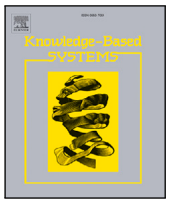
Knowledge-Based Systems 346 (2026) 116115

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Knowledge-injected prompt tuning with semantic regularization for fine-grained image recognition

Qinyu Zhang [a], Xinda Liu [a,*], Qi Liu [a], Pengbo Zhou [b], Guohua Geng [a]

[a] School of Information Science and Technology, Northwest University, Xi'an, 710127, Shaanxi, China
[b] School of Arts and Communication Beijing Normal University, Beijing, 710127, Beijing, China

ARTICLE INFO

Keywords:
Fine-grained image recognition
Prompt tuning
Vision-language models
Few-shot learning
Knowledge-injected

ABSTRACT

Fine-grained image recognition requires models to precisely represent subtle but discriminative local semantic attributes that distinguish visually similar categories. Recent advances in vision-language models demonstrate strong semantic representation capabilities, yet their performance heavily relies on the effectiveness of prompt tuning strategies. However, current prompt tuning methods typically represent each category using a single, globally uniform semantic prompt, which inevitably overlooks the diverse and fine-grained attributes essential for accurately distinguishing visually similar categories. To this end, we propose **Knowledge-Injected Prompt Tuning with Semantic Regularization (KIP)**, a method that injects fine-grained knowledge into prompts to explicitly encode and preserve diverse semantic attributes within the prompt embedding space. Specifically, KIP employs an independent encoding strategy to distinctly represent multiple fine-grained semantic descriptions derived from Large Language Models (LLMs). Moreover, to address the semantic inconsistency caused by hallucinations in LLMs, we introduce a novel semantic regularization mechanism based on Sinkhorn divergence. Extensive experiments across 11 widely used fine-grained image benchmarks demonstrate that KIP significantly enhances performance in cross-dataset transfer, base-to-novel generalization, and few-shot learning tasks. Notably, it achieves substantial improvements in the harmonic mean metric, highlighting its effectiveness and versatility when integrated with existing prompt tuning frameworks. The source code will be released publicly upon acceptance.

1. Introduction

Fine-grained image recognition (FGIR) is a fundamental task in machine learning and computer vision, with broad applications such as rich image captioning [1–5], image generation [6–9], food recognition [10,11], and food recommendation [12].

FGIR is quite challenging since the precise recognition of fine-grained differences requires a large number of annotations from domain experts. Traditional approaches typically rely on large-scale expert annotations to model such fine-grained semantics, making them costly and difficult to scale in real-world scenarios [13]. Nevertheless, FGIR remains practically valuable in many real-world pipelines where subcategory distinctions directly impact downstream retrieval, monitoring, and decision making; moreover, even in the era of large vision–language models, generic pretraining does not always yield reliable fine-grained discrimination, which motivates scalable solutions beyond expert labeling. Recently, vision-language pretraining models such as Contrastive Language-Image Pre-Training (CLIP) have emerged as a promising alternative by leveraging massive image–text pairs to align visual features with natural language supervision, thereby alleviating the dependence on expensive manual labels. Building upon this backbone, prompt tuning [14–17], has gained increasing attention as an efficient strategy to adapt pretrained models to downstream tasks with minimal supervision, offering a practical path to mitigate the annotation bottleneck in FGIR.

Most existing prompt tuning methods are primarily designed for generic tasks, where supervision is typically confined to coarse-grained class labels. Such guidance is insufficient for fine-grained recognition, as it fails to capture the subtle and localized visual variations that distinguish closely related subcategories. A natural direction is to leverage external knowledge from resources such as Wikipedia or LLMs. These sources can provide rich textual descriptions, often covering discriminative visual attributes such as shape, color, and texture. However, directly incorporating such knowledge into prompts introduces two fundamental challenges. First, LLM-derived attributes are diverse and heterogeneous, making them difficult to faithfully represent with a

* Corresponding author.
E-mail address: liuxinda@nwu.edu.cn (X. Liu).

https://doi.org/10.1016/j.knosys.2026.116115
Received 11 December 2025; Received in revised form 4 March 2026; Accepted 29 April 2026
Available online 16 May 2026

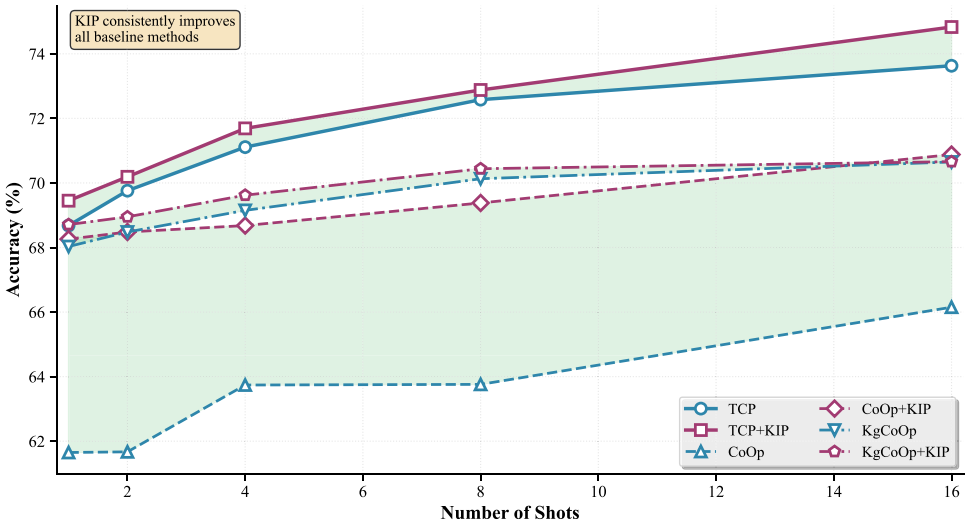


Fig. 1. Performance of base-to-novel generalization across 11 fine-grained datasets with KIP integrated into three baselines at different shot settings. HM: The Harmonic Mean on the base and novel class.

single prompt. Second, LLM-generated descriptions may contain hallucinations and semantic drift, which can misguide prompt optimization and degrade recognition performance. Most prior knowledge-injected prompt tuning methods [18–20] incorporate external knowledge as additional textual inputs or unified anchors, without explicitly handling multi-attribute heterogeneity or mitigating noise in LLM outputs. Consequently, simply injecting more knowledge does not guarantee reliable absorption of fine-grained semantics during training.

The core challenge is therefore not merely to introduce more fine-grained knowledge, but to incorporate it in a structured and robust manner throughout prompt optimization. Without explicit mechanisms, heterogeneous attribute semantics can be diluted, while noisy or hallucinated descriptions may induce semantic drift and harm fine-grained discrimination.

Building on this insight, we propose KIP, a plug-and-play Knowledge-Injected prompt tuning method with semantic regularization for FGIR. KIP is designed for prompt tuning on pretrained vision–language backbones under limited supervision, supporting few-shot learning, base-to-novel generalization, and cross-dataset transfer, while keeping the backbone encoders frozen. KIP formulates knowledge injection as an optimization-time constraint by decoupling multi-attribute encoding and enforcing distribution-level semantic alignment for robust prompt learning. As a plug-and-play module, it integrates seamlessly into existing prompt tuning frameworks and addresses two needs often overlooked in FGIR: independently modeling diverse fine-grained semantics and constraining their influence to prevent semantic drift. In response, we design KIP with two key components. First, we introduce independent multi-attribute encoding that embeds each LLM-derived attribute description separately, preserving semantic specificity and avoiding dilution caused by naive aggregation. Second, we progressively integrate the attribute cues into the prompt learner and apply Sinkhorn-based semantic regularization to align prompt-induced predictions with attribute-induced distributions, thereby mitigating semantic drift under noisy or hallucinated attributes. Together, these components yield semantically grounded and robust prompts, enhancing fine-grained recognition and generalization, as shown in Fig. 1. Compared to large-scale retraining, KIP offers a lightweight alternative for improving fine-grained discrimination via prompt optimization.

In summary, the main contributions of this work include:

- We propose KIP, a plug-and-play framework for knowledge-injected prompt tuning in fine-grained image recognition. KIP decouples multi-attribute encoding and introduces distribution-level semantic alignment as an optimization-time constraint, improving robustness to heterogeneous and noisy attribute descriptions without modifying the backbone architecture.
- We design an **independent encoding strategy** to represent multiple LLM-derived attributes per class as separate semantic vectors. These vectors are then progressively integrated into the prompt space via a stepwise fusion process, enabling precise and independent semantic injection.
- We introduce a **Sinkhorn-based semantic regularization** to align the prompt-induced prediction distribution with attribute-induced distributions. This constraint improves semantic fidelity and mitigates the negative impact of noisy or hallucinated attributes.
- We conduct extensive experiments on 11 widely-used fine-grained image datasets, demonstrating that the KIP framework significantly outperforms state-of-the-art techniques in cross-dataset transfer, base-to-novel generalization, and few-shot learning scenarios.

2. Related works

Prompt tuning was initially introduced in the field of natural language processing (NLP) as a technique to harness the knowledge embedded in large-scale pre-trained language models [21,22]. By designing task-specific prompts, researchers aimed to guide these models toward solving complex downstream tasks without the need for extensive fine-tuning. Inspired by the success of prompt tuning in NLP, the computer vision community began exploring similar strategies to unlock the full potential of vision-language models. Notably, OpenAI's CLIP model [23] pioneered the use of a hand-crafted prompt template—"a photo of [category]"—to generate textual representations for zero-shot image classification. While this approach enabled impressive generalization capabilities, the reliance on manually designed prompts imposed inherent limitations, as they often lacked domain-specific contextual information. The absence of adaptive, task-aware textual representations restricted their ability to capture fine-grained discriminative details, thereby constraining performance in specialized multimodal applications.

To address this, CoOp [14] replaces hand-crafted templates with learnable prompt embeddings, enabling task-specific adaptation. CoCoOp [15] further extends this by conditioning prompts on input images, improving generalization to distributional shifts. Despite their

Table 1
Comparison of representative prompt tuning methods with external knowledge.

Method	Knowledge source	Injection strategy	Constraint
KAPT [18]	Wikipedia descriptions	Knowledge-aware prompts	None
ProText [19]	LLM text templates	Text-only prompt learning	Contextual mapping loss
ATPrompt [20]	LLM attribute pool	Anchored attribute tokens	None
KIP (Ours)	LLM fine-grained attributes	Independent attribute encoding	Sinkhorn OT regularization

success, both approaches operate in a purely data-driven manner and remain agnostic to external knowledge or domain structure.

To enhance prompt robustness, several variants have been proposed. ProGrad [24] constrains prompt updates via gradient guidance to preserve pre-trained semantics. CoPrompt [25] introduces regularization to align prompts under different data distributions. KgCoOp [16] aligns learnable texture embeddings with CLIP features, while LAMM [26] bridges label hierarchies through alignment modules. HisTPT [27] leverages prompt memory during inference for consistency across streams. These works improve prompt stability, but lack explicit modeling of fine-grained semantics.

Recent methods attempt to introduce external semantics to improve prompt expressiveness. KAPT [18] extracts hierarchical attribute trees from Wikipedia to enhance class-level representations. Protext [19] employs LLM-generated attribute groups to supervise prompt composition. ATPrompt [20] incorporates LLM-derived attributes into soft prompts as universal anchors for few-shot generalization. However, these methods primarily focus on knowledge injection, without mechanisms to constrain or regulate semantic consistency during optimization. As a result, hallucinated or redundant concepts may persist, degrading performance in fine-grained settings. Table 1 summarizes representative methods by knowledge source, injection strategy, and regularization mechanism.

Fine-grained image captioning. Fine-grained semantics are equally critical for image captioning, where descriptions must precisely encode subtle attributes and their compositional relations. Attribute Guided Fusion Network strengthens fine-grained captioning through attribute-guided fusion [28]. Hierarchical Region-Context Attention captures fine-grained cues by coordinating region-level evidence with broader contextual signals [29]. ARAFNet further improves caption quality via attribute refinement coupled with attention-based fusion [30]. Although these works mainly optimize decoder-side fusion and attention for generation, they reinforce our premise that fine-grained attribute cues are heterogeneous and should be integrated in a structured manner. In contrast, we focus on prompt optimization for fine-grained recognition by independently encoding LLM-derived attributes and stabilizing their semantics via consistency regularization, offering a complementary route to injecting and consolidating fine-grained knowledge in vision–language models.

In contrast to existing works that primarily inject external semantics into prompts, our approach focuses on the structural organization and regularization of such semantics during optimization. Rather than treating fine-grained knowledge as static textual inputs, we propose a prompt tuning approach that explicitly decouples the encoding, integration, and constraint of fine-grained attributes. This design allows KIP to progressively absorb diverse semantic cues while maintaining internal coherence, offering a controllable pathway to enhance visual-semantic alignment in fine-grained recognition. By addressing both representational diversity and semantic stability, our method complements prior advances and highlights a new direction toward robust and interpretable prompt design in vision-language models.

3. Method

Our goal is to enable vision-language models to distinguish fine-grained categories by effectively injecting external knowledge into the prompt representation space. To achieve this, We propose Knowledge-Injected Prompt Tuning with Semantic Regularization (KIP), a novel prompt tuning method composed of two key components. First, the independent encoding strategy leverages LLMs to generate fine-grained textual descriptions, which are independently encoded and then progressively injected into the prompt space. Then, to guide the stepwise injection process, we introduce a Sinkhorn-based semantic regularization, which aligns the prompt embedding with the distribution of LLM-derived attribute embeddings. The overall architecture is illustrated in Fig. 2. This section begins with a brief review of the CoOp method, followed by a detailed introduction to the proposed KIP framework.

3.1. Preliminaries

Existing vision-language models are predominantly trained using image–text association pairs, which confer strong generalization capabilities for zero-shot image recognition tasks. These models typically include two types of encoders: a visual encoder and a textual encoder. The visual encoder is responsible for mapping a given image into a visual embedding, while the textual encoder extracts the corresponding textual embedding. In the context of prompt tuning, pre-trained visual and textual encoders are kept fixed, and either hand-crafted or learnable prompts are employed to adapt these models to downstream tasks.

KIP follows a similar structural paradigm but introduces significant enhancements by incorporating fine-grained knowledge into the prompt tuning process. This integration allows KIP to effectively leverage category-relevant external knowledge, thereby optimizing the model's performance on fine-grained recognition tasks. Formally, we define the visual encoder and textual encoder as ϕ and θ, respectively. Note that the pretrained visual encoder ϕ and the pretrained textual encoder θ are frozen during training for CoOp. The proposed method follows this same setup to ensure training effectiveness. Given an image I along with its label gt, the visual embedding is extracted with the visual encoder $\phi(\cdot) : \boldsymbol{z} = \phi(I)$.

The completely discrete prompt is relatively sensitive and may affect the generalization performance of downstream tasks. Therefore, CoOp uses a set of continuous context vectors for generating task-related textual embeddings. Specifically, this method introduces m context vectors $\mathbf{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{\mathrm{m}}\}$ as the learnable prompt. The corresponding class token embedding $\mathbf{c}_i$ of the ith class is concatenated with the learnable context vector $\mathbf{V}$ for generating the prompts $\mathbf{V}_i^* = \{\mathbf{c}_i, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{\mathrm{m}}\}$. After that, the textual class embedding $\boldsymbol{y}_i$ is obtained by feeding the learnable prompts $\mathbf{V}_i^*$ into the textual encoder θ, *i.e.*, $\boldsymbol{y}_i = \theta(\mathbf{V}_i^*)$. Therefore, the final textual class embedding for all class is defined as: $\mathbf{Y} = \{\boldsymbol{y}_i\}_{i=1}^{t}$. To optimize the learnable prompts during training, CoOp minimizes the negative log-likelihood between the image features $\boldsymbol{z}$ and the textual class embedding $\mathbf{Y}$:

$$Loss_{ce} = -\sum_{z \in Z} \log \frac{\exp\left(sim\left(\boldsymbol{z}, \boldsymbol{y}_{gt}\right)/\tau\right)}{\sum_{i=1}^{t} \exp\left(sim\left(\boldsymbol{z}, \boldsymbol{y}_i\right)/\tau\right)}, \tag{1}$$

where gt is the corresponding label of the image embedding, $\boldsymbol{Z}$ is the complete set of embeddings, $sim(\cdot,\cdot)$ denotes the cosine similarity, and τ is a learnable temperature parameter. These continuous learnable prompts infer the suitable task-related prompts $\mathbf{Y}$ to boost its generability and discrimination. By introducing fine-grained knowledge embeddings into this process, KIP further enhances the model's capacity to capture detailed, category-specific knowledge.

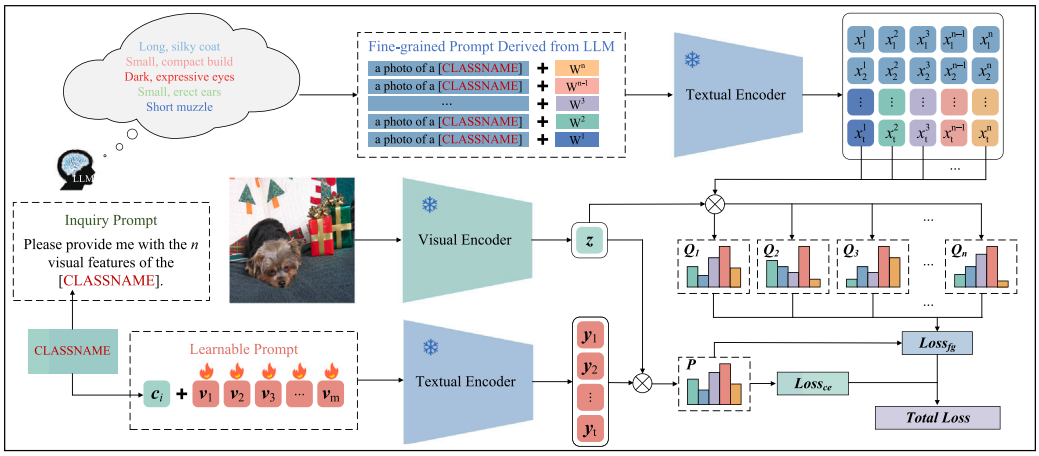


Fig. 2. The framework of the proposed Knowledge-Injected Prompt Tuning (KIP) method. KIP enhances fine-grained recognition by injecting LLM-derived attribute semantics into the prompt space. Class-specific attributes are first extracted via inquiry prompts and encoded independently as semantic vectors. These vectors are then fused into the prompt through a stepwise integration strategy. A Sinkhorn-based regularization aligns the prompt embedding with the attribute distribution, ensuring semantic consistency and suppressing noise.

3.2. Knowledge-injected prompt tuning

The proposed KIP framework consists of two key components: an independent encoding strategy for fine-grained knowledge derived from LLMs, and a Sinkhorn-based Semantic Regularization Loss, as illustrated in Fig. 2. We describe each component in detail below.

3.2.1. Fine-grained knowledge with independent encoding strategy

For a downstream fine-grained image recognition task consisting of t categories, we first propose an approach that integrates CLIP-like hand-crafted prompts with fine-grained prompts to generate the textual class embeddings. Let the name of the ith class be denoted as "CLASSNAME". The CLIP-like hand-crafted prompts are represented as $\mathbf{W}^{CLIP-like} = \{\mathbf{W}_i^*\}_{i=1}^t$, where $\mathbf{W}_i^*$ corresponds to the prompt "a photo of a [CLASSNAME]" for the ith class.

To incorporate fine-grained category-related external knowledge into vision-language pre-training models, fine-grained prompts are a core part of the proposed method. Specifically, we leverage LLMs to generate textual descriptions of fine-grained visual characteristics. By prompting LLMs with queries such as "Please provide me with the n visual features of the [CLASSNAME]", we extract category-specific attributes, including shape, texture, and contextual cues. These features complement CLIP-like prompts by enriching the textual representation with domain-specific knowledge, thereby enhancing discriminative capability in fine-grained recognition tasks. We define the extracted fine-grained visual features as $\mathbf{W}_i^{Fine-grain} = \{\mathbf{W}_i^k\}_{k=1}^n$, where n represents the number of fine-grained features and $\mathbf{W}_i^k$ denotes the kth fine-grained visual feature of the ith class. In practice, we use a fixed prompt template across all classes and request a bounded number of outputs n to obtain concise, visually grounded attribute phrases. This template-based generation provides a controllable knob via n that balances attribute coverage and potential noise. Residual hallucinations or irrelevant attributes may still occur, and we address their impact through the proposed semantic regularization during prompt optimization, as further analyzed in our robustness experiments.

The feature sets generated by LLMs capture a broad range of image attributes, such as color distribution, morphological structures, and local texture details. However, these features are inherently heterogeneous. A naive concatenation of these diverse features with the CLIP-like hand-crafted prompts risks diminishing their discriminative power, as it may overly generalize the representation. This could result in a loss of specificity, as the combined embedding would reflect a global overview of the image rather than its fine-grained visual features, ultimately impairing the performance of the recognition system.

To mitigate this issue, we propose a stepwise process in which each fine-grained description is concatenated individually with its corresponding CLIP-like prompt, rather than concatenating all descriptions at once. Each concatenated pair is then processed separately by the text encoder. Specifically, the textual embedding $\boldsymbol{x}_i^k$ for the kth fine-grained visual feature of the ith class is computed using the text encoder $\theta(\cdot)$, followed by a transformer-based encoder $e(\cdot)$, which processes the sequence of words and outputs the corresponding vectorized textual tokens. Formally, the textual embedding $\boldsymbol{x}_i^k$ is defined as shown in Eq. (2).

$$\boldsymbol{x}_i^k = \theta(e(concate(\mathbf{W}_i^*, \mathbf{W}_i^k))). \tag{2}$$

This design fundamentally mitigates the challenge of feature heterogeneity from two key perspectives. First, processing each fine-grained feature independently through separate text encoder passes ensures the preservation of semantic specificity, preventing the loss of distinctive attribute information. Second, the stepwise encoding strategy mitigates semantic dilution by delaying the fusion of heterogeneous features, thereby enhancing representational integrity and maintaining feature disentanglement.

3.2.2. Semantic regularization loss

The fine-grained textual concepts associated with specific categories, extracted by LLMs, typically exhibit a random distribution across different spatial locations within category images. This discrete spatial distribution characteristic means that simple distribution approximation methods, such as pooling operations, cannot effectively encourage the learnable embedding to capture fine-grained information from different spatial locations of the image. To address this issue, we propose a Semantic Regularization Loss function aimed at minimizing the optimal transport distance between the learnable embedding and each fine-grained knowledge embedding. With this approach, fine-grained knowledge provides precise guidance to the learnable embedding during the context optimization process, enabling the effective integration of multiple sets of fine-grained knowledge within the same embedding space, thereby better reflecting the fine-grained features across different spatial locations within the image.

Specifically, we compute the prediction distribution of the learnable embedding and the prediction distribution of each fine-grained knowledge embedding according to Eqs. (3) and (4).

$$\boldsymbol{P}_i = \frac{\exp\left(sim\left(\boldsymbol{z}, \boldsymbol{y}_i\right)/\tau\right)}{\sum_{i=1}^t \exp\left(sim\left(\boldsymbol{z}, \boldsymbol{y}_i\right)/\tau\right)}, \tag{3}$$

$$\boldsymbol{Q}_i^k = \frac{\exp\left(sim\left(\boldsymbol{z}, \boldsymbol{x}_i^k\right)/\tau\right)}{\sum_{i=1}^{t} \exp\left(sim\left(\boldsymbol{z}, \boldsymbol{x}_i^k\right)/\tau\right)}. \tag{4}$$

Given the logits vectors $\boldsymbol{P}_i$ and $\boldsymbol{Q}_i^k$, we compute the optimal transport distance by first defining the transport plan $\boldsymbol{T} \in \mathbb{R}^{d\times d}$, which specifies how mass is transported between the two embeddings. The cost matrix C, representing the pairwise dissimilarities between $\boldsymbol{P}_i$ and $\boldsymbol{Q}_i^k$, is computed as the Squared Euclidean Distance:

$$C_{ij} = \|\boldsymbol{P}_i - \boldsymbol{Q}_i^k\|_2^2. \tag{5}$$

To prevent the transport plan from being overly sparse and encourage a smoother solution, we apply entropy regularization to the transport plan, defined as:

$$H(\boldsymbol{T}) = -\sum_{ij} T_{ij} \log(T_{ij}). \tag{6}$$

The Sinkhorn divergence is then expressed as:

$$\mathcal{L}_{\text{Sinkhorn}}(\boldsymbol{P}_i, \boldsymbol{Q}_i^k) = \min_{\boldsymbol{T}} \langle C, \boldsymbol{T}\rangle + \lambda \cdot H(\boldsymbol{T}), \tag{7}$$

where $\langle C, \boldsymbol{T}\rangle$ is the inner product between the cost matrix C and the transport plan $\boldsymbol{T}$. Finally, the total fine-grained knowledge fusion loss is computed as the sum of Sinkhorn divergences over all k fine-grained embeddings:

$$\mathcal{L}_{fg} = \sum_{k=1}^{K} \mathcal{L}_{\text{Sinkhorn}}(\boldsymbol{P}_i, \boldsymbol{Q}_i^k). \tag{8}$$

By combining the loss $Loss_{ce}$, the final objective is:

$$Total\,Loss = Loss_{ce} + \lambda Loss_{fg}, \tag{9}$$

where λ is used to balance the effect of $Loss_{fg}$. While $Loss_{fg}$ facilitates the transfer of LLM-derived attribute knowledge to guide prompt optimization, the simultaneous supervision from the cross-entropy loss $Loss_{ce}$ ensures that the learned representations remain anchored in visual-semantic alignment. This interplay establishes a self-reinforcing mechanism in which learnable prompts iteratively integrate domain-specific knowledge while preserving the pre-trained model's fundamental visual-textual correlation capacity. Furthermore, the incorporation of Sinkhorn divergence, leveraging its geometric sensitivity, enables the model to uncover latent correspondences between textual attributes and unlabeled visual patterns in the embedding space. Crucially, by imposing structural consistency in the feature alignment process, Sinkhorn divergence effectively constrains the influence of erroneous or hallucinated attribute knowledge introduced by the LLM, thereby enhancing the robustness and reliability of the learned representations. In particular, we align the prompt-induced and attribute-induced prediction distributions rather than raw embeddings, which makes the constraint less sensitive to heterogeneous attribute phrasing and helps suppress semantic drift during optimization. We adopt Sinkhorn divergence as a geometry-aware, smoothly regularized OT objective that provides stable gradients and soft matching under noisy attribute supervision.

The KIP framework enhances prompt tuning through knowledge anchoring, decoupled alignment, and geometry-aware regularization. LLM-derived attributes serve as semantic priors that constrain optimization toward domain-relevant regions. The optimal transport formulation softly aligns textual attributes with visual embeddings, bridging the modality gap. Together, these mechanisms preserve CLIP's generalization ability while improving fine-grained discrimination.

4. Experiments

To assess the efficacy of KIP, we integrate it into representative prompt-tuning baselines and evaluate it under three widely used FGIR settings: cross-dataset transfer, base-to-novel generalization, and few-shot learning. Since fine-grained benchmarks are often small due to the high cost of expert annotation, these settings reflect realistic low-supervision regimes where lightweight adaptation is most needed.

4.1. Experimental setup

4.1.1. Datasets

To evaluate the effectiveness of the proposed method, we employed 11 fine-grained classification datasets including Dog Breeds [31], Webcar [32], Food-101 [33], FGVC-Aircraft [34], Oxford Flowers [35], Oxford Pets [36], Stanford Cars [37], Stanford Dogs [38,39], Veg200 [40], CUB200 [41] and Fruit92 [40].

4.1.2. Implementation details

To integrate KIP into existing prompt tuning methods, we extended text-based approaches, such as CoOp, KgCoOp, and TCP, by incorporating the plug-and-play KIP method. For consistency across methods, we employed ViT-B/16 as the vision backbone in all experiments, and utilize GPT-4o as our LLM tool for generating categorical features. The original experimental settings for each method were preserved, including the initial learning rate, prompt length, and number of training epochs. Specifically, for KIP, the hyperparameter λ was set to 8.0, and the learning rate was dynamically adjusted using a cosine annealing schedule throughout training. We ensure fair comparison by freezing the vision and text encoders and optimizing only prompt-related parameters for all baselines and their KIP-augmented variants, with identical data splits and a matched training schedule.

4.2. Base-to-novel generalization

In this study, we evaluate the performance of KIP as a plug-and-play module integrated into three baseline prompt tuning methods: CoOp (IJCV 2022), KgCoOp (CVPR 2023), and TCP (CVPR 2024). We further compare our method with several state-of-the-art approaches, including standard prompt tuning baselines such as CLIP (ICML 2021), CoCoOp (CVPR 2023), and ProGrad (ICCV 2023), as well as knowledge-enhanced methods that leverage external information, including KAPT (ICCV 2023), ProText (AAAI 2025), and ATPrompt (ICCV 2025). To assess generalization, we partition each dataset into Base and Novel classes, consistent with the zero-shot learning paradigm, where the Novel classes are entirely disjoint from the Base classes. We employ the Base classes for prompt tuning and evaluate performance on the Novel classes.

Table 2 compares baseline methods with and without the integration of KIP under a 16-shot setting, using ViT-B/16 as the vision backbone. The results show that KIP consistently improves the harmonic mean (HM) performance across all 11 fine-grained datasets when integrated with CoOp-based methods (CoOp, KgCoOp, TCP). Notably, KIP demonstrates superior performance in both Base and Novel class tasks when combined with TCP.

Without KIP, CoOp achieves 3.79% higher accuracy on Base classes, benefiting from a less constrained prompt, which boosts discriminability but risks overfitting. KIP mitigates this, yielding a 9.64% improvement on Novel classes and a higher HM, indicating better balance between recognition accuracy and generalization. While methods like CoCoOp, ProGrad, and KgCoOp reduce overfitting, they still show limited generalization, often underperforming CLIP on Novel classes. In contrast, KIP outperforms CLIP on Novel class generalization across all three backbones, with TCP with KIP achieving the best results: 80.88% accuracy on Base classes and 69.32% on Novel classes.

4.3. Few-shot experiments

To rigorously evaluate the effectiveness of the proposed KIP plugin for enhancing fine-grained recognition tasks, we conducted experiments on several baselines, including CoOp, KgCoOp, and TCP, with and without the integration of KIP. Fig. 3 presents a comparative performance analysis of these methods across 11 fine-grained datasets.

The results demonstrate that the integration of KIP consistently improves both Base and Novel class performance across all datasets.

Table 2
Comparison with SOTA methods on base-to-novel generalization (%) in 16-shots. HM: Harmonic mean.

(a) Average over 11 datasets

Methods	Base	Novel	HM
CLIP [23]	63.75	67.50	64.94
CoCoOp [15]	75.15	66.48	70.11
ProGrad [24]	74.27	58.76	64.88
KAPT [18][a]	–	–	–
ProText [19]	65.14	66.52	65.82
ATPrompt [20]	79.87	60.09	68.58
CoOp [14]	**79.40**	58.34	66.30
+KIP	75.61	**67.98**	**71.07**
KgCoOp [16]	**75.54**	67.32	70.73
+KIP	74.15	**68.93**	**71.11**
TCP [17]	80.8	68.65	73.66
+KIP	**80.88**	**69.32**	**74.07**

(b) Dog breed

Methods	Base	Novel	HM
CLIP [23]	71.80	71.60	71.70
CoCoOp [15]	83.75	74.39	78.79
ProGrad [24]	79.32	63.09	70.28
KAPT [18][a]	–	–	–
ProText [19]	72.55	71.00	71.77
ATPrompt [20]	87.08	67.16	75.83
CoOp [14]	**87.71**	58.89	70.47
+KIP	82.83	**76.40**	**79.49**
KgCoOp [16]	**83.53**	73.29	**78.08**
+KIP	81.06	**74.96**	77.89
TCP [17]	**87.73**	74.88	80.80
+KIP	87.37	**75.30**	**80.89**

(c) Oxford flowers

Methods	Base	Novel	HM
CLIP [23]	72.08	**77.80**	74.83
CoCoOp [15]	94.87	71.75	81.71
ProGrad [24]	95.54	71.87	82.03
KAPT [18][a]	95.00	71.20	81.40
ProText [19]	74.36	78.44	76.35
ATPrompt [20]	**98.02**	69.28	81.18
CoOp [14]	**97.63**	69.55	81.23
+KIP	93.24	**75.18**	**83.24**
KgCoOp [16]	**95.00**	74.73	83.65
+KIP	92.26	**77.31**	**84.13**
TCP [17]	**97.73**	75.57	**85.23**
+KIP	96.72	**75.77**	84.97

(d) Oxford pets

Methods	Base	Novel	HM
CLIP [23]	91.17	97.26	94.12
CoCoOp [15]	95.20	97.69	96.43
ProGrad [24]	95.07	97.63	96.33
KAPT [18][a]	93.13	96.53	94.80
ProText [19]	94.95	**98.00**	96.45
ATPrompt [20]	**95.61**	97.04	96.32
CoOp [14]	94.24	96.66	95.43
+KIP	**95.2**	**97.78**	**96.47**
KgCoOp [16]	94.65	**97.76**	96.18
+KIP	**94.79**	97.72	**96.23**
TCP [17]	94.67	97.20	95.92
+KIP	**95.12**	**97.32**	**96.21**

(e) Stanford cars

Methods	Base	Novel	HM
CLIP [23]	63.37	74.89	68.65
CoCoOp [15]	70.49	73.59	72.01
ProGrad [24]	77.68	68.63	72.88
KAPT [18][a]	69.47	66.20	67.79
ProText [19]	64.54	**76.08**	69.84
ATPrompt [20]	77.27	66.54	71.50
CoOp [14]	**76.2**	69.14	72.49
+KIP	71.63	**74.81**	**73.19**
KgCoOp [16]	**71.76**	75.04	73.36
+KIP	71.04	**75.96**	**73.42**
TCP [17]	**80.80**	74.13	**77.32**
+KIP	79.12	**75.33**	77.18

(f) Fruit92

Methods	Base	Novel	HM
CLIP [23]	53.00	61.10	56.76
CoCoOp [15]	74.58	60.05	66.53
ProGrad [24]	74.82	45.15	56.32
KAPT [18][a]	–	–	–
ProText [19]	54.63	52.70	53.65
ATPrompt [20]	**85.92**	47.21	60.94
CoOp [14]	**85.39**	46.96	60.6
+KIP	78.22	**56.27**	**65.45**
KgCoOp [16]	**76.73**	57.37	65.65
+KIP	73.87	**59.68**	**66.02**
TCP [17]	**85.83**	63.79	**73.19**
+KIP	83.38	**65.15**	73.15

(g) CUB200

Methods	Base	Novel	HM
CLIP [23]	63.90	51.90	57.28
CoCoOp [15]	74.46	51.20	60.68
ProGrad [24]	73.12	38.13	50.12
KAPT [18][a]	–	–	–
ProText [19]	65.07	50.70	56.99
ATPrompt [20]	82.02	40.96	54.64
CoOp [14]	**81.48**	37.37	51.24
+KIP	75.33	**53.94**	**62.87**
KgCoOp [16]	**75.62**	**52.60**	**62.04**
+KIP	73.91	51.98	61.03
TCP [17]	82.10	**52.97**	64.39
+KIP	**82.56**	52.83	**64.43**

(h) Food101

Methods	Base	Novel	HM
CLIP [23]	90.10	91.22	90.66
CoCoOp [15]	90.70	91.70	91.09
ProGrad [24]	90.37	89.59	89.98
KAPT [18][a]	86.13	87.06	86.59
ProText [19]	90.20	**91.98**	91.08
ATPrompt [20]	88.86	88.95	88.90
CoOp [14]	89.44	87.50	88.46
+KIP	**90.68**	**91.83**	**91.25**
KgCoOp [16]	90.50	91.70	91.09
+KIP	**90.52**	**91.72**	**91.17**
TCP [17]	90.57	91.37	90.97
+KIP	**90.72**	**91.88**	**91.30**

(i) Veg200

Methods	Base	Novel	HM
CLIP [23]	44.90	43.80	44.34
CoCoOp [15]	67.19	45.79	54.46
ProGrad [24]	69.92	29.04	41.04
KAPT [18][a]	–	–	–
ProText [19]	44.99	39.20	41.90
ATPrompt [20]	79.85	32.53	46.23
CoOp [14]	**78.19**	30.2	43.57
+KIP	66.94	**44.12**	**53.19**
KgCoOp [16]	**66.85**	44.08	53.13
+KIP	63.39	**48.20**	**54.76**
TCP [17]	83.27	**49.53**	**62.11**
+KIP	**83.52**	49.15	61.88

(j) Web cars

Methods	Base	Novel	HM
CLIP [23]	61.80	73.60	62.45
CoCoOp [15]	67.87	73.55	70.86
ProGrad [24]	52.80	60.38	60.94
KAPT [18][a]	–	–	–
ProText [19]	60.89	72.70	66.27
ATPrompt [20]	67.69	62.27	64.87
CoOp [14]	**68.04**	56.98	65.74
+KIP	67.78	**75.05**	**71.23**
KgCoOp [16]	**68.82**	74.31	**71.46**
+KIP	68.03	**75.19**	71.43
TCP [17]	68.70	72.50	70.55
+KIP	**70.60**	**74.42**	**72.46**

(k) Stanford dogs

Methods	Base	Novel	HM
CLIP [23]	63.75	67.50	64.94
CoCoOp [15]	75.15	66.48	70.11
ProGrad [24]	74.27	58.76	64.88
KAPT [18][a]	–	–	–
ProText [19]	65.14	66.52	65.82
ATPrompt [20]	**79.87**	60.09	68.58
CoOp [14]	**75.80**	58.03	65.74
+KIP	73.58	**67.85**	**70.60**
KgCoOp [16]	71.24	66.12	68.58
+KIP	**71.26**	**69.84**	**70.54**
TCP [17]	75.45	68.79	71.97
+KIP	75.12	69.53	**72.22**

(l) FGVC aircraft

Methods	Base	Novel	HM
CLIP [23]	27.19	36.29	31.09
CoCoOp [15]	33.41	23.71	27.74
ProGrad [24]	40.54	27.57	32.82
KAPT [18][a]	29.67	28.73	29.19
ProText [19]	30.91	34.13	32.44
ATPrompt [20]	39.80	30.74	34.69
CoOp [14]	**39.24**	30.49	34.30
+KIP	36.23	**34.56**	**34.83**
KgCoOp [16]	**36.21**	33.55	34.83
+KIP	35.51	**35.68**	**35.59**
TCP [17]	41.97	34.43	37.83
+KIP	**45.45**	**35.87**	**40.10**

[a] Unavailable public code.

Table 3
Performance comparisons (%) with SOTA methods on fine-grained datasets for cross-dataset transfer task.

Methods	Dog Breeds	Web car	Food 101	FGVC Aircraft	Oxford flower	Oxford Pet	Stanford Cars	Stanford Dogs	Veg 200	CUB 200	Fruit 92	Avg.
CoOp	21.92	24.55	67.85	3.90	29.81	45.17	30.64	23.58	23.97	15.77	33.77	29.18
+KIP	**53.37**	**52.95**	**87.25**	**15.54**	**65.50**	**82.14**	**55.27**	**43.14**	**34.58**	**42.23**	**45.81**	**52.53**
KgCoOp	**54.83**	54.98	82.63	11.30	59.62	78.57	56.83	**49.97**	**35.93**	40.70	45.98	51.94
+KIP	53.92	**59.82**	**83.46**	**19.61**	**68.00**	**79.45**	**58.74**	47.17	35.15	**42.49**	**47.51**	**54.12**
TCP	39.76	49.31	76.24	16.07	57.37	70.79	50.72	37.89	**32.72**	33.56	36.67	45.55
+KIP	**48.87**	**54.53**	**78.68**	**18.29**	**58.24**	**76.80**	**55.18**	**44.14**	32.24	**42.70**	**42.77**	**50.22**

Specifically, the addition of KIP leads to enhanced recognition accuracy for both Base and Novel classes, achieving significant gains over the baseline methods. Furthermore, KIP consistently delivers the highest harmonic mean across all datasets, highlighting its balanced performance across different categories. This consistent trend is observed regardless of dataset size, category granularity, or domain shift, underscoring the method's versatility. These results validate KIP's effectiveness in enhancing the generalization capabilities of baseline methods, demonstrating its robustness across diverse datasets and fine-grained tasks.

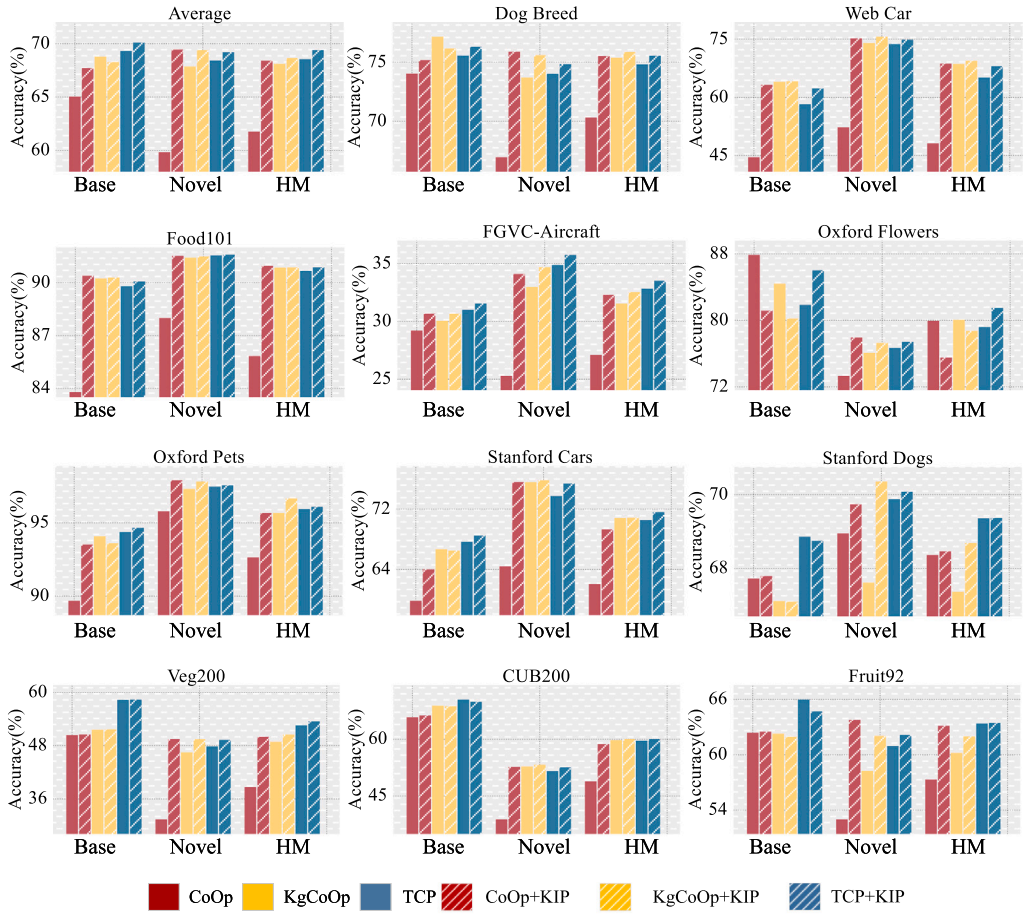


Fig. 3. Performance comparison of different methods in the 1-shot learning setting. Solid bars denote the backbone, while diagonally hatched bars represent the KIP-enhanced variant.

4.4. Cross-dataset transfer

We conduct a comprehensive evaluation of KIP's robustness in distributed offset scenarios, with a particular focus on cross-dataset transfer. Utilizing an ensemble training model that consolidates the training sets of 11 fine-grained datasets, we evaluate the performance of our method across the test sets of these 11 distinct datasets.

Table 3 provides a comprehensive comparison of the performance of baseline approaches including CoOp, KgCoOp, and TCP, both with and without the integration of the proposed KIP framework, evaluated across 11 representative fine-grained image recognition datasets. The empirical results indicate that the incorporation of KIP consistently leads to improved generalization performance across all three baseline methods, highlighting its effectiveness as a complementary enhancement to existing prompt tuning strategies.

Specifically, KIP improves the accuracy of CoOp by an average of 23.35%. Furthermore, significant improvements of 2.18% and 4.67% are observed for KgCoOp and TCP, respectively, which are designed to address generalization issues in prompt tuning methods. Notably, the combination of KgCoOp with KIP yields the best cross-dataset generalization performance.

4.5. Ablation study

4.5.1. Fine-grained feature quantity

Fig. 4 illustrates the effect of Fine-grained Feature Quantity on overall model performance. As the number of fine-grained features increases, performance remains stable initially, peaking at a feature quantity of 8. Beyond this point, however, a gradual performance decline is observed, suggesting a non-linear relationship between feature quantity and model effectiveness.

This trend can be attributed to the design of our independent encoding and stepwise optimization method, which effectively utilizes the specially crafted loss function to enable the learnable embeddings to capture more fine-grained knowledge from local regions of the image, thereby improving performance. However, when the number of features exceeds a certain threshold, the model faces difficulty distinguishing between an excess of text-based feature descriptions that correspond to visually disparate regions. As a result, the model struggles to converge to the optimal solution, leading to a performance decline. Thus, these findings underscore the importance of carefully balancing the quantity of fine-grained features. Beyond the optimal number, the benefits of additional features are outweighed by the complexity they introduce, ultimately hindering model performance.

4.5.2. Hyperparameter λ

To investigate the effect of the hyperparameter λ on the trade-off between the Fine-grained Knowledge Fusion Loss and the overall loss function, we conducted an ablation study using TCP as the backbone while keeping all other hyperparameters fixed. The impact of different values of λ on Base-to-Novel generalization was evaluated under a 16-shot setting, with average results shown in Fig. 5.

As λ increases, performance on the Base classes gradually declines, while performance on the Novel classes improves. This trend suggests that the Fine-grained Knowledge Fusion Loss enhances the model's generalization ability and reduces overfitting. However, as the weight of this loss grows, the model may overly prioritize fine-grained knowledge, potentially leading to the forgetting of generalizable knowledge.

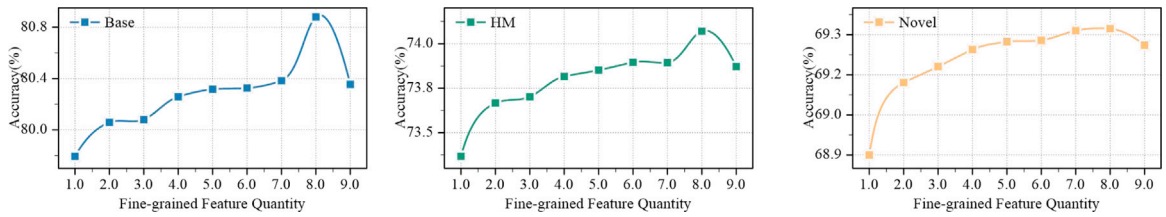


Fig. 4. Effect of the Fine-grained Feature Quantity.

Table 4
Impact of different LLM models on harmonic mean values for two datasets. Blue: Models without Deep Reasoning, Red: Deep Reasoning Models.

	OpenAI GPT-4o	DeepSeek V3	Kimi K1.5 short-CoT	OpenAI o3-mini	DeepSeek R1	Hunyuan T1	Kimi k1.5 long-CoT
Oxford Pets	96.21	96.17	96.03	96.17	**96.23**	96.06	96.25
Fruit92	73.15	**73.98**	73.85	73.24	73.21	73.81	73.12

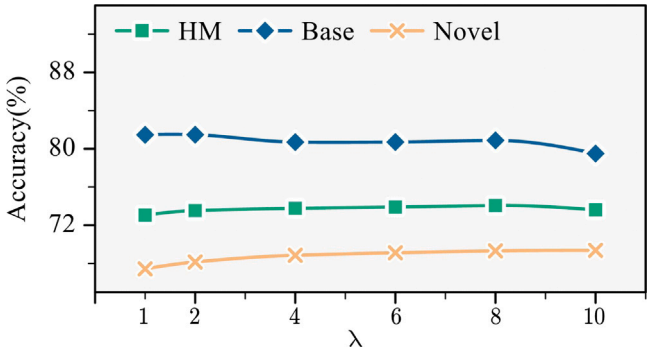


Fig. 5. Effect of λ for 16-shot settings on the base-to-novel generalization. HM: Harmonic mean.

4.5.3. Comparison on LLM integration for hallucination resistance

To assess the efficacy of the independent encoding and stepwise optimization method in mitigating hallucinations in large language model outputs, we conducted experiments on seven different models. These models were categorized into two groups: the standard models and the reasoning models, which are capable of deeper, more complex thought processes. The experiments were performed on the Oxford Pets and Fruit 92 datasets, with harmonic mean results summarized in Table 4.

The results demonstrate that our approach yields robust performance across all tested models, with performance stability observed even in the absence of manually curated clean data. Notably, both reasoning and standard models exhibit comparable performance, highlighting the effectiveness of the independent encoding and stepwise optimization method in minimizing the negative impact of hallucinations on model performance.

4.5.4. Selection of the distance function

Table 5 presents a comparative analysis of Kullback–Leibler (KL) divergence and Sinkhorn divergence across 11 datasets. Both methods guide the learnable distribution to align with frozen fine-grained distributions, facilitating the acquisition of fine-grained knowledge.

The results demonstrate that Sinkhorn divergence consistently outperforms KL divergence in terms of generalization to Novel classes. On average, Sinkhorn achieves a higher accuracy on Novel classes across all 11 datasets compared to KL's. While KL divergence slightly outperforms Sinkhorn on Base class accuracy, Sinkhorn's harmonic mean (HM) of Base and Novel class performance surpasses that of KL, with values of 74.02% and 73.66%, respectively. This indicates that Sinkhorn divergence offers a more balanced performance, effectively improving recognition on Base classes without significantly compromising generalization ability of KIP. These results suggest that Sinkhorn divergence provides superior generalization to unseen categories, making it particularly advantageous in tasks that require robust performance across both Base and Novel classes.

4.5.5. Robustness to noisy attributes

To assess KIP's robustness to noisy LLM-generated attributes, we conducted experiments using Random Replace and Wrong-Class Injection on three datasets: Oxford Flowers, Stanford Cars, and FGVC Aircraft. The results in Table 6 show that Oxford Flowers experiences minimal degradation, while Standford Cars shows moderate degradation. FGVC Aircraft exhibits more significant performance drops under both types of noise.

We hypothesize that these differences arise from CLIP's pretrained representations. Datasets with easily distinguishable features, such as Flower, are less affected by noise, while more complex datasets like Aircraft are more sensitive to attribute noise. These findings demonstrate that KIP remains robust to noise, but performance degradation is more pronounced in domains with subtle distinctions.

4.5.6. Encoding strategies

To assess the effectiveness of KIP's independent encoding strategy, we compared it with two commonly used alternatives: concatenation and averaging. These strategies were evaluated on three baselines: CoOp, KgCoOp, and TCP across 11 fine-grained datasets. The results, shown in Table 7, demonstrate that independent encoding (KIP) consistently outperforms both concatenation and averaging.

These results highlight the advantage of KIP in preserving the distinctiveness of fine-grained attributes. Unlike concatenation and averaging, which tend to dilute attribute semantics, KIP independently encodes each attribute description, ensuring better semantic representation and reducing the risk of semantic overlap.

4.5.7. Backbone generalization

To examine whether the effectiveness of KIP generalizes across different visual backbones, we integrate KIP into CoOp and evaluate under the same 1-shot base-to-novel protocol using two representative backbones, RN50 and ViT-B/32. As reported in Table 8, KIP consistently improves CoOp on both backbones across the 11 fine-grained datasets, demonstrating that the gains are not tied to a specific architecture. In particular, the average HM increases from 57.59 to 60.61 on RN50 and from 60.10 to 63.17 on ViT-B/32, yielding around three points improvement on average. We also observe that improvements are generally reflected on both base and novel splits, indicating that

Table 5
Performance comparisons (%) of different distance functions. KLD:KL divergence.

Avg.	Base	Novel	HM	Dog Breed	Base	Novel	HM	Web Car	Base	Novel	HM
KLD	**80.85**	68.67	73.66	KLD	**87.73**	74.58	80.62	KLD	68.89	71.80	70.31
Sinkhorn	80.73	**69.32**	**74.02**	Sinkhorn	87.37	**75.30**	**80.89**	Sinkhorn	**70.60**	**74.42**	**72.46**
Food 101	Base	Novel	HM	FGVC Aircraft	Base	Novel	HM	Oxford Flowers	Base	Novel	HM
KLD	**90.78**	91.62	91.20	KLD	42.82	34.87	38.44	KLD	**97.98**	75.41	**85.23**
Sinkhorn	90.72	**91.88**	**91.30**	Sinkhorn	**45.45**	**35.87**	**40.10**	Sinkhorn	96.72	**75.77**	84.97
Oxford Pets	Base	Novel	HM	Stanford Cars	Base	Novel	HM	Stanford Dogs	Base	Novel	HM
KLD	94.69	**97.42**	96.04	KLD	**79.8**	74.50	77.06	KLD	74.89	**70.01**	**72.37**
Sinkhorn	**95.12**	97.32	**96.21**	Sinkhorn	79.12	**75.33**	**77.18**	Sinkhorn	**75.12**	69.53	72.22
Veg 200	Base	Novel	HM	CUB 200	Base	Novel	HM	Fruit 92	Base	Novel	HM
KLD	83.33	48.64	61.43	KLD	**82.42**	52.74	**64.32**	KLD	**85.98**	63.78	**73.23**
Sinkhorn	**83.52**	**49.15**	**61.88**	Sinkhorn	80.92	**52.83**	63.93	Sinkhorn	83.38	**65.15**	73.15

Table 6
Sensitivity of datasets to noisy attributes: Maximum degradation due to random replacement and wrong-class injection.

Dataset	Max replace degradation	Max injection degradation
Oxford Flowers	0.25%	0.5%
Stanford Cars	0.5%	0.75%
FGVC Aircraft	3.5%	5.0%

Table 7
Performance comparison of different encoding strategies on 11 datasets. Results are reported as the harmonic mean (HM) of base and novel class performance.

Method	No attribute	Concatenation	Average	KIP
CoOp	66.30	69.55	70.41	**71.07**
KgCoOp	70.73	70.09	70.31	**71.11**
TCP	73.66	73.22	73.50	**74.07**

KIP strengthens fine-grained discrimination while maintaining base-to-novel generalization. Although a few datasets may exhibit minor fluctuations, the overall trend is stable across backbones and datasets, supporting the robustness and generality of KIP.

4.5.8. Prompt length sensitivity

We examine how the prompt length m affects performance by changing the number of learnable prompt tokens in TCP while keeping all other settings fixed. Fig. 6 reports the average HM over 11 fine-grained datasets for $m \in \{1, 2, 4, 8, 16\}$. The results indicate a clear trade-off: performance is stable for short prompts and reaches the best result at $m = 4$, suggesting that a modest prompt capacity is sufficient to capture useful fine-grained cues. However, when the prompt becomes much longer ($m = 16$), the average HM drops sharply to 65.69, which is consistent with harder optimization and increased sensitivity to noise under limited supervision. We therefore set $m = 4$ in the remaining experiments.

4.5.9. Comparison of computational efficiency

We evaluate the computational efficiency of KIP in comparison with the CoOp baseline under an identical training setup. Specifically, we report the wall-clock time per iteration, the total training time per epoch, and the peak GPU memory consumption. The quantitative results are summarized in Table 9. KIP without Sinkhorn introduces negligible computational overhead, requiring 21.28, ms per iteration and 1.10, s per epoch, while maintaining an identical peak GPU memory footprint compared with CoOp. When Sinkhorn alignment is enabled, the time per iteration increases to 25.68, ms and the time per epoch to 1.21, s, whereas the peak GPU memory remains unchanged at 1.073, GB. These results indicate that the additional runtime mainly stems from the Sinkhorn alignment rather than the extra text-encoding operations. Overall, KIP improves performance with only a small increase in training time and no additional memory usage.

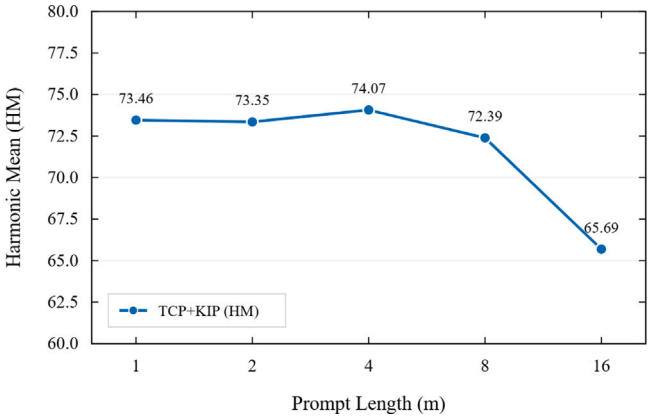


Fig. 6. Effect of prompt length m on TCP with KIP. We report the average HM over 11 fine-grained datasets while varying the number of learnable prompt tokens.

4.6. Visualization study

4.6.1. Grad-CAM analysis

To investigate the impact of fine-grained textual prompts on image feature representations, we computed similarity scores for target images using zero-shot CLIP, CoOp, and CoOp integrated with KIP, a plug-in designed to enrich prompts with fine-grained textual descriptions. Grad-CAM [42] was then applied to backpropagate these similarity scores, generating gradient heatmaps that highlight image regions most sensitive to the target classes. We include zero-shot CLIP as a reference to verify that the visualization pipeline faithfully reflects the saliency patterns of the pretrained model.

During this process, class-specific textual embeddings were used to compute similarity gradients with respect to the visual features. As shown in Fig. 7, zero-shot CLIP exhibits clear activation on salient foreground regions, confirming the reliability of the Grad-CAM procedure. In contrast, under the 1-shot setting, CoOp produces more diffused or shifted activation patterns, suggesting potential overfitting of learnable prompts under extremely limited supervision. When KIP is introduced, the activation maps become more concentrated on semantically relevant foreground regions, indicating that the knowledge-guided regularization helps stabilize prompt learning and alleviate attention drift.

4.6.2. t-SNE visualization

To examine how KIP reshapes the feature space, we visualize test-set visual embeddings with t-SNE on Oxford Pets and Stanford Cars, comparing CoOp with its KIP-enhanced variant. As shown in Fig. 8, without KIP, CoOp produces relatively scattered embeddings with noticeable inter-class overlap, particularly on Stanford Cars where the

Table 8
Backbone generalization on 1-shot base-to-novel evaluation with CoOp. We report Base, New, and HM (%) across 11 fine-grained datasets and the average (AVG).

Backbone	Methods	Dog Breeds	Web Car	Food 101	FGVC Aircraft	Oxford Flower	Oxford Pet	Stanford Cars	Stanford Dogs	Veg 200	CUB 200	Fruit 92	Avg.
RN50	CoOp base	80.27	87.97	72.87	19.73	42.00	47.90	**59.97**	57.20	52.37	65.40	57.80	58.50
	KIP base	**82.80**	**91.03**	**73.67**	**20.80**	**44.27**	**55.33**	58.03	**58.27**	**54.27**	**67.63**	**61.07**	**60.65**
	CoOp new	81.67	94.00	70.17	22.93	32.23	55.50	**60.27**	63.10	49.97	62.80	38.17	57.35
	KIP new	**84.30**	**94.33**	**72.97**	**24.37**	**41.47**	**63.87**	58.67	**66.50**	**55.50**	**65.77**	**43.27**	**61.00**
	CoOp HM	80.96	90.86	71.43	21.19	36.45	51.40	**60.11**	60.00	51.09	64.05	45.97	57.59
	KIP HM	**83.54**	**92.65**	**73.30**	**22.43**	**42.82**	**59.29**	58.34	**62.11**	**54.87**	**66.69**	**50.64**	**60.61**
ViT-B/32	CoOp base	82.03	89.73	**75.27**	**23.77**	45.77	50.30	**61.47**	60.60	58.93	68.47	60.67	61.55
	KIP base	**85.33**	**91.93**	74.43	23.60	**48.13**	**60.23**	60.77	**62.83**	**59.00**	**69.90**	**64.03**	**63.65**
	CoOp new	83.70	**95.20**	67.03	26.90	35.57	56.57	**63.30**	65.67	51.77	65.47	41.37	59.32
	KIP new	**87.17**	94.93	**71.23**	**29.47**	**41.70**	**66.37**	62.60	**69.93**	**55.57**	**68.53**	**47.03**	**63.14**
	CoOp HM	82.85	92.36	70.85	25.19	40.03	53.25	**62.35**	63.03	55.11	66.93	49.18	60.10
	KIP HM	**86.24**	**93.40**	**72.76**	**26.19**	**44.68**	**63.15**	61.66	**66.19**	**57.21**	**69.21**	**54.18**	**63.17**

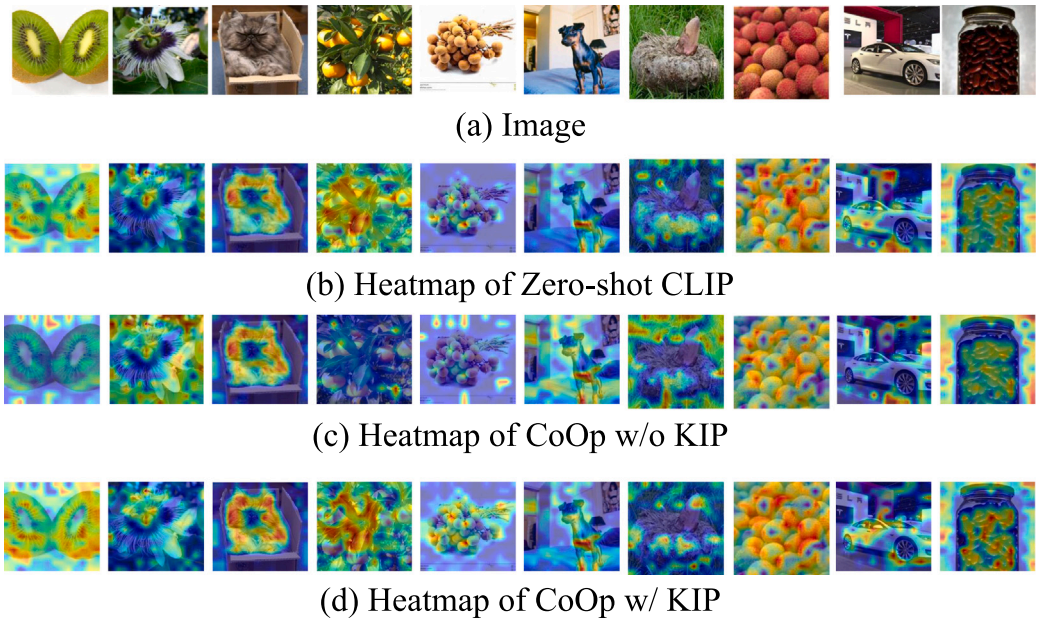


Fig. 7. Grad-CAM visualization of zero-shot CLIP, CoOp, and CoOp with KIP under the 1-shot setting. The comparison highlights differences in foreground localization and attention concentration.

Table 9
Comparison of computational efficiency. We report wall-clock time per iteration, time per epoch, and peak GPU memory under the same training setup.

Method	Time/Iter (ms)	Time/Epoch (s)	Peak GPU Mem (GB)
CoOp	21.11	0.96	1.073
KIP w/o Sinkhorn	21.28	1.10	1.073
KIP w/ Sinkhorn	25.68	1.21	1.073

number of categories is larger and fine-grained differences are subtler. With KIP, the embeddings form tighter intra-class clusters and show clearer inter-class separation on both datasets, with the improvement more evident on Stanford Cars. Overall, these visualizations suggest that KIP encourages a more structured and discriminative embedding geometry in challenging fine-grained settings.

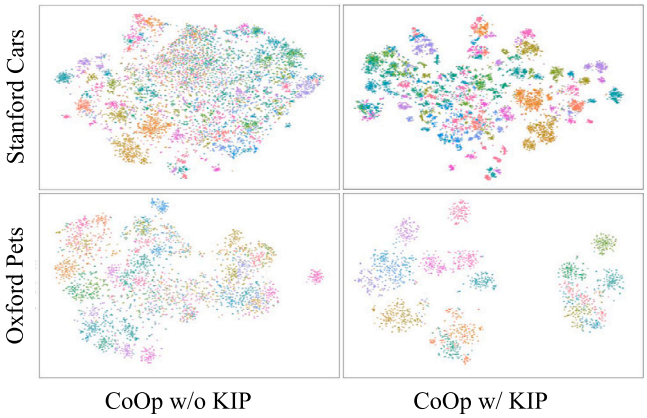


Fig. 8. t-SNE visualizations of visual embeddings on two datasets.

5. Limitation

Despite the promising performance of our model, several limitations warrant attention and further investigation. Firstly, although our model benefits from the incorporation of additional information extracted from LLMs, the fundamental architecture of the CLIP text encoder imposes inherent limitations. Specifically, the encoder is unable to fully capitalize on the extensive information provided by these external sources. While there have been efforts to address this issue [43], the ability of the text encoder to effectively process and utilize the expanded information remains a challenge that needs to be addressed

in future work. Secondly, the model struggles with performance degradation when confronted with an increased number of category features. While the integration of supplementary information intuitively suggests potential improvements in performance, there exists a critical point beyond which the model's effectiveness diminishes. This suggests that further research is needed to develop strategies that can better handle the diminishing returns associated with expanding category features. Lastly, the use of visual prompts has been shown to be a promising strategy for enhancing model performance. However, while initial results are promising, this technique requires further exploration to determine the optimal ways to incorporate visual prompts into the model, and to assess their impact on performance in more complex settings. Moreover, our current study focuses on fine-grained image classification and does not evaluate dense prediction tasks such as object detection or semantic segmentation. Extending KIP to these settings would require adapting knowledge-injected prompts and semantic regularization from image-level to region-level predictions, which we leave for future work. These limitations highlight key areas for future research, which could significantly enhance the overall effectiveness and generalization of the model.

6. Conclusion

The proposed Knowledge-Injected Prompt Tuning (KIP) introduces a novel approach to fine-grained image recognition by systematically integrating extensive textual knowledge into vision-language models. By leveraging LLMs to extract category-specific descriptions and encoding them into a hybrid structure of fixed and learnable continuous prompts, KIP effectively enhances the retention of discriminative fine-grained information. Comprehensive experiments conducted on 11 diverse fine-grained image datasets demonstrate the robustness and generalizability of KIP, particularly in cross-dataset transfer and base-to-novel generalization tasks. Notably, KIP consistently outperforms existing prompt tuning strategies, highlighting its efficacy in bridging the modality gap and improving classification performance in fine-grained visual recognition.

CRediT authorship contribution statement

Qinyu Zhang: Writing – original draft, Visualization, Validation, Methodology. **Xinda Liu:** Writing – review & editing, Funding acquisition. **Qi Liu:** Validation, Formal analysis. **Pengbo Zhou:** Project administration, Methodology, Conceptualization. **Guohua Geng:** Visualization, Validation, Supervision.

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.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2023YFF0906504, the National Natural Science Foundation of China[http://dx.doi.org/10.13039/501100001809] under Grant 62271393, the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University under Grant VRLAB2024C02, the General Projects of the Shaanxi Provincial Department of Science and Technology under Grant 2025JC-YBQN-801, and the General Projects of the Shaanxi Provincial Education Department Research Program under Grant 24JK0675.

Data availability

Data will be made available on request.

References

[1] W. Zhou, C. Song, D. Chen, T. Su, H. Hu, C. Shan, From multi-scale grids to dynamic regions: Dual-relation enhanced transformer for image captioning, Knowl.-Based Syst. 311 (2025) 113127.
[2] C. Cai, S. Wang, K.-H. Yap, Y. Wang, Top-down framework for weakly-supervised grounded image captioning, Knowl.-Based Syst. 287 (2024) 111433.
[3] J. Zhang, Z. Fang, H. Sun, Z. Wang, Adaptive semantic-enhanced transformer for image captioning, IEEE Trans. Neural Netw. Learn. Syst. 35 (2) (2024) 1785–1796.
[4] Y. Yu, Y. Kim, Y.M. Ro, Advancing causal intervention in image captioning with causal prompt, IEEE Trans. Neural Netw. Learn. Syst. (2024).
[5] S. Chang, P. Ghamisi, Changes to captions: An attentive network for remote sensing change captioning, IEEE Trans. Image Process. 32 (2023) 6047–6060.
[6] K. Zhang, Z. Yuan, T. Huang, Diverse and tailored image generation for zero-shot multi-label classification, Knowl.-Based Syst. 299 (2024) 112077.
[7] X. He, F. Yang, F. Liu, G. Lin, Few-shot image generation via style adaptation and content preservation, IEEE Trans. Neural Netw. Learn. Syst. 36 (7) (2025) 12326–12337.
[8] Y. Xu, W. Yu, P. Ghamisi, M. Kopp, S. Hochreiter, Txt2Img-MHN: Remote sensing image generation from text using modern Hopfield networks, IEEE Trans. Image Process. 32 (2023) 5737–5750.
[9] H. Tan, X. Liu, B. Yin, X. Li, DR-GAN: Distribution regularization for text-to-image generation, IEEE Trans. Neural Netw. Learn. Syst. 34 (12) (2022) 10309–10323.
[10] Y. Liu, W. Min, S. Jiang, Y. Rui, Convolution-enhanced bi-branch adaptive transformer with cross-task interaction for food category and ingredient recognition, IEEE Trans. Image Process. 33 (2024) 2572–2586.
[11] X. Xia, Y. Shi, P. Li, X. Liu, J. Liu, H. Men, FBANet: An effective data mining method for food olfactory EEG recognition, IEEE Trans. Neural Netw. Learn. Syst. 35 (10) (2023) 13550–13560.
[12] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, S. Jiang, Large scale visual food recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (8) (2023) 9932–9949.
[13] J. Choi, S. Lee, M. Lee, S. Lee, H. Shim, Fine-grained image-text correspondence with cost aggregation for open-vocabulary part segmentation, 2025, arXiv preprint arXiv:2501.09688.
[14] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348.
[15] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Conditional prompt learning for vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16816–16825.
[16] H. Yao, R. Zhang, C. Xu, Visual-language prompt tuning with knowledge-guided context optimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6757–6767.
[17] H. Yao, R. Zhang, C. Xu, TCP: Textual-based class-aware prompt tuning for visual-language model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23438–23448.
[18] B. Kan, T. Wang, W. Lu, X. Zhen, W. Guan, F. Zheng, Knowledge-aware prompt tuning for generalizable vision-language models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15670–15680.
[19] M.U. Khattak, M.F. Naeem, M. Naseer, L. Van Gool, F. Tombari, Learning to prompt with text only supervision for vision-language models, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 4230–4238.
[20] Z. Li, Y. Song, M.-M. Cheng, X. Li, J. Yang, Advancing textual prompt learning with anchored attributes, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 3618–3627.
[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.
[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
[23] 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.
[24] B. Zhu, Y. Niu, Y. Han, Y. Wu, H. Zhang, Prompt-aligned gradient for prompt tuning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15659–15669.
[25] S. Roy, A. Etemad, Consistency-guided prompt learning for vision-language models, in: International Conference on Learning Representations, 2024, pp. 6485–6498.
[26] J. Gao, J. Ruan, S. Xiang, Z. Yu, K. Ji, M. Xie, T. Liu, Y. Fu, Lamm: Label alignment for multi-modal prompt learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 1815–1823.
[27] J. Zhang, J. Huang, X. Zhang, L. Shao, S. Lu, Historical test-time prompt tuning for vision foundation models, in: The Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024.

[28] M.B. Hossen, Z. Ye, A. Abdussalam, F.E. Wahab, Attribute guided fusion network for obtaining fine-grained image captions, Multimedia Tools Appl. 84 (13) (2025) 11907–11941.
[29] M.A. Hossain, Z. Ye, M.B. Hossen, M.A. Rahman, M.S. Islam, M.I. Abdullah, Hierarchical region-context attention for image captioning, Eng. Appl. Artif. Intell. 168 (2026) 114014.
[30] M.B. Hossen, Z. Ye, M.S. Hossain, M.I. Hossain, ARAFNet: an attribute refinement attention fusion network for advanced visual captioning, Digit. Signal Process. 162 (2025) 105155.
[31] D.-N. Zou, S.-H. Zhang, T.-J. Mu, M. Zhang, A new dataset of dog breed images and a benchmark for finegrained classification, Comput. Vis. Media 6 (2020) 477–487.
[32] Z. Sun, Y. Yao, X.-S. Wei, Y. Zhang, F. Shen, J. Wu, J. Zhang, H.T. Shen, Webly supervised fine-grained recognition: Benchmark datasets and an approach, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10602–10611.
[33] L. Bossard, M. Guillaumin, L. Van Gool, Food-101 – mining discriminative components with random forests, in: ECCV, 2014, pp. 446–461.
[34] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, A. Vedaldi, Fine-grained visual classification of aircraft, 2013, arXiv preprint arXiv:1306.5151.
[35] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Indian Conference on Computer Vision, Graphics & Image Processing, 2008.
[36] O.M. Parkhi, A. Vedaldi, A. Zisserman, C.V. Jawahar, Cats and dogs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2012, pp. 3498–3505.
[37] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D object representations for fine-grained categorization, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 554–561.
[38] A. Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei, Novel dataset for fine-grained image categorization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vol. 5, 2011, p. 2.
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[40] S. Hou, Y. Feng, Z. Wang, Vegfru: A domain-specific dataset for fine-grained visual categorization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 541–549.
[41] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200–2011 dataset, 2011.
[42] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[43] B. Zhang, P. Zhang, X. Dong, Y. Zang, J. Wang, Long-clip: Unlocking the long-text capability of clip, in: European Conference on Computer Vision, Springer, 2024, pp. 310–325.