SARE: Sample-wise Adaptive Reasoning for Training-free Fine-grained Visual Recognition

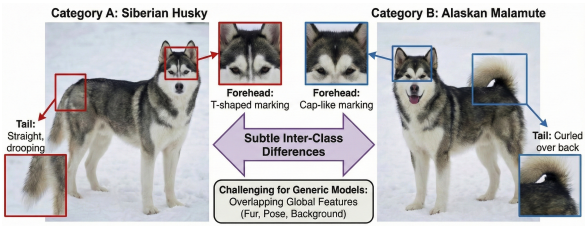
Anonymous ACL submission

Abstract

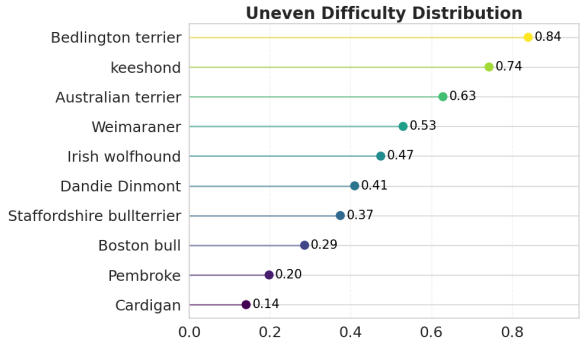
Recent advances in Large Vision–Language Models (LVLMs) have enabled training-free Fine-Grained Visual Recognition (FGVR). However, effectively exploiting LVLMs for FGVR remains challenging due to the inherent visual ambiguity of subordinate-level categories. Existing methods predominantly adopt either retrieval-oriented or reasoning-oriented paradigms to tackle this challenge, but both are constrained by two fundamental limitations: (1) They apply the same inference pipeline to all samples without accounting for uneven recognition difficulty, thereby leading to suboptimal accuracy and efficiency; (2) The lack of mechanisms to consolidate and reuse error-specific experience causes repeated failures on similar challenging cases. To address these limitations, we propose **SARE**, a **S**ample-wise **A**daptive **RE**asoning framework for training-free FGVR. Specifically, SARE adopts a cascaded design that combines fast candidate retrieval with fine-grained reasoning, invoking the latter only when necessary. In the reasoning process, SARE incorporates a self-reflective experience mechanism that leverages past failures to provide transferable discriminative guidance during inference, without any parameter updates. Extensive experiments across 14 datasets substantiate that SARE achieves state-of-the-art performance while substantially reducing computational overhead.

1 Introduction

Recent advances in Large Vision-Language Models (LVLMs) have led to remarkable progress across a wide range of visual recognition tasks (Wang et al., 2024; Yang et al., 2025; Wang et al., 2025; Chen et al., 2025; Comanici et al., 2025; OpenAI, 2023). However, despite their strong general capabilities, LVLMs remain weak on Fine-Grained Visual Recognition (FGVR) (Geigle et al., 2024; Zhang et al., 2024b; Peng et al., 2026; He et al., 2025). FGVR aims to distinguish visually similar sub-categories, such as different bird species or dog breeds (Wei et al., 2021; Kim and Ji, 2024). Unlike general visual recognition where coarse visual features often suffice, FGVR presents a unique challenge: decisive and discriminative cues are subtle, highly localized, and easily confounded (Wei et al., 2021; Peng et al., 2026). As shown in Figure 1a, two dog breeds may only differ in forehead pattern and tail shape. This challenge is further exacerbated by large intra-class variations caused by changes in pose, background, etc.



(a) The challenge of FGVR.



(b) CLIP normalized scores corresponds to a 70% Top-1 Acc.

Figure 1: (a) Fine-Grained Visual Recognition challenge: sub-categories (e.g., Siberian Husky vs. Alaskan Malamute) exhibit subtle visual differences, requiring attention to localized discriminative features such as forehead patterns and tail shape. (b) For a fixed overall retrieval accuracy (70%), different sub-categories require substantially different confidence thresholds, indicating that identical confidence scores imply varying reliability across categories.

In light of these challenges, a variety of methods have been proposed to improve model's performance in FGVR. These approaches can be broadly categorized into two paradigms, according to how they employ LVLMs: i) **Retrieval-oriented methods** prompt the LVLM to generate fine-grained textual descriptions and perform multimodal matching. These methods are efficient and scalable, but rely primarily on global representations and struggle to capture localized discriminative cues required to distinguish visually similar sub-categories (Bianchi et al., 2024; Xie et al., 2025). ii) **Reasoning-oriented methods** instead formulate recognition as a multi-choice visual question-answering (VQA) (Kabir et al., 2024) problem by presenting candidate labels as context. While capable of deeper analysis, as the candidate space grows, the model's attention becomes increasingly diffused, leading to unstable reasoning behaviors such as hallucination and performance degradation (Wei, 2024; Zhang et al., 2024c; Wu et al., 2025; He et al., 2025).

In spite of different strategies, both paradigms share two fundamental limitations. First, they apply the same inference pipeline to all samples regardless of their difficulty. As shown in Figure 1b, even with a 70% top-1 retrieval accuracy, the confidence scores vary significantly across categories, indicating that sample difficulty is highly uneven. Applying uniform processing wastes computation on easy cases while providing insufficient analysis for hard ones. Moreover, for simple samples, excessive reasoning may introduce overthinking and hallucination (Zhong et al., 2024; Li et al., 2025a), actually degrading performance. Second, existing methods are designed to be stateless: unlike human experts who learn from past mistakes to refine their judgment criteria, these methods cannot accumulate experience across instances, leading to repeated mistakes in similar scenarios.

To address the aforementioned limitations, we draw inspiration from how humans learn and handle complex visual tasks. On the one hand, Humans rely on a dual-system mechanism (Li et al., 2025c) to cope with varying recognition difficulty: System 1 performs rapid, intuitive judgments based on salient cues and is effective for clear, easily distinguishable cases, whereas System 2 engages in deliberate, nuanced reasoning to resolve ambiguous or highly similar examples. On the other hand, humans continuously distill experience from errors, forming transferable decision criteria that prevent repeated mistakes in similar situations.

Building on these insights, we propose **SARE**, a **S**ample-wise **A**daptive **RE**asoning framework designed for training-free FGVR. First, we conceptualize the image-text retrieval and matching process as **System 1**: it operates rapidly and intuitively, leveraging global visual-semantic alignment to handle straightforward samples efficiently. Subsequently, we formulate the LVLM-based nuanced reasoning process as **System 2**: it engages in deliberate, step-by-step analysis to resolve samples where subtle inter-class differences demand focused attention on discriminative details.

The synergy between the dual-system is crucial. System 1 solely cannot handle hard samples with subtle visual differences, while System 2 applied uniformly to all samples not only incurs prohibitive computational costs but also introduces overthinking on easy cases, potentially causing hallucination and degraded accuracy. To this end, SARE introduces a statistics-based trigger that dynamically determines whether System 2 reasoning is required for each sample. This trigger accounts for three factors: the model's confidence on the current sample, the historical difficulty of its category, and the ambiguity among candidate options, thereby routing each sample to the appropriate processing depth.

To support both efficient retrieval and robust reasoning, we construct three lightweight knowledge bases from a small set with labeled category names: (1) A multimodal prototype library containing textual and visual prototypes for fast retrieval; (2) A statistical retrieval library that records category-level retrieval statistics, enabling category-aware and robust trigger decisions; (3) An experience library derived from model self-reflection to guide reasoning in challenging scenarios. When System 2 is activated, it receives the Top-K_c candidates from System 1 along with guidance from the experience library, enabling the LVLM to focus on truly discriminative fine-grained cues and behave more like a domain expert. In a nutshell, main contributions are summarized as:

- We propose SARE, a sample-wise adaptive reasoning framework for training-free FGVR. SARE departs from uniform inference by dynamically adapting its inference strategy to sample difficulty, improving efficiency and accuracy without any parameter updates.
- SARE effectively synergizes fast retrieval (System 1) with nuanced reasoning (System

2) via a dynamic trigger mechanism. Furthermore, it incorporates a self-reflective experience module that learns from past errors to distill reusable discriminative guidance, enabling the system to continuously improve through accumulated experience.

- Extensive experiments across 14 datasets demonstrate the superiority of SARE.Notably, it outperforms leading training-free methods by over 8% and training-based baselines by 1.64%. Meanwhile, SARE also exhibits superior robustness on out-of-distribution tasks and reduces computational overhead.

2 Methodology

We present SARE, a sample-wise adaptive reasoning framework tailored to training-free FGVR. SARE mainly addresses two fundamental issues identified in prior work: the highly uneven difficulty of fine-grained instances and the stateless inference paradigm that fails to leverage past errors. As illustrated in Figure 2, SARE explicitly decouples fast perception from nuanced reasoning. Each sample is first processed by a lightweight retrieval module(System 1), and a statistics-driven trigger adaptively determines whether additional reasoning is required. For ambiguous cases, SARE invokes experience-guided reasoning(System 2), where reusable discriminative experience distilled from historical failures is injected to support reliable fine-grained decisions.

2.1 Knowledge Base Construction

To support sample-wise adaptive inference without parameter updates, SARE constructs lightweight and reusable knowledge offline from a k_{shot} subcategory support set $\mathcal{D}_{\text{kshot}}$. This knowledge provides how perception, uncertainty estimation, and reasoning are performed at inference time. Specifically, we build three complementary knowledge bases with non-overlapping roles.

Multimodal Prototype Library. Fast perception requires a compact and stable representation that supports reliable early decisions. Purely visual similarity is often unstable under large intra-class variations, while textual semantics provide complementary global structure. We therefore construct multimodal prototypes to support efficient retrieval.

Specifically, for category $c \in D_{kshot}$, we compute: (i) a visual prototype by averaging CLIP image embeddings; (ii) a textual prototype encoded from an LVLM-generated fine-grained description of the category. These prototypes form the visual prototype set P_v and textual prototype set P_t, respectively. They are then used exclusively by the fast retrieval module (System 1) without any additional model overhead.

Statistical Retrieval Library. While System 1 produces similarity-based confidence scores, their reliability varies significantly across subcategories (Zhang et al., 2022), rendering static thresholds unsuitable (Figure 1b). To calibrate these variations, we construct a statistical retrieval library that captures class-conditional retrieval behavior. We perform retrieval on D_{kshot} using the multimodal prototypes P_v and P_t. For category c, we record its retrieval count n_c and the number of successful matches N, forming a class-conditional retrieval history. These statistics serve as class-conditional priors for uncertainty estimation and are used by the dynamic trigger in Formula 6.

Self-Reflective Experience Library. To mitigate repeated errors on challenging cases, we introduce a self-reflective experience library that consolidates past errors into reusable discriminative guidance. We formalize each inference on the D_{kshot} as a trajectory:

$$\mathcal{T} = (I, \mathcal{C}, \tau, \hat{y}, y), \tag{1}$$

where I denotes the query image, $\mathcal{C}$ is the retrieved candidate set, τ is the reasoning path, and $\hat{y}$ and y are the predicted and ground-truth labels, respectively. When an error occurs (i.e., $\hat{y} \neq y$), a retrospective analysis is triggered to identify overlooked discriminative cues from τ. These cues are abstracted into compact, structured decision rules and stored as experience entries $\mathcal{E} = \{e_i\}$. To keep the experience library compact and informative, we apply a lightweight maintenance policy that merges semantically complementary rules, filters redundant ones, and replaces low-value entries with more informative experience. During inference, relevant experience entries are retrieved as contextual guidance for fine-grained reasoning, enabling the model to avoid repeating similar mistakes without any parameter updates. For details, please refer to the Appendix B.2.

2.2 Sample-wise Adaptive Inference

FGVR datasets exhibit highly uneven sample difficulty: many instances can be resolved using coarse visual cues, while others require careful inspection of subtle, localized attributes. Applying a

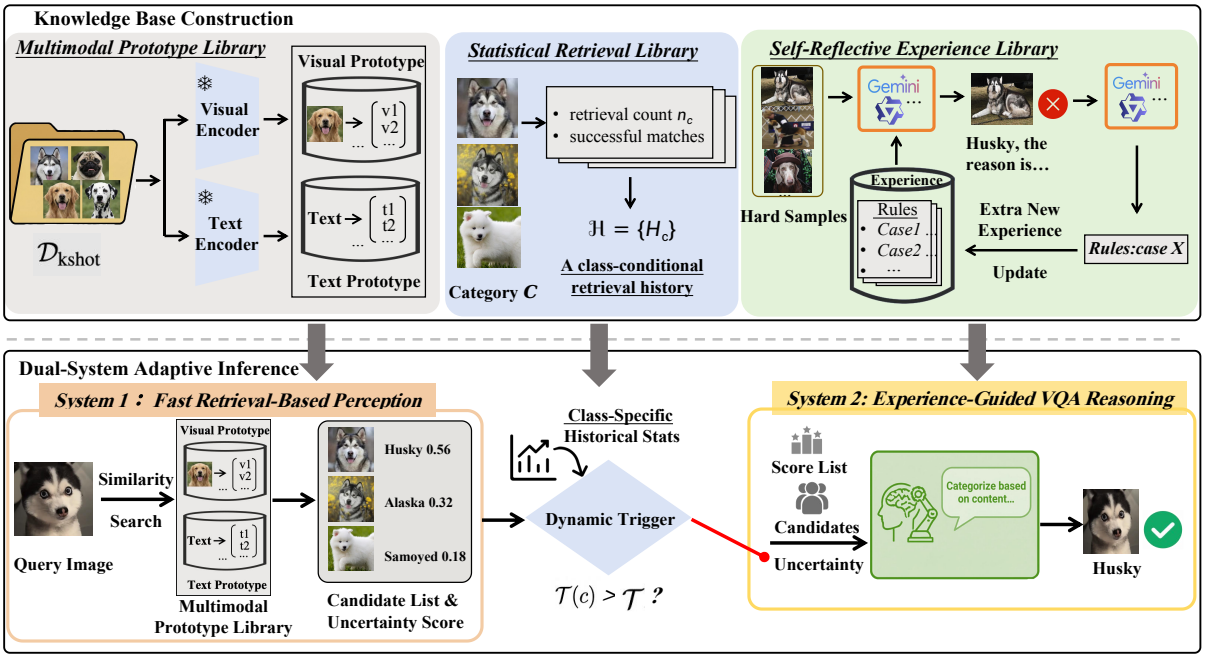


Figure 2: **Overview of SARE**. The framework performs fast prototype-based retrieval to generate candidate categories for a query image, followed by a class-conditional trigger that adaptively invokes fine-grained reasoning when retrieval confidence is insufficient. Experience distilled from past errors is injected as contextual guidance, enabling accurate and efficient training-free FGVR.

uniform pipeline is therefore structurally inefficient and risks either wasting computation on easy samples or under-analyzing hard ones. SARE adopts a sample-wise adaptive inference mechanism, inspired by the observation that LVLMs often achieve low Top-1 but high Top-K_c accuracy on fine-grained tasks (Yang et al., 2026). This suggests that retrieval can reliably narrow down candidates, while reasoning should be reserved for genuinely ambiguous cases. This observation motivates a clear separation between fast perceptual filtering and deliberate fine-grained reasoning.

System 1: Fast Retrieval-based Perception. System 1 is a lightweight perception module that performs nearest-neighbor matching between the query image and the multimodal prototypes P_v and P_t. It involves only similarity computation and incurs minimal computational overhead.

System 1 serves two purposes: (1) it enables early exit for samples with clear decision boundaries; and (2) it produces a compact Top-K_c candidate set $\mathcal{C} = \{c_1, \ldots, c_K\}$ along with similarity-based confidence scores. These outputs provide the necessary signals for the subsequent trigger to assess reliability and decide whether further reasoning is required.

System 2: Experience-guided Nuanced Reasoning. System 2 is activated only for ambiguous cases with low retrieval confidence and is formulated as a VQA inference. Specifically, System 2 takes the input image I, the Top-K_c candidate set $\mathcal{C}$, and a small set of retrieved experience entries $\mathcal{E}$ as contextual input to the LVLM. The prediction is formalized as:

$$\hat{y} = f(I, \mathcal{C}, \mathcal{E}). \tag{2}$$

By constraining the candidate space and injecting structured experience, System 2 is guided to attend to truly discriminative fine-grained cues, enabling precise reasoning in similar ambiguous scenarios without resorting to repeated model invocations or costly multi-step prompting. We provide the complete prompt templates and implementation details in Appendix B to ensure reproducibility.

2.3 Statistics-based Dynamic Trigger

The dynamic trigger serves as the decision-making core of SARE, managing the transition from System 1 to System 2. Its primary objective is to selectively allocate expensive computational resources to genuinely ambiguous cases while preserving the efficiency of rapid perception for simple samples.

Given the Top-K_c retrieved category candidates from System 1, our goal is to assess whether the

top-1 prediction is sufficiently reliable. In practice, System 1 may fail under three common conditions: (1) calibration mismatch between visual and textual similarities; (2) category-dependent confidence bias from VLM embeddings (Figure 1b); (3) high ambiguity among candidates. To address cross-modal calibration mismatch, we first normalize visual and textual similarities using temperature-scaled Softmax to obtain distributions P_{img} and P_{text}, which are linearly fused as:

$$P_{\text{fuse}} = \lambda P_{img} + (1 - \lambda) P_{text}, \tag{3}$$

where λ is set to 0.3. Since probability values can still be poorly calibrated across modalities, we further incorporate Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) to exploit robust rank-level agreement. For each category $c \in C$, we compute an RRF score based on its rank $r_m(c)$:

$$\mathcal{R}(c) = \sum_{m \in \{v,t\}} \frac{1}{\kappa + r_m(c)}, \tag{4}$$

where $r_m(c)$ denotes the rank of category c within modality m, $\kappa = 60$ is a smoothing constant. This metric is robust to score outliers as it relies solely on rank positions. The final retrieval confidence for category c is defined as

$$\hat{p}_c = P_{\text{fuse}}(c) + \beta \cdot \mathcal{R}(c), \tag{5}$$

where $\beta = 0.1$. Based on the fused confidence, we define a trigger score for the Top-1 candidate as

$$\mathcal{G}(c) = \hat{p}_c - \eta \sqrt{\frac{\log N}{2n_c}} - \alpha H(\mathbf{p}_c), \tag{6}$$

where each term captures a distinct source of uncertainty in System 1. Specifically, $\hat{p}_c$ reflects the fused retrieval confidence. The second term, derived from Hoeffding's inequality (Sec. 2.1), penalizes categories with limited historical retrieval evidence (n_c), preventing overconfident decisions on statistically unreliable candidates. The third term $H(\mathbf{p}_c)$ measures ambiguity among the Top-K_c candidates: high entropy indicates that System 1 fails to clearly separate competing hypotheses, even when absolute confidence is high.

If $\mathcal{G}(c)$ exceeds a predefined threshold θ, the prediction from System 1 is accepted; otherwise, System 2 is activated for fine-grained reasoning. This statistics-driven trigger allows SARE to selectively allocate computation to ambiguous cases, effectively decoupling inference cost from recognition difficulty.

3 Experiments

3.1 Setup

Datasets. We carefully select 14 datasets to systematically assess its fine-grained discrimination ability, general recognition performance, and robustness under distribution shifts: (1) Fine-grained recognition (7 datasets): We select widely-used benchmarks that cover diverse domains and subtle category distinctions, including CUB-200-2011 (Wah et al., 2011), Stanford Dogs (Khosla et al., 2011), Stanford Cars (Krause et al., 2013), FGVC-Aircraft (Maji et al., 2013), Oxford-IIIT Pet (Parkhi et al., 2012), Oxford 102 Flowers (Nilsback and Zisserman, 2008) and Birdsnap (Berg et al., 2014). (2) General recognition (5 datasets) Beyond fine-grained tasks, we evaluate general visual recognition on Food-101 (Bossard et al., 2014), ImageNet-1K (Deng et al., 2009), Describable Textures Dataset (DTD) (Cimpoi et al., 2014), SUN397 (Xiao et al., 2014), and UCF101 (Soomro et al., 2012). (3) Robustness evaluation (2 datasets): To probe performance under distribution shifts, we incorporate challenging ImageNet variants: ImageNet-V2 (Recht et al., 2019), and ImageNet-Sketch (Wang et al., 2019). Detailed dataset descriptions are provided in Appendix A.1.

Baseline. We compare SARE with several representative SoTA FGVR methods, selected to ensure fair and comprehensive coverage of both training-free and training-based paradigms. (i) Training-free methods contain retrieval-based baselines, such as Sus-X-LC (Udandarao et al., 2023), FineR (Xu et al., 2024), E-FineR (Demidov et al., 2025), AWT (Zhu et al., 2024), ProtoMM (Zhu et al., 2025b), CascadeVLM (Wei, 2024) and SCAN (Yang et al., 2026), as well as reasoning-oriented frameworks including UniFGVR (Guo et al., 2025), AutoSEP (Hong et al., 2025), and MCQA (Atabuzzaman et al., 2025). (ii) Training-based methods are also selected as reference baselines, including two of the most representative cross-modal alignment methods, i.e., ternary contrastive learning FineDefics (He et al., 2025) and volume alignment VT-FSL (Li et al., 2025b).

Implementation details. We use Qwen2.5-VL-7B (Wang et al., 2024) as the LVLM backbone for textual prototype generation, experience construction, and System 2 reasoning. For System 1, CLIP-B/32 (Radford et al., 2021) encodes visual inputs and textual prototypes for multimodal prototype construction and candidate retrieval. We

Method	Fine-Grained Visual Recognition (FGVC)							General Recognition					Avg.	Δ
	Dogs	Flowers	CUB	Cars	Pets	Air.	Bird.	Food	DTD	IN-1k	SUN	UCF		
Backbone														
CLIP-B/32	52.09	63.44	53.26	62.38	81.66	23.67	42.37	80.32	44.26	63.67	62.32	64.13	57.80	-2.80
Qwen2.5-VL-7B	65.90	68.50	43.70	75.31	85.12	54.01	54.48	69.90	68.83	68.40	58.11	73.30	65.46	+4.86
Training-free LVLM Methods														
(Retrieval-based Methods)														
SuS-X-LC	62.35	72.17	55.02	66.32	86.89	27.96	46.83	84.32	52.34	66.87	68.02	66.22	62.94	+2.34
FineR	52.88	67.67	56.75	65.36	83.24	24.96	43.69	81.31	49.76	68.65	65.72	67.26	60.60	+0.00
E-FineR	52.72	68.96	57.89	65.89	82.96	25.66	44.32	81.56	50.27	69.02	66.96	67.74	61.16	+0.56
RAR	53.29	67.99	57.01	66.01	84.37	24.39	43.87	81.65	49.57	70.26	67.09	67.01	61.04	+0.44
AWT	68.35	76.13	59.54	70.03	90.76	30.93	48.63	85.58	55.20	71.32	70.33	69.12	66.33	+5.73
ProtoMM	68.63	77.40	60.04	69.92	91.29	31.02	48.79	85.89	56.38	72.76	70.67	70.62	66.95	+6.35
SCAN	69.98	77.86	64.83	77.96	89.78	39.68	53.57	86.37	62.65	72.14	72.83	73.71	70.11	+9.51
(Reasoning-based Methods)														
AutoSEP	67.82	70.63	63.19	68.12	86.49	58.76	60.34	84.75	61.47	71.45	71.92	67.53	69.37	+8.77
CascadeVLM	64.54	74.23	60.26	79.26	86.17	29.18	56.27	84.51	53.24	73.98	69.96	69.05	66.72	+6.12
UniFGVC	65.23	**95.84**	78.80	94.60	90.92	61.10	66.25	82.30	73.90	81.10	76.60	80.90	78.96	+18.36
Training-based LVLM Methods														
FineDefics	73.36	87.81	55.46	85.01	84.73	64.08	45.72	84.26	63.53	71.41	68.90	75.97	71.69	+11.09
VT-FSL	**84.63**	93.64	**90.98**	92.75	88.92	80.31	76.57	**91.32**	76.31	**89.75**	**81.98**	85.37	86.04	+25.44
Ours (SARE)	84.29	88.31	90.76	**99.34**	**95.38**	**83.02**	**87.30**	88.02	**83.10**	85.06	80.05	**87.58**	**87.68**	**+27.08**

Table 1: Comparison with state-of-the-art methods on 7 fine-grained and 5 general datasets. We highlight the **best** results in bold and the second-best results with underlining. "Avg." denotes the average accuracy across all 12 datasets. Δ indicates the performance gain compared to FineR.

set the number of candidate categories $K_c = 10$, experience $\mathcal{E} = 8$ and $k_{shot} = 3$. We emphasize that all methods' settings are strictly consistent. All experiments were conducted on a single NVIDIA RTX A800 GPU with fixed random seeds.

3.2 Main Results

Best overall performance. As shown in Table 1, SARE reaches an average accuracy of 87.68%, substantially outperforming all training-free baselines and surpassing strong training-based methods such as FineDefics by 15.88% and VT-FSL by 1.64%. Unlike prior approaches that exhibit pronounced performance fluctuations across domains (e.g., UniFGVC), SARE consistently exhibits high performance on both fine-grained and general datasets, indicating robust cross-domain generalization. The improvements are particularly notable on highly ambiguous fine-grained datasets (e.g., Aircraft, Birdsnap and Dogs), where recognition relies on subtle inter-class differences. These gains underscore the advantage of experience-guided reasoning, as the distilled experience help the model focus on truly decisive attributes. To further analyze performance under varying difficulty levels, we selecte two datasets based on their intrinsic difficulty (Table 4): Stanford Cars (relatively easy) and Stanford Dogs (challenging). As illustrated in Figure 3 and Figure 8 (In Appendix), SARE not only improves accuracy but also achieves a superior accuracy–efficiency trade-off, dynamically allocating more computation to challenging samples while reducing effort for easier ones.

4 Analysis

4.1 Effectiveness of Dynamic Trigger

Efficient difficulty adaptation. As shown in Figure 4, we analyze the performance of the dynamic trigger. SARE dynamically adjusts the activation rate of System 2 according to sample difficulty. When recognition is easy, the trigger rate remains low. As difficulty increases, the trigger rate rises accordingly, allocating more samples to fine-grained reasoning. This strong correlation indicates that the dynamic trigger effectively learns *when* additional reasoning is necessary, rather than blindly increasing computation. Table 5 shows that samples identified as requiring only System 1 achieve very high accuracy, demonstrating that the trigger maintains both efficiency and reliable recognition.

4.2 Ablation Studies

Component Effectiveness. As shown in Table 2, relying solely on retrieval-based System 1 yields

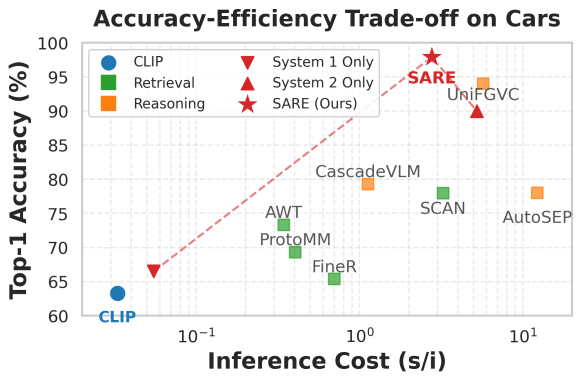


Figure 3: Comparison of SARE against baselines on Stanford Cars dataset. SARE achieves the optimal balance, significantly outperforming baselines in accuracy with lower inference overhead.

System 1	System 2	Experience	Dogs	CUB
✓			58.24	73.02
	✓		78.61	78.43
✓	✓		76.01	81.36
	✓	✓	81.34	88.21
✓	✓	✓	**84.29**	**90.76**

Table 2: Ablation study of key components in SARE.

Method	ImageNet-V2 Acc. (%)	ImageNet-V2 Δ	ImageNet-Sketch Acc. (%)	ImageNet-Sketch Δ
$SARE_{tgt}$	84.43	-	72.49	-
$SARE_{src}$	83.72	**-0.69**	71.96	**-0.53**

Table 3: Comparison of SARE with experience constructed on ImageNet-1K ($SARE_{src}$) versus experience constructed on target-domain experience ($SARE_{tgt}$).

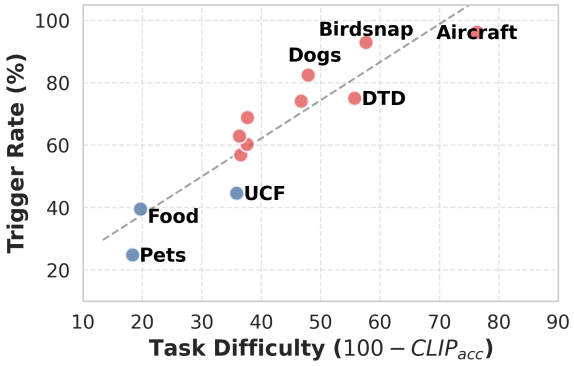


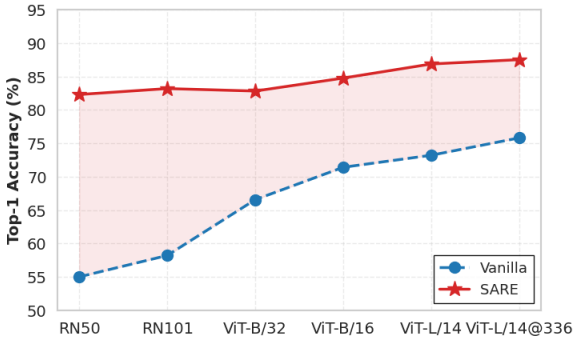
Figure 4: The proportion of samples triggering System 2 across datasets with varying recognition difficulty. The x-axis measures dataset-level difficulty using $100\% - CLIP_{Top-1}$ accuracy.

an accuracy of 58.24%, indicating that global semantic alignment struggles with subtle inter-class differences. Incorporating reasoning-based System 2 boosts performance to 76.65%. This confirms that VQA-based reasoning captures fine-grained details via attribute localization. Finally, the experience library brings a further gain, reaching the best accuracy of 84.29%. This demonstrates that experience act as effective in-context demonstrations, guiding the LVLM to discriminative regions and enhanced fine-grained recognition capability.

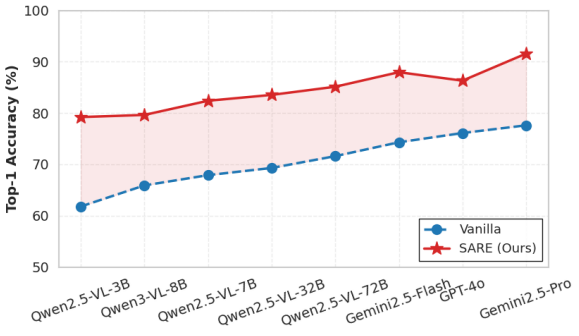
Hyperparameter Analysis. In Appendix C, we analyze SARE's sensitivity to three key hyperparameters: the candidate set size K_c, the number of retrieved experience entries $\mathcal{E}$, and the k-shot size of D_{kshot}. As shown in Figures 7 and Figure 9, SARE remains stable across a wide range of values, achieving the best performance at $K_c = 10$ and $\mathcal{E} = 8$, and exhibiting strong robustness to variations in k_{shot}. In Appendix D, we evaluate the impact of different sampling strategies for k_{shot} in D_{kshot}, The results show that SARE maintains stable performance across six distinct strategies.

4.3 Performance on different Backbones

Strong model-agnostic robustness. To assess the model-agnostic property of SARE, we conduct controlled experiments by separately varying the visual backbone (VLM) and the reasoning backbone (LVLM). Specifically, when evaluating different VLMs, we fix the LVLM as Qwen2.5-VL-7B; when evaluating different LVLMs, we fix the visual encoder as CLIP ViT-B/32. As shown in Figure 5a and Figure 5b, SARE consistently delivers strong performance across all evaluated backbones, demonstrating its robustness and model-agnostic applicability. Across different VLMs, SARE yields stable performance improvements regardless of backbone capacity. Notably, the relative gains are more pronounced for weaker visual backbones (e.g., RN50 and RN101) and gradually diminish as the base representations become stronger. This trend indicates that SARE effectively compensates for limited perceptual capacity rather than relying on architecture-specific characteristics. A similar pattern is observed for LVLMs. Although larger models achieve higher absolute accuracy, SARE consistently enhances performance across different model scales, with more substantial gains on smaller LVLMs. Overall, these results indicate that SARE provides complementary reasoning benefits rather than relying on backbone-specific characteristics, demonstrating strong robustness and model-agnostic applicability.



(a) VLM Backbone Performance



(b) LVLM Backbone Performance

Figure 5: Performance of SARE across different backbone architectures. SARE consistently enhances performance across all backbones, with larger relative gains on lower-performing models.

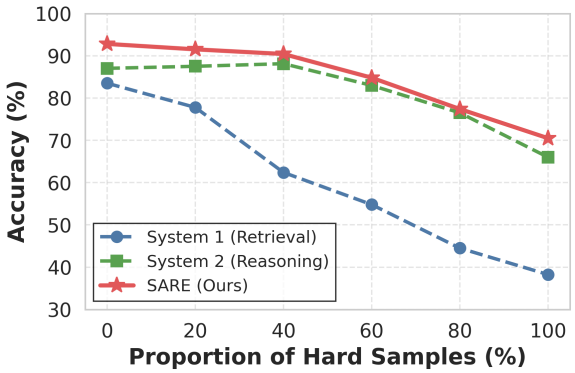


Figure 6: Behavior of different inference strategies under increasing sample difficulty on Stanford Dogs.

4.4 Transferability Analysis

Experience is transferable. To examine this, we perform cross-domain evaluations on ImageNet-V2 and ImageNet-Sketch. We build the experience library on ImageNet-1K and directly apply it to the target domains, denoted as $SARE_{src}$, and compare it with a target-domain upper bound where the experience library is constructed directly from the target dataset, denoted as $SARE_{tgt}$. As shown in Table 3, $SARE_{src}$ achieves performance comparable to $SARE_{tgt}$, with only marginal accuracy degradation, even under significant distribution and style shifts. This indicates that the learned experience captures transferable discriminative cues, enabling the model to resolve a family of similar challenging cases rather than overfitting to domain-specific visual patterns.

4.5 Behavioral Analysis

Mitigating overthinking & underthinking. To analyze how different inference strategies behave under varying sample difficulty, we perform a controlled study on the Stanford Dogs dataset by splitting it into *easy* and *hard* subsets based on retrieval correctness, and then creating five test sets with gradually increasing proportions of hard samples, from 0% to 100%. As shown in Figure 6, System 1's overall performance steadily declines with higher difficulty, reflecting its limited ability to handle highly ambiguous instances. System 2, by contrast, shows a counterintuitive accuracy increase when hard samples constitute from 0 to 40% of the test set, which we attribute to overreasoning on simpler examples that can sometimes degrade performance. SARE, however, consistently achieves the highest accuracy across all difficulty levels, demonstrating its effectiveness in avoiding overthinking on easy samples while effectively addressing the challenging ones.

4.6 Qualitative Analysis and Case Studies

A qualitative case study illustrating how experience is constructed and used to guide model reasoning is presented in Appendix E. Figure 11 provides a heatmap highlighting how SARE influences and guides the model's decisions.

5 Conclusion

We present SARE, a sample-wise adaptive reasoning framework for training-free FGVR. It adaptively integrates fast retrieval (System 1) with fine-grained VQA-style reasoning (System 2) via a sample-wise trigger. The trigger invokes reasoning only when retrieval is unreliable. A self-reflective experience mechanism abstracts past errors into transferable discriminative rules, guiding reasoning toward key discriminative cues without any parameter updates. SARE consistently achieves SoTA results across 14 datasets while reducing time costs, and the rationality of SARE's design has been verified through extensive experiments.

Limitations

Despite demonstrating state-of-the-art performance and efficiency in training-free FGVR, we emphasize that SARE has some limitations. First, although deriving transferable guidance from past errors is a promising design, further improving the quality of this experience would require additional self-reflection, inevitably introducing extra computational overhead. Second, while the trigger mechanism is effective, SARE adopts a conservative triggering strategy for very challenging datasets (e.g., Stanford Dogs, FGVC-Aircraft), resulting in a higher-than-expected activation rate. We plan to explore more accurate and efficient dynamic triggers in future work. Finally, SARE primarily enhances LVLMs from a textual context perspective, future work could explore complementary strategies that leverage visual features to further improve fine-grained recognition.

References

Md Atabuzzaman, Andrew Zhang, and Chris Thomas. 2025. Zero-shot fine-grained image classification using large vision-language models. *arXiv preprint arXiv:2510.03903*.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2018.

Lorenzo Bianchi, Fabio Carrara, Nicola Messina, and Fabrizio Falchi. 2024. Is clip the main roadblock for fine-grained open-world perception? *Preprint*, arXiv:2404.03539.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, pages 446–461.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *Preprint*, arXiv:2305.05176.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *Preprint*, arXiv:2501.17811.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, and Marcel Blistein. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. 2024. Vocabulary-free image classification. *Preprint*, arXiv:2306.00917.

Alessandro Conti, Massimiliano Mancini, Enrico Fini, Yiming Wang, Paolo Rota, and Elisa Ricci. 2025. On large multimodal models as open-world image classifiers. *Preprint*, arXiv:2503.21851.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Dmitry Demidov, Zaigham Zaheer, Omkar Thawakar, Salman Khan, and Fahad Shahbaz Khan. 2025. Vocabulary-free fine-grained visual recognition via enriched contextually grounded vision-language model. *arXiv preprint arXiv:2507.23070*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Peng Gao, Shijie Geng, Renrui Zhang, Tianshu Ma, Rongyao Fang, Yong Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language modeling with feature adapters. *International Journal of Computer Vision (IJCV)*.

Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024. African or european swallow? benchmarking large vision-language models for fine-grained object classification. *Preprint*, arXiv:2406.14496.

Hongyu Guo, Kuan Zhu, Xiangzhao Hao, Haiyun Guo, Ming Tang, and Jinqiao Wang. 2025. Unifgvc: Universal training-free few-shot fine-grained vision classification via attribute-aware multimodal retrieval. *arXiv preprint arXiv:2508.04136*.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. 2025. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. *arXiv preprint arXiv:2501.15140*.

Yunqi Hong, Sohyun An, Andrew Bai, Neil Y.C. Lin, and Cho-Jui Hsieh. 2025. Unlabeled data improves fine-grained image zero-shot classification with multimodal llms. *arXiv preprint arXiv:2506.03195*.

Raihan Kabir, Naznin Haque, Md Saiful Islam, and Marium-E-Jannat. 2024. A comprehensive survey on visual question answering datasets and algorithms. *Preprint*, arXiv:2411.11150.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. *Preprint*, arXiv:2210.03117.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO.

Jeonghwan Kim and Heng Ji. 2024. Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 6187–6207. Association for Computational Linguistics.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 554–561.

Hari Chandana Kuchibhotla, Sai Srinivas Kancheti, Abbavaram Gowtham Reddy, and Vineeth N Balasubramanian. 2025. Efficient vocabulary-free fine-grained visual recognition in the age of multimodal llms. *arXiv preprint arXiv:2505.01064*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

Jiansheng Li, Xingxuan Zhang, Hao Zou, Yige Guo, Renzhe Xu, Yilong Liu, Chuzhao Zhu, Yue He, and Peng Cui. 2025a. Counts: Benchmarking object detectors and multimodal large language models under distribution shifts. *Preprint*, arXiv:2504.10158.

Wenhao Li, Qiangchang Wang, Xianjing Meng, Zhibin Wu, and Yilong Yin. 2025b. Vt-fsl: Bridging vision and text with llms for few-shot learning. *Preprint*, arXiv:2509.25033.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, and 2 others. 2025c. From system 1 to system 2: A survey of reasoning large language models. *Preprint*, arXiv:2502.17419.

Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. 2023. Learning customized visual models with retrieval-augmented knowledge. *Preprint*, arXiv:2301.07094.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024b. Rar: Retrieving and ranking augmented mllms for visual recognition. *Preprint*, arXiv:2403.13805.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

M. Jehanzeb Mirza, Mengjie Zhao, Zhuoyuan Mao, Sivan Doveh, Wei Lin, Paul Gavrikov, Michael Dorkenwald, Shiqi Yang, Saurav Jha, Hiromi Wakaki, Yuki Mitsufuji, Horst Possegger, Rogerio Feris, Leonid Karlinsky, and James Glass. 2025. Glov: Guided large language models as implicit optimizers for vision language models. *Preprint*, arXiv:2410.06154.

Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. 2024. Sparse attention vectors: Generative multimodal model features are discriminative vision-language classifiers. *arXiv preprint arXiv:2412.00142*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2025. Routellm: Learning to route llms with preference data. *Preprint*, arXiv:2406.18665.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and 1 others. 2012. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505.

Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibo Yin, and Hulingxiao He. 2026. A survey on fine-grained multimodal large language models. *Chinese Journal of Electronics*. Accepted.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, pages 5389–5400.

Sheng Shen, Chunyuan Li, Xiaowei Hu, Jianwei Yang, Yujia Xie, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, Kurt Keutzer, Trevor Darrell, Anna Rohrbach, and Jianfeng Gao. 2022. *Preprint*, arXiv:2204.09222. [link].

Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, and Ninghao Liu. 2025. Enhancing cognition and explainability of multimodal foundation models with self-synthesized data. *Preprint*, arXiv:2502.14044.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *Preprint*, arXiv:1212.0402.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. 2023. Sus-x: Training-free name-only transfer of vision-language models. *Preprint*, arXiv:2211.16198.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Canshi Wei. 2024. Enhancing fine-grained image classifications via cascaded vision language models. *Preprint*, arXiv:2405.11301.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. 2021. Fine-grained image analysis with deep learning: A survey. *Preprint*, arXiv:2111.06119.

Xiu-Shen Wei, Jianxin Wu, and Quan Cui. 2019. Deep learning for fine-grained image analysis: A survey. *Preprint*, arXiv:1907.03069.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. 2025. When language overrules: Revealing text dominance in multimodal large language models. *Preprint*, arXiv:2508.10552.

Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. 2014. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119:3–22.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. 2025. Fg-clip: Fine-grained visual and textual alignment. *Preprint*, arXiv:2505.05071.

Zhicai Xu, Yanqing Lai, Tao Chen, Kai Cui, Yazhou Wang, and Chee-Meng Vong. 2024. Democratizing fine-grained visual recognition with large language models. *arXiv preprint arXiv:2401.13837*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yutong Yang, Lifu Huang, Yijie Lin, Xi Peng, and Mouxing Yang. 2026. Endowing vision-language models with system 2 thinking for fine-grained visual recognition. In *AAAI*.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2022. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Preprint*, arXiv:2110.08263.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *Preprint*, arXiv:2111.03930.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. 2024a. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. *Preprint*, arXiv:2406.07057.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024b. Why are visually-grounded language models bad at image classification? *Preprint*, arXiv:2405.18415.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024c. Why are visually-grounded language models bad at image classification? *Preprint*, arXiv:2405.18415.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. *Preprint*, arXiv:2407.00569.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. *Preprint*, arXiv:2203.05557.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Shiding Zhu, Wenhui Dong, Jun Song, Yingbo Wang, Yanan Guo, and Bo Zheng. 2025a. Fila: Fine-grained vision language models. *Preprint*, arXiv:2412.08378.

Xingyu Zhu, Shuo Wang, Beier Zhu, Miaoge Li, Yunfan Li, Junfeng Fang, Zhicai Wang, Dongsheng Wang, and Hanwang Zhang. 2025b. Dynamic multimodal prototype learning in vision-language models. *Preprint*, arXiv:2507.03657.

Yuhan Zhu, Yuyang Ji, Zhiyu Zhao, Gangshan Wu, and Limin Wang. 2024. Awt: Transferring vision-language models via augmentation, weighting, and transportation. *Preprint*, arXiv:2407.04603.

Appendix

An anonymized code repository is provided: https://anonymous.4open.science/r/SARE-8E56/.

The appendix is organized as follows:

- Section A summarizes statistics and descriptions of the 14 datasets used in our experiments.
- Section B details the prompt templates for Knowledge Base construction, System 2 reasoning, and the Self-Reflective mechanism.
- Section C presents analysis of key hyperparameters, including candidate count (K_c), experience capacity ($\mathcal{E}$), and k_{shot} size.
- Section D evaluates the effect of different k-shot support set sampling strategies on performance.
- Section E provides qualitative case studies illustrating how the Experience Library guides model reasoning.
- Section F provides related work for this work.

A Experimental Setup

A.1 Dataset Details

We follow the few-shot experimental protocol established in FineR. We utilize the official test split of each dataset for performance evaluation to report the results. Instead of using the full training set, we construct the few-shot support set (used for Knowledge Base construction in SARE) by randomly sampling K images per category from the official training split. In our main experiments, we set the shot number $K = 3$ and ensure strict data separation where $\mathcal{D}_{\text{support}} \cap \mathcal{D}_{\text{test}} = \emptyset$. Table 6 summarizes the detailed statistics of the 14 datasets used in our experiments, including the number of categories and the size of the official test sets used for evaluation. For FGVC benchmarks (e.g., Stanford Dogs), we follow the standard splits adopted in prior work, and results on UCF101 are reported on Split-1.

B Prompt Templates

B.1 Knowledge Base and Reasoning Prompt

This section details the prompt templates used for (1) generating the initial multimodal knowledge base and (2) performing the final System 2 adaptive reasoning. Note that the prompts for the *Self-Reflective Experience Construction* are detailed separately in the subsequent Appendix B.2 due to their iterative nature.

Textual Prototype

```
You are an expert taxonomist specializing in fine-grained visual recognition.
Input:
• Category: {category_name}
• Reference: [Set of Support Images]
Task: Generate a comprehensive and discriminative description that captures the key visual characteristics that distinguish this category from other similar categories.
Focus on:
1. Distinctive physical features
2. Color patterns and markings
3. Size and proportions
4. Behavioral characteristics (if applicable)
5. Unique identifying traits
Constraint: The description should be concise but informative, suitable for fine-grained visual recognition task.
```

System 2 Inference

```
You are a fine-grained recognition expert. Your task is to identify the specific sub-category of the provided image.
Context Provided:
1.Candidate Classes (highly likely to contain the correct option):
{candidate_text}
2. Expert Guidance (Retrieved Experience):
{experience_context}
Task: Please analyze the image step by step and provide:
1. Your reasoning chain (Chain-of-Thought) based on the visual evidence and expert guidance.
2. Your final prediction (only the category name).
Output Format:
Reasoning: [your step-by-step reasoning]
Prediction: [category name]
```

B.2 The implementation of Self-Reflection

The construction of the Experience Library is not a single-step generation but a closed-loop process involving reasoning, analysis, and strategy update. We detail the specific prompts for each phase below.

Dataset	Aircraft	Birdsnap	Dogs	DTD	CUB	SUN	IN-1K	Cars	Flowers	UCF	Food	Pets
Trigger Rate (%)	96.19	92.92	82.44	75.03	74.08	68.82	62.90	60.24	56.85	44.57	39.49	24.78

Table 4: **Analysis of Trigger Rates.** The percentage of samples activating the System 2 varies significantly based on dataset difficulty.

Dataset	Aircraft	Birdsnap	Dogs	DTD	CUB	SUN	IN-1K	Cars	Flowers	UCF	Food	Pets
Sys-1 Acc. (%)	88.93	96.96	91.52	100	95.78	91.98	99.97	99.37	97.94	97.74	92.81	95.43

Table 5: Reliability of System 1 Decisions. We report the recognition accuracy on the subset of samples that the dynamic trigger determined *did not* require System 2 reasoning. The consistently high accuracy across datasets demonstrates that our trigger effectively identifies easy samples where retrieval-based perception is sufficient.

Dataset	# Cls.	# Test Size	Domain
Fine-Grained Visual Recognition			
CUB-200-2011	200	5,794	Birds
Stanford Dogs	120	8,580	Dogs
Stanford Cars	196	8,041	Cars
FGVC-Aircraft	100	3,333	Aircraft
Oxford-IIIT Pet	37	3,669	Animals
Oxford 102 Flowers	102	6,149	Flowers
Birdsnap	500	2,443	Birds
General Visual Recognition			
Food-101	101	25,250	Food
ImageNet-1K	1k	50,000	Objects
DTD (Textures)	47	1,880	Textures
SUN397	397	19,850	Scenes
UCF101	101	3,783	Actions
Out-of-Domain Datasets			
ImageNet-V2	1k	10,000	Objects
ImageNet-Sketch	1k	50,889	Sketches

Table 6: Detailed statistics of the 14 datasets used in our experiments. To save space, # Cls. denotes the number of categories.

Step 1: Initial Self-Belief Reasoning

```
You are an expert in fine-grained visual recognition. Please follow these steps to identify the object:
1. Observe: Look at the overall object and identify its coarse category.
2. Localize: Identify the most discriminative local parts.
3. Compare: Recall visual characteristics of candidate subcategories.
4. Decide: Choose the most likely class based on the details.

Constraint: Answer ONLY with the final class name.
```

If the prediction is incorrect ($\hat{y} \neq y$), the specific failure diagnosis is triggered:

Step 2: Specific Failure Diagnosis

```
You are an expert in fine-grained visual recognition. Analyze this specific failure case where the model incorrectly predicted '{predicted_category}' but the correct answer is '{true_category}'.
Context:
• Model's Reasoning: {model_reasoning}
• Top Candidates: {candidates_info}
• Definition ({true_category}): {correct_category_desc}
• Definition ({predicted_category}): {predicted_category_desc}
Task: Focus ONLY on the visual evidence in this image.
1. Locate the specific region where the visual feature contradicts the prediction.
2. Compare this feature against the category definitions provided.
3. Identify the exact visual attribute (e.g., tail shape, beak color) that caused confusion.
Constraint: Do not generalize yet. Provide a detailed diagnosis of this specific image instance. Output format: Visual Evidence and Direct Cause.
```

Next, the system distills this specific diagnosis into a generalized rule:

Step 3: Abstraction & Generalization

```
You are a knowledge engineer. Your task is to distill a specific failure diagnosis into a universal, abstract rule to guide future predictions.
Input Data:
• Conflict: {true_category} vs.
```

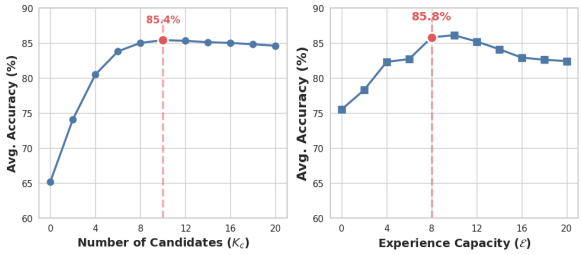


Figure 7: **Hyperparameter Sensitivity Analysis.** We analyze the impact of candidate count (K_c) and experience capacity ($\mathcal{E}$) on average accuracy across seven fine-grained benchmarks. The **line plots** reveal a clear inverted-U trend for both parameters: performance improves as context increases, peaks at the optimal settings ($K_c = 10, \mathcal{E} = 8$), and slightly declines beyond this point due to diminishing marginal utility and the introduction of semantic noise.

```
{predicted_category}
• Diagnosis: {step2_diagnosis_output}
Task: Formulate a concise, high-level
verification rule.
1. Abstract: Remove references to "this
   image". Focus on the concept.
2. Actionable: The rule should be a direct
   instruction for what to check.
3. Discriminative: Clearly distinguish the
   two categories.
Constraint: Return ONLY the rule text (under
30 words).
Example Output: "To distinguish Husky from
Malamute, check the tail curvature: Husky
tails are straight, while Malamute tails curl
over the back."
```

Finally, the model integrates the new rule to update its internal strategy:

Step 4: Self-Belief Strategy Update

```
Based on the failure analysis and new
insights, update the Self-Belief strategy.
Input:
• Current Strategy: {current_self_belief}
• New Rule/Insight: {failure_analysis}
Task: Update the strategy to:
1. Maintain the core recognition framework.
2. Add specific guidance for handling similar
   difficult cases.
3. Emphasize discriminative features that
   were previously overlooked.
Constraint: Provide only the updated
Self-Belief strategy without additional
explanation.
```

C Hyperparameter Analysis

In this section, we investigate the sensitivity of SARE to three critical hyperparameters: the number of retrieved candidates (K_c), the capacity of the Experience Library ($\mathcal{E}$) and the number of small labeled set (k_{shot}). We evaluate the accuracy across Stanford Dogs (Khosla et al., 2011). As illustrated in Figure 7, when K_c is small, the candidate set is unlikely to contain the correct label, which directly limits the upper bound of recognition accuracy. As K_c increases, performance improves and reaches its peak around $K_c = 10$, where the candidate set already includes the correct category with high probability. Further increasing K_c does not yield additional gains, as the inclusion of more candidates no longer improves coverage but instead introduces redundant alternatives. Similarly, when the experience capacity $\mathcal{E}$ is too small, System 2 lacks sufficient contextual and historical guidance to effectively correct fine-grained reasoning errors. In contrast, excessively large values of $\mathcal{E}$ introduce irrelevant semantic noise, which disperses the LVLM's attention and leads to a degradation in performance. Figure 9 demonstrate that SARE exhibits remarkable robustness to the size of the support set. Specifically, the accuracy fluctuates within a narrow margin (83%–86%) across the range of $k_{shot} = 1$ to $k_{shot} = 10$. This indicates that our framework can effectively distill discriminative knowledge even from minimal examples, without being overly sensitive to the exact number of reference images. This verified its practicability and stability in scenarios with few samples

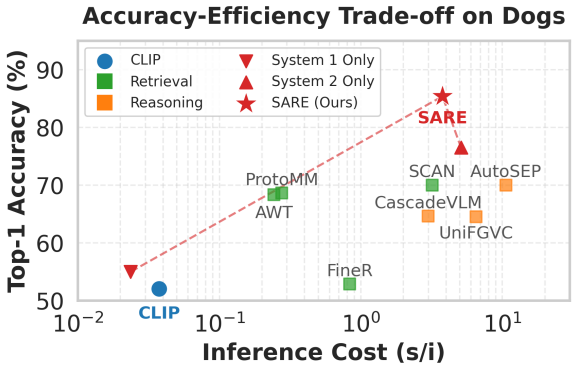


Figure 8: Comparison of SARE against baselines on Stanford Dogs dataset.

D Sensitivity Analysis

To evaluate the sensitivity of SARE to the selection of the k_{shot} of D_{kshot}, we conduct experiments on Stanford Dogs using six sampling strategies: Random, Centroid, Boundary, Entropy, Diversity, and Confusion.

Sampling Strategy	Top-1 Acc (%)
Random	84.29
Diversity	84.59
Confusion	83.56
Centroid	85.15
Entropy	85.44
Boundary	85.68

Table 7: **Impact of Sampling Strategies.** Performance of SARE on Stanford Dogs under different k-shot sampling strategies.

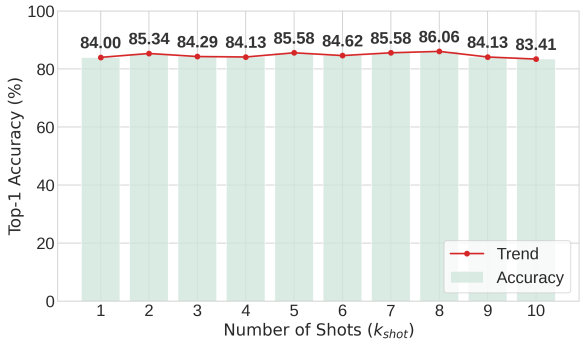


Figure 9: Sensitivity Analysis of Shot Number k_{shot}. The results demonstrate that SARE maintains robust performance across different few-shot settings. As shown in the figure, the accuracy fluctuation is minimal as k varies from 1 to 10, indicating that our method is not overly sensitive to the exact size of the support set.

As shown in Table 7, SARE exhibits stable performance across all sampling strategies, with accuracy varying within a narrow range (approximately 2%), indicating low sensitivity to the specific composition of the support set. Notably, strategies that prioritize informative or difficult samples (Boundary and Entropy) achieve slightly higher accuracy than representative or random sampling. This suggests that while SARE is robust to different sampling choices, constructing the Experience Library from more challenging samples can provide more effective supervisory signals for fine-grained reasoning.

E Case Study

In this section, we provide a detailed visualization of how the Experience Library is constructed and applied to guide fine-grained recognition. As shown in Figure 10, the left panel illustrates the *reflection phase*, where the model distills generalized discriminative rules from past misclassifications (e.g., confusing an Afghan Hound with a Collie or Golden Retriever), such as prioritize morphological features over coat color. The right panel shows the *adaptive inference phase*, where a visually ambiguous sample (e.g., a Black-and-tan Coonhound) is initially misclassified by System 1 due to misleading cues. System 2 then retrieves the relevant experience rule, refocuses on critical morphological attributes (ears), and correctly identifies the category. This example demonstrates how the Experience Library captures reusable knowledge from prior errors, enabling the model to handle complex and ambiguous scenarios effectively.

F Related Work

Fine-Grained Visual Recognition (FGVR) has recently shifted from training-centric paradigms to training-free methods that leverage Large Vision-Language Models (LVLMs) (Wei et al., 2019, 2021; Peng et al., 2026; Zhang et al., 2024b; Geigle et al., 2024; Conti et al., 2025). Training-based adaptations, such as prompt learning (Zhou et al., 2022b,a; Khattak et al., 2023) and adapter tuning (Gao et al., 2024; He et al., 2025; Kuchibhotla et al., 2025; Li et al., 2025b; Shi et al., 2025), improve discrimination through domain-specific parameter updates but require substantial annotation efforts and are prone to catastrophic forgetting. As a result, recent research emphasizes harnessing the intrinsic capabilities of frozen LVLMs, broadly categorized into retrieval-oriented perception and reasoning-oriented cognition.

Retrieval-Oriented Paradigms. These methods formulate FGVR as a feature-matching or knowledge-retrieval task, prioritizing efficiency. Early works such as SuS-X-LC (Udandarao et al., 2023), FineR (Xu et al., 2024), and E-FineR (Demidov et al., 2025) generate hierarchical attribute descriptions to construct robust textual prototypes and bridge the modality gap. Beyond text-based enhancement, recent approaches leverage external knowledge bases or visual caches. For example, K-Lite (Shen et al., 2022) and REACT (Liu et al., 2023) retrieve external definitions to enrich CLIP's text encoder. To address distribution shifts and noisy samples, Tip-Adapter (Zhang et al., 2021) builds a training-free key-value cache, AWT (Zhu et al., 2024) uses optimal transport for robust alignment, and ProtoMM (Zhu et al., 2025b) and RAR (Liu et al., 2024b) retrieve diverse visual examples to support decision-making. Other methods, like CaSED (Conti et al., 2024) and UniFGVR (Guo et al., 2025), relax fixed-label

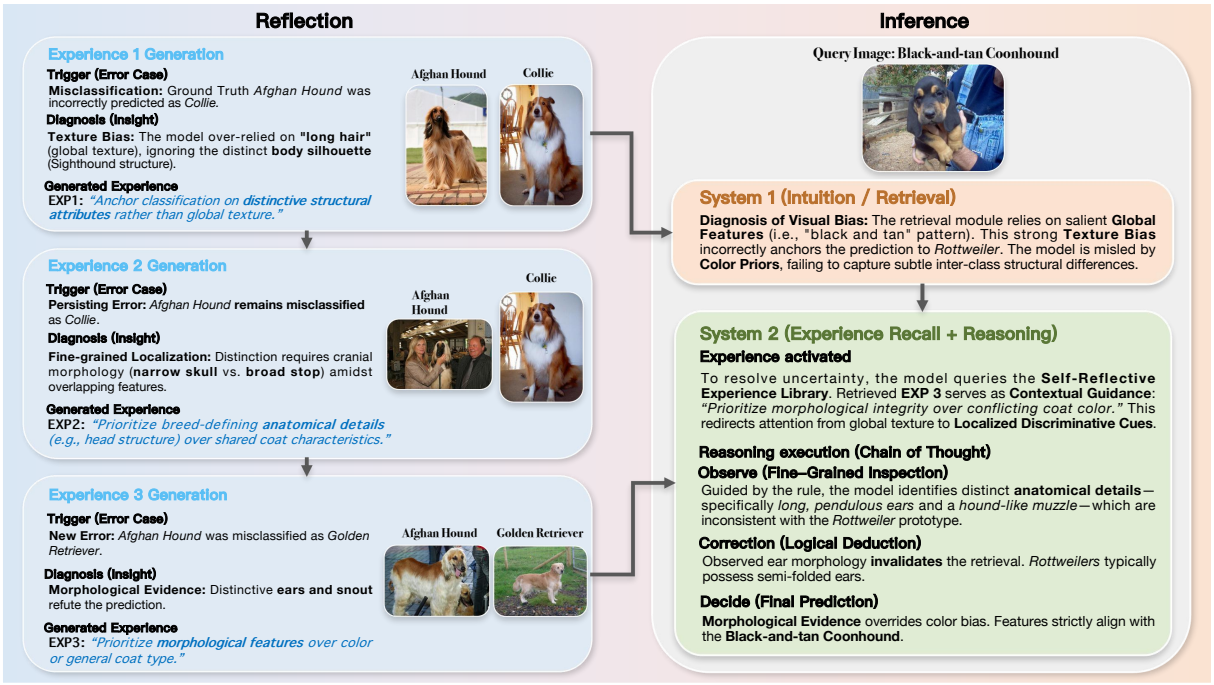


Figure 10: **Visualization of Experience Generation and Reuse. Left (Reflection):** The model analyzes past misclassifications (e.g., confusing an Afghan Hound with a Collie or Golden Retriever) to distill a generalized decision rule: "Prioritize morphological features over color." **Right (Inference):** When encountering a visually ambiguous *Black-and-tan Coonhound*, System 1 is misled by the coat color and retrieves *Rottweiler*. However, System 2 retrieves the accumulated rule, corrects the focus to morphological details (ears), and successfully identifies the correct category.

constraints and enrich semantic coverage, capturing intra-class variations. Despite their efficiency, these approaches rely on static, global similarity matching and often fail on challenging samples with occlusions or subtle local differences, as they lack mechanisms for dynamic local feature alignment.

Reasoning-Oriented Paradigms. These approaches exploit LVLMs' reasoning capabilities to perform in-depth analysis. AutoSEP (Hong et al., 2025) and GLOV (Mirza et al., 2025) optimize prompts automatically to elicit domain-specific knowledge, while MCQA (Atabuzzaman et al., 2025) formulates FGVR as multi-turn Question-Answering to focus on discriminative parts. From a representation perspective, SAV (Mitra et al., 2024) uses sparse attention vectors from generative models as discriminative classifiers.

However, reasoning-based methods face intrinsic challenges. First, including too many candidate categories can dilute context, diverting attention from key visual cues and causing inference drift (Leng et al., 2024; Liu et al., 2024a). Second, applying heavy reasoning uniformly across all samples incurs substantial computational overhead (Chen et al., 2023; Ong et al., 2025). Finally, these methods are typically stateless, preventing accumulation of reusable experience and leading to repeated failures in similar scenarios (Shinn et al., 2023).

The most closely related methods to ours are CascadeVLM (Wei, 2024) and SCAN (Yang et al., 2026), which explore retrieval-reasoning integration. CascadeVLM adopts a cascaded framework but relies on a simple static confidence threshold to connect retrieval and reasoning, making it sensitive to uncalibrated LVLM scores (Tian et al., 2023; Zhang et al., 2024a). SCAN endows LVLMs with System 2 reasoning for fine-grained recognition, but applies it indiscriminately, ignoring easy samples and incurring unnecessary computation (Chen et al., 2023; Zhu et al., 2025a).

To address these limitations, we propose SARE, a dual-system framework with sample-adaptive triggering and a self-reflective mechanism for experience accumulation, enabling selective reasoning and improved efficiency.

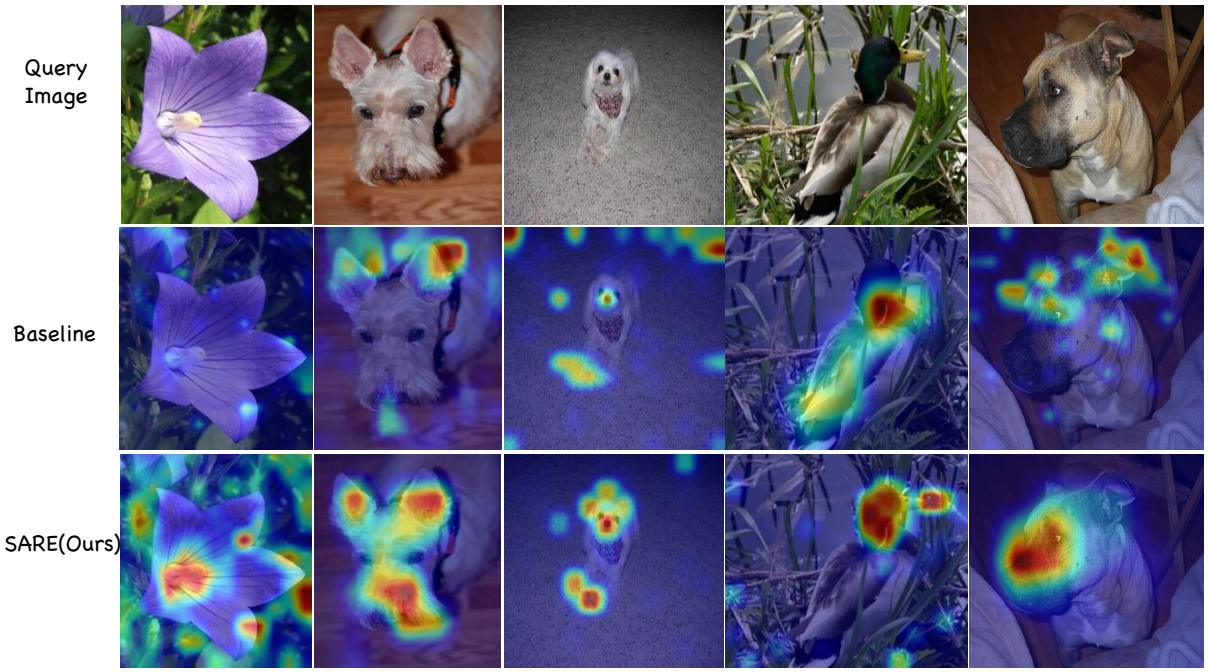


Figure 11: Visualization of attention heatmaps.