Endowing Vision-Language Models with System 2 Thinking for Fine-Grained Visual Recognition

Yutong Yang[1], **Lifu Huang**[2], **Yijie Lin**[1], **Xi Peng**[1,3], **Mouxing Yang**[1]*

[1]College of Computer Science, Sichuan University
[2]College of Computing & Data Science, Nanyang Technological University
[3]National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation, Sichuan University

Abstract

Vision-Language Models (VLMs) excel at extracting salient visual features from query images, thus exhibiting promising visual recognition performance. However, VLMs would encounter significant degradation in fine-grained scenarios due to their deficiency in distinguishing nuanced differences among candidate categories. As a remedy, we draw inspiration from the "System 1 & System 2" cognitive theory of humans, paving the way to achieve fine-grained recognition for VLMs. To be specific, we observe that VLMs naturally align with System 1, quickly identifying candidate categories but leaving easily-confused ones unresolved. Based on the observation, we propose System-2 enhanCed visuAl recogNition (SCAN), a novel plug-and-play approach that makes VLMs aware of nuanced differences. In brief, SCAN first specifies and abstracts the discriminative attributes for the confused candidate categories and query images by resorting to off-the-shelf large foundation models, respectively. After that, SCAN adaptively integrates the salient visual features from System 1 with the nuanced differences derived from System 2, resolving confusion in candidates with estimated uncertainty. Extensive experiments on eight widely used fine-grained recognition benchmarks against 10 state-of-the-art baselines verify the effectiveness and superiority of SCAN.

Code — github.com/XLearning-SCU/2026-AAAI-SCAN

Introduction

Pre-trained on web-scale images with corresponding alt-texts, Vision-Language Models (VLMs) (Radford et al. 2021; Zhai et al. 2023; Huang et al. 2024; Lin et al. 2024) exhibit impressive visual recognition capability, achieving remarkable performance on downstream tasks, including but not limited to visual recognition (Liang and Davis 2025), clustering (Li et al. 2024b), video reasoning (Lin et al. 2024), person re-identification (Lu et al. 2025), and image-text retrieval (Ding et al. 2025). Despite their success, VLMs would encounter heavy performance degradation once query images deviate from the pre-training distribution (Xie et al. 2025; Du et al. 2025; Li et al. 2024a), particularly in fine-grained scenarios where the category of query image might never even emerge in the pre-training data (Yang et al. 2024).

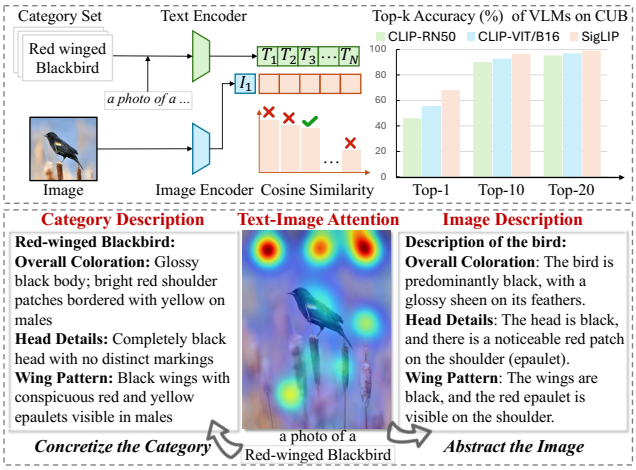


Figure 1: Our observation and key idea. (a) **Observation**: While VLMs fail to distinguish easily-confused categories (*i.e.*, undesirable top-1 accuracy), they are capable of identifying a reasonable subset of candidate categories (*i.e.*, promising top-20 accuracy), suggesting that VLMs could function as System 1 of humans for visual recognition. (b) **Key Idea**: Motivated by the observation, we aim to endow VLMs with System-2 thinking to make them aware of the nuanced differences among the above candidate subset. To this end, we concretize and abstract the category names and query image into discriminative attributes, which facilitates the nuanced reasoning in two folds. On the one hand, concretizing the category name into detailed attributes could supplement VLMs with knowledge about rare categories. On the other hand, abstracting the image into detailed attributes not only avoids the interference of irrelevant information but also bridges the granularity gap between image and category.

To enhance the visual recognition capability of VLMs in target distribution, numerous methods have been proposed (Shu et al. 2022; Xiao et al. 2025; Zanella et al. 2025; Tian et al. 2024). Among them, learning-based approaches adapt VLMs into new scenarios by introducing additional trainable modules (*e.g.*, prompt tokens, adapters), which are typically optimized by resorting to a few la-

*Corresponding author. Email: yangmouxing@gmail.com

beled images (Zhou et al. 2022) or self-supervised strategies on unlabeled images (Shu et al. 2022). To eliminate the training overhead, cache-based methods perform adaptation by associating query images with class prototypes stored in the cache, where each prototype is derived from pre-collected relevant images annotated either manually or via pseudo-labeling (Karmanov et al. 2024; Zhang et al. 2024). Although achieving promising performance, their success heavily relies on labeled or reference target data for model tuning or cache construction, thus limiting their applicability in recognizing fine-grained categories, where such data and annotations are often scarce.

Different from most existing approaches that exhaustively adapt VLMs on target distribution, humans tackle the complex visual tasks through a dual-system framework, *i.e.*, System 1 & System 2 (Li et al. 2025). In the case of fine-grained visual recognition, System 1 conducts intuitive-driven judgments based on salient visual features, rapidly identifying possible categories while preserving easily-confused options. After that, System 2 engages in deliberate thinking by analyzing nuanced differences among the candidate categories. Motivated by this, we aim to imitate the above cognitive pattern of humans and make VLMs aware of the nuance differences, thus enhancing fine-grained visual recognition capability. As illustrated in Fig. 1, we observe that although VLMs struggle in capturing nuances, they are still able to identify a reasonable set of candidate categories. In other words, VLMs can naturally serve as a counterpart to System 1. Therefore, our goal becomes endowing the current VLMs with the System 2 thinking.

Based on the above discussions and observations, we propose a novel approach, termed System-2 enhanCed visuAl recogNition (SCAN), to improve the recognition ability of VLMs across diverse fine-grained scenarios. In brief, SCAN consists of two core modules: the Nuance Reasoning (NR) module to infer nuances among the candidate categories, and the Uncertainty-aware Integration (UI) module to take the best of System-1 identification and System-2 Thinking. Specifically SCAN first employs the NR module to specify and abstract the discriminative attributes for the candidate categories and query images, respectively. As a result, the information granularity of the image and the candidate category could be aligned, and the nuanced differences between the query image and the candidate categories could be naturally uncovered. After that, the UI module estimates the identification uncertainty in the candidate categories and accordingly integrate the salient visual features from System 1 with the nuanced differences from System 2 in a dynamic mechanism. The main contributions and novelties of this work could be summarized as follows:

- Inspired by cognitive science, we propose System-2 enhanCed visuAl recogNition (SCAN), a novel approach for fine-grained recognition. To the best of our knowledge, this work could be one of the first studies to endow VLMs with System-2 thinking and thus make VLMs aware of nuanced differences.
- Different from most existing VLM-enhanced studies that require labeled or reference target data for adaptation, the proposed SCAN imitates the cognition pattern of human beings and could directly recognize the query image in a zero-shot manner.
- Extensive experiments on eight fine-grained recognition benchmarks verify the effectiveness and superiority of SCAN. Furthermore, we exhibit the generalizability of SCAN, demonstrating that it could serve as a plug-and-play solution to enhance different VLMs.

Related Work

In this section, we briefly review three related topics for VLM recognition capacity enhancement, *i.e.*, learning-based VLM enhancement, cache-based VLM enhancement, and LLM-based VLM enhancement.

Learning-based VLM Enhancement

Learning-based VLM enhancement approaches usually incorporate additional trainable modules to enhance the adaptability of VLMs. One of the most representative works is prompt tuning, which replaces manually-crafted prompts with learnable tokens and optimizes them using a small set of labeled samples (Zhou et al. 2022; Qi et al. 2025). To mitigate the need for labeled data, test-time prompt tuning approaches have been proposed, where learnable prompt tokens are updated based on prediction consistency across test-time inputs (Shu et al. 2022). Another line of research replaces learnable tokens with lightweight linear adapters, which are similarly updated using a few labeled samples for adaptation (Gao et al. 2024).

Despite the success of these learning-based approaches, they still rely on labeled data or introduce additional training overhead. In contrast, our method seeks to enhance VLMs in a zero-shot setting, eliminating the need for extra supervision or training.

Cache-based VLM Enhancement

To alleviate training overhead, cache-based methods have recently emerged as an efficient adaptation strategy for VLMs. For example, Tip-Adapter(Zhang et al. 2022) constructs a key-value cache by storing features of a small number of labeled samples and performs category inference by treating the test image as a query and retrieving the most relevant information in the cache. To reduce the label dependency for cache construction, TDA(Karmanov et al. 2024) and DMN (Zhang et al. 2024) infer pseudo labels for the reference samples, thus building and updating the cache in an unsupervised manner.

While these approaches enable VLMs to generalize to the target domain, their performance heavily relies on the quality of the cached features, making them less effective in data-scarce or noisy environments. In contrast, our approach allows for fine-grained recognition on online query images without requiring reference data from the target domain.

LLM-based VLM Enhancement

With the rapid advancement of large foundation models, recent studies have explored integrating LLMs to enhance VLMs by generating category-specific textual descriptions.

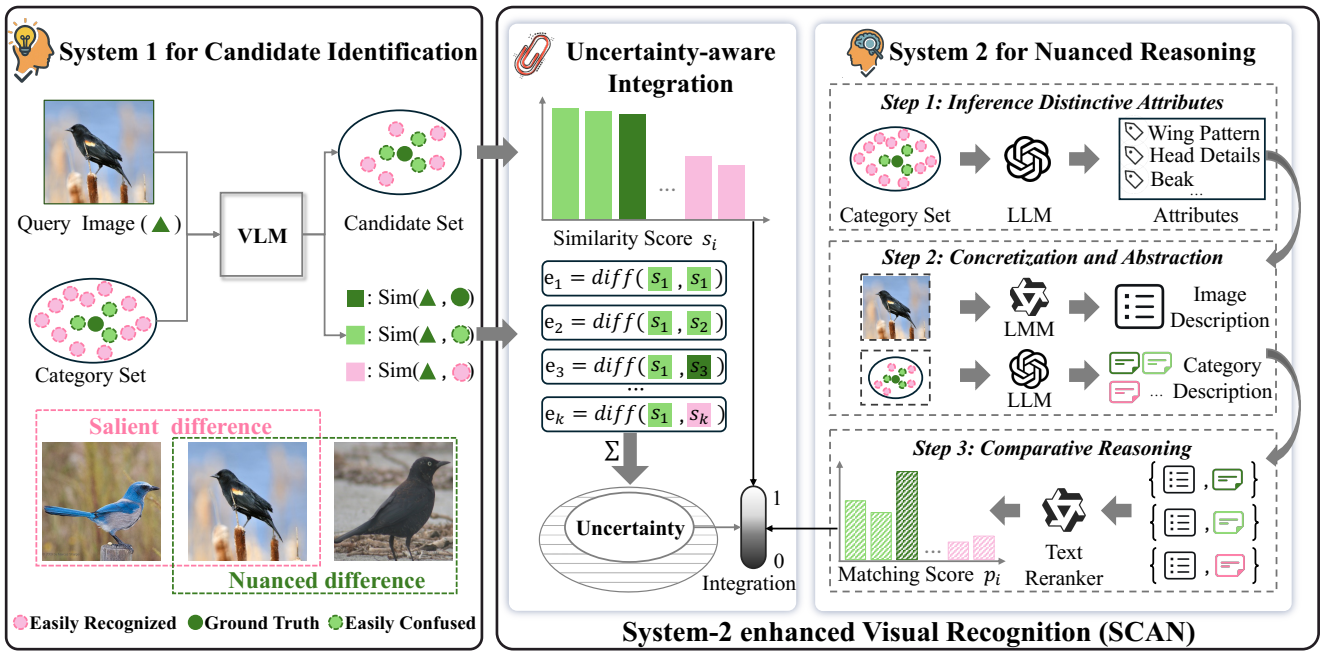


Figure 2: Overview of SCAN. SCAN consists of two key components: the Nuanced Reasoning (NR) module and the Uncertainty-aware Integration (UI) module. In our method, VLMs play the role of System 1 to identify a reasonable set of candidate categories. After that, SCAN first employs the NR module to infer a set of discriminative attributes via an off-the-shelf LLM. Based on these attributes, both visual and textual modalities are aligned into a shared attribute space, transforming the recognition task into semantic comparisons in the textual space. To this end, the NR module employs a pre-trained text re-ranker model to assess the fine-grained distinctions. Finally, SCAN utilizes the UI module to dynamically integrate the recognition results from System 1 and System 2 based on the difference-based uncertainty estimation strategy.

For example, DCLIP (Menon and Vondrick 2022) leverages GPT-3 (Brown et al. 2020) to enrich category names with attribute-level information, improving the expressiveness of textual prompts. HIE (Ren, Su, and Liu 2023) employs LLMs to generate discriminative descriptions across hierarchical category names, which are then used for category clustering and hierarchical reasoning. CuLP (Pratt et al. 2023) replaces manually crafted prompts with LLM-generated ones, making the method more applicable to real-world zero-shot recognition scenarios. ProAPO (Qu et al. 2025) further scales the idea by generating large prompt sets and adaptively optimizing them to identify the task-optimal prompts.

In summary, most existing current LLM-based VLM Enhancement works shared the same underlying insight, *i.e.*, improve the richness and quality of the text prompt for VLMs by leveraging external knowledge encoded in LLMs. While sharing some technical similarities, our method significantly differs from them in the following two main aspects. First, rather than simply enriching category descriptions, our method performs both concretization and abstraction on the candidate categories and query images, thereby aligning the information granularity of both. Second, our approach goes beyond enhancing text prompts by endowing VLMs with System-2 thinking, enabling them to recognize subtle differences among the easily-confused categories and achieving significant improvements in fine-grained recognition performance.

Method

In this section, we present the proposed method, System-2 enhanCed visuAl recogNition (SCAN). We begin by using CLIP (Radford et al. 2021) as a showcase to illustrate how Vision-Language Models (VLMs) naturally function as System 1 in identifying candidate categories for fine-grained visual recognition. Then, we introduce the two key modules of SCAN. In brief, as shown in Fig. 2, the Nuanced Reasoning (NR) module serves as System 2 to enhance System 1's awareness of the nuanced differences among candidate categories, while the Uncertainty-aware Integration (UI) module could dynamically integrate both System 1 and System 2.

System-1 for Candidate Identification

VLMs, such as CLIP, typically employ a dual-encoder architecture comprising a vision encoder E_v and a text encoder E_t, which embed visual and textual inputs into a shared representation space. For a given image x and a set of categories $C = \{c_1, c_2, \ldots, c_N\}$, the similarity between x and each category c_i is computed as follows:

$$s(x, c_i) = \cos(E_v(x), E_t(f(c_i))), \quad (1)$$

where $f(\cdot)$ converts the category name into the corresponding textual prompt (*e.g., "a photo of a [category name]"*),

and $\cos(\cdot)$ denotes the cosine similarity. Then, the category with the highest similarity is selected as the prediction result of VLMs.

While VLMs exhibit strong performance in general recognition tasks, they often struggle in fine-grained scenarios due to their limited ability to capture subtle visual differences between similar categories (Xie et al. 2025). Fortunately, as observed in Fig. 1, although VLMs fail to precisely predict the correct category, they are still capable of filtering out irrelevant categories and narrowing C into a reasonable subset C^* with k candidate categories. Formally,

$$C^* = \mathrm{Top_k}(C), \tag{2}$$

where $\mathrm{Top_k}(\cdot)$ returns the top k categories ranked by similarity scores $s(x, c_i), \forall c_i \in C$.

As discussed in the Introduction, the above candidate identification process of VLMs could be regarded as System-1 thinking from the perspective of cognitive theory (Li et al. 2025). However, System 1 lacks the capacity for nuanced reasoning within the candidate subset C^*, prohibiting it from accurate fine-grained recognition. As a remedy, we propose SCAN, a System-2-inspired framework to enhance VLMs with nuanced reasoning capabilities. In the following, we will elaborate on the modules of SCAN.

Nuanced Reasoning

The Nuanced Reasoning (NR) module of SCAN plays the role of System 2, performing comprehensive and detailed comparisons and reasoning based on key differentiating attributes of the candidate subset C^*. In the following, we will detail the three key steps of the NR module.

Discriminative Attributes Inference. To make VLMs aware of nuanced differences among C^*, it is essential to first identify a set of discriminative attributes that could characterize the fine-grained differences. To this end, we leverage the off-the-shelf Large Language Model (LLM) to automatically derive a set of discriminative attributes $A = \{a_1, a_2, \cdots, a_n\}$ from the category names. Notably, to avoid the redundant attribute inference for the candidate categories of each individual query image, we prompt the LLM to infer attributes for all categories in C at once. Formally,

$$A = \mathrm{LLM}(C, \mathrm{prompt}_A), \tag{3}$$

where prompt_A denotes the system prompt for attribute inference.

After inferring the attributes, one might simply use these attributes as prompts to enhance the recognition capability, like existing LLM-based studies for VLM enhancement (Menon and Vondrick 2022; Qu et al. 2025). However, such a straightforward approach yields only limited improvement, as verified in Fig. 1, which could stem from the difference in granularity between the visual and textual modalities. Specifically, images are rich in low-level visual details, whereas category names are abstract and often underspecified. To bridge this gap, we propose a concretization and abstraction mechanism that aligns the semantic granularity of both modalities at the attribute level. Formally,

- **Concretization:** Based on the discriminative attribute set A, we leverage an LLM to expand each category name c_i into an attribute-aligned category description d_{c_i} based on the intrinsic knowledge of LLM. Formally,

$$d_{c_i} = \mathrm{LLM}(c_i, \mathrm{prompt}_c, A), \tag{4}$$

  where $c_i \in C^*$ and prompt_c denotes the system prompt used in the concretization process.

- **Abstraction:** Similar to Eq. 4, we leverage a Large Multimodal Model (LMM) to abstract the query image x into an attribute-level textual description, *i,e,*

$$d_x = \mathrm{LMM}(x, \mathrm{prompt}_x, A), \tag{5}$$

  where prompt prompt_x is the system prompt designed to instruct the model to extract visual cues corresponding to A, thus facilitating the elimination of irrelevant information (*e.g.*, background elements in the image).

Clearly, by aligning the query image and category names in a shared attribute space, this approach supports more interpretable and fine-grained difference reasoning. Due to the space limitation, the details of the aforementioned prompts would be presented in the supplementary materials.

Comparative Reasoning. Thanks to the above abstraction and concretization mechanism, the recognition task now could be transformed into the semantic difference comparison between d_{c_i} and d_x. To this end, we resort to the off-the-shelf textual reranker model, which evaluates the relevance between d_{c_i} and d_x, $\forall c_i \in C^*$, and outputs the corresponding similarity score $p(x, c_i)$. In brief, $p(x, c_i)$ reflects the matching degree between the image description and each candidate category description, based on attribute-level comparisons in the textual space.

Uncertainty-aware Integration

Given the similarity scores $s(x, c_i)$ and $p(x, c_i)$ from VLMs and the NR module, respectively, the final recognition results could be obtained by combining the effects of both as follows:

$$c^* = \underset{c_i \in C^*}{\mathrm{argmax}} \left(\alpha \cdot s(x, c_i) + (1 - \alpha) \cdot p(x, c_i)\right), \tag{6}$$

where α is a coefficient to balance the contributions of VLMs and the NR module. Instead of simply fixing α as a constant, it is more desirable to adaptively adjust it according to the identification uncertainty of VLMs among C^*. In other words, VLMs is expected to contribute less when C^* contains highly-similar categories and yields very close similarity scores $s(x, c_i)$ for $c_i \in C^*$, exhibiting relatively high uncertainty of VLMs, and vice versa.

To this end, we propose a novel uncertainty estimation strategy tailored for fine-grained recognition, enabling SCAN to dynamically balance the contributions of System 1 and System 2. To be specific, we utilize the similarity score differences among the candidate categories as evidence for VLMs' identification. To unify the measurement of such evidence, we take the $\mathrm{top}-1$ similarity score as the reference and compute the relative difference with the remaining candidates as follows:

$$\mathrm{diff}_j = s_{h_1} - s_{h_j}, \quad j = 1, 2, \ldots, k, \tag{7}$$

Method	Settings (**AF RF LR**)	Flower	CUB	Food	Pet	Aircraft	Car	Dog	SUN	Average	Δ
CLIP ResNet-50 Backbone											
CLIP	✓✓✓	61.20	46.00	78.59	83.62	15.75	53.95	52.19	58.49	56.22	-
CLIP$_{PE}$	✓✓✓	66.06	46.50	80.84	85.80	17.13	54.45	53.66	59.92	58.04	+ 1.82
CoOp (IJCV'22)	× ✓×	61.55	48.60	75.59	87.00	15.12	55.32	58.40	58.15	57.46	+1.24
TPT (NeurIPS'22)	✓✓×	62.69	50.00	74.88	84.49	17.58	58.46	55.48	61.46	58.13	+1.90
Tip-Adapter (ECCV'22)	×× ✓	73.00	48.24	77.41	86.15	18.57	57.62	56.72	61.19	59.86	+3.63
DCLIP (ICLR'23)	✓✓✓	65.68	49.13	79.65	83.10	16.92	53.86	54.94	61.04	58.04	+1.81
TDA (CVPR'24)	✓× ✓	68.74	50.22	77.75	86.18	17.61	57.78	56.31	62.53	59.64	+3.41
DMN (CVPR'24)	✓× ✓	67.93	48.75	76.70	86.78	22.77	60.02	-	64.39	61.04	+4.82
BCA (CVPR'25)	✓× ✓	66.30	-	77.19	85.58	19.89	58.13	-	63.38	61.74	+5.52
ProAPO (CVPR'25)	× ✓✓	75.10	50.70	81.80	88.70	21.10	58.00	-	63.70	62.72	+6.50
CLIP+SCAN	✓✓✓	77.57	64.41	86.76	89.94	38.49	77.37	68.81	72.06	71.92	+15.70
CLIP$_{PE}$+SCAN	✓✓✓	**79.29**	**64.92**	**87.42**	**91.06**	**40.11**	**77.97**	**70.94**	**72.40**	**73.01**	**+16.79**
CLIP ViT-B/16 Backbone											
CLIP	✓✓✓	67.70	54.97	88.23	88.20	23.91	63.72	60.55	62.71	63.74	-
CLIP$_{PE}$	✓✓✓	71.34	55.37	88.72	89.13	24.24	64.68	62.81	65.21	65.18	+1.44
CoOp (IJCV'22)	× ✓×	68.71	52.10	85.30	89.14	18.47	64.51	64.50	64.15	63.36	-0.38
TPT (NeurIPS'22)	✓✓×	68.98	56.75	84.67	87.79	24.78	66.87	62.28	65.50	64.70	+0.95
Tip-Adapter (ECCV'22)	×× ✓	83.23	57.75	86.10	89.10	27.66	66.91	65.36	65.68	67.72	+3.97
DCLIP (ICLR'23)	✓✓✓	70.62	57.75	88.50	86.92	24.96	63.25	63.47	66.85	65.29	+1.54
TDA (CVPR'24)	✓× ✓	71.42	57.59	86.14	88.63	23.91	67.28	65.85	67.62	66.05	+2.30
DMN (CVPR'24)	✓× ✓	74.49	56.71	85.08	92.04	30.03	67.96	-	70.18	68.07	+4.32
BCA (CVPR'25)	✓× ✓	73.12	-	85.97	90.43	28.59	66.86	-	68.41	68.89	+5.14
ProAPO (CVPR'25)	× ✓✓	**83.67**	58.89	89.14	**92.39**	27.39	67.31	-	67.34	69.44	+5.69
CLIP+SCAN	✓✓✓	79.88	**66.70**	89.46	89.88	44.01	79.21	71.99	73.08	74.27	+10.53
CLIP$_{PE}$+SCAN	✓✓✓	80.89	66.20	**89.69**	91.06	**45.15**	**79.49**	**73.65**	**73.61**	**74.96**	**+11.22**

Table 1: Comparisons with state-of-the-art methods on 8 fine-grained datasets regarding the top-1 accuracy. The best and second best results are marked in **bold** and underline. Here, **AF**, **RF**, **LR** denote annotation-free, reference-free, learning-free methods, respectively. The experimental results of the compared methods are obtained either from their respective papers or reproduced based on their publicly available code.

where $h_1, \cdots, h_k$ denote the indices of top $-k$ categories in C^* sorted by descending similarity, *i.e.*, $s_{h_1} \geq s_{h_2} \geq \cdots \geq s_{h_k}$.

As discussed above, diff_j could serve as a proxy for the VLM's evidence in distinguishing candidate categories. Following prior work on evidential deep learning in general visual recognition (Sensoy, Kaplan, and Kandemir 2018; Han et al. 2022; Du et al. 2023), we transform the above evidence values into the parameters of a Dirichlet distribution and derive the uncertainty as:

$$\text{Uncertainty}(C^*) = \frac{k}{S} = \frac{k}{\sum_{j=1}^{k}(\text{diff}_j \cdot \tau + 1)}, \quad (8)$$

where τ is a temperature parameter controlling the sensitivity to similarity differences and S is the total evidence implying the uncertainty. As illustrated by Fig. 4, compared to vanilla uncertainty modeling methods (Ma et al. 2025), our nuance-aware design leads to a more precise characterization of uncertainty in fine-grained scenarios.

Finally, the integration coefficient α in Eq. 6 could be obtained based on the estimated uncertainty, *i.e.*,

$$\alpha = 1 - \text{Uncertainty}(C^*). \quad (9)$$

Experiments

In this section, we verify the effectiveness of SCAN through extensive experiments. We begin by presenting the implementation details of SCAN. Then, we validate the performance superiority of SCAN by conducting extensive experiments on 8 fine-grained recognition datasets, comparing with 10 state-of-the-art methods. Moreover, we conduct a series of ablation studies and analysis studies to provide a comprehensive understanding of SCAN. Due to the space limitation, we present more experimental details and results in the supplementary materials.

Implementation Details

Unless otherwise stated, we utilize GPT-4.1-mini (Achiam et al. 2023) as the default LLM for discriminative attributes inference in Eq. 3 and category name concretization in Eq. 4. In addition, we employ Qwen2.5-VL-32B (Bai et al. 2025) to perform image abstraction in Eq. 5. For the comparative reasoning step, we adopt Qwen3-Reranker-8B (Zhang et al. 2025) to assess the semantic similarity between image descriptions and category descriptions. As for the number of candidate categories, k is fixed as 20 across all experiments for simplicity. The temperature parameter τ in Eq. 8 is consistently set to 40. All experiments are conducted on Ubuntu

20.04 with NVIDIA RTX 4090 GPUs.

Comparison with State-of-the-Art Methods

Datasets. We evaluate the proposed SCAN on eight widely-used fine-grained image recognition datasets, covering several domains (*e.g.*, plants, animals, scenes, vehicles). To be specific, the datasets include Flowers102 (Nilsback and Zisserman 2008), CUB200 (Wah et al. 2011), Food101 (Bossard, Guillaumin, and Van Gool 2014), Oxford Pets (Parkhi et al. 2012), Aircraft (Maji et al. 2013), Stanford Cars (Krause et al. 2013), Stanford Dogs (Khosla et al. 2011), and SUN397 (Xiao et al. 2010).

Baseline Methods. We compare our SCAN against 10 state-of-the-art (SOTA) baselines, including the vanilla CLIP model (Radford et al. 2021) with a hand-crafted prompt (*i.e.*, "a photo of a {category name}") or with prompt ensembling (denoted as CLIP$_{PE}$), the training-based VLM-enhanced methods (CoOp (Zhou et al. 2022), TPT (Shu et al. 2022)), cache-based VLM-enhanced methods (Tip-Adapter (Zhang et al. 2022), TDA (Karmanov et al. 2024), DMN (Zhang et al. 2024), BCA (Zhou et al. 2025)), and LLM-based VLM-enhanced methods (DCLIP (Menon and Vondrick 2022), ProAPO (Qu et al. 2025)). All models are evaluated on two widely adopted backbones: CLIP-ResNet50 and CLIP-ViT-B/16.

Results. As shown in Table 1, our method consistently achieves remarkable performance improvements in various datasets. Specifically, SCAN achieves the absolute improvements of 15.70% and 10.53% for CLIP-ResNet50 and CLIP-ViT-B/16 in terms of the average top-1 accuracy, respectively. Moreover, our SCAN exhibits significant performance superiority in a fully zero-shot setting, outperforming other VLM-enhanced methods, which typically rely on labeled or reference target data from the target domain or require additional tuning for adaptation.

Ablation and Analytic studies

In this section, we conduct comprehensive ablation and analytic studies to further investigate the effectiveness of SCAN. If not specified, all experiments in the section are conducted on the CUB and Dog datasets using CLIP$_{PE}$ with ViT-B/16.

Method	CUB	Dog	Average
(System 1) CLIP$_{PE}$	55.37	62.81	59.09
(System 1) CLIP$_{PE}$ + NR-C	50.03	58.66	54.34
(System 2) NR	50.79	63.46	57.12
(System 1 + 2) CLIP$_{PE}$ + NR	65.99	73.08	69.53
(System 1 + 2 + UI) SCAN	**66.20**	**73.65**	**69.92**

Table 2: Ablation study. Here, "NR-C" denotes the concretization operation in the Nuanced Reasoning module, and the Default settings are marked in gray.

Ablation Studies. As shown in Table 2, to investigate the importance of each component of SCAN, we design the following method variants and accordingly draw some conclusions. To be specific: i) "(System 1) CLIP$_{PE}$" refers to the result of vanilla System 1; ii) "(System 1) CLIP$_{PE}$ + NR-C" represents enhancing CLIP only with the concretization operation (Eq. 4), where category names are expanded into detailed descriptions and used as alternative textual prompts for CLIP, similar to LLM-enhancement methods (Menon and Vondrick 2022). The degraded results indicate that CLIP is incapable of directly benefiting from the detailed category descriptions. iii) "(System 2) NR" represents that using the NR module only, *i.e.*, performing comparative reasoning over all category in C instead of C^*. The degraded results highlight the importance of adopting System 1 as the basis. iv) "(System 1+2) CLIP$_{PE}$+NR" denotes capsulating Systems 1 and 2 in a vanilla way, *i.e.*, setting $\alpha = 0.5$ in Eq. 6. The inferior results demonstrate that it is more desirable to take the best of both System 1 and System 2 in an adaptive manner.

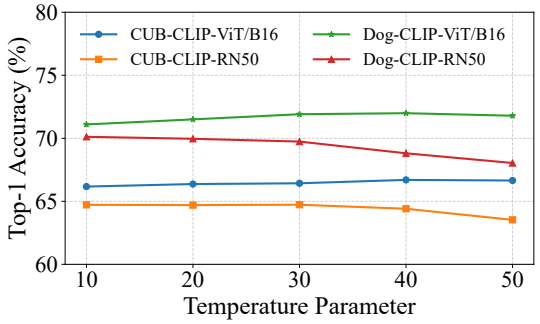
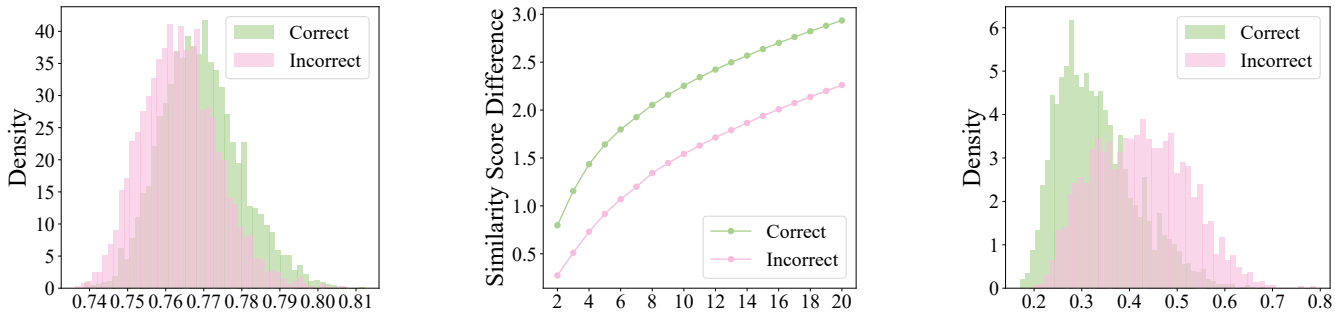


Figure 3: Sensitivity analysis of the hyper-parameter τ in Eq. 8.

Sensitivity of Temperature τ. In the UI module of SCAN, the similarity score differences used as evidence are typically small in magnitude. As a remedy, we introduce a temperature parameter τ to scale the evidence. To investigate the sensitivity of our method to τ, we investigate the performance of SCAN by increasing τ from 10 to 50 with an interval of 10. From Fig. 3, one could observe that SCAN performs stably with different choices of τ, which shows its robustness to the parameter.

Effectiveness of the Difference-based Uncertainty Estimation. Besides the quantitative results in Table 2, we further conduct qualitative analysis to investigate the effectiveness of the difference-based uncertainty estimation strategy introduced in the UI module. We first follow the vanilla uncertainty modeling approach (Sensoy, Kaplan, and Kandemir 2018), which directly regards the vanilla similarity scores $s(x, c_i)$ as evidence. As illustrated in Fig. 4(a), this method results in highly overlapping uncertainty distributions for correct and incorrect predictions, making it less effective in fine-grained scenarios. To further analyze this issue, we compute the average similarity score difference between the top-1 and other candidate categories for correctly and incorrectly predicted samples. As shown in Fig. 4(b), when VLMs are able to select the correct category based on salient visual features, the similarity gap is significantly larger than in failure cases where the model is confused. Motivated by this observation, we propose to use similarity



(a) (b) (c)

Figure 4: Uncertainty distributions of correct and incorrect predictions under different modeling strategies: (a) Using vanilla similarity as evidence leads to indistinguishable distributions; (b) Using similarity differences as evidence (Eq. 7) results in a considerable margin between correct and wrong predictions. (c) Modeling based on similarity differences yields well-separated uncertainty profiles.

score differences as the evidence (*i.e.*, Eq. 7) for uncertainty estimation. As demonstrated in Fig. 4(c), the resulting probability density distributions show a much clearer separation between correct and incorrect predictions compared to the vanilla strategy, indicating the effectiveness of our dedicated uncertainty estimation for fine-grained recognition.

Generalization of SCAN. As our SCAN is a general framework that could ideally endow various VLMs with the capacity of fine-grained recognition in a plug-and-play manner, it is necessary to explore the effect beyond the CLIP model families. To this end, we choose the recently published strong VLMs, *i.e.*, SigLIP (Zhai et al. 2023) with another base model. Similar to the enhancement for CLIP, we use SigLIP as System 1 in this experiment and adopt our SCAN upon it. The results are summarized in Table 3, where one could observe that SCAN boosts the average top-1 accuracy across CUB and Dog by 4.80% (from 66.91% to 71.71%), demonstrating that the effectiveness of SCAN is architecture-agnostic. Due to space limitations, the complete experimental results of SCAN on all eight datasets using SigLIP are provided in the supplementary material.

Method	CUB	Dog	Average	Δ
SigLIP	68.04	65.78	66.91	-
SigLIP + SCAN	**71.28**	**72.14**	**71.71**	+4.80

Table 3: Generalization investigation of SCAN on SigLIP.

Necessity of Test-time Robustness Enhancement. Recent advances in large multimodal models (LMMs) have significantly improved fine-grained visual understanding. To rigorously evaluate the effectiveness of our SCAN, we construct a balanced test subset by randomly sampling three test images per category on the CUB and Dog datasets. We compare SCAN against the two LMMs used in our approach, *i.e.*, Qwen2.5-VL-32B (Bai et al. 2025) and GPT-4.1-mini (Achiam et al. 2023), along with the widely-used LLaVA-v1.6-34B (Liu et al. 2023) in the academic community. As shown in Table 4, despite being trained on massive datasets across diverse domains, existing LMMs still underperform on fine-grained recognition benchmarks. The results highlight the fact that beyond designing stronger pre-trained models, it is equally important to focus on enhancing their test-time robustness, especially in challenging, fine-grained scenarios.

Method	CUB	Dog	Average	Δ
SCAN(Ours)	**66.33**	**71.38**	**68.85**	-
Qwen2.5-VL-32B	49.66	61.33	55.49	-13.36
LLaVA-v1.6-34B	7.83	18.61	13.22	-55.63
GPT-4.1-mini	54.66	63.05	58.85	-10.00

Table 4: Comparison with state-of-the-art LMMs. The Δ column reports the performance gap relative to SCAN.

Conclusion

In this work, we propose SCAN, a novel test-time scaling framework inspired by cognitive science to endow VLMs with the capacity to perform fine-grained visual recognition. By imitating the "System 1 & System 2" cognition pattern of human beings, SCAN enhances VLMs with a System-2 thinking process that allows for the awareness of nuanced differences, resolving easily-confused categories in fine-grained scenarios. Through extensive experiments on eight widely-used fine-grained recognition benchmarks, we verify that SCAN not only significantly improves the fine-grained recognition performance of VLMs, but also offers a plug-and-play solution that can be seamlessly integrated into various VLM architectures. Moreover, even compared to the LMMs pre-trained on numerous data, SCAN also exhibits impressive performance superiority. The results highlight the potential of SCAN as a test-time scaling method that supplements pre-training, particularly in domains where pre-training data cannot cover all fine-grained variations. In the future, we plan to expand SCAN to other tasks such as visual-text retrieval and language-guided grounding, hoping to expand the boundaries of "System 1 & System 2" cognition framework into a wider range of applications.

Acknowledgments

This work was supported in part by NSFC under Grant 624B2099, 62176171, 62472295, U24B20174; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303, CJ202403; in part by Sichuan Science and Technology Planning Project under Grant 24NSFTD0130; and in part by Baidu Scholarship.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

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

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Ding, G.; Lu, Y.; Hu, P.; Yang, M.; Lin, Y.; Peng, X.; et al. 2025. Visual Abstraction: A Plug-and-Play Approach for Text-Visual Retrieval. In *ICML*.

Du, S.; Fang, Z.; Lan, S.; Tan, Y.; Günther, M.; Wang, S.; and Guo, W. 2023. Bridging trustworthiness and open-world learning: An exploratory neural approach for enhancing interpretability, generalization, and robustness. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8719–8729.

Du, S.; Fang, Z.; Tan, Y.; Wang, C.; Wang, S.; and Guo, W. 2025. OpenViewer: Openness-Aware Multi-View Learning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, 16389–16397.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2): 581–595.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted multi-view classification with dynamic evidential fusion. *TPAMI*, 45(2): 2551–2566.

Huang, Z.; Yang, M.; Xiao, X.; Hu, P.; and Peng, X. 2024. Noise-robust vision-language pre-training with positive-negative learning. *TPAMI*.

Karmanov, A.; Guan, D.; Lu, S.; El Saddik, A.; and Xing, E. 2024. Efficient test-time adaptation of vision-language models. In *CVPR*, 14162–14171.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F.-F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Computer Vision and Pattern Recognition Workshop*.

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

Li, H.; Hu, P.; Zhang, Q.; Peng, X.; Liu, X.; and Yang, M. 2024a. Test-time Adaptation for Cross-modal Retrieval with Query Shift. *ICLR*.

Li, X.; Pan, Y. P.; Sun, Y.; Sun, Q. S.; Tsang, I. W.; and Ren, Z. 2024b. Fast unpaired multi-view clustering. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*.

Li, Z.-Z.; Zhang, D.; Zhang, M.-L.; Zhang, J.; Liu, Z.; Yao, Y.; Xu, H.; Zheng, J.; Wang, P.-J.; Chen, X.; et al. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Liang, T.; and Davis, J. 2025. Making Better Mistakes in CLIP-Based Zero-Shot Classification with Hierarchy-Aware Language Prompts. *arXiv preprint arXiv:2503.02248*.

Lin, Y.; Zhang, J.; Huang, Z.; Liu, J.; Wen, Z.; and Peng, X. 2024. Multi-granularity correspondence learning from long-term noisy videos. *ICLR*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *NeurIPS*, 36: 34892–34916.

Lu, Y.; Yang, M.; Peng, D.; Hu, P.; Lin, Y.; and Peng, X. 2025. LLaVA-ReID: Selective Multi-image Questioner for Interactive Person Re-Identification. *arXiv preprint arXiv:2504.10174*.

Ma, H.; Chen, J.; Zhou, J. T.; Wang, G.; and Zhang, C. 2025. Estimating LLM Uncertainty with Evidence. *arXiv preprint arXiv:2502.00290*.

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

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

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. x.

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

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 15691–15701.

Qi, Z.; Pan, Y.; Meng, L.; Zhou, S.; Yu, H.; Li, X.; and Meng, X. 2025. Global Prompt Refinement with Non-Interfering Attention Masking for One-Shot Federated Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Qu, X.; Gou, G.; Zhuang, J.; Yu, J.; Song, K.; Wang, Q.; Li, Y.; and Xiong, G. 2025. Proapo: Progressively automatic prompt optimization for visual classification. In *CVPR*, 25145–25155.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Ren, Z.; Su, Y.; and Liu, X. 2023. ChatGPT-powered hierarchical comparisons for image classification. *NeurIPS*, 36: 69706–69718.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *NeurIPS*, 31.

Shu, M.; Nie, W.; Huang, D.-A.; Yu, Z.; Goldstein, T.; Anandkumar, A.; and Xiao, C. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 35: 14274–14289.

Tian, Y.; Yang, M.; Li, Y.; Liu, D.; Ren, X.; Peng, X.; and Lv, J. 2024. An empirical study of parameter efficient fine-tuning on vision-language pre-train model. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*.

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

Xiao, Z.; Yan, S.; Hong, J.; Cai, J.; Jiang, X.; Hu, Y.; Shen, J.; Wang, Q.; and Snoek, C. G. 2025. Dynaprompt: Dynamic test-time prompt tuning. *arXiv preprint arXiv:2501.16404*.

Xie, S.; Lingjing, L.; Zheng, Y.; Yao, Y.; Tang, Z.; Xing, E. P.; Chen, G.; and Zhang, K. 2025. SmartCLIP: Modular Vision-language Alignment with Identification Guarantees. In *CVPR*, 29780–29790.

Yang, M.; Li, Y.; Zhang, C.; Hu, P.; and Peng, X. 2024. Test-time adaptation against multi-modal reliability bias. In *ICLR*.

Zanella, M.; Fuchs, C.; De Vleeschouwer, C.; and Ben Ayed, I. 2025. Realistic test-time adaptation of vision-language models. In *CVPR*, 25103–25112.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *ICCV*, 11975–11986.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, 493–510. Springer.

Zhang, Y.; Li, M.; Long, D.; Zhang, X.; Lin, H.; Yang, B.; Xie, P.; Yang, A.; Liu, D.; Lin, J.; Huang, F.; and Zhou, J. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*.

Zhang, Y.; Zhu, W.; Tang, H.; Ma, Z.; Zhou, K.; and Zhang, L. 2024. Dual memory networks: A versatile adaptation approach for vision-language models. In *CVPR*, 28718–28728.

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

Zhou, L.; Ye, M.; Li, S.; Li, N.; Zhu, X.; Deng, L.; Liu, H.; and Lei, Z. 2025. Bayesian test-time adaptation for vision-language models. In *CVPR*, 29999–30009.