BEYOND FIXED SAMPLING: UNCERTAINTY-AWARE RETRIEVAL AND SPATIALLY-GROUNDED REASONING FOR FINE-GRAINED VISUAL RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent fine-grained visual recognition methods increasingly adopt a retrieve-then-reason paradigm powered by multimodal large language models. However, this paradigm is often hindered by two critical bottlenecks: **Rigid Selection Inflexibility** and **Ungrounded Speculation.** Specifically, conventional top-k retrieval fails to adapt to varying sample difficulties, frequently introducing irrelevant noise or missing crucial candidates. Furthermore, holistic image processing during reasoning lacks fine-grained spatial grounding, which often triggers hallucinations and yields unverified visual evidence. To overcome these issues, we propose **P-Gaze**, a framework designed to refine both the retrieval and reasoning trajectories. First, we introduce Dynamic-p, an uncertainty-aware candidate selection mechanism that adapts the probability budget to sample-level confidence and determines the candidate set through cumulative-probability selection. Second, we develop a Discriminative Attribute Anchor (DAA) mechanism for the reasoning phase. By synergizing spatially-grounded evidence with decoupled class attributes, DAA guides the model to anchor onto discriminative local regions while filtering out non-essential semantic features. Comprehensive experiments demonstrate that P-Gaze establishes new state-of-the-art benchmarks across 11 fine-grained datasets, achieving a significant average accuracy improvement of 2.18%.

1 INTRODUCTION

While generic image classification has achieved remarkable success (Dosovitskiy et al., 2021; Krizhevsky et al., 2012), Fine-Grained Visual Recognition (FGVR) remains challenging due to the high inter-class similarity among subcategories. CLIP-based methods (Radford et al., 2021; Cai et al., 2025a) have become the de facto standard, predicting classes via image-category similarity (Fig. 1(a)). More recently, Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Liu et al., 2023) have brought extensive world knowledge and reasoning capabilities to various downstream tasks (Zhang et al., 2024a; Chen et al., 2024). Since discriminating among hundreds of classes at once remains difficult for MLLMs, several studies (Yang et al., 2026; Wei, 2024) adopt a "retrieval-then-reasoning" pipeline (Fig. 1(b)): a retrieval stage, typically via top-k sampling, produces a candidate set, which an MLLM then examines for final discrimination. The candidate set is thus the interface between the two stages: it determines which classes the reasoner can still recover, and which classes it must tell apart.

This raises a question that has received little attention: *is a fixed top-k a suitable interface for fine-grained retrieval?* Fig. 2 examines retrieval results at two levels. At the dataset level (Fig. 2(a)), ground-truth (GT) ranks on Caltech101 (Fei-Fei et al., 2007) concentrate at the very top, so a small k suffices. On Aircraft (Maji et al., 2013), they spread far beyond $k = 5$: a small k loses recall, while a larger k floods most samples with redundant candidates. The same tension reappears within a single dataset (Fig. 2(b)). A confident sample with a peaked distribution needs only one candidate, whereas an ambiguous sample with a flat, long-tailed distribution needs many, so a shared k over-retrieves for the former and truncates the GT for the latter. We refer to this mismatch between a fixed candidate count and heterogeneous sample difficulty as **Rigid Selection Inflexibility**.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

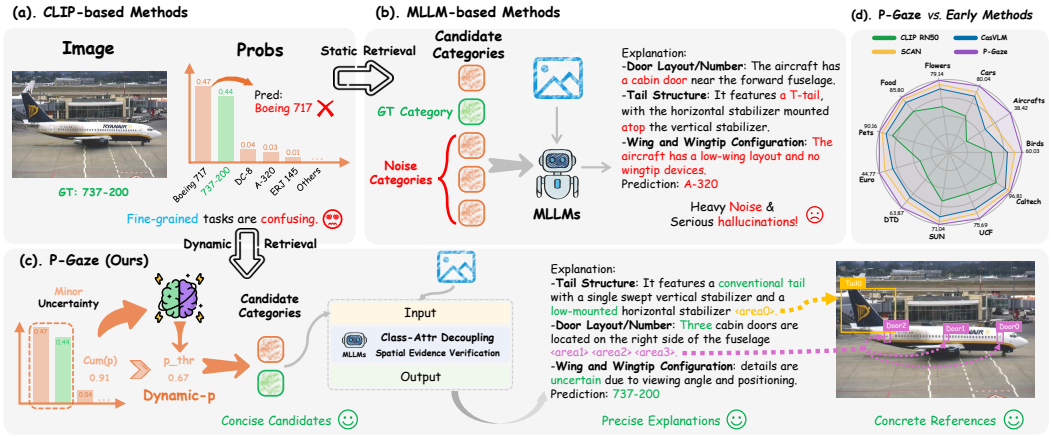


Figure 1: **Motivation and Overview of P-Gaze. (a-b)** Existing methods struggle with static retrieval noise and severe reasoning hallucinations. **(c)** our P-Gaze employs dynamic sampling and spatial evidence verification to produce accurate and visually grounded explanations. **(d)** P-Gaze consistently outperforms early methods across various fine-grained visual recognition datasets.

A natural remedy is top-p (nucleus) sampling (Holtzman et al., 2019), which retains the smallest set of candidates whose cumulative probability reaches a threshold p, letting the candidate count follow the shape of each distribution. Top-p, however, only relocates the rigidity from the count to the threshold. Every sample must cover the same probability mass, although missing the GT is far less likely when the leading prediction clearly dominates its competitors than when it barely stands out. Consistently, the best-performing threshold varies across datasets (Tab. 3), so the dataset-level mismatch resurfaces as threshold sensitivity. Retrieval should therefore adapt not only the number of candidates, but also the probability mass they are required to cover.

Adapting the candidate set, however, does not make recognition easier; it shifts the difficulty to the reasoning stage. The retained candidates are, by construction, the classes the retriever finds hardest to separate, and for ambiguous samples they are also the most numerous. Existing methods (He et al., 2025; Wei, 2024) nevertheless prompt the MLLM to inspect the entire image and reason over a flat list of attributes for all candidates. Among such similar classes, most attributes are shared, and the few discriminative cues are buried in redundant descriptions. More critically, without tying each claim to a specific image region, the MLLM tends to "see" the attributes it expects rather than those actually present, producing unverified evidence and hallucinations (Liu et al., 2024b). We term this failure mode **Ungrounded Speculation**.

To address both bottlenecks, we propose **P-Gaze**, a training-free framework that progressively narrows the recognition focus: from the full label space to a sample-specific candidate set, from candidate descriptions to their distinguishing attributes, and from these attributes to the image regions that confirm them (Fig. 1(c)). For retrieval, **Dynamic-p** conditions the probability budget on sample-level confidence. The highest-ranked class's average probability advantage over its competitors determines how much of the residual probability mass is included, and cumulative probabilities then translate this budget into a candidate set. Confident samples thus receive compact sets while ambiguous ones retain broader coverage, under a single parameter setting shared across datasets.

For reasoning, a Discriminative Attribute Anchor process makes the comparison both discriminative and verifiable. Because the candidate set now varies from sample to sample, so do the attributes that separate its members; **Contrastive Attribute Decoupling** therefore contrasts the selected candidates jointly, deriving attributes conditioned on the current comparison set. **Spatially-Grounded Evidence Verification** follows a locate-then-verify procedure to extract evidence from localized parts, and the prediction is made by aligning this evidence with the decoupled attributes.

The main contributions of this work could be summarized as follows:

- We introduce **Dynamic-p**, an uncertainty-aware candidate-selection mechanism that adapts the probability budget to each sample's mean confidence margin and uses cumulative probabilities to determine the candidate set, extending retrieval beyond a shared count or probability threshold.

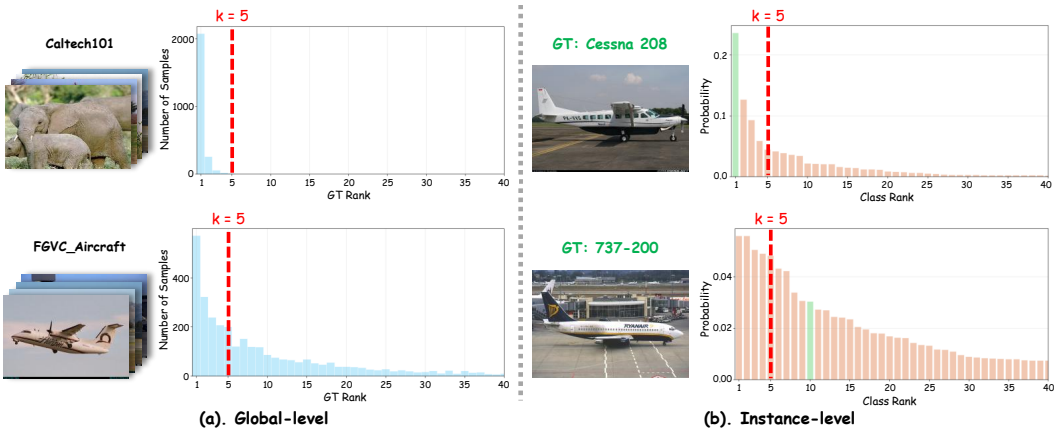


Figure 2: Distributions of retrieval results at the global and instance levels. **(a) Global-level**: Histograms of Ground Truth (GT) rank on the Aircraft (Maji et al., 2013) (top) and Caltech (Fei-Fei et al., 2007) (bottom) datasets. **(b) Instance-level**: Predicted class probability distributions for specific samples, where the green bars indicate the GT classes. The red dashed lines denote the recall boundary at $k = 5$.

- We develop a candidate-conditioned reasoning procedure that couples **Contrastive Attribute Decoupling** with **Spatially-Grounded Evidence Verification**, aligning jointly contrasted class descriptions with evidence extracted through a locate-then-verify process.
- Extensive experiments across eleven fine-grained visual recognition datasets demonstrate the state-of-the-art performance of our **P-Gaze** framework.

2 Related Work

CLIP in FGVR. CLIP (Radford et al., 2021; Zhai et al., 2023) aligns images and texts via contrastive learning on massive image-text pairs, achieving robust performance across various downstream tasks (Cai et al., 2025b; Zhou et al., 2022; Cai et al., 2026). In the context of FGVR, earlier studies (Pratt et al., 2023; Menon & Vondrick, 2023; Roth et al., 2023) focus on enriching or diversifying category prompts to capture subtle intra-class visual variations effectively. For instance, CuPL (Pratt et al., 2023) leverages GPT-4o (Hurst et al., 2024) to generate 20 detailed descriptions per category, aggregating predictions through prompt ensembling. While effective, these approaches primarily operate on global image representations, which often fail to fully exploit the local visual evidence critical for fine-grained distinction.

Consequently, recent research (Cai et al., 2025a; Shiming et al., 2025; Silva et al., 2025) has increasingly emphasized localized understanding. By leveraging attention maps or region-level features, these methods explicitly explore discriminative parts—such as bird wings or aircraft cabins—to enable finer distinctions. This strategy typically yields superior classification performance and interpretability, highlighting that fine-grained recognition benefits from the synergy between the robust global alignment provided by CLIP and precise local evidence discovery.

Attribute Description in FGVR. Attributes serve as a bridge between visual content and categorical labels. A category consists of a fixed set of attributes, where an image can be viewed as an integration of multiple attribute cues (Shi et al., 2025). Building upon this insight, several methods for FGVR (He et al., 2025; Liu et al., 2024a) treat attribute modeling as a primary objective. Specifically, Finedefics (He et al., 2025) learns attribute descriptions to enhance the correspondence between visual evidence and class semantics, improving both recognition accuracy and explanation faithfulness. Recent benchmarks (Geigle et al., 2024; Yu et al., 2025) further incorporate attribute awareness as a core evaluation dimension, requiring models to perceive attributes within local regions. Closest to our attribute design, FuDD (Esfandiarpoor & Bach, 2024) generates pairwise differential descriptions between ambiguous classes for follow-up CLIP classification. We likewise rely on comparative attributes, but contrast a sample-adaptive candidate set jointly and verify the resulting attributes against explicitly localized visual evidence.

MLLM in FGVR. The rapid development of Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Liu et al., 2023) has driven a paradigm shift toward retrieve-then-reason pipelines. Within FGVR, recent methods leverage this two-stage architecture to first identify ambiguous candidate categories and subsequently prompt MLLMs to extract discriminative clues (Kumar et al., 2025; Wei, 2024; Yang et al., 2026; Liu et al., 2026). For example, CasVLM (Wei, 2024) utilizes CLIP as an uncertainty estimator, forwarding only high-uncertainty samples to an MLLM for fine-grained discrimination. Similarly, SCAN (Yang et al., 2026) adopts a "fast-slow" dual system to balance the efficiency of CLIP with the reasoning capabilities of MLLMs. Despite these advancements, existing methods rely on a fixed number of retrieved candidates, and their reasoning outputs often contain hallucinations that undermine interpretability and trustworthiness.

Adaptive Candidate Selection. Selecting a variable-size set from a ranked distribution has also been studied in LLM decoding and set-valued prediction. Nucleus sampling (Holtzman et al., 2019) truncates at a fixed cumulative probability, while later decoding strategies let the truncation depend on the distribution, e.g., through its entropy (Hewitt et al., 2022) or a threshold proportional to the top token's probability (Nguyen et al., 2025). Conformal prediction builds class sets from cumulative probabilities with calibrated coverage (Romano et al., 2020; Angelopoulos et al., 2021), at the cost of a held-out calibration split. Dynamic-p shares the cumulative-probability view but serves a different purpose: supplying an MLLM reasoner with a candidate set that balances GT recall against reasoning burden, with a per-sample budget derived from retrieval confidence.

3 PRELIMINARY

3.1 RETRIEVAL

In zero-shot classification tasks for vision-language models (VLMs) such as CLIP (Radford et al., 2021), the input comprises a query image x and a complete category set $C = \{c_1, c_2, ..., c_N\}$. The model employs an image encoder E_v and a text encoder E_t to extract feature representations, and computes the probability distribution $P(C|x)$ of the image x over each category c_i. The logits and final probabilities are computed according to the following equations:

$$\text{logits}_i = cos(E_v(x), E_t(f(c_i))),$$

$$p_i = \frac{e^{\text{logits}_i/\tau}}{\sum_{j=1}^{N} e^{\text{logits}_j/\tau}},$$

where τ denotes the temperature parameter. To efficiently select a candidate subset from the large-scale global category space, a standard practice is to implement the following sampling strategies.

Top-k Sampling constructs a fixed-size candidate subset by retaining the k categories with the highest predicted probabilities. Given the distribution $P(C|x)$, all categories are sorted in descending order of p_i, yielding $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(N)}$, where (i) denotes the index of the category with the i-th largest probability. The candidate subset is then defined as $C_k = \{c_{(1)}, c_{(2)}, \ldots, c_{(k)}\}$. Although this strategy is simple and efficient, it is insensitive to the shape of the probability distribution. A fixed k may retain noisy low-probability categories when the model is highly confident, yet prematurely exclude plausible candidates when the distribution is relatively flat.

Top-p Sampling, also known as **nucleus sampling** (Holtzman et al., 2019), dynamically adjusts the size of the candidate subset based on cumulative probability mass. Following the same descending order, the method identifies the smallest index k_p such that:

$$k_p = \min \left\{ m \in \{1, \ldots, N\} \mid \sum_{i=1}^{m} p_{(i)} \geq p \right\},$$

where $p \in (0, 1]$ is a predefined cumulative probability threshold. The resulting candidate subset $C_p = \{c_{(1)}, c_{(2)}, \ldots, c_{(k_p)}\}$ is referred to as the nucleus set. Since k_p is determined by the shape of the probability distribution rather than being a fixed constant, top-p sampling exhibits a degree of adaptivity: it tends to select fewer candidates when the model concentrates high probability on a small number of categories, and more candidates when predictions are dispersed.

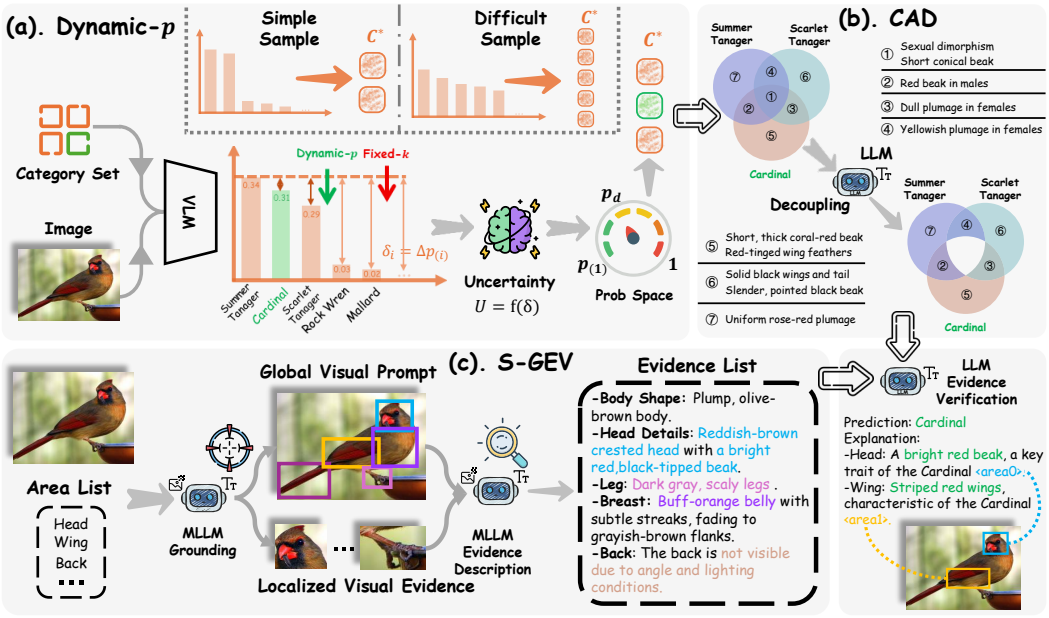


Figure 3: **Architecture of P-Gaze.** Our framework advances the retrieval-then-reasoning paradigm through three key stages. (a) **Dynamic-p**: Dynamically adjusts the retrieval based on sample uncertainty to balance recall and precision. (b) Contrastive Attribute Decoupling (**CAD**): Distills discriminative attributes by contrasting descriptions to eliminate redundancy. (c) Spatially-Grounded Evidence Verification (**S-GEV**): actively localizes visual evidence to ground the reasoning process.

3.2 Reasoning

Once the candidate subset $C' \subseteq C$ is obtained through retrieval, the reasoning stage selects the final predicted class from C'. A prevalent approach employs a multimodal large language model (MLLM) to perform classification within the candidate set. Given a query image x and C', the MLLM receives a prompt $\mathcal{T}(C')$ that enumerates the candidate classes and instructs the model to identify the correct one. The prediction is expressed as $\hat{c} = \mathcal{M}(x, \mathcal{T}(C'))$, where $\mathcal{M}$ denotes the MLLM. The model produces the final class name either through autoregressive decoding or by computing the conditional log-likelihood for each candidate and selecting the highest-scoring one.

4 Method

We present P-Gaze for explainable Fine-Grained Visual Recognition (FGVR), as illustrated in Fig. 3. We first use Dynamic-p (Sec. 4.1) to adapt the retrieval probability budget to sample-level confidence and select the candidate classes for subsequent reasoning. The subsequent Discriminative Attribute Anchor process couples two complementary operations. Contrastive Attribute Decoupling (Sec. 4.2) jointly contrasts the selected class descriptions, while Spatially-Grounded Evidence Verification (Sec. 4.3) extracts localized visual evidence for comparison with those descriptions. All stages use pretrained models without task-specific parameter training.

4.1 Dynamic Candidate Selection

As discussed in Sec. 1, fixed top-p adapts the candidate count while retaining a shared probability budget. Dynamic-p makes this budget sample-dependent by relating it to the confidence margin between the leading prediction and competing classes. Given the ranked class probabilities, we first compute these margins:

$$\delta_i = p_{(1)} - p_{(i)}, \quad i \in \{2, 3, \ldots, N\}, \tag{1}$$

where $N \geq 2$ is the total number of classes. We summarize these margins using a sample-level ambiguity score U, which serves as the uncertainty proxy in Dynamic-p:

$$U = \frac{N-1}{\sum_{i=2}^{N}(\beta\delta_i + 1)}. \tag{2}$$

Here, $\beta > 0$ is a fixed probability-margin coefficient. This score is an inverse transformation of the scaled mean probability margin: a smaller average advantage over competing classes produces a larger U.

We use U to determine the cumulative probability budget:

$$p_d = p_{(1)} + (1 - p_{(1)}) \cdot (1 - e^{-\alpha U}), \tag{3}$$

where $\alpha > 0$ controls the sensitivity of budget allocation. The exponential mapping determines the fraction of probability mass beyond the top-ranked class that the budget includes. The final candidate set is the shortest ranked prefix reaching this budget:

$$k_d = \min\left\{ m \in \{1, \ldots, N\} \mid \sum_{i=1}^{m} p_{(i)} \geq p_d \right\}, \tag{4}$$

$$C^* = \{c_{(1)}, c_{(2)}, \ldots, c_{(k_d)}\}. \tag{5}$$

Dynamic-p thus combines confidence-dependent budget allocation with cumulative-probability selection, allowing the candidate count to reflect how competing probability mass is distributed. We provide the algebraic interpretation and properties of this formulation in Sec. B.2.

4.2 Contrastive Attribute Decoupling

Following the retrieval of the optimized candidate set C^*, the reasoning phase aims to identify the correct category by verifying visual evidence. In FGVR, categories within the candidate set C^* typically share a common semantic hierarchy and exhibit high visual similarity, differing only in subtle features (e.g., distinguishing bird species by forked versus fan-shaped tails). When descriptions are generated independently for each category, they tend to be dominated by shared characteristics, resulting in semantic redundancy that obscures inter-class distinctions during subsequent reasoning.

To address this, Contrastive Attribute Decoupling (CAD) organizes discriminative attributes relative to the selected candidate set. Specifically, we leverage the extensive world knowledge of an LLM (Comanici et al., 2025) to generate a comprehensive visual description $d_{c_{(i)}}$ for each candidate $c_{(i)} \in C^*$. In the contrastive decoupling phase, all initial descriptions $D_{\text{init}} = \{d_{c_{(i)}}\}_{c_{(i)} \in C^*}$ are jointly provided to the LLM, which performs comparative analysis to explicitly filter out shared features and retain the attributes most useful for distinguishing the current candidates:

$$D_{\text{decoupled}} = \text{LLM}(D_{\text{init}}, \mathcal{T}_{\text{decouple}}), \tag{6}$$

where $\mathcal{T}_{\text{decouple}}$ denotes the prompt that guides the contrastive decoupling process. The resulting set $D_{\text{decoupled}} = \{d'_{c_{(i)}}\}_{c_{(i)} \in C^*}$ provides candidate-specific comparison cues for alignment with the localized evidence extracted by S-GEV.

4.3 Spatially-Grounded Evidence Verification

Current MLLMs often suffer from hallucinations when inferring categories globally, primarily due to the lack of explicit spatial constraints. To mitigate this, we propose Spatially-Grounded Evidence Verification (S-GEV), a mechanism that enforces a rigorous "locate-then-verify" protocol. This process orchestrates two coordinated phases: active visual probing and evidence-attribute alignment.

Active Visual Probing. Rather than processing the image globally, we instruct the MLLM to actively localize discriminative parts based on a predefined list L_{area}. Guided by a grounding prompt $\mathcal{T}_{\text{ground}}$, the model isolates key regions, yielding a set of bounding boxes $B = \text{MLLM}(x, L_{area}, \mathcal{T}_{\text{ground}})$. To extract reliable evidence, we employ a synergistic strategy that combines global context with local details. Specifically, we construct a visually prompted global image x_{global} (annotated with bounding boxes) to verify the *existence* of attributes (answering "is it there?")

Table 1: **Zero-Shot Classification Benchmarks.** Comparison with state-of-the-art methods on eleven benchmarks under zero-shot setting. Results use CLIP ResNet-50 and ViT-B/16 backbones. * denotes results reproduced using the official codebase, with Qwen3-VL-8B as the MLLM. † denotes results quoted from the original papers under their original settings. Best and second-best are **bold** and underlined respectively. FG Avg. and All Avg. denote averages over six fine-grained and all eleven datasets.

Method	CUB	Air	Cars	Flo	Food	Pets	Euro	DTD	SUN	UCF	Cal	FG Avg.	All Avg.
CLIP ResNet-50 Backbone													
CLIP (Radford et al., 2021)	44.70	15.90	53.90	60.20	74.00	83.20	28.60	38.80	58.00	56.90	84.50	55.32	54.43
DCLIP (Menon & Vondrick, 2023)	49.13	16.92	53.86	65.68	79.65	83.10	34.96	41.78	61.04	56.97	88.24	58.06	57.39
CuPL (Pratt et al., 2023)	48.70	18.90	55.30	67.00	80.10	86.10	38.20	49.50	63.10	63.30	88.30	59.35	59.86
FuDD† (Esfandiarpoor & Bach, 2024)	49.26	19.92	56.77	68.76	80.95	87.52	39.42	43.51	-	-	-	60.53	-
TDA (Karmanov et al., 2024)	50.22	17.61	57.78	68.74	77.75	86.18	42.11	43.74	62.53	64.18	89.70	59.71	60.05
DMN (Zhang et al., 2024b)	48.75	22.77	60.02	67.93	76.70	86.78	**48.72**	50.41	64.39	65.34	90.14	60.49	62.00
BCA (Zhou et al., 2025)	-	19.89	58.13	66.30	77.19	85.58	42.12	48.58	63.38	60.45	89.70	61.42	61.13
ProtoMM† (Zhu et al., 2025)	-	21.21	60.10	69.95	77.31	87.92	42.35	50.53	65.74	65.34	89.94	-	-
LaZSL (Shiming et al., 2025)	52.45	20.16	59.16	67.52	80.76	88.55	42.59	49.93	63.17	73.73	87.34	61.43	62.31
CasVLM*(τ=0.75) (Wei, 2024)	51.45	26.64	58.41	72.11	80.12	83.57	43.33	59.16	66.46	68.23	90.26	62.05	63.61
CasVLM*(τ=0) (Wei, 2024)	56.09	29.61	74.83	76.70	84.73	88.36	43.16	59.10	70.56	71.87	95.82	68.39	68.26
RAR* (Liu et al., 2026)	55.77	28.53	73.69	76.71	83.03	89.48	40.93	59.46	69.08	72.11	95.90	67.87	67.70
SCAN* (Yang et al., 2026)	**60.57**	35.51	75.18	77.50	85.49	87.74	43.02	59.50	70.04	72.40	94.83	70.33	69.25
P-Gaze (Ours)	60.03	**38.42**	**80.04**	**79.14**	**85.80**	**90.16**	44.77	**63.87**	**71.04**	**75.69**	**96.81**	**72.27**	**71.43**
CLIP ViT-B/16 Backbone													
CLIP (Radford et al., 2021)	55.30	24.40	65.50	70.60	88.40	89.00	51.00	44.30	62.50	67.70	93.20	65.53	64.72
DCLIP (Menon & Vondrick, 2023)	57.75	24.96	63.25	70.62	88.50	86.92	51.32	45.59	66.85	65.95	93.27	65.33	65.00
CuPL (Pratt et al., 2023)	57.20	26.60	66.10	73.90	88.90	91.20	55.70	53.80	69.00	70.30	94.30	67.32	67.91
TDA (Karmanov et al., 2024)	57.59	23.91	67.28	71.42	86.14	88.63	58.00	47.40	67.62	70.66	94.24	65.83	66.63
DMN (Zhang et al., 2024b)	56.71	30.03	67.96	74.49	85.08	92.04	**59.43**	55.85	70.18	65.34	90.14	67.72	67.93
BCA (Zhou et al., 2025)	-	28.59	66.86	73.12	85.97	90.43	56.63	53.49	68.41	67.59	94.69	68.99	68.58
AWT† (Zhu et al., 2024)	60.20	29.22	69.93	75.07	85.54	92.53	58.61	55.56	70.58	72.51	95.54	68.75	69.57
ProtoMM† (Zhu et al., 2025)	-	31.02	69.92	77.40	85.89	91.90	56.11	56.38	70.78	71.76	95.70	-	-
ABS (Cai et al., 2025a)	61.01	31.38	67.80	73.60	89.69	**92.64**	56.62	54.26	69.51	72.21	94.89	69.35	69.42
LaZSL (Shiming et al., 2025)	60.30	29.49	68.67	74.76	**89.70**	87.40	56.89	54.08	69.37	71.42	94.44	68.39	68.77
CasVLM*(τ=0.75) (Wei, 2024)	56.16	31.20	66.36	74.10	85.84	88.47	51.64	59.22	68.03	69.07	94.20	67.02	67.66
CasVLM*(τ=0) (Wei, 2024)	56.80	34.71	78.14	77.75	85.90	87.82	48.41	59.57	71.98	73.75	96.71	70.19	70.14
RAR* (Liu et al., 2026)	56.73	33.60	79.17	76.73	85.68	86.72	46.32	60.11	71.07	74.09	96.75	69.77	69.73
SCAN* (Yang et al., 2026)	**61.97**	40.62	79.01	77.53	86.02	89.42	53.26	61.95	72.08	73.46	96.87	72.43	72.02
P-Gaze (Ours)	61.32	**41.61**	**80.11**	**81.92**	86.53	89.87	53.60	**65.91**	**72.36**	**75.82**	**97.21**	**73.56**	**73.30**

and a set of cropped local patches x_{local} to capture fine-grained *details* (answering "what is it?"). The MLLM processes these dual inputs to generate a precise visual evidence list D_{evidence}:

$$D_{\text{evidence}} = \text{MLLM}(x_{\text{global}}, x_{\text{local}}, B, \mathcal{T}_{\text{desc}}). \tag{7}$$

This step serves as an evidence filter: the model is instructed to omit attributes that are absent in x_{global} or visually indistinct in x_{local}, reducing unsupported claims.

Evidence-Attribute Alignment. Finally, we formulate the inference step as a logical cross-examination. The language model aligns the verified visual evidence list D_{evidence} with the decoupled semantic profiles $D_{\text{decoupled}}$ obtained from Eq. 6:

$$y^*, \mathcal{E}, B_{\text{final}} = \text{LLM}(D_{\text{evidence}}, D_{\text{decoupled}}, \mathcal{T}_{\text{verify}}). \tag{8}$$

By matching the verified visual facts against the discriminative semantic criteria, our framework identifies the most consistent category y^*, providing a natural language rationale $\mathcal{E}$ and supporting spatial coordinates B_{final}. This links the prediction and its rationale to localized visual evidence. Full area lists L_{area} and each prompt templates are provided in the Appendix C.

5 Experiments

Datasets and Evaluation Metrics. We evaluate P-Gaze and compare to previous work on a benchmark of 11 datasets including fine-grained ones like **CUB** (Welinder et al., 2010) (CUB, 200 classes), **FGVC Aircrafts** (Maji et al., 2013) (Air, 100 classes), **Stanford Cars** (Krause et al., 2013) (Cars, 196 classes), **Flowers 102** (Nilsback & Zisserman, 2008) (Flo, 102 classes), **Food101** (Bossard et al., 2014) (Food, 101 classes) and **Oxford Pets** (Parkhi et al., 2012) (Pets, 37 classes). It also

includes datasets like **EuroSAT** (Helber et al., 2019) (Euro, 10 classes), **DTD** (Cimpoi et al., 2014) (DTD, 47 classes), **Sun397** (Xiao et al., 2010) (SUN, 397 classes), **UCF101** (Soomro et al., 2012) (UCF, 101 classes) and **Caltech101** (Fei-Fei et al., 2007) (Cal, 100 classes). Following prior work (Yang et al., 2026), we report top-1 accuracy(%) and aggregate via FG Avg. (six fine-grained) and All Avg. (all eleven).

Implementation Details. We adopt CLIP (Radford et al., 2021) ViT-B/16 as the default retrieval backbone and additionally report results with CLIP ResNet-50, consistent with prior work (Yang et al., 2026). For the reasoning stage, Qwen3VL-8B (Bai et al., 2025) serves as the MLLM for visual grounding and evidence extraction, while Qwen3-8B (Yang et al., 2025) handles CAD and final verification. Following SCAN (Yang et al., 2026), the initial category attribute descriptions for CAD are generated offline via Gemini-2.5 Pro (Comanici et al., 2025), leveraging its broad world knowledge, and cached for reuse. The hyperparameter for Dynamic-p is set to $\alpha = 6$ uniformly across all datasets and backbones. The probability-margin scaling coefficient β in Eq. 2 is empirically chosen as 40. Dynamic-p computes the ambiguity score over the complete set of N classes.

5.1 Comparison with State-of-the-Art Methods

We compare P-Gaze with (i) *CLIP-based methods*, including CLIP (Radford et al., 2021), description-enhanced (DCLIP (Menon & Vondrick, 2023), CuPL (Pratt et al., 2023), FuDD (Esfandiarpoor & Bach, 2024)), cache-based (TDA (Karmanov et al., 2024), DMN (Zhang et al., 2024b), BCA (Zhou et al., 2025)), multimodal matching and prototype (AWT (Zhu et al., 2024), ProtoMM (Zhu et al., 2025)), and local-feature methods (ABS (Cai et al., 2025a), LaZSL (Shiming et al., 2025)); and (ii) *MLLM-powered retrieve-then-reason methods*: CasVLM (Wei, 2024), RAR (Liu et al., 2026), and SCAN (Yang et al., 2026). CasVLM(τ=0), which retrieves a fixed set of candidates before MLLM reasoning, serves as our direct baseline. For FuDD, we quote the ResNet-50 results of its $k = |C|$ variant with the original GPT-3.5 descriptions. More details are given in Sec. C.1.

Table 2: **Component ablation** on three fine-grained datasets (ViT-B/16). ✔ and ✘ denote enabled and disabled modules.

Dynamic-p	S-GEV	CAD	CUB	Air	Cars
✘	✘	✘	56.87	35.68	75.51
✔	✘	✘	58.32	36.99	78.05
✘	✔	✘	58.65	36.63	78.06
✘	✘	✔	58.08	35.86	76.14
✔	✔	✘	59.92	41.10	79.38
✔	✘	✔	58.75	38.19	77.85
✘	✔	✔	59.87	37.74	78.90
✔	✔	✔	**61.32**	**41.61**	**80.11**

As shown in Tab. 1, P-Gaze achieves the best FG Avg. and All Avg. under both backbones, surpassing the strongest competitor SCAN by 1.94%/2.18% with ResNet-50 and 1.13%/1.28% with ViT-B/16. The margin is larger with the weaker ResNet-50 retriever (e.g., +4.86% over SCAN on Cars), whose GT is less often ranked first and for which a fixed candidate count is more likely to be mis-sized, precisely the regime that adaptive selection targets. The gains, however, are not uniform, and where they concentrate is informative. They are largest on datasets whose classes differ in structurally localizable parts. On Aircraft (ViT-B/16), P-Gaze improves over its direct baseline CasVLM(τ=0) by **6.90%** (41.61% vs. 34.71%), since aircraft variants are distinguished by engine configuration, wing geometry, and tail shape, precisely the cues that CAD isolates and S-GEV localizes. Conversely, on EuroSAT, where satellite scenes are characterized by global texture rather than parts, the cache-based DMN remains stronger, marking the boundary of region-level reasoning.

5.2 Ablation Studies

Does adapting the probability budget help? Tab. 3 compares retrieval strategies within the same CLIP+MLLM pipeline. The best top-p setting (64.89%) exceeds the best top-k setting (63.43%), supporting the benefit of letting the count follow the distribution, yet the optimal p still shifts across datasets (e.g., 0.6 on Aircraft vs. 0.4 on DTD). More candidates are not a remedy either: passing all classes to the MLLM drops accuracy to 61.07%, indicating that redundant candidates actively hurt reasoning. Dynamic-p resolves this tension with a single shared setting, outperforming every fixed k and p on all six datasets and reaching 65.56% on average. Its Avg. k is equally telling: without

Table 3: **Retrieval Strategy Analysis.** Comparison of retrieval sampling strategies on the CLIP + MLLM baseline. We evaluate fixed Top-k and Top-p against our proposed **Dynamic-p**. Avg. k denotes the mean number of retained candidates per sample. Results demonstrate that Dynamic-p achieves the highest classification accuracy while maintaining an efficient candidate set size.

Method	CUB	Air	Flo	Euro	DTD	Cal	Avg.
best-k (upper bound)	73.11	57.34	85.21	75.49	79.55	98.58	78.21
$k = 1$	54.44	24.36	70.56	48.63	43.20	92.09	55.55
$k = 5$	54.80	34.71	77.75	48.41	59.57	96.71	61.99
$k = 10$	55.45	36.91	80.23	46.75	63.18	96.75	63.21
$k = 20$	54.56	37.10	81.08	46.86	64.18	96.82	63.43
$p = 0.2$	55.11	31.20	74.58	50.85	55.32	94.00	60.18
Avg. k	*1.85*	*3.08*	*1.42*	*1.35*	*2.48*	*1.16*	*1.89*
$p = 0.4$	56.40	37.11	79.50	51.73	64.95	95.25	64.16
Avg. k	*4.32*	*6.88*	*2.98*	*2.29*	*6.17*	*1.98*	*4.10*
$p = 0.6$	56.08	38.86	80.59	52.41	64.78	96.63	64.89
Avg. k	*9.37*	*12.88*	*6.74*	*3.65*	*12.33*	*4.68*	*8.28*
$p = 0.8$	54.21	38.52	81.16	49.94	64.30	96.93	64.18
Avg. k	*22.11*	*24.41*	*16.76*	*5.68*	*22.46*	*13.58*	*17.50*
MLLM (full classes)	44.11	37.29	80.76	44.06	64.08	96.12	61.07
dynamic-p	**56.47**	**39.79**	**81.51**	**53.03**	**65.60**	**96.96**	**65.56**
Avg. k	*8.20*	*18.56*	*6.99*	*4.76*	*17.44*	*4.72*	*10.11*

any dataset-level prior, it retains 4.72 candidates on Caltech101 but 18.56 on Aircraft, recovering from instance-level confidence the dataset-level difficulty gap in Fig. 2(a).

How do retrieval and reasoning interact? Tab. 2 ablates the three components on three fine-grained datasets (ViT-B/16). Each component helps on its own and the full P-Gaze is best throughout, but the most revealing pattern lies in how they combine. On Aircraft, where Dynamic-p retains the largest candidate sets, Dynamic-p and S-GEV individually add only 1.31% and 0.95%, whereas together they add 5.42% (35.68%→41.10%). This is the coupling anticipated in Sec. 1: a broader set of confusable candidates pays off only when the reasoner can verify its evidence locally. On CUB and Cars, the two gains overlap rather than compound, suggesting that this coupling matters most when the GT is frequently ranked deep. CAD likewise benefits from the other two components: its gain grows when added on top of them (e.g., from 1.21% to 1.40% on CUB and from 0.18% to 0.51% on Aircraft), as attribute contrast is most useful once the candidates and evidence are refined. End-to-end latency is analyzed in Sec. C.2.

The appendix further reports hyperparameter sensitivity, scaling with model size, stability, S-GEV grounding quality, and both successful and failure cases.

6 Conclusion

We revisited the retrieve-then-reason pipeline for fine-grained visual recognition from the perspective of its interface, the candidate set. Fixed top-k and top-p retrieval impose a shared count or probability budget on samples of very different difficulty, a Rigid Selection Inflexibility, while holistic reasoning over the resulting confusable candidates invites Ungrounded Speculation. P-Gaze addresses both by progressively narrowing the recognition focus: Dynamic-p adapts the probability budget to sample-level confidence, and the Discriminative Attribute Anchor process contrasts the selected candidates and anchors their distinguishing attributes to localized visual evidence. Experiments on eleven benchmarks show state-of-the-art average accuracy, and the ablations indicate that adaptive retrieval and grounded reasoning are most effective in combination. Two limitations remain: Dynamic-p inherits the calibration of the retrieval model under a fixed temperature, and the reasoning stage relies on the grounding model, whose localization can fail under occlusion or atypical viewpoints (Sec. F). We hope this perspective encourages future work to design the two stages jointly rather than in isolation.

AI USE STATEMENT

Large language models (LLMs) appear in this work in two distinct roles. First, they are components of the proposed method and of our evaluation protocol: Qwen3-VL-8B performs visual grounding and evidence extraction, Qwen3-8B performs contrastive attribute decoupling and final verification, Gemini-2.5 Pro generates the initial category descriptions offline, and Gemini 3 Pro serves as an automated evaluator of hallucination (Sec. E.1). These uses are fully specified in the paper, and all reported results come from our own experimental runs.

Second, we used an LLM-based assistant during manuscript preparation. It reviewed the experimental analysis and structure of the paper.

All AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Anastasios Nikolas Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, the Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

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

Lincan Cai, Jingxuan Kang, Shuang Li, Wenxuan Ma, Binhui Xie, Zhida Qin, and Jian Liang. From local details to global context: Advancing vision-language models with attention-based selection. In *Forty-second International Conference on Machine Learning*, 2025a.

Lvpan Cai, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. Me-fas: Multimodal text enhancement for cross-domain face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2025b.

Lvpan Cai, Haowei Wang, Jiayi Ji, YanShu ZhouMen, Yiwei Ma, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. Zooming in on fakes: A novel dataset for localized ai-generated image detection with forgery amplification approach. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the

frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Google DeepMind. Gemini 3 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf, November 2025. Published November 2025, accessed March 12, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Reza Esfandiarpoor and Stephen H. Bach. Follow-up differential descriptions: Language models resolve ambiguities for image classification. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/17eca81974d2fd34f458916e0bfd820d-Abstract-Conference.html.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding*, 106(1):59–70, 2007.

Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2653–2669, 2024.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=p3NKpom1VL.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

John Hewitt, Christopher D. Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Raja Kumar, Arka Sadhu, and Ram Nevatia. Dive-k: Differential visual reasoning for fine-grained image recognition. *arXiv preprint arXiv:2511.18305*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *ICLR*, 2024a.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024c.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Rar: Retrieving and ranking augmented mllms for visual recognition. *IEEE Transactions on Image Processing*, 2026.

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

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *International Conference on Learning Representations*, 2025.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15691–15701, 2023.

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

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, 2020.

Karsten Roth, Jae Myung Kim, Andrew Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15746–15757, 2023.

Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, and Ninghao Liu. Enhancing cognition and explainability of multimodal foundation models with self-synthesized data. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chen Shiming, Duan Bowen, Khan Salman, and Khan Fahad Shahbaz. Interpretable zero-shot learning with locally-aligned vision-language model. In *ICCV*, 2025.

Sathira Silva, Eman Ali, Chetan Arora, and Muhammad Haris Khan. microclip: Unsupervised clip adaptation via coarse-fine token fusion for fine-grained image classification. *arXiv preprint arXiv:2510.02270*, 2025.

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

Canshi Wei. Enhancing fine-grained image classifications via cascaded vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1857–1871, 2024.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. In *Forty-second International Conference on Machine Learning*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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

Hong-Tao Yu, Xiu-Shen Wei, Yuxin Peng, and Serge Belongie. Benchmarking large vision-language models on fine-grained image tasks: A comprehensive evaluation. *arXiv preprint arXiv:2504.14988*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28718–28728, 2024b.

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

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian test-time adaptation for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29999–30009, 2025.

Xingyu Zhu, Shuo Wang, Beier Zhu, Miaoge Li, Yunfan Li, Junfeng Fang, Zhicai Wang, Dongsheng Wang, and Hanwang Zhang. Dynamic multimodal prototype learning in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2501–2511, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_Dynamic_Multimodal_Prototype_Learning_in_Vision-Language_Models_ICCV_2025_paper.html.

Yuhan Zhu, Yuyang Ji, Zhiyu Zhao, Gangshan Wu, and Limin Wang. AWT: Transferring vision-language models via augmentation, weighting, and transportation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2d4eaf042567f1c03c086103cc154c1f-Abstract-Conference.html.

A Appendix Contents

This appendix provides supplementary materials organized as follows:

- **Appendix B** provides a comprehensive analysis of the Dynamic-p mechanism, including ambiguity score comparisons (Sec. B.1), the algebraic interpretation and probability-budget analysis (Sec. B.2), and per-sample candidate distribution visualizations (Sec. B.3).
- **Appendix C** presents additional experimental results, covering descriptions of compared methods (Sec. C.1), efficiency–accuracy trade-off analysis (Sec. C.2), S-GEV module Analysis (Sec. C.3), scaling law experiments across retrieval and reasoning backbones (Sec. C.4), hyperparameter sensitivity analysis (Sec. C.5), and model stability analysis (Sec. C.6).
- **Appendix D** details the complete prompt templates (Sec. D.1) and dataset-specific area lists (Sec. D.2) used in our framework.
- **Appendix E** provides qualitative evaluations, including automated hallucination assessment (Sec. E.1), qualitative case studies (Sec. E.2), and failure case analysis (Sec. E.3).
- **Appendix F** discusses limitations of the current framework and directions for future work (Sec. F).

B Analysis of Dynamic-p

B.1 Ambiguity Score Comparison

Tab. 4 organizes the analysis into two stages: the standalone uncertainty statistic and its use in probability-budget allocation. In the upper group, "Uncertainty (Ours)" denotes the score U in Eq. 2, a mean-margin-based ambiguity score used as an uncertainty proxy. We compare it with Shannon entropy $H = -\sum_{i=1}^{N} p_i \log p_i$ and the Gini index $G = 1 - \sum_{i=1}^{N} p_i^2$ through their Spearman rank correlations with the ground-truth rank in the retrieval list. This measures association with retrieval difficulty rather than uncertainty calibration.

The lower group examines the subsequent budget-allocation stage, where the uncertainty statistic is used in a mapping that determines how much residual probability mass to include. These two groups therefore reflect the progression from a confidence summary to a candidate-selection budget, rather than interchangeable definitions of uncertainty. In Dynamic-p, this progression is expressed by $\lambda(U) = 1 - e^{-\alpha U}$ and $p_d = p_{(1)} + \lambda(U)(1 - p_{(1)})$; the resulting budget is then converted into a candidate count through cumulative-probability selection.

As derived in Sec. B.2, when all classes participate and the class count is fixed, U is a strictly decreasing function of $p_{(1)}$. Its ordering therefore corresponds to decreasing top-1 confidence. A strictly increasing transformation of U alone preserves this ordering, so the stage-wise comparison should not be interpreted as evidence that such a transformation by itself improves Spearman correlation.

B.2 Mean-Margin Interpretation and Adaptive Probability Budget

Mean Probability Advantage. Let $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(N)}$ be the ranked class probabilities, with $N \geq 2$ and $\sum_{i=1}^{N} p_{(i)} = 1$. The mean margin between the highest-ranked class and all remaining classes is

$$\bar{\delta} = \frac{1}{N-1} \sum_{i=2}^{N} \big(p_{(1)} - p_{(i)}\big). \tag{9}$$

Rewriting Eq. 2 gives

$$U = \frac{N-1}{(N-1) + \beta \sum_{i=2}^{N} \delta_i} = \frac{1}{1 + \beta \bar{\delta}}. \tag{10}$$

Table 4: **Spearman Correlation (ρ) of Different Uncertainty Metrics and Dynamic Threshold Formulations.** The upper group compares standalone uncertainty metrics; the lower group evaluates full dynamic threshold formulations $p_d = p_{(1)} + \lambda \cdot (1 - p_{(1)})$ with different λ functions. **Bold**: best.

Metric	CUB	Air	Euro	Cal	DTD	UCF
Entropy	0.2268	0.2959	0.5410	0.4310	0.3414	0.1382
Gini Index	0.3534	0.3066	0.5588	0.4591	0.4435	0.2072
Uncertainty U (Ours)	0.4019	0.2817	0.5601	0.4821	0.4718	0.2279
$\lambda \cdot (1 - p_{(1)})$						
$\lambda = e^{\text{certainty}}$	0.4441	**0.3071**	0.5690	0.4920	**0.5091**	**0.2537**
$\lambda = 1 - e^{-\text{certainty}}$	**0.4448**	0.3050	**0.5693**	**0.4927**	0.5083	0.2532

Thus, U is an inverse transformation of the highest-ranked class's scaled average probability advantage over its competitors. The coefficient $\beta > 0$ sets the scale of the probability margins relative to the additive unit term; we empirically choose $\beta = 40$ and keep it fixed across all experiments. It changes the magnitude of U and hence the allocated probability budget, while preserving the ordering induced by the mean margin. A smaller mean margin yields a larger ambiguity score. This is a deterministic confidence proxy computed from the retrieval probabilities, rather than a fitted evidential model or an estimate of epistemic uncertainty.

Simplification and Information Content. Because the probabilities sum to one,

$$\sum_{i=2}^{N} \delta_i = (N-1)p_{(1)} - \big(1 - p_{(1)}\big) = Np_{(1)} - 1. \tag{11}$$

Consequently,

$$\bar{\delta} = \frac{Np_{(1)} - 1}{N-1}, \qquad U = \frac{N-1}{(N-1) + \beta(Np_{(1)} - 1)}. \tag{12}$$

When all classes participate, the score depends only on the top-1 probability and the class count. The mean-margin form provides an interpretation of this confidence transformation; it does not introduce information beyond these two quantities. In particular, redistributing probability among the remaining classes while keeping $p_{(1)}$ and N fixed leaves U unchanged.

Range and Boundary Cases. Since $1/N \le p_{(1)} \le 1$, the mean margin lies in $[0, 1]$, and

$$\frac{1}{1+\beta} \le U \le 1. \tag{13}$$

The uniform distribution gives $\bar{\delta} = 0$ and $U = 1$. At the opposite endpoint, $p_{(1)} = 1$ gives $\bar{\delta} = 1$ and $U = 1/(1+\beta)$. Thus, $U = 0$ is not an attainable state. For fixed N and β, the score decreases monotonically with top-1 confidence:

$$\frac{\partial U}{\partial p_{(1)}} = -\frac{\beta N(N-1)}{\big[(N-1) + \beta(Np_{(1)} - 1)\big]^2} < 0. \tag{14}$$

From Ambiguity to Probability Budget. We express the adaptive threshold as

$$p_d = p_{(1)} + \lambda(U)\big(1 - p_{(1)}\big), \qquad \lambda(U) = 1 - e^{-\alpha U}, \quad \alpha > 0. \tag{15}$$

Here, $1 - p_{(1)}$ is the residual probability mass outside the highest-ranked class, and $\lambda(U)$ specifies the fraction of this mass included in the budget. The mapping is increasing and saturating:

$$\lambda'(U) = \alpha e^{-\alpha U} > 0, \qquad \lambda''(U) = -\alpha^2 e^{-\alpha U} < 0. \tag{16}$$

A larger ambiguity score therefore allocates a larger fraction of the residual mass. The exponential form is a monotone, saturating design choice; these properties do not establish its optimality over alternative mappings.

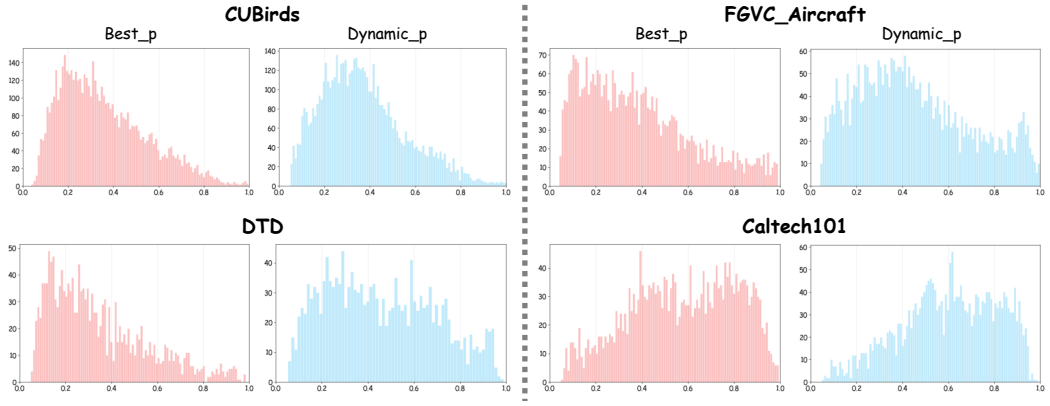


Figure 4: **Per-sample p value distributions** of Best-p (oracle) and Dynamic-p on CUB, Aircraft, DTD, and Caltech.

The range of U implies

$$1 - e^{-\alpha/(1+\beta)} \leq \lambda(U) \leq 1 - e^{-\alpha}, \tag{17}$$

and hence

$$1 - (1 - p_{(1)})e^{-\alpha/(1+\beta)} \leq p_d \leq 1 - (1 - p_{(1)})e^{-\alpha}. \tag{18}$$

In particular, $p_{(1)} \leq p_d \leq 1$. When $p_{(1)} = 1$, the residual mass is zero, so $p_d = 1$ and the first class alone reaches the budget. This boundary behavior follows from the vanishing residual mass, not from U approaching zero.

From Probability Budget to Candidate Count. Although the budget depends on $p_{(1)}$ and N, Eq. 4 uses the full cumulative ranked distribution to construct C^*. Two samples with identical top-1 confidence and class count receive the same budget, but can require different prefix lengths when their remaining probability mass is distributed differently. Dynamic-p therefore separates the role of the confidence summary in setting the budget from the role of the cumulative probabilities in determining how many candidates are retained.

For a prescribed budget p_d, this prefix has minimum cardinality among sets whose model-probability mass reaches the budget:

$$C^* \in \arg\min_{S \subseteq C} |S| \quad \text{subject to} \quad \sum_{c \in S} p(c \mid x) \geq p_d. \tag{19}$$

Indeed, the top-m classes have at least as much probability mass as any other set of m classes. If their cumulative mass falls below p_d, no set of that size can satisfy the constraint; the first prefix reaching the budget is therefore of minimum size. This property holds for a given budget and does not prove that the mapping used to choose p_d is optimal.

The excluded model-probability mass satisfies

$$1 - \sum_{i=1}^{k_d} p_{(i)} \leq 1 - p_d = (1 - p_{(1)})e^{-\alpha U}. \tag{20}$$

Discrete selection may overshoot the budget, so this is an upper bound rather than an equality. It is also a statement about model probabilities, not a calibrated ground-truth coverage guarantee. Likewise, a bounded probability budget does not imply a fixed bound on candidate count: diffuse distributions may require many classes to accumulate the specified mass.

B.3 Candidate Distribution Visualization

To provide an intuitive understanding of the adaptive behavior of Dynamic-p, we visualize the per-sample p value distributions. As shown in Fig. 4, the Best-p distributions exhibit diverse shapes across datasets, yet no clear correlation emerges between dataset difficulty and the distributional pattern: for instance, the challenging Aircraft dataset and the relatively easier Caltech dataset both

display broad, non-trivial distributions. This confirms that the optimal threshold is fundamentally governed by instance-level characteristics rather than dataset-level statistics, further justifying the necessity of per-sample adaptive selection.

Comparing the two strategies, the Dynamic-p distributions consistently shift rightward relative to Best-p, reflecting a slightly more conservative allocation that retains a broader candidate set to safeguard ground-truth recall. Importantly, within the primary mass region, the shape of the Dynamic-p distribution closely mirrors that of Best-p on each dataset—both exhibiting similar modality and spread patterns. This shape consistency demonstrates that Dynamic-p effectively approximates the oracle per-sample threshold without access to ground-truth information, capturing the underlying difficulty structure of each sample through uncertainty estimation alone.

C Additional Experiments

C.1 Description of Compared Methods

For completeness, we provide a brief description of each compared method referenced in the main paper.

CLIP-based Methods. CLIP (Radford et al., 2021) performs zero-shot classification by computing cosine similarity between image and text embeddings. **DCLIP** (Menon & Vondrick, 2023) enriches category prompts with LLM-generated visual descriptions to improve class discriminability. **CuPL** (Pratt et al., 2023) generates multiple descriptive prompts per category via GPT and aggregates predictions through ensembling. **TDA** (Karmanov et al., 2024) constructs a training-free adapter using a cached key-value store from few-shot examples. **DMN** (Zhang et al., 2024b) builds dual memory networks that jointly exploit visual and textual prototypes for enhanced classification. **BCA** (Zhou et al., 2025) applies Bayesian calibration to adapter-based methods to improve prediction reliability. **ABS** (Cai et al., 2025a) leverages attention maps to extract local image features for finer-grained CLIP-based classification. **LaZSL** (Shiming et al., 2025) employs interpretable local attention to bridge zero-shot learning with fine-grained visual evidence.

MLLM-powered Methods. CasVLM (Wei, 2024) uses CLIP as an uncertainty estimator, cascading ambiguous samples to an MLLM for fine-grained discrimination. **RAR** (Liu et al., 2026) applies MLLM-based re-ranking to refine initial CLIP retrieval results. **SCAN** (Yang et al., 2026) adopts a "fast-slow" dual-system architecture that balances CLIP's efficiency with MLLM's reasoning depth.

C.2 Efficiency and Accuracy Comparison

Tab. 5 reports the per-image latency, accuracy, and interpretability capabilities on Aircraft (ViT-B/16). All methods share the same retrieval cost (8.82 ms); the differences lie in the reasoning stage. Generating explanations dominates the latency: the baseline jumps from 176 ms (label-only) to 3110 ms (8B with explanation) or 7910 ms (72B), and SCAN requires 6290 ms. Our full pipeline takes 7218 ms, where CAD and S-GEV are executed in a parallel-friendly manner to avoid excessive additional overhead. Compared with the 72B baseline, P-Gaze is both faster (7218 vs. 7910 ms) and more accurate (**41.61%** vs. 39.12%). Although P-Gaze remains slower than the current SOTA SCAN, this latency gap is a deliberate trade-off for producing more faithful explanations and spatially-grounded region references, while still achieving higher classification accuracy.

C.3 S-GEV Module Analysis

We compare three grounding paradigms for S-GEV: Grounding DINO (Liu et al., 2024c), SAM3 (Carion et al., 2025), and MLLM-native grounding via Qwen3VL-8B (Bai et al., 2025). As visualized in Fig. 5, Grounding DINO and SAM3 frequently produce missed or erroneous detections on structurally localized parts (e.g., "beak", "wing bars", "tail shape"). This is because both models are trained on generic object-level or segmentation-level annotations and lack semantic priors for fine-grained attribute-specific regions. In contrast, Qwen3VL-8B leverages its broad world knowledge to correctly interpret and localize such domain-specific parts, yielding substantially more reliable grounding results. We therefore adopt Qwen3VL-8B as the default grounding backbone for S-GEV.

Table 5: **Efficiency, accuracy, and interpretability comparison on Aircraft** (ViT-B/16). **Ret.**: retrieval latency per image; **Inf.**: reasoning-stage latency per image; **Pred.**: outputs a class label; **Expl.**: generates a natural-language explanation; **Ref.**: provides spatially-grounded region references.

Method	Ret.(ms)	Inf.(ms)	Acc.(%)	Pred.	Expl.	Ref.
CLIP	8.82	-	24.40	✔	✘	✘
Baseline ($k=5$, 8B)	8.82	176.25	34.71	✔	✘	✘
Baseline ($k=5$, w/ expl., 8B)	8.82	3110	36.47	✔	✔	✘
Baseline ($k=5$, w/ expl., 72B)	8.82	7910	39.12	✔	✔	✘
SCAN	8.82	6290	40.62	✔	✔	✘
P-Gaze (Ours)	8.82	7218	**41.61**	✔	✔	✔

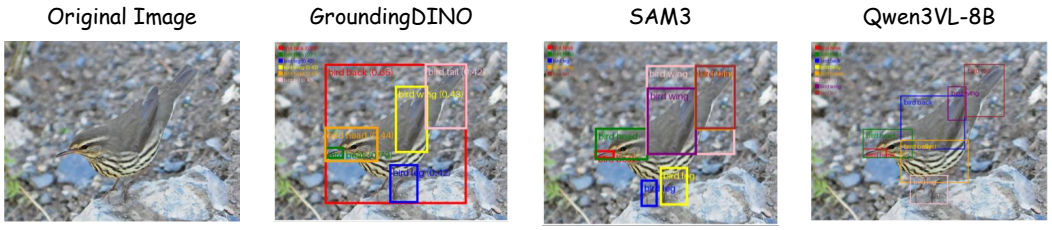


Figure 5: **Grounding model comparison.** Grounding DINO and SAM3 exhibit missed or incorrect detections on fine-grained structural parts, while Qwen3VL-8B reliably localizes attribute-specific regions.

Table 6: Performance comparison on S-GEV and Pipeline Error.

Model	IoU	Recall	Precision
GroundingDINO	0.45	0.70	0.55
SAM3	0.53	0.80	0.60
S-GEV Stage 1	0.57	0.85	0.86
S-GEV Stage 2	0.59	0.92	0.89

Beyond the qualitative comparison above, we further provide a quantitative evaluation to assess the grounding accuracy of S-GEV and analyze whether the cascaded pipeline introduces cumulative errors. We randomly sample 100 images across all datasets and manually annotate ground-truth local regions for each attribute-specific part. Following multi-instance referring detection protocols, we adopt IoU, Recall, and Precision as evaluation metrics; results are reported in Tab. 6. At the region level, Grounding DINO and SAM3 suffer from notable missed detections and false positives, leading to substantially lower Recall (0.70 / 0.80), respectively. In contrast, S-GEV Stage 1 (MLLM-native grounding) achieves performance benefiting from the model's richer semantic priors for fine-grained structural parts. Notably, Stage 2—which further filters irrelevant regions through perceptual verification—raises IoU to 0.59 and Precision to 0.89, demonstrating that the cascaded design does not amplify errors but instead *progressively improves* overall grounding reliability. However, we acknowledge that the overall performance of our model can be further boosted with the continuous improvement of grounding performance.

C.4 Impact of Model Scale

Tab. 7 examines the scalability of P-Gaze along two dimensions. For the retrieval backbone (Tab. 7a), upgrading to CLIP-L/14 or the domain-specialized FG-CLIP (Xie et al., 2025) yields consistent gains, indicating that Dynamic-p effectively amplifies stronger retrieval models through more precise uncertainty estimation. For the reasoning MLLM (Tab. 7b), scaling from Qwen3VL-4B to 32B brings steady improvements across all datasets.

C.5 Hyperparameter Sensitivity

Fig. 6 evaluates the sensitivity to α, which controls the fraction of residual probability mass included in the adaptive budget. For fixed retrieval probabilities, increasing α cannot decrease p_d or the

Table 7: **Scalability analysis of P-Gaze. (a)** Impact of stronger retrieval backbones (CLIP-L/14, FG-CLIP). **(b)** Impact of scaling the MLLM (Qwen3VL series, 4B to 32B). P-Gaze benefits consistently from stronger backbones in both stages.

(a) Impact of Retrieval Model

Ret. Model	CUB	Air	Cars
CLIP-B/16	61.32	41.61	80.11
CLIP-L/14	61.41	42.20	80.74
FG-CLIP (Xie et al., 2025)	**61.57**	**43.14**	**81.52**

(b) Impact of Inference Model Size

Infer. Model	CUB	Air	Cars
4B	59.93	40.02	79.48
8B	61.32	41.61	80.11
32B	**63.05**	**42.00**	**80.83**

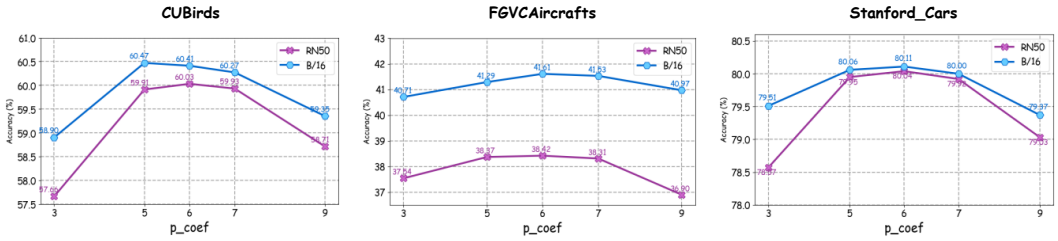


Figure 6: Sensitivity analysis of the hyperparameter α in Dynamic-p across datasets.

selected prefix length. The reported accuracy curves characterize the empirical sensitivity to this choice; the monotone, saturating mapping alone does not guarantee an accuracy plateau.

Tab. 8 reports variants that restrict the class set used to aggregate probability margins. For a retained prefix of M classes, $2 \le M \le N$, the corresponding statistic is

$$U_M = \frac{M-1}{(M-1)+\beta\sum_{i=2}^{M}(p_{(1)}-p_{(i)})} = \frac{M-1}{(M-1)+\beta(Mp_{(1)}-S_M)}, \qquad S_M = \sum_{i=1}^{M} p_{(i)}. \tag{21}$$

The ratio M/N specifies a fraction of ranked classes, not a cumulative probability threshold. Truncation limits the contribution of the lowest-ranked alternatives to the mean-margin statistic. In this formulation, the final candidate set is still selected from the full cumulative distribution. The default rule uses $M = N$; retained-prefix variants examine sensitivity to the aggregation set rather than define an additional requirement of Dynamic-p.

Table 8: **Category Ratio Ablation.** We evaluate performance across three datasets under different ratios of retained categories.

Ratio	**CUB**	**Aircraft**	**Cars**
0.50	58.66	40.01	78.29
0.75	60.95	41.22	79.86
0.95	61.30	41.57	80.01
1.00	**61.32**	**41.61**	**80.11**

C.6 Stability Analysis

Since our framework involves stochastic components—most notably the autoregressive decoding of the MLLM during evidence extraction and reasoning—it is important to verify that the final performance is stable across independent runs. To this end, we execute the complete P-Gaze pipeline five times on three fine-grained datasets (CUB, Aircraft, and Cars) using ViT-B/16 as the retrieval backbone, with identical hyperparameters ($\alpha = 6$) and infrastructure across all runs. As reported in Tab. 9, the standard deviations are consistently negligible (all $\le 0.12\%$), confirming that the stochastic variation introduced by MLLM decoding has minimal impact on the final classification accuracy.

D PROMPT TEMPLATES AND AREA LISTS

We provide the complete prompt templates and dataset-specific area lists used in P-Gaze. Fig. 10 presents the four prompt templates, and Fig. 7 shows the target area lists L_{area} for all eleven datasets.

D.1 PROMPT TEMPLATES

As shown in Fig. 10, P-Gaze employs four prompt templates, each corresponding to a specific stage of the reasoning pipeline:

Table 9: Mean and standard deviation of P-Gaze on three representative fine-grained datasets.

CUB	Aircraft	Cars
61.35 ± 0.08	41.55 ± 0.05	80.15 ± 0.12

(1) Contrastive Attribute Decoupling. This prompt instructs the LLM to perform contrastive analysis over the initial category descriptions within the candidate set C^*. It explicitly requires the model to remove shared, non-discriminative characteristics and retain only the 2–3 most distinctive visual traits per category, yielding the decoupled description set $D_{\text{decoupled}}$.

(2) Active Visual Probing. This prompt guides the MLLM to localize discriminative regions in the query image based on the target area list L_{area}. The model is instructed to only ground instances that are clearly visible and report bounding box coordinates in a structured JSON format, producing the region set B.

(3) Evidence Description. This prompt directs the MLLM to generate visual evidence descriptions from the query image. We design different variants tailored to different dataset characteristics.

(4) Evidence-Attribute Alignment. This prompt formulates the final inference as a cross-examination between the visual evidence D_{evidence} and the decoupled category descriptions $D_{\text{decoupled}}$. The model is required to analyze key features and explicitly cite the corresponding spatial `<area>` tags from the evidence description, producing the predicted category y^* along with an interpretable reasoning chain $\mathcal{E}$.

D.2 DATASET-SPECIFIC AREA LISTS

The target area list L_{area} is a core component of the Active Visual Probing stage in S-GEV, defining which semantic regions the MLLM should localize for each dataset. As shown in Fig. 7, we define L_{area} for all eleven datasets. Each list is curated based on domain knowledge to capture the most informative regions for the corresponding visual domain. For fine-grained datasets, the lists target structurally localizable parts: for example, CUB includes eight anatomical regions (beak, head, back, belly, breast, leg, wing, tail). For general recognition datasets, the lists are tailored to general visual elements: for example, EuroSAT targets land-cover components (vegetation, water bodies, buildings, roads, agricultural fields) and Cal specifies general cues (human, object, animal). By providing dataset-specific area lists across all benchmarks, every dataset benefits from the spatially-grounded evidence verification pipeline.

E QUALITATIVE ANALYSIS

E.1 AUTOMATED EVALUATION OF HALLUCINATION MITIGATION

To quantitatively assess the hallucination mitigation capability of P-Gaze, we employ Gemini 3 Pro (DeepMind, 2025) as an automated evaluator. Compared to subjective human annotation, leveraging a commercial MLLM API for hallucination detection is both more scalable and more consistent, as verifying whether a stated visual attribute physically exists in an image is a relatively straightforward perceptual task well-suited to frontier MLLMs. Specifically, we randomly sample 100 images from each of the six fine-grained test sets (CUB, Aircraft, Cars, Flowers, Pets, Food), yielding 600 evaluation instances in total. For each image, we provide Gemini 3 Pro with the original image alongside the reasoning outputs of both the baseline (CasVLM τ=0 with holistic prompting) and P-Gaze, and instruct it to score each output along three dimensions on a 1–5 Likert scale: (i) *Factual Correctness*—whether each stated visual attribute is physically present in the image; (ii)

Table 10: **Automated hallucination evaluation.** Scores (1–5 Likert scale) are averaged over 600 samples across six fine-grained datasets. Higher is better.

Method	Correctness	Consistency	Faithfulness
SCAN	2.61	3.43	2.85
P-Gaze (Ours)	**4.17**	**4.52**	**4.25**

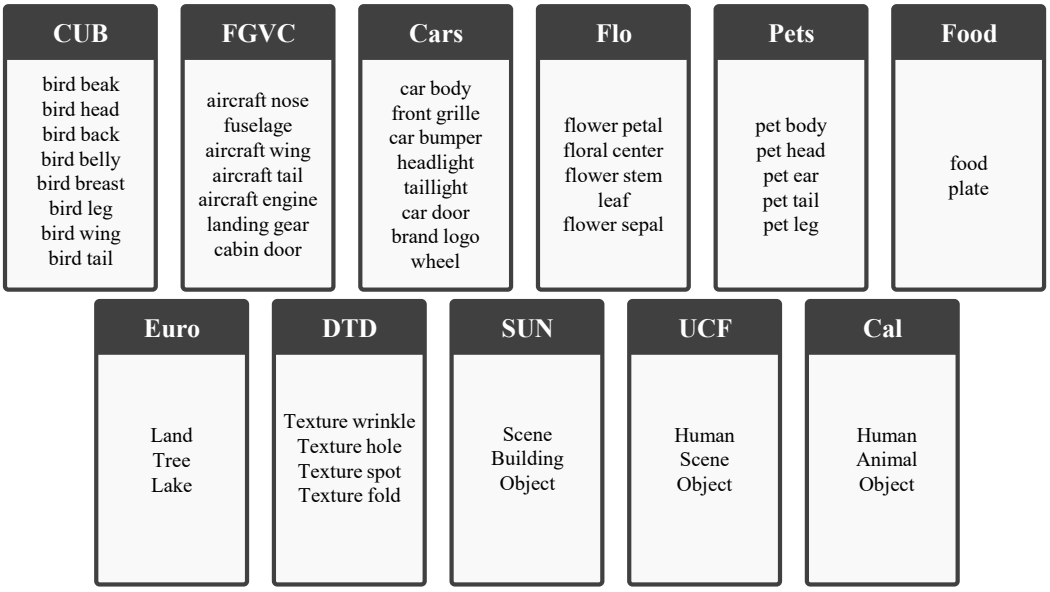


Figure 7: **Dataset-specific target area lists L_{area}.** We define structured region vocabularies for all eleven datasets to guide the spatial grounding in S-GEV.

Spatial Consistency—whether the described regions correspond to the actual locations in the image; and (iii) *Reasoning Faithfulness*—whether the logical chain from evidence to conclusion is valid and free of unsupported inferences. As shown in Tab. 10, P-Gaze achieves significantly higher scores across all three dimensions, with the most pronounced improvement in Spatial Consistency. These automated evaluations support the effectiveness of spatially-grounded evidence verification in reducing unsupported spatial claims and improving evidence-based reasoning.

E.2 Qualitative Case Study

Fig. 8 visualizes the end-to-end reasoning process of P-Gaze on representative fine-grained recognition examples. For each input image, the framework first localizes discriminative regions through S-GEV, then aligns the extracted visual evidence with contrastively decoupled attributes to derive the final prediction. Unlike baseline methods that generate holistic, ungrounded descriptions—often producing plausible-sounding but factually incorrect reasoning—P-Gaze anchors every reasoning step to specific spatial locations in the image, making the classification rationale directly verifiable by visual inspection. The examples demonstrate that P-Gaze not only achieves more accurate predictions but also produces reasoning chains that are transparent and auditable. This spatial grounding significantly enhances both interpretability and trustworthiness, a critical advantage in deployment scenarios where practitioners need to understand and validate the model's decisions.

E.3 Failure Case Analysis

We analyze representative failure cases in Fig. 9 and identify two principal failure modes. First, *extraneous element intrusion*: when multiple objects are in close proximity (left), the grounding stage may inadvertently capture regions beyond the target subject—e.g., including a nearby business jet's empennage alongside the target airliner's tail—thereby contaminating the evidence and mislead-

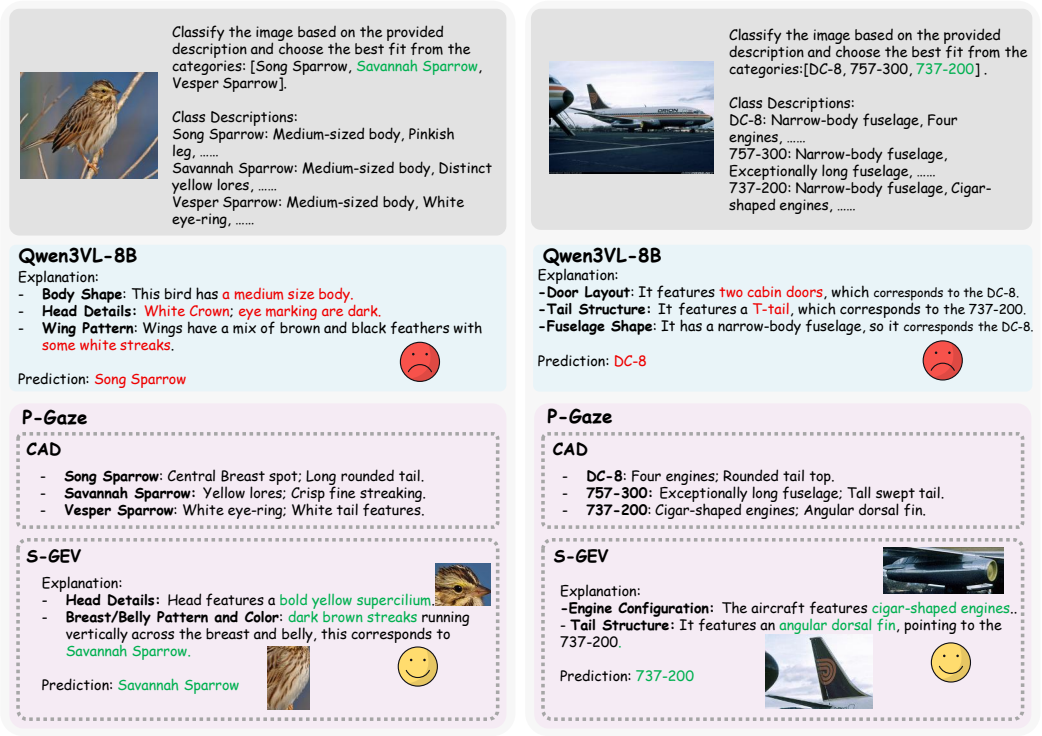


Figure 8: **Qualitative comparison.** For each example, we visualize retrieval candidate set, spatially-grounded evidence regions, decoupled discriminative attributes, and a complete reasoning chain. By explicitly anchoring every reasoning step to verifiable visual evidence, P-Gaze yields accurate predictions supported by highly interpretable and grounded explanations. Conversely, baseline models tend to produce holistic yet ungrounded descriptions, frequently suffering from hallucinated attributes.

ing the alignment. Second, *domain bias from atypical viewpoints*: our pipeline implicitly assumes canonical external viewpoints where structural parts are visible. When the input is a cabin-interior photograph (right), all predefined area attributes (tail, engine, landing gear) become unobservable, S-GEV returns "Uncertainty" for every attribute, and the alignment stage lacks any discriminative evidence for reliable prediction.

F Limitation and Future Work

Our framework has two main limitations. First, Dynamic-p derives its adaptive threshold from a probability distribution computed under a fixed temperature τ; learnable or instance-adaptive temperature scaling could better characterize uncertainty and further refine candidate selection. Second, the reasoning stage is bounded by the grounding model's spatial perception: when localization fails due to occlusion, atypical viewpoints, or extraneous background elements (see Sec. E.3), errors propagate through the entire evidence-attribute alignment pipeline. Incorporating grounding confidence to adaptively fall back to holistic reasoning could mitigate this bottleneck.

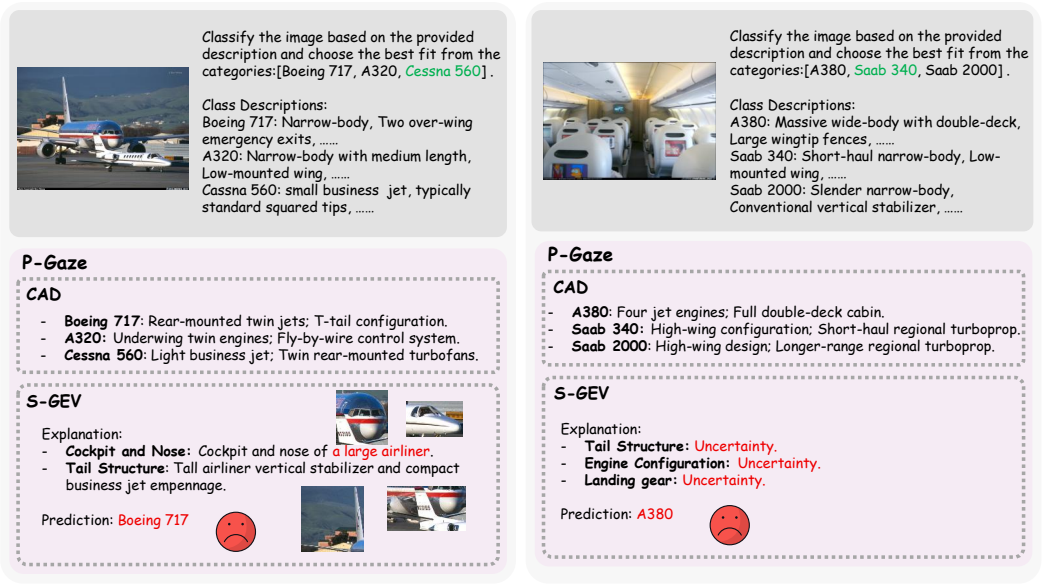


Figure 9: **Representative failure cases of P-Gaze.** Left: grounding introduces background elements (another aircraft's empennage), corrupting the evidence. Right: a cabin-interior viewpoint causes all external structural attributes to be unobservable, leading to complete evidence failure.

Prompt for Contrastive Attribute Decoupling

You are an expert in `{dataset_type}` identification, skilled at distilling complex and redundant species descriptions into their essential diagnostic features. Please refine the provided `{dataset_type}` descriptions by removing redundancy and highlighting discriminative traits.

****Objectives**:**
1. Remove redundant information: Eliminate generic or shared characteristics that appear across most species and offer little value for visual differentiation.
2. Preserve discriminative features: Retain only the 2-3 most distinctive visual traits per species, especially those that differ meaningfully across the given set.

****Input Data****: `{input_data}`

****Refined descriptions**:**

Prompt for Active Visual Probing

Locate and report instances belonging to the following categories: `{image_ground_area}`.

****RULES**:**
1. ONLY ground instances that are **CLEARLY VISIBLE** and **VISUALLY SIGNIFICANT.**
2. Report bbox coordinates in JSON format.
For Example:

```
[
    {{
        'bbox_2d': [x1, y1, x2, y2],
        'label': 'xxx'
    }}, ......
]
```

Prompt for Evidence Description

full:
Please describe the `{dataset_type}` in the image. Ignore irrelevant background information. The goal is to support fine-grained classification based on visible features of the `{dataset_type}`. Only describe visible and distinguishable features. Stick strictly to the following format without adding or changing sections.

****Important**:**
1. If a required feature is not clearly visible in the image, simply write "**uncertain**". Do not invent or guess any detail that is not visible.
2. Keep each attribute description to no more than 20 words.
3. Follow the format exactly as shown below: `{dataset_template}`.

local_area:
You are a `{dataset_type}` experts. Attribute to analyze: `{attribute_name}`

****Images**:**
1. Context image (Whole bird with **RED BOX** marking the region).
2. Detail crops (High-resolution views of the marked region).

****Task**:**
Describe the visual appearance of the `{attribute_name}` in ONE concise sentence.
Constraints:
- Do NOT start with "`The image shows`" or "`In the crop`". Start directly with the description.
- If the attribute is occluded, blurry, or not visible, output **EXACTLY**: "Uncertain due to the angle and lighting."
- Keep it under 20 words."

Prompt for Evidence-Attribute Alignment

Task:
Classify the image based on the provided description and choose the best fit from the categories.

****Image Description****:`{caption}`

****Categories****:`{refined_categories}`

****Instructions****:
Analyze the description to determine the category. In your Analysis, describe the key features that justify your choice. Crucially, whenever you mention a specific feature, you **MUST** append the corresponding <area> tag found in the description immediately after it.

****Final Output Format**:**
Analysis: `[Brief description of key features with their corresponding <area> tags]`
Answer: `[Category Name]`

Figure 10: **Prompt templates used in P-Gaze.** We design four prompts for Contrastive Attribute Decoupling, Active Visual Probing, Evidence Description (`full` and `local_area` variants), and Evidence-Attribute Alignment, respectively.