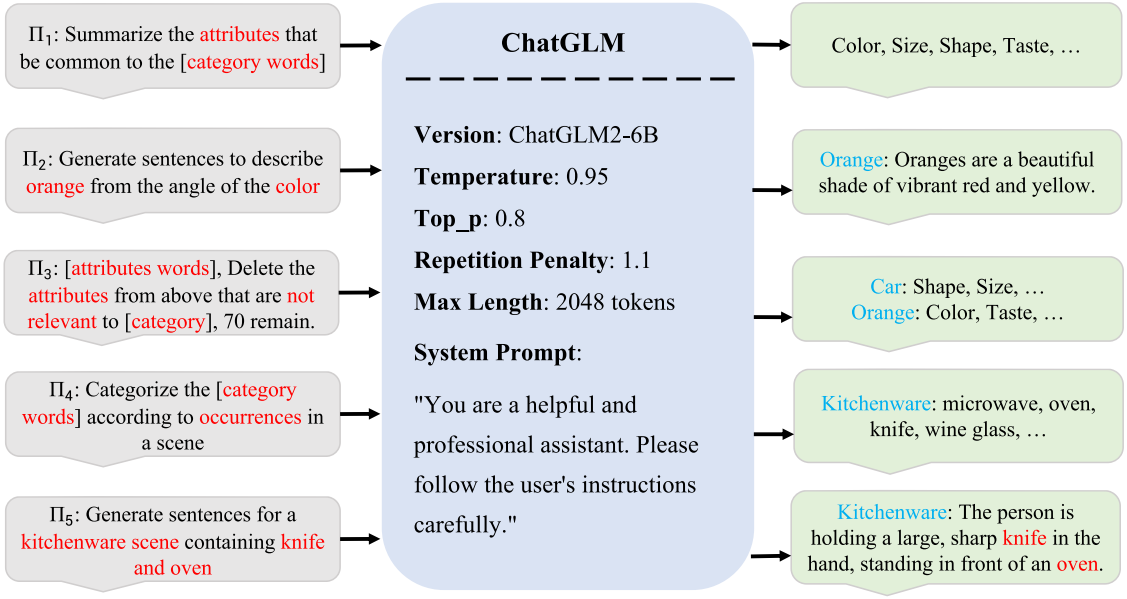

Fig. 3. An example of the designed questions and their corresponding answers from ChatGLM. We also provide the configuration details for the ChatGLM model used. More detailed examples can be found in the supplementary materials.

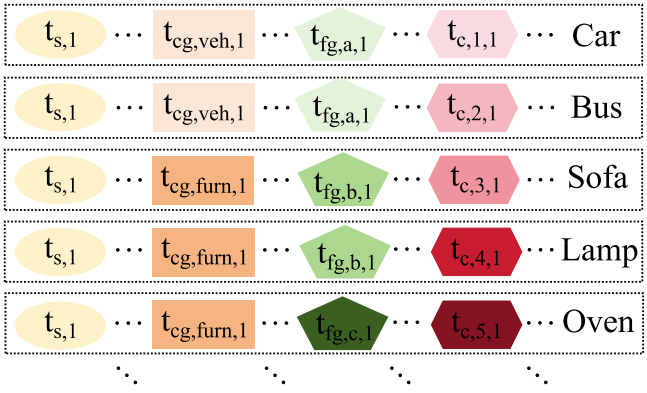

Fig. 4. An illustration of the proposed four-level hierarchical prompts. Different shapes are used to conceptually represent the distinct token types: $\mathbf{t}_s$ (circles) denotes global shared tokens; $\mathbf{t}_{cg}$ (rectangles) denotes coarse-grained partial-shared tokens shared by super-categories (e.g., 'vehicle'); $\mathbf{t}_{fg}$ (pentagons) denotes fine-grained partial-shared tokens shared by co-occurrence cliques (e.g., 'sofa', 'lamp'); and $\mathbf{t}_c$ (hexagons) denotes the final category-specific tokens.

category co-occurrence graph $G = (V, E)$ based on sentence-level statistics from the training set, where V is the set of all categories. An undirected edge (Y_i, Y_j) is added if their co-occurrence frequency exceeds a certain threshold τ. We then identify all maximal cliques in G. A maximal clique is a subset of categories where every category co-occurs frequently with every other category in the subset, and this subset is not part of a larger clique. Each identified maximal clique is defined as a fine-grained co-occurrence group and is assigned a unique learnable token $\mathbf{t}_{fg,k}$ shared only by its members.

(4) **Category-specific Tokens ($\mathbf{t}_c$):** These tokens are unique to each individual category, capturing its most distinct and granular features.

This hierarchical structure can be represented as an illustrative matrix $\mathbf{P}^h$. Each row represents the prompt for a specific category, composed of blocks of tokens from each hierarchical level:

$$\mathbf{P}^h = \begin{bmatrix} \mathbf{t}_{s,1} & \cdots & \mathbf{t}_{cg,\text{veh},1} & \cdots & \mathbf{t}_{fg,a,1} & \cdots & \mathbf{t}_{c,1,1} & \cdots & Y_1 \text{ (car)} \\ \mathbf{t}_{s,1} & \cdots & \mathbf{t}_{cg,\text{veh},1} & \cdots & \mathbf{t}_{fg,a,1} & \cdots & \mathbf{t}_{c,2,1} & \cdots & Y_2 \text{ (bus)} \\ \mathbf{t}_{s,1} & \cdots & \mathbf{t}_{cg,\text{furn},1} & \cdots & \mathbf{t}_{fg,b,1} & \cdots & \mathbf{t}_{c,3,1} & \cdots & Y_3 \text{ (sofa)} \\ \mathbf{t}_{s,1} & \cdots & \mathbf{t}_{cg,\text{furn},1} & \cdots & \mathbf{t}_{fg,b,1} & \cdots & \mathbf{t}_{c,4,1} & \cdots & Y_4 \text{ (lamp)} \\ \mathbf{t}_{s,1} & \cdots & \mathbf{t}_{cg,\text{furn},1} & \cdots & \mathbf{t}_{fg,c,1} & \cdots & \mathbf{t}_{c,5,1} & \cdots & Y_5 \text{ (oven)} \\ & \vdots & & \vdots & & \vdots & & \vdots & \vdots \end{bmatrix}$$

Here, the matrix illustrates that each hierarchical level consists of multiple tokens (indicated by '⋯'). For example, the shared block contains tokens $\mathbf{t}_{s,1}, \ldots, \mathbf{t}_{s,L_s}$ which are shared by all categories. In the coarse-grained block, Y_1 (car) and Y_2 (bus) share the "vehicle" tokens ($\mathbf{t}_{cg,\text{veh},1}, \ldots, \mathbf{t}_{cg,\text{veh},L_{cg}}$), while Y_3 (sofa), Y_4 (lamp), and Y_5 (oven) share the "furniture" tokens ($\mathbf{t}_{cg,\text{furn},1}, \ldots, \mathbf{t}_{cg,\text{furn},L_{cg}}$). Similarly, in the fine-grained block, categories within a clique share tokens. Y_1 and Y_2 are shown to be in clique "a" (sharing $\mathbf{t}_{fg,a,1}, \ldots, \mathbf{t}_{fg,a,L_{fg}}$), while Y_3 and Y_4 are in clique "b" (sharing $\mathbf{t}_{fg,b,1}, \ldots, \mathbf{t}_{fg,b,L_{fg}}$). Y_5 (oven), belonging to a different clique "c", has its own set of fine-grained tokens ($\mathbf{t}_{fg,c,1}, \ldots, \mathbf{t}_{fg,c,L_{fg}}$). Finally, each category i has its own block of unique, category-specific tokens ($\mathbf{t}_{c,i,1}, \ldots, \mathbf{t}_{c,i,L_c}$). This multi-level structure allows the model to learn features at global, super-category, co-occurrence, and instance-specific levels simultaneously. We present an illustration of this concept in Fig. 4.

Global Learning. Global learning aims to learn global hierarchical prompts to grasp the discriminative ability of global features. Let $\mathbf{P}^h_{g,i}$ be the global hierarchical prompts of the i-th category. We initialize $\mathbf{P}^h_{g,i}$ randomly and then feed it to the text encoder Ψ_t of CLIP to generate the global category embedding $\mathbf{G}_i \in \mathbb{R}^d$:

$$\mathbf{G}_i = \Psi_t(\mathbf{P}^h_{g,i}), \quad i \in \{1, 2, \cdots, N\} \tag{6}$$

where $d = 512$ denotes the embedding dimension. Meanwhile, for each text description $\mathbf{r} \in \mathbb{D}$, we extract its global feature $\mathbf{T}^G \in \mathbb{R}^d$:

$$\mathbf{T}^G = \Psi_t(\mathbf{r}) + \xi, \tag{7}$$

where ξ is a Gaussian noise adding to the text features to simulate the alignment inconsistency between images and texts, which enhances inference performance when inputting images. The cosine similarity between the global feature of text description and the global category embedding, denoted as S^G_i, is calculated by

$$S^G_i = \langle \mathbf{T}^G, \mathbf{G}_i \rangle, \quad i \in \{1, \cdots, N\} \tag{8}$$

Local Learning. Local learning aims to learn local hierarchical prompts to grasp the discriminative ability of fine-grained features. Let $\mathbf{P}^h_{l,i}$ be the local hierarchical prompts of the i-th category, whose structure is identical to global prompts but with different parameters. We initialize $\mathbf{P}^h_{l,i}$ randomly and feed it to the encoder Ψ_t of CLIP to generate the local category embedding $\mathbf{L}_i \in \mathbb{R}^d$:

$$\mathbf{L}_i = \Psi_t(\mathbf{P}^h_{l,i}), \quad i \in \{1, 2, \cdots, N\} \tag{9}$$

For the text description $\mathbf{r} \in \mathbb{D}$, we extract its local features $\mathbf{T}_i^L \in \mathbb{R}^{N_r \times d}$ by a modified text encoder $\widetilde{\Psi}_t$:

$$\mathbf{T}^L = \widetilde{\Psi}_t(\mathbf{r}) + \xi, \tag{10}$$

where ξ is a Gaussian noise, N_r is the number of tokens in $\mathbf{r}$, $\widetilde{\Psi}_t$ denotes that we preserve the sequential token features of the entire sentence instead of only the $\langle EOS \rangle$ token features (global features). The category-aware similarity between the sequential local features of text description and the local category embedding is calculated in a weighted manner:

$$S_i^L = \sum_{j=1}^{N_r} \frac{\exp(\mathbf{s}_{i,j})}{\sum_{j=1}^{N_r} \exp(\mathbf{s}_{i,j})} \cdot \mathbf{s}_{i,j}, \quad i \in \{1, \cdots, N\} \tag{11}$$

where $\mathbf{s}_{i,j} = \left\langle \mathbf{L}_i, \mathbf{T}_j^L \right\rangle$ is the similarity between the i-th local class embedding and j-th token (column) of local features.

3.4. Training objectives

For the global similarity $\mathbf{S}^G = [S_1^G, \cdots, S_N^G]^\top$ and local similarity $\mathbf{S}^L = [S_1^L, \cdots, S_N^L]^\top$ of each text description, we adopt two loss functions to optimize the corresponding learnable prompts.

Ranking Loss. We utilize the ranking loss to assess the disparity between classification scores and ground-truth labels. Specifically, the ranking loss for global and local learning is calculated separately:

$$\begin{aligned} \mathcal{L}_{rank} &= \mathcal{L}_{global} + \mathcal{L}_{local} \\ \mathcal{L}_{global} &= \textstyle\sum_{i \in \mathbb{Y}^+, j \in \mathbb{Y}^-} \max(0, m - S_i^G + S_j^G) \\ \mathcal{L}_{local} &= \textstyle\sum_{i \in \mathbb{Y}^+, j \in \mathbb{Y}^-} \max(0, m - S_i^L + S_j^L) \end{aligned} \tag{12}$$

where $m = 1$ is the margin controlling how much higher the similarity score with the positive classes $\mathbb{Y}^+$ is than with the negative classes $\mathbb{Y}^-$.

Order Loss. Due to the potential noise introduced by text generated from ChatGLM, which could mislead the optimization process, we introduce an anchor to the learned prompts. Specifically, we anchor the learned prompts using hand-crafted prompts, such as *a photo of [category]*. The rationale behind this is that human-defined prompts have demonstrated mediocre zero-shot performance; thus, the resulting order of all categories for a given image can be considered reasonable to some extent. We posit that maintaining this order in the learned prompts helps mitigate the impact of noisy inputs, thereby enhancing the overall effectiveness of the model.

Specifically, for learnable global and local prompts and hand-crafted prompts, denoted as G, L, H, we extract their category embedding using Eq. (6), denoted by $\mathbf{G}_k, k \in \{G, L, H\}$, followed by similarity calculation between different categories: $\mathbf{D}^k = \mathbf{G}_k \times \mathbf{G}_k^\top, k \in \{G, L, H\}$. The order loss is then calculated by the Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{order} = \mathrm{KL}(\mathbf{D}^G, \mathbf{D}^H) + \mathrm{KL}(\mathbf{D}^L, \mathbf{D}^H) \tag{13}$$

Finally, the overall loss is given by

$$\mathcal{L} = \mathcal{L}_{rank} + \lambda_1 \cdot \mathcal{L}_{order} \tag{14}$$

3.5. Inference

Thanks to the large-scale image-text contrastive pre-training of CLIP, text features have been well-aligned to the image features of the same semantic meanings. As a result, our prompts learned from text descriptions can be applied to images during inference.

Specifically, with the learned hierarchical prompts, we extract the global and local class embeddings using Eq. (6) and Eq. (9), respectively, denoted as $\mathbf{T}^G = [\mathbf{T}_1^G, \cdots, \mathbf{T}_N^G]^\top$ and $\mathbf{T}^L = [\mathbf{T}_1^L, \cdots, \mathbf{T}_N^L]^\top$. For an input image $\mathbf{x}$, we extract the global and local image features by

$$\mathbf{I}^G = \Psi_i(\mathbf{x}) \in \mathbb{R}^d, \qquad \mathbf{I}^L = \widetilde{\Psi}_i(\mathbf{x}) \in \mathbb{R}^{N_I \times d} \tag{15}$$

where $\mathbf{I}^G$ is the global image feature, $\mathbf{I}^L$ is the local image feature, $\widetilde{\Psi}_i$ is the modified image encoder of CLIP that keeps the dense image features as output, and N_I is the length of flatted dense image features. Finally, we calculate the similarities between category embeddings and image features by

$$\mathbf{S} = \lambda_2 \cdot \mathbf{S}^G + (1 - \lambda_2) \cdot \mathbf{S}^L \tag{16}$$

where λ_2 is a parameter to weigh how much contribution of the global prompts and local prompts. $\mathbf{S}^G = [\langle \mathbf{I}^G, \mathbf{T}_1^G \rangle, \cdots, \langle \mathbf{I}^G, \mathbf{T}_N^G \rangle]^\top$ is the global similarity. $\mathbf{S}^L = [S_1^L, \cdots, S_N^L]^\top$ is the local similarity, where

$$S_i^L = \textstyle\sum_{j=1}^{N_I} \frac{\exp(\mathbf{s}_{i,j}/\tau)}{\sum_{j=1}^{N_I} \exp(\mathbf{s}_{i,j}/\tau)} \cdot \mathbf{s}_{i,j}$$

and $\mathbf{s}_{i,j} = \left\langle \mathbf{T}_i^L, \mathbf{I}_j^L \right\rangle$ is the cosine similarity between the i-th local category embedding and the j-th token (column) of local image features.

For multi-label prediction, we select the top-k categories with the highest similarity scores in $\mathbf{S}$ as the predicted labels for each image.

4. Experiments

4.1. Datasets and evaluation metrics

Datasets. We conduct experiments on the MS-COCO [41], VOC2007 [42] and NUS-WIDE [12] datasets for evaluation. For all three datasets, no training data is used, and the testing is performed on the testing or validation sets. MS-COCO is a widely used multi-label dataset for image recognition, which contains 80 categories with 82,081 training images and 40,504 validation images. VOC2007 contains 20 object categories with a total of 5011 images for training and validation, and 4952 images for testing. NUS-WIDE contains 81 categories with 161,789 images for training and 107,859 images for testing.

Metrics. We use the conventional evaluation metrics, including the mean of class-average precision (mAP) and the overall F1 score at Top-3 predictions.

4.2. Implementation details

ChatGLM Query Setup. We queried ChatGLM via a locally deployed instance based on the official GitHub implementation from THUDM/ChatGLM, using default parameters of ChatGLM2-6B without modification. No explicit stop criteria were applied during generation; instead, the number of required responses was specified within the prompt, and text generation naturally terminated after those responses were produced. All interactions were performed offline to ensure consistency and reproducibility.

CLIP Model Setup. Our model is built upon the official CLIP implementation using a ResNet-50 as the vision encoder, with a fixed input image resolution of 224×224. ResNet-50 is the default and most widely used option in CLIP, offering a strong balance between representation quality and computational efficiency. Using ResNet-50 also facilitates direct comparison with prior prompt learning approaches (e.g., DualCoOp [5], TaI-DPT [6]) that share the same setup. Crucially, during the entire prompt tuning process, both the vision and text encoders of the pre-trained CLIP model are kept completely frozen. No modifications are made to their architectures. The only trainable parameters in our framework are the hierarchical prompt embeddings.

Hierarchical Prompt Tuning. The local and global hierarchical prompts are composed of 32 learnable context tokens each. Both sets of prompt embeddings are randomly initialized. During training, these learnable prompts are optimized to align with CLIP's rich feature space. The local prompt embeddings are specifically trained to match the local features produced by the frozen text encoder, guided by the $\mathcal{L}_{local}$ loss component. Concurrently, the global prompt embeddings are aligned with the global features from the text encoder via the $\mathcal{L}_{global}$ loss. At inference time, the optimized local and global prompt embeddings are used to compute similarity scores against the corresponding local and global visual features extracted by the frozen vision encoder. These two

Table 1
Results of different text descriptions on MS-COCO, VOC2007, and NUS-WIDE. † means that results are from TaI-DPT [6]. Note that the hand-crafted prompts used in this paper are designed specifically for each category.

Knowledge	Training	MS-COCO		VOC2007		NUS-WIDE	
		F1	mAP	F1	mAP	F1	mAP
Hand-crafted prompts†	×	–	49.7	–	77.3	–	37.4
Hand-crafted prompts	×	45.4	67.7	52.1	88.4	30.4	43.3
Image Captions	✓	46.4	71.4	46.6	84.7	**43.4**	44.1
Coarse Attributes	✓	48.2	68.3	52.3	89.5	36.6	46.0
Fine-grained Attributes	✓	48.8	68.9	52.4	89.7	36.8	46.1
Ours	✓	**56.4**	**72.1**	**58.5**	**90.4**	40.0	**47.0**

sets of scores are then integrated via a weighted sum, with the parameter λ_2 in Eq. (16) set to 0.65, to produce the final prediction.

Training Hyperparameters. Similar to DualCoOp [5] and TaI-DPT [6], We train the model for 10 epochs using a learning rate of 0.002. A learning rate decay of 0.1 is applied at the 2nd and 5th epochs. The balancing parameter λ_1 in our training loss function Eq. (14) is set to 0.2.

4.3. Ablation studies

Effectiveness of different text descriptions. To evaluate the acquired text descriptions from ChatGLM by our method, we employ different inquiry strategies to obtain different text descriptions for comparison, including: **(1)** "*Hand-crafted prompts*": the inference is directly performed using hand-crafted prompts without prompt tuning; **(2)** "*Image Captions*": the human-written image captions from existing datasets are used for prompt tuning; **(3)** "*Coarse Attributes*": the text descriptions of object attributes with noise are used for prompt tuning, generated by Eq. (1); **(4)** "*Fine-grained Attributes*": the text descriptions of filtered object attributes are used for prompt tuning, generated by Eq. (3).

Table 1 shows the results of different text descriptions on the MS-COCO, VOC2007, and NUS-WIDE datasets. We have the following observations: **(1)** Our method achieves substantial improvements over "Hand-crafted prompts", with F1-score increases of 11.0, 6.4, and 9.6 percentage points on the three datasets, respectively. This clearly demonstrates the significant contribution of knowledge extracted from ChatGLM in enhancing the zero-shot performance of CLIP for multi-label image recognition. **(2)** Our method yields absolute gains in F1-score of 7.6, 6.1, and 3.2 percentage points over the "Fine-grained Attributes" descriptions on the three datasets, respectively. This superiority emphasizes that considering the relationships between objects captures more discriminative information to enhance multi-label recognition; **(3)** Our method performs better than "Image Captions" on most metrics, suggesting that ChatGLM provides more comprehensive knowledge than human-written image captions; We observe that the "Image Captions" baseline achieves a higher F1 score as its conservative label predictions lead to increased precision at the evaluation threshold. Conversely, our LLM-based semantic attribute generation expands semantic coverage, thereby enhancing recall while causing a slight decrease in precision; **(4)** The performance of "Coarse Attributes" is lower than that of "Fine-grained Attributes", confirming that the presence of noise within the coarse attributes of objects degrades the performance.

Effectiveness of different prompts. To evaluate the proposed hierarchical prompts, we employ different types of prompts for comparison, including: **(1)** "*Hand-crafted prompts*": these prompts are initialized using a basic template, *"The image contains [category]"*. They then undergo an iterative, manual refinement process. Specifically, for categories that perform below a set Average Precision (AP) threshold 0.5, we augment the prompts with fine-grained descriptions based on the objects' distinct visual features. This tuning is performed iteratively for up to 10 rounds, and the final prompt selected for each category is the one that achieves its peak AP score during this process. **(2)** "*Category-specific prompts*": the learnable tokens of prompts are specific for each category; **(3)** "*Shared prompts*": the learnable tokens of prompts are shared across all categories.

Table 2
Results of different prompts on MS-COCO.

Prompts	F1	mAP
Hand-crafted prompts	45.4	67.7
Category-specific prompts	53.9	71.1
Shared prompts	54.4	71.8
Hierarchical prompts (Ours)	**56.4**	**72.1**

Table 3
Results of different components on MS-COCO.

Local learning	Order loss	F1	mAP
×	×	56.2	69.8
✓	×	55.1	71.7
×	✓	**56.4**	70.2
✓	✓	**56.4**	**72.1**

Table 4
Results of generating training data using different LLMs on MS-COCO.

LLM	F1	mAP
Llama2	55.3	71.7
ChatGLM	**56.4**	**72.1**

Table 5
Results of different numbers of different token types in hierarchical prompts on MS-COCO. S: shared token; P-S #1: partial-shared token over more categories (within coarse subgroups); P-S #2: partial-shared token over fewer categories (within more fined subgroups); C-S: category-specific token.

S	P-S #1	P-S #2	C-S	F1	mAP
8	8	8	8	55.6	72.0
12	8	6	6	55.4	72.0
16	8	4	4	**56.4**	**72.1**
20	4	4	4	55.7	72.0
20	6	4	2	54.7	71.9

Table 2 shows the results of different types of prompts on MS-COCO. Our hierarchical prompts outperform all other methods, demonstrating the effectiveness of incorporating inter-category relationships into prompts. Moreover, compared to hand-crafted prompts, learnable prompts (i.e., category-specific, shared, and hierarchical prompts) achieve much better results. Interestingly, "Shared prompts" outperforms "Category-specific prompts", indicating the better generalization ability of shared prompts.

Effectiveness of order loss. To evaluate the effectiveness of the order loss in Eq. (13), we remove it for comparison. The results on MS-COCO are shown in Table 3, verifying the advantage of the order loss in mitigating the noise of text descriptions acquired from ChatGLM.

Effectiveness of local learning. To evaluate the effectiveness of the local learning of hierarchical prompts, we remove it for comparison. The results on MS-COCO are shown in Table 3, highlighting the significant impact of focusing on image sub-regions in multi-label recognition.

Effectiveness of different LLMs. To evaluate the effectiveness of the data generation using different LLMs, we replace ChatGLM with Llama2 [43], a 7B version with similar parameters of ChatGLM, and the results on the MS-COCO dataset are shown in Table 4. We observe similar performance when using the different LLMs, demonstrating the generalities of the proposed knowledge acquirement approach.

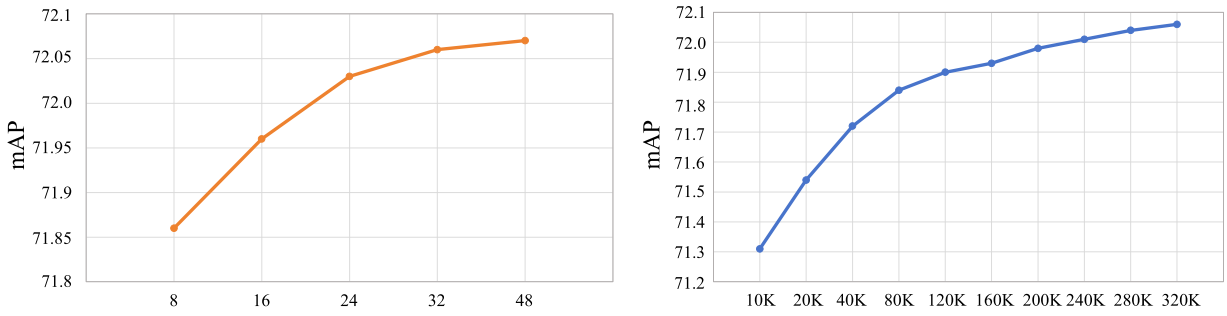

Fig. 5. (a) Results of the different number of tokens on MS-COCO. (b) Results of different numbers of generated training text descriptions using ChatGLM on MS-COCO.

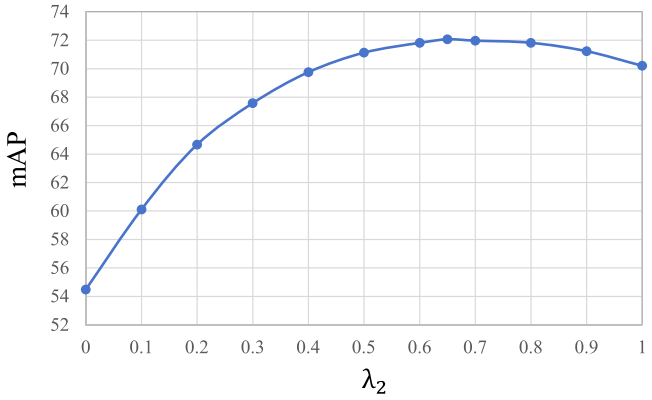

Fig. 6. Analysis of the effect of loss weights between global and local prompts, *i.e.* λ_2 of Eq. (16), on MS-COCO.

4.4. Parameter analysis

Number of different types of tokens. The performances of different numbers of different token types in hierarchical prompts on MS-COCO are shown in Fig. 5 (a). We observe that the performance increases as the token number increases, but a number larger than 32 brings a negative impact. Moreover, we also analyze the composition of different types of tokens in Table 5. Note that the partial-shared tokens are split into two parts: "P-S #1" are coarse parts that tokens are shared over more categories than those of "P-S #2". We observe similar mAP scores for different configurations, indicating that our method is robust to the numbers of different token types.

Number of generated training text descriptions. To analyze the impact of the amount of generated training text descriptions, we conducted experiments with varied numbers of generated training text descriptions, and the results are presented in Fig. 5 (b). We observe that when the number of text descriptions increases, the performance first increases and then gradually decreases, which demonstrates that more training text descriptions can describe more scenes, thus improving the performance of image recognition, but also may bring redundant, noisy information that misleads the prompt learning.

Weights of global and local prompts. To evaluate the contributions of the global and local prompts to the final performance, we conducted experiments with varied weights assigned to each branch, as depicted in Fig. 6. We observe that as the weight allocated to the global prompts increases, the performance initially rises, peaking at a weight of 0.65, then gradually declines. This trend demonstrates that the global branch plays a more critical role, but the local branch is also necessary.

4.5. Comparison with state-of-the-art methods

To provide a comprehensive understanding of how our image-free framework compares with existing approaches, we review representative state-of-the-art methods across different annotation regimes and briefly outline their core innovations. Under the *fully labeled* setting, **SRN** [44] learns spatial regularization through attention maps to capture object locations more precisely, **ML-GCN** [13] utilizes a graph convolutional network to explicitly model label co-occurrence dependencies, and **ASL** [18]) designs an asymmetric loss function that mitigates label imbalance and enhances hard-positive learning. For *partially labeled* methods, **SARB** [22] fuses semantic-aware representations from annotated and unannotated categories, **SST** [23] propagates structured semantic relationships via graph-based transfer, **DualCoOp** [5] proposes dual context prompts to provide positive and negative cues for CLIP adaptation with limited labels, and **DualCoOp++** [31] extends this idea with evidential prompts and dual adapters to strengthen feature aggregation. In the more sparse *one-labeled* scenario, **LL-R** [45] introduces a large-loss mechanism to emphasize rare positive labels, while **G^{2}NetPL** [24] formulates the learning process as a cooperative game for robust label inference under extreme sparsity. Among *unlabeled image* methods, **LSAN** [46] employs spatial attention within the Inception framework for weakly supervised learning, **WAN** [47] leverages geographical priors and presence-only cues for scene-level reasoning, **Curriculum** [48] applies curriculum learning to progressively refine predictions from easy to hard samples, and **Naive AN** [49] extracts weakly supervised visual patterns to exploit partial annotations. Recently, text-driven approaches such as **TaI-DPT** [6] and **TaI++** [7] treat human-written image captions as surrogates for visual data, with **TaI++(pseudo)** further augmenting training by generating pseudo-captions via an LLM. In contrast, our *image-free* method relies solely on LLM-generated descriptions of object attributes and inter-category relationships. These structured textual representations are used to perform hierarchical prompt tuning on CLIP, enabling multi-label image recognition without requiring any annotated images or captions.

Table 6 shows the comparison results on VOC2007, MS-COCO, and NUS-WIDE. We have observations as follows: **(1)** Our method outperforms all the unsupervised methods using unlabeled training images, underscoring the superiority of comprehensive knowledge of objects stored in ChatGLM; **(2)** Our method exhibits a slight performance advantage over TaI++(label), which is trained on human-written image captions. These results suggest that ChatGLM has the ability to emulate human understanding, further validating the effectiveness of our method; **(3)** Our method outperforms the TaI++(pseudo), which also uses the LLM to generate training texts. These results show the superiority of our Knowledge Acquisition strategy and the tailored Hierarchical Prompt Learning approach; **(4)** The performance of our method significantly drops compared to the fully labeled methods, probably due

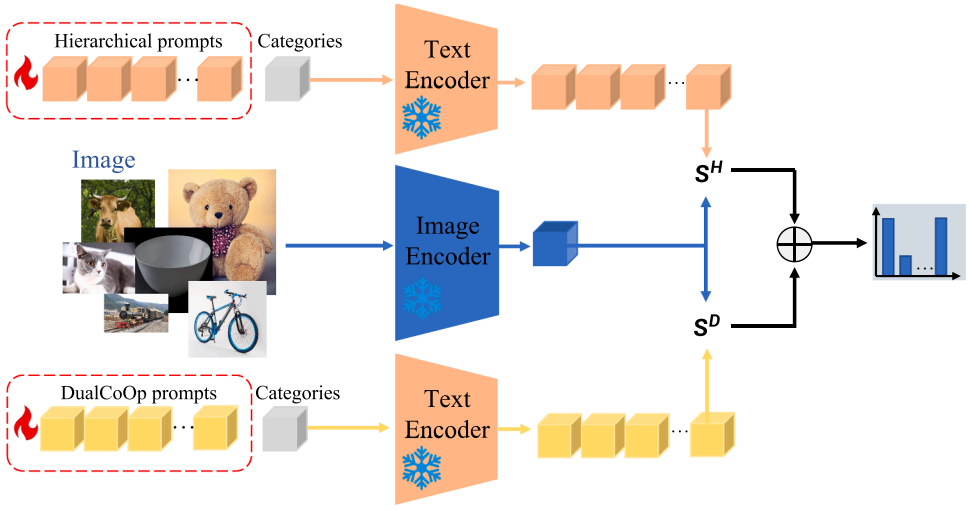

Fig. 7. An illustration of integrating our hierarchical prompts with DualCoOp [5] prompts learned from images. S^D and S^H are scores calculated between the corresponding prompts and the input image.

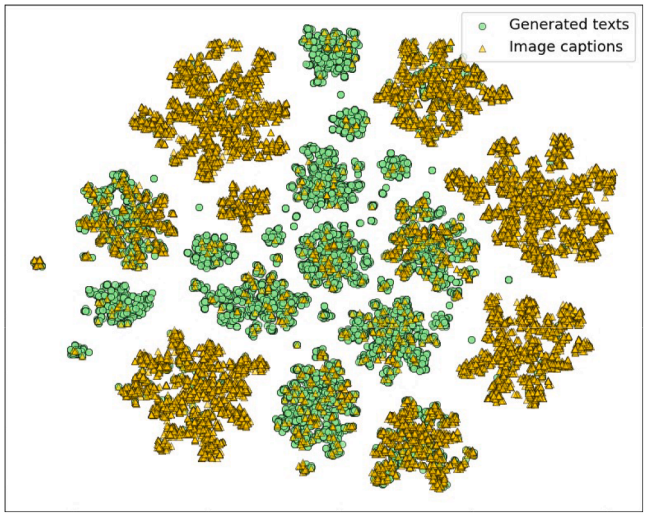

Fig. 8. Feature distributions of LLM-generated text descriptions and the human-written image captions using t-SNE [50].

Table 6
Comparison results (mAP) with the existing methods on MS-COCO, VOC2007, and NUS-WIDE.

Method	Annotation	MS-COCO	VOC2007	NUS-WIDE
SRN [44]	Fully labeled Image	77.1	–	62.0
ML-GCN [13]		83.0	94.0	–
ASL [18]		**86.6**	**94.6**	**65.2**
SARB [22]	Partially labeled Image	71.2	83.5	–
SST [23]		68.1	81.5	–
DualCoOp [5]		78.7	90.3	–
DualCoOp++ [31]	One labeled Image	**81.4**	**92.7**	–
LL-R [45]		**72.6**	**90.6**	47.4
G^{2}NetPL [24]		72.5	89.9	**48.5**
LSAN [46]	Unlabeled Image	**65.5**	**87.9**	**41.3**
WAN [47]		63.9	86.2	40.1
Curriculum [48]		63.2	83.1	39.4
Naive AN [49]		65.1	86.5	40.8
TaI-DPT [6]	Unlabeled Caption	65.1	88.3	46.5
TaI++(label) [7]	Unlabeled Caption	70.9	89.7	44.3
TaI++(pseudo) [7]	LLM	70.8	90.0	46.0
Ours	LLM	**72.1**	**90.4**	**47.0**

Table 7
Statistical results on MS-COCO. Results are averaged over five independent runs for our model. p-values are obtained using the Wilcoxon signed-rank test.

Method	F1 (%)	mAP (%)
Hand-crafted prompts (Baseline)	45.41	67.70
Hierarchical prompts (Ours)	**56.37** ± **0.28**	**72.07** ± **0.03**
p-value (Wilcoxon)	0.0313	0.0313

to the domain gap between the training data of CLIP and the target data of the specific task.

4.6. Statistical analysis

To verify the robustness of our improvements, we conducted statistical significance testing on the MS-COCO dataset. Each result of our model was averaged over five independent runs with different random seeds. The Wilcoxon signed-rank test was used to compare our hierarchical prompts with the hand-crafted prompts. As shown in Table 7, both F1 and mAP improvements are statistically significant ($p < .05$), indicating that the performance gain of our method is consistent and not due to random variation.

4.7. Incorporating with prompt tuning from annotated images

While our prompt tuning relies on text descriptions generated by the LLM, we extend its capability by incorporating prompt tuning from annotated images, thereby enhancing multi-label image recognition. Similar to TaI-DPT [6], we adopt a late fusion strategy that simply combines the scores of prompts learned from images with those from our hierarchical prompts learned from LLM-generated text descriptions. This integration is illustrated in Fig. 7. By leveraging the late fusion strategy, our hierarchical prompts can seamlessly integrate with any existing methods that learn prompts from annotated images, offering enhanced versatility and performance.

To evaluate the effectiveness of incorporating our method with prompt tuning from annotated images, we follow the experiment configurations in previous work [5]. Specifically, we integrate our hierarchical prompt tuning method with DualCoOp [5], which utilizes partially labeled training images (only some labels are known). To achieve this, we reproduce DualCoOp [5] on the MS-COCO dataset, where the partial-label data is generated by randomly masking out labels of the fully annotated data during training.

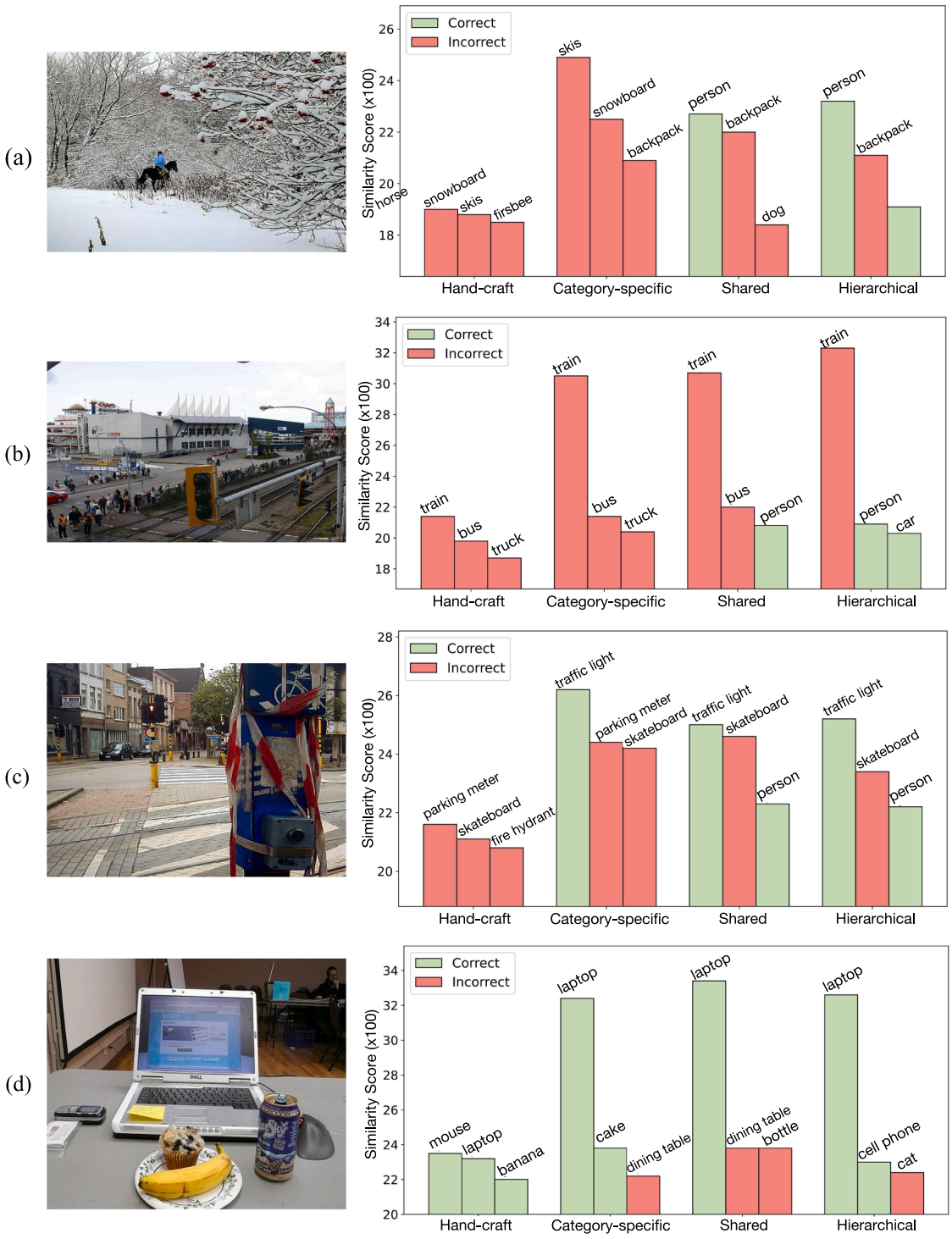

Fig. 9. Visualization of top-3 predicted categories by different prompts.

From the results in Table 8, consistent improvements can be achieved when integrating our hierarchical prompt tuning method with DualCoOp [5], which further demonstrates the effectiveness of our method and its potential applicability within existing frameworks.

4.8. Qualitative analysis

Features Analysis. In Fig. 8, we employ t-distributed Stochastic Neighbor Embedding (t-SNE) [50] to visualize the feature distributions of both LLM-generated text descriptions and human-written image captions, which reveals a significant overlap between the two types of text descriptions, suggesting a deep, intrinsic similarity in how both descriptions characterize images. Such overlap explains why prompt tuning with LLM-generated text can be effective, as these machine-generated descriptions share core characteristics with their human-written counterparts. Moreover, the LLM-generated text descriptions occupy a broader range of the feature space compared to the image captions. This indicates that our LLM-generated descriptions are not only

Table 8
Results of integrating our hierarchical prompts with DualCoOp [5] (demoted as ◇) on partial-label MS-COO dataset, the ∗ means our reproduction.

Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
◇ [5]	78.7	80.9	81.7	82.0	82.5	82.7	82.8	83.0	83.1	81.9
◇* [5]	79.1	80.9	81.5	**82.1**	82.4	82.7	82.9	83.1	83.4	82.0
◇ + Ours	**79.9**	**81.5**	**82.1**	**82.1**	**82.9**	**83.2**	**83.3**	**83.5**	**83.8**	**82.5**

aligned with human perspectives but also extend beyond them. This broader coverage may enhance prompt tuning by providing more comprehensive descriptions of the visual content.

Cases Analysis. Fig. 9 illustrates the top-3 category predictions of different prompts. Notably, the hierarchical prompts achieve better performance, especially on smaller objects, as shown in (a), (b), and (c). However, as depicted in (b), an instance of incorrect top-1 prediction arises in the category labeled "train", likely due to an excessive emphasis on global image features, resembling a train station. Conversely, as shown in (d), the hand-crafted prompts demonstrate superior performance, probably due to the meticulous design of hand-crafted prompts integrating certain human prior knowledge. For instance, when designing the prompt for the "mouse" category, we use the term "computer mouse", aligning more closely with its contextual usage to improve the performance.

5. Conclusion

We have presented an image-free framework for multi-label image recognition, which leverages enriched text descriptions powered by Large Language Models (LLMs), such as ChatGLM, to effectively adapt Vision–Language Models (VLMs) like CLIP through hierarchical prompt tuning. By first querying ChatGLM with well-designed questions and then learning hierarchical prompts with contextual relationships between categories, our method achieves competitive performance without requiring any annotated images or image captions. Specifically, it outperforms previous text-only baselines by up to **2.0%** in mAP on MS-COCO and **1.0%** in mAP on VOC2007, validating the effectiveness of combining LLM-generated linguistic knowledge and CLIP's cross-modal alignment.

Strengths. The proposed framework is (1) *data-efficient*, requiring no manually annotated images or captions; (2) *generalizable*, as evidenced by consistent performance across datasets and with different LLMs (e.g., Llama2); and (3) *integrative*, enabling seamless combination with existing prompt-tuning frameworks such as DualCoOp to further boost performance. The results on three public benchmarks demonstrate that the synergy between LLMs and VLMs offers a scalable and low-cost direction for visual understanding under data-scarce conditions.

Limitations. Although the proposed framework demonstrates strong performance, several limitations remain. First, the text descriptions generated by LLMs may contain noise or hallucinated details, especially for ambiguous or fine-grained attributes, which can slightly impact prompt learning. Second, there may be domain shifts between LLM's general-world knowledge and the specific visual distributions of the target classes, potentially reducing recognition accuracy in specialized scenarios. Third, while hierarchical prompt tuning improves results, the reasons why certain prompt structures work better are not fully understood, leaving the interpretability of prompt effectiveness underexplored. Finally, the approach relies on the assumption that CLIP's visual–text alignment generalizes well to new domains; significant misalignment could degrade overall performance.

Potential Applications. This framework can serve as a universal solution for visual recognition in low-resource domains such as *medical imaging diagnosis*, *remote sensing analysis*, and *industrial inspection*, where labeled visual data are difficult or expensive to obtain but textual knowledge is abundant. Moreover, its image-free nature provides a new paradigm for integrating future multimodal large models into cross-modal reasoning systems without additional data collection or fine-tuning.

CRediT authorship contribution statement

Shuo Yang: Conceptualization, Methodology, Writing – original draft, Writing – review & editing; **Zirui Shang:** Project administration, Methodology, Data curation; **Yongqi Wang:** Methodology, Investigation, Data curation; **Derong Deng:** Data curation; **Hongwei Chen:** Data curation; **Xinxiao Wu:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization; **Qiyuan Cheng:** Validation, Resources.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the Shenzhen Science and Technology Program under Grant No. JCYJ20241202130548062, the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006, and the Key Research Platforms and Projects of the Guangdong Provincial Department of Education under Grant No. 2023ZDZX1034.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.patcog.2025.112986.

References

[1] A. Ben-Cohen, N. Zamir, E. Ben-Baruch, I. Friedman, L. Zelnik-Manor, Semantic diversity learning for zero-shot multi-label classification, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 640–650.
[2] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[3] Z. Guo, R. Zhang, L. Qiu, X. Ma, X. Miao, X. He, B. Cui, CALIP: zero-shot enhancement of clip with parameter-free attention, in: AAAI Conference on Artificial Intelligence (AAAI), vol. 37, 2023, pp. 746–754.
[4] R. Abdelfattah, Q. Guo, X. Li, X. Wang, S. Wang, CDUL: CLIP-driven unsupervised learning for multi-label image classification, in: IEEE International Conference on Computer Vision (ICCV), 2023, pp. 1348–1357.
[5] X. Sun, P. Hu, K. Saenko, DualCoOp: fast adaptation to multi-label recognition with limited annotations, Advances in Neural Information Processing Systems (NeurIPS) vol. 35 (2022) 30569–30582.
[6] Z. Guo, B. Dong, Z. Ji, J. Bai, Y. Guo, W. Zuo, Texts as images in prompt tuning for multi-label image recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2808–2817.
[7] X. Wu, Q. Jiang, Y. Yang, Y. Wu, Q. Chen, J. Lu, TAI++ Text as image for multi-label image classification by co-learning transferable prompt, in: International Joint Conference on Artificial Intelligence (IJCAI), 2024, pp. 5226–5234.
[8] S. Yang, Z. Shang, Y. Wang, D. Deng, H. Chen, Q. Cheng, X. Wu, Data-free multi-label image recognition via LLM-powered prompt tuning, (2024). arXiv preprint arXiv:2403.01209.
[9] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, S. Yan, HCP: a flexible CNN framework for multi-label image classification, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 38 (9) (2015) 1901–1907.
[10] W. Liu, I. Tsang, On the optimality of classifier chain for multi-label classification, Advances in Neural Information Processing Systems (NeurIPS) vol. 28 (2015).
[11] I. Misra, C. Lawrence Zitnick, M. Mitchell, R. Girshick, Seeing through the human reporting bias: visual classifiers from noisy human-centric labels, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2930–2939.
[12] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: a real-world web image database from National University of Singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval, 2009, pp. 1–9.

[13] Z.-M. Chen, X.-S. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5177–5186.
[14] J. Zhang, Q. Wu, C. Shen, J. Zhang, J. Lu, Multilabel image classification with regional latent semantic dependencies, IEEE Trans. Multimed. (IEEE TMM) 20 (10) (2018) 2801–2813.
[15] V.O. Yazici, A. Gonzalez-Garcia, A. Ramisa, B. Twardowski, W.J. van de, Orderless recurrent models for multi-label classification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13440–13449.
[16] B.-B. Gao, H.-Y. Zhou, Learning to discover multi-class attentional regions for multi-label image recognition, IEEE Trans. Image Process. 30 (2021) 5920–5932.
[17] J. Ye, J. He, X. Peng, W. Wu, Y. Qiao, Attention-driven dynamic graph convolutional network for multi-label image recognition, in: European Conference on Computer Vision (ECCV), Springer, 2020, pp. 649–665.
[18] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 82–91.
[19] D. Huynh, E. Elhamifar, A shared multi-attention framework for multi-label zero-shot learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8776–8786.
[20] Y. Liu, J. Wen, C. Liu, X. Fang, Z. Li, Y. Xu, Z. Zhang, Language-driven cross-modal classifier for zero-shot multi-label image recognition, in: Forty-first International Conference on Machine Learning, 2024, pp. 1–11.
[21] R. Abdelfattah, X. Zhang, Z. Wu, X. Wu, X. Wang, S. Wang, PLMCL: partial-label momentum curriculum learning for multi-label image classification, in: European Conference on Computer Vision (ECCV), Springer, 2022, pp. 39–55.
[22] T. Pu, T. Chen, H. Wu, L. Lin, Semantic-aware representation blending for multi-label image recognition with partial labels, in: AAAI Conference on Artificial Intelligence (AAAI), vol. 36, 2022, pp. 2091–2098.
[23] T. Chen, T. Pu, H. Wu, Y. Xie, L. Lin, Structured semantic transfer for multi-label recognition with partial labels, in: AAAI Conference on Artificial Intelligence (AAAI), vol. 36, 2022, pp. 339–346.
[24] R. Abdelfattah, X. Zhang, M.M. Fouda, X. Wang, S. Wang, G2NetPL: Generic game-theoretic network for partial-label image classification, British Machine Vision Conference (BMVC) (2022).
[25] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, S.-N. Lim, Visual prompt tuning, in: European Conference on Computer Vision (ECCV), Springer, 2022, pp. 709–727.
[26] B. Zhu, Y. Niu, Y. Han, Y. Wu, H. Zhang, Prompt-aligned gradient for prompt tuning, in: IEEE International Conference on Computer Vision (ICCV), 2023, pp. 15659–15669.
[27] M. Singha, H. Pal, A. Jha, B. Banerjee, AD-CLIP: adapting domains in prompt space using CLIP, in: IEEE International Conference on Computer Vision (ICCV), 2023, pp. 4355–4364.
[28] K. Sohn, H. Chang, J. Lezama, L. Polania, H. Zhang, Y. Hao, I. Essa, L. Jiang, Visual prompt tuning for generative transfer learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19840–19851.
[29] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, H. Li, Tip-adapter: training-free adaption of clip for few-shot classification, in: European Conference on Computer Vision (ECCV), Springer, 2022, pp. 493–510.
[30] U. Upadhyay, S. Karthik, M. Mancini, Z. Akata, ProbVLM: probabilistic adapter for frozen vison-Language models, in: IEEE International Conference on Computer Vision (ICCV), 2023, pp. 1899–1910.
[31] P. Hu, X. Sun, S. Sclaroff, K. Saenko, DualCoOp++: fast and effective adaptation to multi-label recognition with limited annotations, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 46 (5) (2023) 3450–3462.
[32] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E.H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, Trans. Mach. Learn. Res. (2022).
[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems (NeurIPS) vol. 33 (2020) 1877–1901.
[34] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, L. Wang, An empirical study of GPT3 for few-shot knowledge-based VQA, in: AAAI Conference on Artificial Intelligence (AAAI), vol. 36, 2022, pp. 3081–3089.
[35] T. Gupta, A. Kembhavi, Visual programming: compositional visual reasoning without training, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14953–14962.
[36] J. Yang, X. Chen, S. Qian, N. Madaan, et al., LLM-grounder: open-vocabulary 3D visual grounding with large language model as an agent, in: International Conference on Robotics and Automation (ICRA), 2024, pp. 7694–7701.
[37] Z. Ren, Y. Su, X. Liu, ChatGPT-powered hierarchical comparisons for image classification, in: Advances in Neural Information Processing Systems (NeurIPS), 2023.
[38] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, J. Tang, GLM: general language model pretraining with autoregressive blank infilling, in: ACL, 2022, pp. 320–335.
[39] Y. Wang, D. He, F. Li, X. Long, Z. Zhou, J. Ma, S. Wen, Multi-label classification with label graph superimposing, in: AAAI Conference on Artificial Intelligence (AAAI), vol. 34, 2020, pp. 12265–12272.
[40] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vision (IJCV) 130 (9) (2022) 2337–2348.
[41] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: European Conference on Computer Vision (ECCV), Springer, 2014, pp. 740–755.
[42] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vision (IJCV) 88 (2010) 303–338.
[43] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: open foundation and fine-tuned chat models, (2023). arXiv preprint arXiv:2307.09288.
[44] F. Zhu, H. Li, W. Ouyang, N. Yu, X. Wang, Learning spatial regularization with image-level supervisions for multi-label image classification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5513–5522.
[45] Y. Kim, J.M. Kim, Z. Akata, J. Lee, Large loss matters in weakly supervised multi-label classification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14156–14165.
[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.
[47] O. Mac Aodha, E. Cole, P. Perona, Presence-only geographical priors for fine-grained image classification, in: IEEE International Conference on Computer Vision (ICCV), 2019, pp. 9596–9606.
[48] T. Durand, N. Mehrasa, G. Mori, Learning a deep convnet for multi-label classification with partial labels, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 647–657.
[49] K. Kundu, J. Tighe, Exploiting weakly supervised visual patterns to learn from partial annotations, Advances in Neural Information Processing Systems (NeurIPS) vol. 33 (2020) 561–572.
[50] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).