# OPENSIGHT: EXPLAINABLE OPEN-SET FINE-GRAINED RECOGNITION WITH MLLMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal large language models (MLLMs) have demonstrated strong general capabilities. Still, they remain unreliable in open-set fine-grained visual recognition, where the correct decision is sometimes to reject all provided candidate categories. These tasks often rely on subtle visual distinctions to tell different categories apart. Through controlled analysis, we first find that perception is not the dominant reason for failed recognition. After extracting visual features, models often make decisions based on memorized category priors rather than grounded visual cues, leading to unfaithful reasoning. To this end, we introduce OpenSight, a reasoning-guided reinforcement-learning pipeline that curates samples with verified, attribute-grounded reasoning rationales to enhance fine-grained reasoning in MLLMs. By explicitly identifying hard negatives via subtle visual contradictions, OpenSight enables the model to ground its textual explanations in specific visual evidence. Importantly, the acquired reasoning pattern can be generalized to domains that are visually distinct from the training domain. Across six fine-grained open-set benchmarks, our approach improves the average accuracy by 4.1% and 13.0%, and the harmonic mean by 16.8% and 13.6% over standard GRPO-tuned and base Qwen2.5-VL, respectively. Qualitatively, we show that OpenSight facilitates more evidence-linked explanations and more reliable abstentions.

## 1 INTRODUCTION

Multimodal large language models (MLLMs) have recently demonstrated impressive capabilities in perception and reasoning across diverse visual and textual tasks Huang et al. (2023); Yin et al. (2024); Qwen et al. (2025); Chen et al. (2024). Despite their generality, they still face fundamental challenges in fine-grained visual recognition, especially in open-set settings. When none of the provided options match the visual input, MLLMs often fail to reliably refuse all options Kirichenko et al. (2025). Such limitations undermine their trustworthiness in high-stakes domains, such as biodiversity, where subtle distinctions and awareness of novelty are critical for distinguishing and discovering categories Huang et al. (2025); Omiye et al. (2024).

While previous work attributed MLLMs' failures in fine-grained recognition to poor visual grounding or insufficient world knowledge Tong et al. (2024); Khan et al. (2025b), we argue that the core limitation lies in the reasoning process itself. Specifically, *textual reasoning does not match the visual evidence*. Even when MLLMs can attend to the correct image regions, their explanations often rely on memorized textual priors rather than grounded visual cues, leading to hallucinated predictions.

A well-suited testbed for this shortfall is biological species identification, which requires distinguishing among visually similar categories based on subtle traits such as beak shape, wing bars, or tail length. Moreover, the open nature of biodiversity means novel or unlisted species are frequent, demanding models that reason about fine-grained differences yet confidently abstain when no known option matches. These properties make species classification a realistic and rigorous environment for evaluating MLLMs' open-set awareness.

We first empirically analyze MLLMs' reasoning behaviors and reveal accurate visual perception of various concepts. However, in the decision reasoning step, these visual cues are not effectively leveraged to produce the prediction. Building on these findings, we propose a framework, named



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.

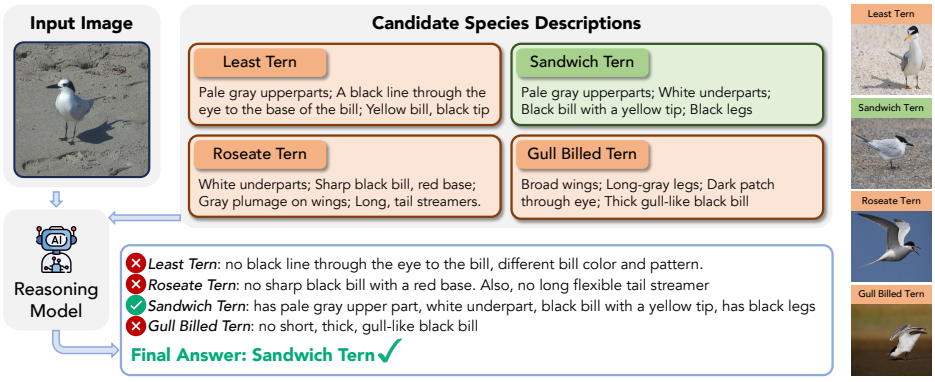



Figure 1: Our model's fine-grained reasoning process. Given a query image and candidate categories with textual descriptions, the model generates grounded reasoning for each option, verifying visual traits and identifying contradictions. It then selects the matching species or rejects all options with "None of the above." We note that the candidate species images shown above are not provided to the model; they are displayed only to illustrate the difficulty of fine-grained recognition among similar species.

**OpenSight**, to explicitly reinforce grounded reasoning through (i) collecting high-quality reasoning exemplars grounded in visual evidence, and (ii) reasoning-guided GRPO to optimize reasoning quality and refusal behavior. Our framework naturally extends to novel-species recognition, where the model must justify both acceptance and rejection decisions through grounded reasoning.

After training with images from the CUB dataset Wah et al. (2011), we evaluate the acquired model on three bird datasets (CUB Wah et al. (2011), Birdsnap Berg et al. (2014), iNaturalist Birds Van Horn et al. (2018)) and three non-bird fine-grained benchmarks (Flowers Nilsback & Zisserman (2008), DTD Cimpoi et al. (2014), Cars Krause et al. (2013)). Although trained only on bird recognition, the model generalizes to other fine-grained tasks: our reasoning-guided GRPO improves average accuracy across the six benchmarks by 4.1% over standard GRPO trained on the same images, and by 13.0% over the base Qwen2.5-VL model. These results show that guiding the reasoning process substantially enhances generic fine-grained discrimination.

Beyond higher accuracy, our model exhibits qualitatively better reasoning, producing faithful and interpretable justifications, as well as more reliable refusals ("none of the above"). We further show qualitative reasoning examples to prove that OpenSight facilitates the effective use of visual evidence in text reasoning. Our key contributions are:

- We introduce a controlled and rigorous benchmark to evaluate MLLMs' reasoning reliability under fine-grained and open-set conditions with species classification.
- We develop OpenSight, including a reasoning-oriented data curation pipeline and reasoning-guided GRPO objectives, enabling MLLMs to learn to generate grounded, evidence-based explanations.
- We adapt and train Qwen2.5-VL Qwen et al. (2025) for this task, achieving significantly improved performance across six fine-grained evaluation datasets. Moreover, the model is equipped with explicit refusal behavior ("none of the above") and fine-grained visual reasoning capabilities.

## 2 Related Work

**Fine-Grained Recognition** distinguishes visually similar categories using subtle cues Wei et al. (2022); Xu et al. (2022); Jing et al. (2024); Wah et al. (2011); Berg et al. (2014); Van Horn et al. (2018). FOCI Geigle et al. (2024) extends this setting to MLLMs through multiple-choice evaluation and negative mining, but remains closed-set Radford et al. (2021); Naeem et al. (2023). We instead test fine-grained text-image matching when the correct candidate may be absent, requiring both discrimination and reliable abstention.

**MLLM Reasoning and Explainability.** Recent MLLMs show strong visual-language understanding and reasoning Qwen et al. (2025); Chen et al. (2024); Yao et al. (2024); Lu et al. (2024); Touvron et al. (2023); Team et al. (2025); OpenAI (2024b); AI (2024); Jiang et al. (2024), aided by pre-training, supervised fine-tuning, and RLHF Ouyang et al. (2022). Yet they often respond overconfidently to incorrect or ambiguous options Kirichenko et al. (2025). Explainability has likewise progressed from saliency maps Selvaraju et al. (2019); Petsiuk et al. (2018) to textual multimodal reasoning Qwen et al. (2025); Chen et al. (2024); OpenAI (2024a), but generated explanations can hallucinate or remain weakly grounded Bai et al. (2025); Zellers et al. (2019). OpenSight therefore requires explanations to cite verifiable visual traits and evaluates their alignment with image evidence.

**Open-Set Recognition** Scheirer et al. (2013) extends closed-set classification by requiring models to classify known samples while rejecting unseen classes Bendale & Boult (2016); Vaze et al. (2022); Oza & Patel (2019), explored across classification, detection, and segmentation Dhamija et al. (2020); Hwang et al. (2021); Zhou et al. (2023). In web-scale MLLMs such as Qwen2.5-VL Qwen et al. (2025), InternVL Chen et al. (2024), and DeepSeek-VL Lu et al. (2024), however, "unseen" becomes ambiguous since these models are trained on vast multimodal data. The challenge thus shifts from detecting novel classes to verifying whether textual descriptions are semantically consistent with visual evidence, reframing open-set reasoning as calibrated refusal and grounded justification Geifman & El-Yaniv (2017); Kirichenko et al. (2025); Mukhoti et al. (2020). We evaluate MLLMs in a constrained setting where the model must reject all options when none align with the visual input.

## 3 Problem Definition and Exploration

We use the widely adopted bird species recognition as the controlled environment for our analysis. Given an image $I$ of a bird and a set of four candidate species $\{S_i\}_{i=1}^4$ and their corresponding descriptions $\{D_i\}_{i=1}^4$, the objective is to predict the correct species $S^*$ or output *"None of the above"* if no match exists. [1]

The model generates a textual response

$$R = \{\text{Reasoning: } r_1, \ldots, r_T, \text{ Answer: } a\},$$

where the reasoning sequence $\{r_t\}_{t=1}^T$ contains the explanation, and $a$ denotes the answer among $\{S_i\}_{i=1}^4 \cup \{\text{None}\}$.

The task thus requires both (1) visual-textual understanding and (2) reasoning grounded in factual attributes of the candidate species. An ideal model should correctly identify a species by associating visual cues in the image with ground-truth linguistic attributes of the corresponding species.

**Visual Perception.** It is a common blame that MLLMs cannot accurately localize visual concepts. We accordingly conduct an experiment to analyze the perceptual ability of the base Qwen2.5-VL model. Specifically, we formulate the task as a series of binary questions that do not explicitly disclose species names based on the CUB dataset Wah et al. (2011) and its annotated attributes (details in Appendix):

> For each body part ("eye," "wing," "body," "leg," "neck," "bill"), we ask:
> "Is the color of the {part} {color}? Only answer Yes or No."

We then study whether the model can attend to these visual parts necessary for reliable bird species recognition. Each query is posed twice for each part, once with the correct color and once with an incorrect one, to ensure that the model's responses are visually grounded rather than memorized. The model answers "Yes" correctly 75.2% of the time and "No" correctly 87.5% of the time, indicating a strong ability to perceive body-part colors.

To confirm these responses are visually grounded rather than driven by memorized priors, we run a no-image experiment posing the same questions without visual input. When forced to choose "Yes"

[1] We follow the four-candidate evaluation protocol established in prior studies Geigle et al. (2024); He et al. (2025).

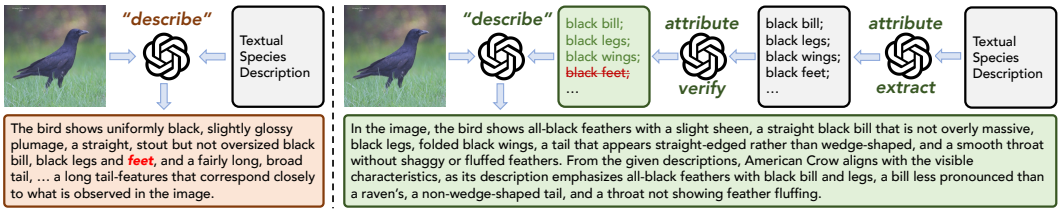



Figure 3: Reasoning collection directly from GPT-5 (left) vs our pipeline (right). GPT-5 hallucinates about the invisible feet, while our method mitigates the hallucination via explicit verification. More examples in Appendix.

or "No," the model answers "No" in 82% of cases; when unconstrained, it never answers "Yes" and instead consistently returns "I'm sorry, but you haven't provided any context or image for me to refer to," suggesting minimal reliance on attribute memorization.

We additionally evaluate object-level localization on the CUB-200-2011 Wah et al. (2011) test set using Qwen2.5-VL, achieving an average IoU of 89.6%, indicating strong spatial grounding. Attention visualizations for color-related tokens (e.g., "red" when asking "What is the color of the eye?") in fig. 2 show the model focuses precisely on the corresponding region, such as the beak or eye. Together, these findings suggest Qwen2.5-VL yields reliable perception and is not the major bottleneck for fine-grained recognition, motivating our focus on the reasoning process. We next study how to improve MLLM fine-grained reasoning.



Figure 2: Attention Visualization when asked about the color of the beak (left) and the color of the eye (right).

## 4 OpenSight

We surmise that improving the fine-grained open-set capabilities of MLLMs requires explicitly grounding their reasoning process in visual evidence. Towards this, we leverage species descriptions from Khan et al. (2025a) for birds to collect corresponding reasoning annotations, as detailed below. For non-bird fine-grained datasets, we prompt GPT-5 to generate class descriptions, which are subsequently verified by human annotators to ensure accuracy and consistency. During inference, the model integrates visual evidence with these textual descriptions, treating them as structured linguistic attributes to guide decision-making.

### 4.1 Reasoning Collection

We prompt GPT-5 with a species image and textual descriptions to generate reasoning. On 200 randomly sampled CUB-200 images covering all species, the model often hallucinates traits described in text but absent from the image (fig. 3, left).

To address this, we propose a structured reasoning generation pipeline grounded explicitly in visual evidence, consisting of three steps. (1) Attribute Extraction: We first extract key visual attributes for each species from its textual description by prompting GPT-5. (2) Attribute Verification: For each training image, GPT-5 then determines which of the extracted attributes from the species description are visually observable in the image. (3) Reasoning Generation: Finally, GPT-5 produces a concise, structured explanation for why the image belongs to the target species, conditioned on the species description but constrained to use only the verified visual attributes from Step 2.

This explicit filtering of textual attributes through visual verification before generating explanations in our pipeline enforces faithful grounding in observable cues and mitigates text-induced hallucina-

tions. As illustrated in fig. 3 (right), this structured approach effectively eliminates the hallucinations commonly observed under direct prompting.

For each image, we also generate reasoning explaining why it does not belong to any of the four hard-negative species identified via CLIP similarity. This follows the same procedure shown in Fig.3 (right), except GPT-5 reasons about why the image fails to match each hard-negative description. This contrastive supervision helps the model distinguish fine-grained inter-species differences and learn cases where no correct option is present. Our pipeline prevents the model from introducing attributes neither present in the image nor specified in the description, ensuring both positive and negative reasoning are grounded in verifiable visual evidence.

**Quantitative Assessment.** We compare our structured reasoning pipeline with direct prompting on how faithfully each method grounds its reasoning in the species' textual descriptions. For both cases, we randomly select 200 samples, extract all morphological parts mentioned in the reasoning using GPT-5. We then compute the intersection-over-union (IoU) between the parts extracted from the reasoning and those from the species description. This metric reflects how closely the acquired reasoning aligns with the textual evidence defining each species. We observe that direct prompting achieves an IoU of 42.4, whereas our reasoning-guided approach attains 75.2, indicating that our method produces reasoning that is significantly more consistent with the descriptive attributes. This confirms that the structured collection process effectively constrains the model to rely on visual-textual evidence rather than hallucinated or generic priors.

We further conduct a human evaluation on the 200 randomly sampled reasoning outputs. Human annotators compare explanations generated by our pipeline against those produced via direct prompting, focusing on groundedness and susceptibility to hallucination. Human evaluators consistently preferred the explanations generated by our pipeline over those produced via direct prompting by 28%. For the remaining samples, both approaches were judged to be of comparable quality, with no clear preference. More details are provided in section C.2.

### 4.2 Reasoning-Guided GRPO

We adopt **Group Relative Preference Optimization (GRPO)** Shao et al. (2024) to align a pre-trained MLLM $\pi_\theta$ with the fine-grained recognition task. We optimize the model(policy) $\pi_\theta$ by maximizing expected reward under the model's own rollouts R:

$$\max_{\theta} \mathbb{E}_{R\sim\pi_\theta(\cdot|I,\{D_i\})}\big[r(R)\big], \tag{1}$$

where $r(R)$ denotes the scalar reward assigned to the answer.

**Baseline.** In the baseline setting, $r(R)$ consists of two binary rewards: 1) whether the predicted answer $a$ matches the ground-truth label $y^*$, 2) whether $R$ follows a pre-defined format. While this setup often induces emergent reasoning, it lacks explicit supervision of reasoning quality, which can lead to hallucinations and inadequate explanations. Details on the GRPO background are provided in section B.

**Reasoning reward.** To encourage faithful and human-aligned reasoning, we extend GRPO with a structured composite reward. Each training instance is represented as

$$(I, \{D_i\}_{i=1}^4, R^*, y^*),$$

where $R^*$ is our collected reasoning(section 4) and $y^*$ is the correct label (or "None of the above").

The overall reward is defined as the sum of three individual reward components:

$$r_{\text{total}}(R) = r_{\text{format}}(R) + r_{\text{answer}}(a, y^*) + r_{\text{reason}}(R, R^*) \tag{2}$$

**Format Reward.** To guide the model towards producing a response in the desired format, we require the output to follow the format `<think>reasoning process here</think><answer>final answer</answer>`. We then apply regular expression matching to verify whether the model output follows the desired format.

$$r_{\text{format}}(R) = \begin{cases} 1, & \text{if } R \text{ follows the format,} \\ 0, & \text{otherwise.} \end{cases}$$

**Answer Reward.** Answer reward $r_{\text{answer}}$ to evaluate the final prediction of policy model is

$$r_{\text{answer}}(R, y^*) = \mathbb{1}[a = y^*],$$

where $a$ denotes the final answer extracted from the model's response $R$, and $y^*$ is the ground-truth label. This reward assigns a value of 1 if the predicted answer matches the ground truth and 0 otherwise, thereby directly encouraging the policy to produce correct final predictions.

**Reasoning Reward.** Let $\hat{R}$ denote the reasoning portion of the model's output. We employ a frozen LLM-based judge $J(\cdot, \cdot)$ that evaluates the semantic and factual alignment between $\hat{R}$ and the ground-truth reasoning $R^*$:

$$r_{\text{reason}}(\hat{R}, R^*) = J(\hat{R}, R^*) \in [0, 1]. \tag{3}$$

The judge assigns a scalar score based on three criteria: (i) **Factual correctness** whether the stated attributes are accurate and consistent with the ground-truth reasoning; (ii) **Logical coherence** whether the explanation forms a clear and sound argument linking visual attributes to the predicted class; and (iii) **Semantic alignment** the degree of overlap in meaning and key discriminative cues between $\hat{R}$ and $R^*$. The final score, normalized to $[0, 1]$, provides a soft supervision signal that encourages faithful, structured, and semantically aligned reasoning.

#### 4.2.1 Implementation Details

We initialize $\pi_\theta$ from the pretrained **Qwen2.5-VL-8B** model. For each input $(I, \{D_i\})$, multiple reasoning–answer rollouts are sampled, scored using the three reward components, and aggregated via GRPO to update $\theta$. Training is performed for 2,200 steps on four A100 GPUs, with a batch size of 4 per device, $G = 8$, gradient accumulation of 2 steps, $\beta = 0.04$, and a LoRA rank of 64. We used Qwen3-4B-Instruct-2507 Team (2025) as our reward model judge.

## 5 Experimental Results

In this section, we describe our experimental setup (datasets, evaluation metrics, and baselines), then present results and an ablation analysis on the contribution of each component.

### 5.1 Experimental Setup

**Benchmarks.** We train with GRPO using only the CUB-200-2011 Wah et al. (2011) training split and evaluate on six fine-grained recognition benchmarks grouped into bird and non-bird datasets.

The bird benchmarks are CUB-200-2011 Wah et al. (2011) (200 species; 5,794 test images), Birdsnap Berg et al. (2014) (500 species; 2,443 test images), and the Aves subset of the iNaturalist Van Horn et al. (2018) validation set (1,323 species; 13,230 images). The non-bird benchmarks are Stanford Cars Krause et al. (2013) (196 classes; 8,041 test images), DTD Cimpoi et al. (2014) (47 texture classes; 1,692 test images), and Oxford Flowers-102 (102 flower classes).

**Evaluation Benchmark.** We retrieve the nearest neighbors for each test image using CLIP Radford et al. (2021) embeddings. For each species, we randomly sample $\mathcal{P}\%$ of the images and construct four candidate options from their top five most similar samples, ensuring that the correct species is not among them. This represents the open-set(novel species) part of the benchmark. For the remaining $100 - \mathcal{P}\%$ of the images, we ensure that the correct species appears within the top four options. This strategy guarantees the presence of challenging distractors (hard negatives) while preserving fine-grained inter-species similarity, making the recognition task both realistic and discriminative. We show results for $\mathcal{P}$ as 20 and 30.

**Metrics.** For evaluation, we use Overall Accuracy across the entire test set and the Harmonic mean of accuracy over the open-set(novel) and closed set.

**Compared models.** In our experiments, we evaluate several MLLMs on our proposed task. Specifically, we assess Qwen2.5-VL-7B-Instruct Qwen et al. (2025), InternVL3-8B Chen et al. (2024), Idefics-2 Laurençon et al. (2023), Phi-4 Microsoft et al. (2025), and MiniCPM-V2.6 Yao et al. (2024). We also include FineDefics He et al. (2025), which is built on top of Idefics-2 Laurençon

| | $\mathcal{P} = 20$ | | | | | | $\mathcal{P} = 30$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CUB** | | **Birdsnap** | | **iNaturalist Birds** | | **CUB** | | **Birdsnap** | | **iNaturalist Birds** | |
| | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean |
| Qwen 2.5-VL | 43.8 | 35.8 | 38.0 | 35.9 | 32.2 | 34.2 | 39.7 | 35.0 | 34.5 | 31.9 | 30.3 | 33.1 |
| Finedefics | 25.3 | 0.0 | 24.0 | 0.0 | 20.0 | 0.0 | 22.6 | 0.0 | 20.4 | 0.0 | 17.8 | 0.0 |
| Idefics-2 | 37.4 | 0.0 | 27.3 | 0.0 | 21.4 | 0.0 | 33.7 | 0.0 | 23.9 | 0.0 | 18.9 | 0.0 |
| Intern3-VL | 40.8 | 19.6 | 38.4 | 23.9 | 33.4 | 30.8 | 37.6 | 23.4 | 32.3 | 20.2 | 31.1 | 30.0 |
| MiniCPM-V-2.6 | 38.9 | 23.4 | 28.3 | 23.3 | 24.6 | 22.6 | 35.3 | 19.9 | 25.1 | 19.5 | 23.3 | 22.2 |
| Phi4 | 38.5 | 11.9 | 25.6 | 19.0 | 24.6 | 13.1 | 34.0 | 11.5 | 23.4 | 17.6 | 22.8 | 13.2 |
| Qwen GRPO | 58.7 | 33.3 | 50.1 | 28.6 | 43.2 | 37.6 | 53.2 | 34.8 | 45.1 | 29.9 | 39.9 | 36.4 |
| Ours | **65.7** | **47.5** | **52.8** | **41.3** | **46.7** | **48.8** | **60.5** | **47.1** | **49.4** | **41.4** | **44.0** | **46.8** |

Table 1: Birds Recognition Results: We report the overall accuracy and the harmonic mean of accuracy on the closed and open-set test sets. Qwen-GRPO represents Qwen2.5-VL trained with standard GRPO. Detailed results are presented in Appendix.

| | $\mathcal{P} = 20$ | | | | | | $\mathcal{P} = 30$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Flowers** | | **DTD** | | **Cars** | | **Flowers** | | **DTD** | | **Cars** | |
| | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean |
| Qwen 2.5-VL | 66.4 | 44.5 | 55.9 | 34.4 | 53.9 | 11.5 | 59.0 | 45.1 | 48.8 | 32.0 | 47.0 | 11.4 |
| Finedefics | 60.3 | 0.0 | 20.0 | 0.0 | 58.3 | 0.0 | 54.3 | 0.0 | 16.0 | 0.0 | 52.8 | 0.0 |
| Idefics-2 | 44.9 | 0.0 | 39.4 | 30.1 | 52.3 | 0.8 | 39.7 | 0.0 | 36.5 | 31.7 | 45.4 | 0.6 |
| Intern3-VL | 48.3 | 21.2 | 59.4 | 44.0 | 37.2 | 24.3 | 50.7 | 22.7 | 52.1 | 40.0 | 33.5 | 25.1 |
| MiniCPM-V-2.6 | 40.5 | 29.8 | 54.1 | 14.9 | 46.8 | 15.0 | 46.7 | 27.6 | 47.5 | 22.9 | 40.6 | 14.8 |
| Phi4 | 43.4 | 33.3 | 45.6 | 44.2 | 30.0 | 11.1 | 41.0 | 32.8 | 38.8 | 39.3 | 25.7 | 10.4 |
| Qwen GRPO | 72.0 | 29.8 | 60.5 | 34.9 | 62.0 | 9.2 | 63.1 | 28.1 | 52.8 | 33.3 | 53.7 | 10.0 |
| Ours | **75.2** | **47.4** | **62.1** | **49.2** | **66.2** | **42.5** | **66.5** | **45.4** | **56.6** | **49.0** | **58.0** | **41.1** |

Table 2: Non-Birds Recognition Results: We report the overall accuracy and the harmonic mean of accuracy on the closed and open-set test sets for three datasets.

et al. (2023) specifically for fine-grained recognition and uses CUB-200 as part of its training data. In addition, we include the Qwen2.5-VL model trained using standard GRPO for comparative analysis. We compare with reasoning-based models in section D.2.

### 5.2 Results

**Recognition.** In the upper part of table 1, we present all baselines using their standard base variants across three datasets under $\mathcal{P} = 20$ and $\mathcal{P} = 30$. Both Idefics-2 and Finedefics fail to handle the open-set case where none of the candidates match the image, while other baselines show some ability to predict "none of the above." Qwen-2.5-VL achieves the highest harmonic mean, indicating better calibration in rejecting incorrect options. We select it as our primary backbone to compare GRPO and our reasoning-guided approach.

Both standard GRPO and our Reason-Guided GRPO build on Qwen-2.5-VL. While GRPO improves overall accuracy, it offers limited gains in harmonic mean; our method achieves consistent, significant improvements in both accuracy and harmonic mean across settings. Strong gains on CUB-200 are expected since training used it, but our method also delivers substantial improvements on Birdsnap and iNaturalist, demonstrating robustness and generalization. Furthermore, in table 2 we show the training paradigm generalizes to domains substantially different from birds.

**Non-Birds Recognition.** We report results on three distinct benchmarks: Stanford Cars, DTD, and Oxford-Flowers in Tab.2. Our reasoning-guided approach reasons effectively over images using textual descriptions, yielding a significant gain over the base model. Since we train only on CUB Wah et al. (2011), these results highlight the generalization of our approach. Across all three benchmarks, our method consistently outperforms both the base model and its standard GRPO variant, whose gains are not uniform across benchmarks. This demonstrates that explicitly structured, visually grounded reasoning is key to fine-grained discrimination and open-set robustness in MLLMs.

**Reasoning.** To assess the quality of the model's reasoning process, we evaluate the reasoning generated by each model. Specifically, for each test image, we provide GPT-5-mini with the input image, four candidate species descriptions, the correct answer, and the model's generated reasoning. GPT-5-mini is then instructed to rate the reasoning process along four dimensions: Completeness, Visual and Textual Evidence, Coverage and Coherence, on a scale from 1 to 5.

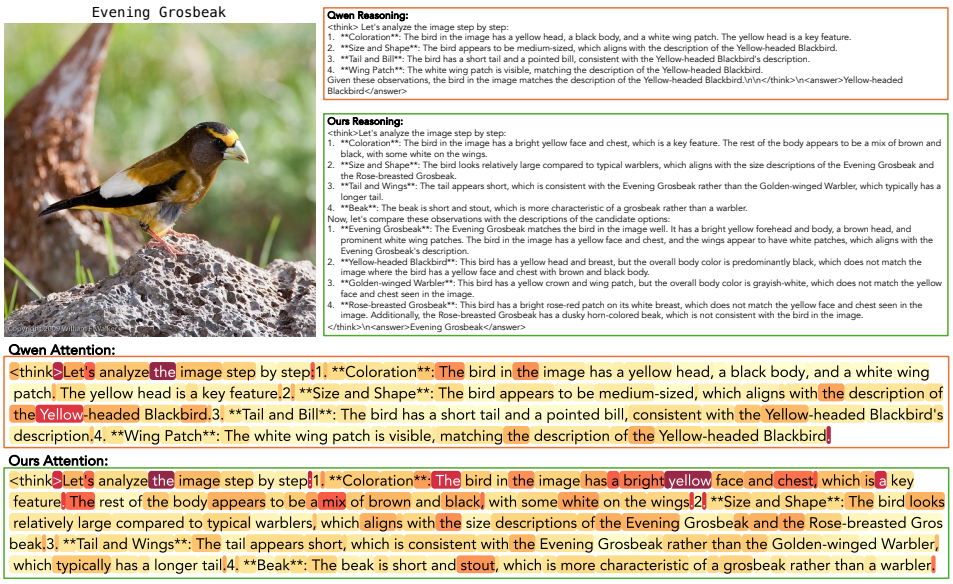



Figure 5: Baseline Qwen2.5-VL hallucinates an incorrect answer, while our model uses grounded, step-by-step reasoning to correctly identify the species. Additional examples with species descriptions as input are presented in the Appendix

- Completeness: Considers reasoning for all candidate options.
- Visual and Textual Evidence: Uses image traits and species descriptions.
- Coverage: Identifies relevant visual attributes.
- Coherence: Presents clear, logically consistent reasoning.

This evaluation measures not only the answer correctness but also how the model reasons, capturing the interpretability and consistency of the underlying thought process.

To complement this automatic evaluation, we conduct a human study on 100 examples sampled from each of the CUB, Birdsnap, and iNaturalist datasets. For each example, three annotators perform pairwise comparisons to select the more visually grounded reasoning between our method and a baseline. Our reasoning is consistently preferred across all backbones, with win rates ranging from roughly 58% to 77%; detailed human-evaluation results are provided in section C.1.

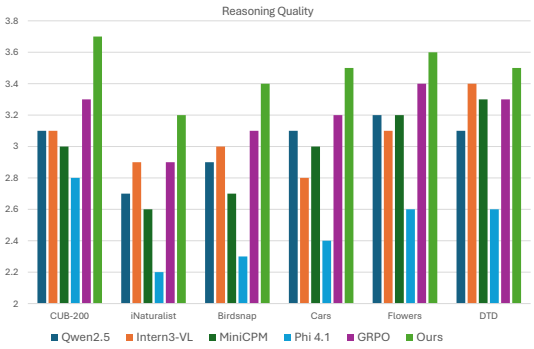



Figure 4: Reasoning Results: We report the average scores across the four dimensions for each dataset.

As shown in fig. 4, our model achieves the highest reasoning quality across all three datasets. Both the GRPO-based model and our model demonstrate superior reasoning compared to the baselines. We omit Idefics-2 and Finedefics, as they fail to produce reasoning for their predictions and score 1.1 and 1, respectively.

Our model produces coherent, visually grounded reasoning by connecting image evidence with candidate descriptions. In the example in fig. 5 (top), it compares concrete visual attributes step by step, whereas Qwen2.5-VL relies on high-level or memorized priors and predicts incorrectly. We further compare their attention patterns in fig. 5 (bottom), separating the output into visual description and reasoning. During reasoning, our model attends to multiple discriminative attributes—including beak shape and face, body, and wing colors—indicating systematic visual-textual comparison. In

contrast, Qwen2.5-VL focuses on the selected species name without verifying relevant attributes, revealing a label-driven shortcut that increases hallucinations and ungrounded predictions.

### 5.3 Ablation Study

We conduct a detailed ablation to evaluate the contribution of each component: the training objective (SFT vs. GRPO), open-set data, and our reasoning reward. We use $\mathcal{P} = 20$ and $= 30$ (section 5.1) on iNaturalist Van Horn et al. (2018) as the primary basis for comparison.

**SFT vs. GRPO with Open-Set Data.** We first establish the baseline and necessity of our core components.

Table 3: SFT vs GRPO vs GRPO+Open-Set. In this experiment, GRPO uses $r_{\text{format}}$ and $r_{\text{answer}}$.

| Model | Open-Set Data | $\mathcal{P} = 20$ | | $\mathcal{P} = 30$ | |
|---|---|---|---|---|---|
| | | Acc. | Har. Mean | Acc. | Har. Mean |
| Qwen 2.5VL | 0 | 32.2 | 34.2 | 30.3 | 33.2 |
| Qwen 2.5VL-SFT | 0 | 37.6 | 0.2 | 34.0 | 0.2 |
| Qwen 2.5VL-GRPO | 0 | **43.2** | **37.6** | **39.9** | **36.4** |
| Qwen 2.5VL-SFT | 100 | 26.8 | 23.5 | 25.5 | 21.4 |
| Qwen 2.5VL-GRPO | 100 | **42.4** | **45.7** | **39.9** | **43.4** |

• Baseline Models: The base Qwen 2.5VL model in table 3 (Row 1) shows poor accuracy with a Harmonic Mean of 34.2. Standard SFT (Row 2) improves accuracy (37.6) but completely fails at open-set rejection, collapsing the Harmonic Mean to 0.2, indicating the model overfits to closed-set classes and loses the ability to abstain. Standard GRPO instead improves both accuracy and abstention.

• GRPO vs. SFT: With open-set data included, the SFT model on 100% open-set data (Row 4) overfits on rejection, dropping accuracy but improving Harmonic Mean. GRPO (Row 5), trained on the same data with $r_{\text{format}}$ and $r_{\text{answer}}$, performs even better, reaching a 47.4 Harmonic Mean with a slight accuracy dip. These experiments confirm GRPO is more effective than standard SFT.

**Impact of the Reasoning Reward.** Our key contribution is the reasoning reward with open-set data (table 4). The critical comparison is between Row 1 (GRPO) and Row 6 (GRPO + open-set data + $r_{\text{reason}}$). Our full model (Row 6) achieves the highest performance (46.7% Accuracy, 48.8 Harmonic Mean at $\mathcal{P} = 20$), versus 43.2% accuracy and 37.6 Harmonic Mean for standard GRPO.

Table 4: GRPO vs Reasoning Guided GRPO

| Model | Reward | Open-Set Data | $\mathcal{P} = 20$ | | $\mathcal{P} = 30$ | |
|---|---|---|---|---|---|---|
| | | | Acc. | Har. Mean | Acc. | Har. Mean |
| Qwen 2.5VL | $r_{\text{format}}, r_{\text{answer}}$ | 0 | 43.2 | 37.6 | 39.9 | 36.4 |
| Qwen 2.5VL | $r_{\text{format}}, r_{\text{answer}}$ | 100 | 42.4 | 45.7 | 39.9 | 43.4 |
| Qwen 2.5VL | $r_{\text{format}}, r_{\text{answer}}, r_{\text{reason}}$ | 0 | 42.0 | 7.9 | 38.7 | 8.5 |
| Qwen 2.5VL | $r_{\text{format}}, r_{\text{answer}}, r_{\text{reason}}$ | 100 | **46.3** | **47.4** | **43.3** | **44.9** |
| Qwen 2.5VL | $r_{\text{format}}, r_{\text{answer}}, r_{\text{vlm}}$ | 50 | 45.6 | 42.5 | 42.8 | 40.8 |
| Qwen 2.5VL | $r_{\text{format}}, r_{\text{answer}}, r_{\text{reason}}$ | 50 | **46.7** | **48.8** | **44.0** | **46.8** |

**Open-Set Data**: In Rows 4 and 5, we vary the percentage of open-set data during training. Trained with only 50% open-set data, our model still outperforms the model without the reasoning reward (Row 2), which used 100%. This shows $r_{\text{reason}}$ is a powerful, data-efficient training signal that improves the model's ability to ground its decisions.

**Comparison to $r_{\text{vlm}}$**: Finally, we test the necessity of our curated reasoning data by replacing $r_{\text{reason}}$ with a generic $r_{\text{vlm}}$ reward, generated by a VLM "judge" (Qwen2.5-VL Qwen et al. (2025)) that scores reasoning without access to our ground-truth explanations. As shown in the second last row of table 4, this substitution degrades performance substantially, with the Harmonic Mean dropping to 42.5. Reasoning quality also drops: the VLM-judge reward scores 3.4 on CUB-200, versus 3.7 for our method and 3.3 for standard GRPO (section 5.2). These findings validate our design: the targeted $r_{\text{reason}}$ reward built on curated reasoning data is decisively more effective than a generic VLM reward, in both overall performance and reasoning fidelity.

## 6 Conclusion

We studied the reasoning limitations of MLLMs in fine-grained and open-set visual recognition using biodiversity recognition as a controlled testbed. Our results show that, despite strong perceptual abilities, existing models struggle to align visual evidence with textual reasoning. To address this, we introduced OpenSight, a reasoning-guided framework combining visually grounded training data, GRPO, and an explicit reasoning reward. Across multiple benchmarks, OpenSight improves accuracy, interpretability, abstention, and reasoning faithfulness over standard GRPO. These findings establish reasoning alignment as a key step toward trustworthy MLLMs in open-world settings.

## ETHICS STATEMENT

This work uses established image benchmarks: CUB-200-2011, Birdsnap, iNaturalist, Stanford Cars, DTD, and Oxford Flowers-102, and introduces no new image collection. Dataset use and any release of derived annotations must follow the licenses of the respective sources. Human participation was limited to blind judgments of generated descriptions and reasoning; results are reported only in aggregate, and annotator identities are not released. Fine-grained recognition systems can reproduce geographic, taxonomic, and photographic biases in their source data. The work aims to improve grounded reasoning and abstention, but these mechanisms do not eliminate hallucinations or guarantee calibrated uncertainty.

## AI USE STATEMENT

Generative AI tools were used as components of the research methodology. GPT-5 generated non-bird class descriptions and the attribute-grounded positive and hard-negative reasoning annotations used by the pipeline; human annotators verified the generated class descriptions and evaluated subsets of the reasoning. GPT-5 was also used to extract morphological parts for the grounding analysis. GPT-5-mini evaluated generated reasoning, Qwen3-4B-Instruct served as the reasoning-reward judge during training, and Qwen2.5-VL served as the generic VLM judge in an ablation. The paper documents these uses, inputs, evaluation criteria, and human checks in Sections 4-5.2 and the appendix.

## REPRODUCIBILITY STATEMENT

Section 3 defines the task and controlled perception analyses; Section 4 describes the reasoning-collection pipeline and its human evaluation; and Section 4.2 specifies the GRPO objective, reward components, backbone, training duration, hardware, batch configuration, rollout count, KL coefficient, LoRA rank, and reward model. The experimental section reports dataset splits, hard-negative and open-set construction, baselines, metrics, and ablations. The appendix further details the perception and human-evaluation protocols and provides additional quantitative and qualitative results. To enable exact replication, we will release the code, data, and models trained with appropriate license.

## REFERENCES

Google AI. Gemini: Google's multimodal ai model. *Google AI Research*, 2024. https://fireflies.ai/blog/gemini-vs-gpt-4.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey, 2025. URL https://arxiv.org/abs/2404.18930.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles, 2025. URL https://arxiv.org/abs/2503.17352.

Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 1021–1030, 2020.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, et al. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025. doi: 10.48550/arXiv.2502.13595. URL https://arxiv.org/abs/2502.13595.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

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

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*, 2025. URL https://openreview.net/forum?id=p3NKpom1VL.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models, 2023. URL https://arxiv.org/abs/2302.14045.

Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1175–1184, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. Fineclip: Self-distilled region-based clip for better fine-grained understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27896–27918. Curran Associates, Inc., 2024. doi: 10.52202/079017-0875. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3122aaa22b2fe83f9cead1a696f65ceb-Paper-Conference.pdf.

Faizan Farooq Khan, Jun Chen, Youssef Mohamed, Chun-Mei Feng, and Mohamed Elhoseiny. Neural catalog: Scaling species recognition with catalog of life-augmented generation, 2025a. URL https://arxiv.org/abs/2505.05635.

Faizan Farooq Khan, Yousef Radwan, Eslam Abdelrahman, Abdulwahab Felemban, Aymen Mir, Nico K. Michiels, Andrew J. Temple, Michael L. Berumen, and Mohamed Elhoseiny. Fishnet++: Analyzing the capabilities of multimodal large language models in marine biology, 2025b. URL https://arxiv.org/abs/2509.25564.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel Bell. Abstentionbench: Reasoning LLMs fail on unanswerable questions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=OkHC30LLpO.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. URL https://arxiv.org/abs/2306.16527.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. 2020.

Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2MVFormer: Large Language Model Generated Multi-View Document Supervision for Zero-Shot Image Classification . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15169–15179, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01456. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01456.

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

Jesutofunmi A. Omiye, Haiwen Gui, Shawheen J. Rezaei, James Zou, and Roxana Daneshjou. Large language models in medicine: The potentials and pitfalls: A narrative review. *Annals of Internal Medicine*, 177(2):210–220, February 2024. ISSN 1539-3704. doi: 10.7326/m23-2772. URL http://dx.doi.org/10.7326/M23-2772.

OpenAI. Gpt-4o: Enhanced multimodal language model. *OpenAI Research*, 2024a. https://openai.com/index/hello-gpt-4o/.

OpenAI. Gpt-4v: Multimodal language model with vision capabilities. *OpenAI Research*, 2024b. https://openai.com/index/gpt-4/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2307–2316, 2019.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018. URL https://arxiv.org/abs/1806.07421.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Kimi Team, Yifan Bai, Yiping Bao, et al. Kimi k2: Open agentic intelligence, 2025. URL https://arxiv.org/abs/2507.20534.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, June 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

S Vaze, K Han, A Vedaldi, and A Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International Conference on Learning Representations (ICLR)*, 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Jul 2011.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4oYxzssbVg.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-Grained Image Analysis With Deep Learning: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(12):8927–8948, December 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3126648. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3126648.

Shu-Lin Xu, Faen Zhang, Xiu-Shen Wei, and Jianhua Wang. Dual attention networks for few-shot fine-grained recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3): 2911–2919, Jun. 2022. doi: 10.1609/aaai.v36i3.20196. URL https://ojs.aaai.org/index.php/AAAI/article/view/20196.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 11 2024. ISSN 2095-5138. doi: 10.1093/nsr/nwae403. URL https://doi.org/10.1093/nsr/nwae403.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Zhongxiang Zhou, Yifei Yang, Yue Wang, and Rong Xiong. Open-set object detection using classification-free object proposal and instance-level contrastive learning. *IEEE Robotics and Automation Letters*, 8(3):1691–1698, 2023.

# APPENDIX: OPENSIGHT: EXPLAINABLE OPEN-SET FINE-GRAINED RECOGNITION WITH MLLMS

The appendix material provides:

- Sec. A covers details about the visual perception experiment.
- Sec. B provides the necessary background.
- Sec. C describes the human evaluation setup.
- Sec. D presents additional results comparing our method with baselines.
- Sec. E shows examples from our data collection pipeline.
- Sec. F presents reasoning examples from our approach and the Qwen2.5-VL baseline.

## A VISUAL PERCEPTION

To construct the visual perception evaluation split, we leveraged the exhaustive part-level attribute annotations provided by the CUB dataset Wah et al. (2011). For every image in the test set, we parsed the annotations to identify the ground-truth color for each visible body part. To formulate the negative query (where the target answer is "No"), we sampled an incorrect color directly from the CUB dataset's existing color vocabulary for that specific anatomical part. This guaranteed that the distractors remained domain-realistic rather than trivially easy. Instances where a specific part was heavily occluded or lacked a definitive color annotation in the ground truth were excluded from that part's evaluation subset.

The final reported scores were calculated as the macro-average accuracy across all valid part-color queries within the entire test set. For the controlled "no-image" ablation, the evaluation pipeline, generation parameters, and prompt templates were kept strictly identical to the main experiment; the only modification was the complete omission of the visual tokens from the model's context window, forcing it to rely entirely on its pre-trained semantic priors.

## B GRPO BACKGROUND

**GRPO.** Group Relative Policy Optimization (GRPO) Shao et al. (2024) is a variant of Proximal Policy Optimization (PPO) Schulman et al. (2017) for reinforcement learning. GRPO directly compares groups of candidate responses. This design eliminates the dependency on a critic model, thereby simplifying the training process. Given a question-answer pair $(q, a)$, the policy $\pi_{\theta_{\text{old}}}$ generates $G$ distinct candidate responses $o = o_1, \ldots, o_G$ through policy sampling. Then, verifiable rewards $r_1, \ldots, r_G$ are calculated for each response. GRPO normalizes these rewards by computing their mean and standard deviation, and then evaluates the relative quality of the responses accordingly.

$$A_i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^G)}{\text{std}(\{r_i\}_{i=1}^G) + \delta}, \tag{4}$$

where $\delta$ is a small constant (e.g., $10^{-8}$ ) for numerical stability, and $A_i$ denotes the relative quality of the $i$-th response within the group $G$. GRPO promotes higher-scoring answers within each group while regularizing the policy $\pi_\theta$ against the reference policy $\pi_{\text{ref}}$ via a KL-divergence penalty $D_{\text{KL}}(\cdot\|\cdot)$. This leads to maximizing the final objective below:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}\left[\{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(q)\right] \tag{5}$$

$$\frac{1}{G}\sum_{i=1}^{N}\{\min[s_1 \cdot A_i, s_2 \cdot A_i] - \beta \mathbb{D}_{KL}[\pi_\theta \| \pi_{\text{ref}}]\} \tag{6}$$

$$s_1 = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} \tag{7}$$

$$s_2 = \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1+\epsilon, 1-\epsilon\right) \tag{8}$$

where $\beta$ is a regularization coefficient that prevents excessive deviation from the reference policy during optimization.

**LLM Reward Model.** Large Language Models (LLMs) are utilized as sophisticated "judge" models for evaluation. Unlike models such as BERT, which may lack the necessary semantic understanding for complex tasks, a modern LLM is employed to perform a nuanced, reference-based comparison Enevoldsen et al. (2025). Specifically, the LLM judge evaluates a candidate response $o_i$ by comparing it against a pre-collected set of "ground truth reasoning" data associated with the input $q$. The reward $r_i$ is therefore a score that reflects the semantic and logical alignment of the candidate response with this ground truth. For the GRPO objective, these alignment scores from the LLM judge serve as the verifiable rewards $r_1, \ldots, r_G$.

# C Evaluation Setup

## C.1 Comparsion against baselines

. We randomly sampled 100 examples each from CUB, Birdsnap, and iNat, and asked three independent annotators to perform pairwise comparisons between the reasoning generated by our method and that of each baseline. Annotators selected the explanation they considered more visually grounded while remaining blind to the source model. As shown in Fig. X, our method is consistently preferred across all baselines, receiving the majority of votes in every comparison. The largest gains are observed over Phi-4.1 and MiniCPM, where our explanations are preferred in approximately 78% and 73% of comparisons, respectively. Even against the strongest competitor, GRPO Qwen, our method is still favored in roughly 58% of cases. These results demonstrate that incorporating structured visual reasoning leads to explanations that are more faithfully grounded in image evidence, rather than relying on generic or language-prior-driven justifications.

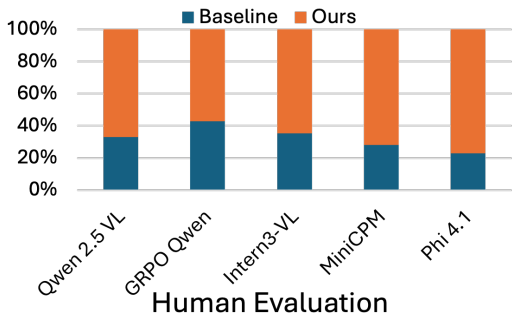



Figure 6: Human evaluation on reasoning quality.

## C.2 Evaluating generated reasoning

. To conduct the human evaluation of the generated reasoning, we recruited 10 independent participants. Each annotator was assigned 20 distinct image samples, yielding a total of 200 randomly sampled reasoning outputs across the study. For each sample, the evaluators were presented with the target image alongside the textual explanations generated by two distinct methods: our structured reasoning pipeline and the direct prompting baseline. These texts were displayed side-by-side in a blind setup. The human annotators were tasked with comparing the explanations, focusing specifically on groundedness and susceptibility to hallucination. They were asked to select the explanation they felt was more accurately grounded in the visual evidence and contained fewer hallucinations, or to explicitly indicate if both approaches produced outputs of comparable quality. The final preference scores were determined by averaging annotators' selections across the entire set of 200 samples.

## C.3 Reasoning reward without open-set data

Adding $r_{\text{reason}}$ without open-set data (table 4, Row 3) lowers the harmonic mean to 7.9, although overall accuracy (42.0) is on par with GRPO (43.2). Separating the two splits shows that closed-

set accuracy in fact increases (51.4 vs. 46.1), while open-set accuracy collapses (4.3 vs. 31.8): the model rarely predicts "None of the above". This is expected from the reward's design. $r_{\text{reason}}$ rewards alignment with the ground-truth reasoning $R^*$, and without open-set samples every $R^*$ explains why one candidate matches, so the reward systematically favors committing to a candidate. Because it is a dense signal, it dominates the group-normalized advantage whenever answer rewards tie, pushing the policy toward the reasoning distribution of the training data, the same failure seen with SFT at 0% (HM 0.2, table 3). Once open-set samples supply rejection reasoning, $r_{\text{reason}}$ yields the best results ( table 4, Row 4: 47.4 HM vs. 45.7 for GRPO), and with only 50% open-set data it still exceeds GRPO trained on 100% (48.8 vs. 45.7). The reasoning reward and open-set reasoning data are therefore complementary: the reward amplifies the reasoning the data contains.

# D Additional Results

## D.1 Hard-Negative Evaluation

| | $\mathcal{P}=20$ | | | | | | $\mathcal{P}=30$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CUB** | | **Birdsnap** | | **iNaturalist Birds** | | **CUB** | | **Birdsnap** | | **iNaturalist Birds** | |
| | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean |
| Qwen 2.5-VL Qwen et al. (2025) | 74.2 | 68.3 | 73.4 | 69.8 | 67.0 | 65.6 | 67.8 | 65.6 | 67.4 | 66.6 | 61.0 | 61.9 |
| Finedefics He et al. (2025) | 46.6 | 0.0 | 24.0 | 0.0 | 25.0 | 0.0 | 42.2 | 0.0 | 28.3 | 0.0 | 22.3 | 0.0 |
| Idefics-2 Laurençon et al. (2023) | 37.9 | 0.0 | 36.3 | 0.0 | 28.8 | 0.2 | 33.5 | 0.0 | 30.7 | 0.0 | 25.3 | 0.0 |
| Intern3-VL Chen et al. (2024) | 76.7 | 65.3 | 77.5 | 69.1 | 70.4 | 65.8 | 69.0 | 62.4 | 69.2 | 64.6 | 60.0 | 59.9 |
| MiniCPM-V-2.6 Yao et al. (2024) | 56.1 | 47.1 | 56.4 | 52.9 | 50.0 | 47.5 | 48.9 | 44.1 | 49.9 | 49.6 | 46.0 | 45.9 |
| Phi4 Microsoft et al. (2025) | 51.9 | 34.9 | 50.6 | 39.6 | 48.1 | 33.9 | 47.1 | 35.7 | 44.9 | 40.5 | 43.6 | 32.2 |
| Qwen GRPO | 86.4 | 78.8 | 85.4 | 80.9 | 80.9 | 51.2 | 78.3 | 76.2 | 75.5 | 74.7 | 70.9 | 71.2 |
| Ours | **90.1** | **86.5** | **89.4** | **87.9** | **83.1** | **83.7** | **81.9** | **82.2** | **80.9** | **82.9** | **75.3** | **78.1** |

Table 5: Recognition Results: We report the overall accuracy and the harmonic mean of accuracy on the closed and open-set test sets. The negatives are chosen randomly, rather than being hard negatives. Qwen-GRPO represents Qwen2.5-VL trained with standard GRPO.

To rigorously evaluate the model's discriminative ability, we introduce a hard-negative evaluation setting in Tab. 1, where CLIP is used to retrieve hard negatives rather than random negatives. This forces the model to distinguish between highly similar species, effectively transforming the task into a fine-grained recognition problem. For completeness, Tab. 5 reports results under the standard random-negative setting. While random negatives primarily evaluate general semantic understanding, the hard-negative setting provides a more stringent assessment of the model's ability to resolve subtle visual differences between closely related classes.

## D.2 Comparison with Reasoning Models

To evaluate the effectiveness of our approach against recent reasoning-focused vision-language models, we compare OpenSight with OpenThinker Deng et al. (2025) and Rethinker Wang et al. (2025) on three representative datasets. As shown in Tab. 7, OpenSight consistently achieves the best performance across all datasets and metrics, improving both closed-set accuracy and harmonic mean. These results demonstrate that our domain-specific reasoning framework remains effective even when compared to models developed specifically for reasoning.

# E Qualitative Data Samples

In Fig. 7, we illustrate two examples in which directly prompting GPT results in hallucinated attributes. In the first case, GPT incorrectly describes a "clear white eye ring" that is not present in the

| | $\mathcal{P}=20$ | | | | | | $\mathcal{P}=30$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CUB** | | **Birdsnap** | | **iNaturalist Birds** | | **CUB** | | **Birdsnap** | | **iNaturalist Birds** | |
| | Closed-Set | Open-Set | Closed-Set | Open-Set | Closed-Set | Open-Set | Closed-Set | Open-Set | Closed-Set | Open-Set | Closed-Set | Open-Set |
| Qwen 2.5-VL Qwen et al. (2025) | 47.6 | 28.7 | 39.6 | **32.9** | 30.3 | 39.3 | 42.1 | 29.9 | 36.3 | 28.4 | 27.3 | 42.2 |
| Finedefics He et al. (2025) | 31.6 | 0.0 | 31.2 | 0.0 | 25.1 | 0.0 | 28.3 | 0.0 | 26.5 | 0.0 | 22.2 | 0.0 |
| Idefics-2 Laurençon et al. (2023) | 46.9 | 0.0 | 35.4 | 0.0 | 26.7 | 0.0 | 42.2 | 0.0 | 31.1 | 0.0 | 23.6 | 0.0 |
| Intern3-VL Chen et al. (2024) | 48.1 | 12.3 | 44.9 | 16.3 | 34.9 | 27.6 | 43.0 | 16.1 | 37.7 | 13.8 | 31.8 | 28.5 |
| MiniCPM-V-2.6 Yao et al. (2024) | 44.5 | 15.9 | 31.1 | 18.6 | 25.7 | 20.1 | 40.8 | 13.2 | 28.1 | 14.9 | 23.9 | 20.7 |
| Phi4 Microsoft et al. (2025) | 46.3 | 6.8 | 29.0 | 14.1 | 28.6 | 8.5 | 40.7 | 6.7 | 26.4 | 13.2 | 26.3 | 8.8 |
| Qwen GRPO | 67.9 | 22.1 | **59.5** | 18.8 | 46.1 | 31.8 | 60.9 | 24.4 | 52.3 | 20.9 | **41.8** | 32.3 |
| Ours | **73.4** | **35.1** | 59.2 | 31.7 | **45.1** | **53.2** | **66.6** | **36.4** | **54.1** | **33.6** | 41.6 | **53.5** |

Table 6: Recognition Results: We report the accuracy on the closed and open-set across the three benchmarks with $P$=20, 30. Qwen-GRPO represents Qwen2.5-VL trained with standard GRPO.

| | $\mathcal{P} = 20$ | | | | | | $\mathcal{P} = 30$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CUB** | | **Birdsnap** | | **iNaturalist Birds** | | **CUB** | | **Birdsnap** | | **iNaturalist Birds** | |
| | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean | Acc. | Har. Mean |
| Openthinker Deng et al. (2025) | 55.4 | 30.5 | 41.8 | 24.4 | 36.8 | 31.0 | xx | xx | 38.2 | 27.6 | 33.7 | 31.0 |
| Rethinker Wang et al. (2025) | 50.7 | 46.4 | 45.5 | 43.2 | 39.5 | 41.2 | 46.4 | 45.2 | 40.9 | 39.3 | 36.7 | 38.9 |
| Ours | **65.7** | **47.5** | **52.8** | **41.3** | **46.7** | **48.8** | **60.5** | **47.1** | **49.4** | **41.4** | **44.0** | **46.8** |

Table 7: Performance comparison with reasoning-based vision-language models. OpenSight achieves the highest accuracy and harmonic mean across all evaluated datasets.

image. In the second, it hallucinates and describes the tail, even though it is not visible. In contrast, our pipeline references only the visual evidence actually present in the image, substantially reducing such hallucinations.

In Fig. 8 and Fig. 9 we show two examples of complete reasoning that include both positive and negative reasoning.



In the image, the bird shows olive upperparts, whitish underparts with a slight olive wash on the breast, ***a clear white eye ring***, and bright white wing bars on darker wings; it also has a relatively broad bill with a pale yellowish lower mandible and a medium-length tail. From the given descriptions, Acadian Flycatcher aligns with the visible characteristics, as its description emphasizes olive upperparts, whitish underparts with an olive wash, a prominent white eye ring, white wing bars, and a wide bill with dark above and yellowish below, which correspond closely to what is observed in the image.

In the image, the bird shows olive upperparts, whitish underparts, two bright white wing bars on darker wings, a dusky tail, and a broad bill with a yellowish-orange lower mandible. From the given descriptions, Acadian Flycatcher aligns with the visible characteristics, as its description emphasizes olive upperparts, whitish underparts, contrasting white wing bars on darker wings and tail, and a notably wide, bicolored bill.



The bird belongs to White-necked Raven. The image shows predominantly black bird with a striking, large white patch on the nape, which is the most notable characteristics described for its species. Its bill is deep and strongly arched with a distinctive tip, exactly as noted for White-necked Raven. **The tail also appears relatively short, consistent with the description of a shorter tail than the other ravens.**

In the image, the bird shows a large white patch of feathers on the nape, glossy black plumage over the neck, throat, breast, and wings, and a deep, arched bill with a distinctly pale bill tip. From the given descriptions, White-necked Raven aligns with the visible characteristics, as its description emphasizes a white-feathered nape contrasting with black plumage and a heavy arched bill bearing a white tip and subtle gloss on the throat, breast, and neck.

Figure 7: Reasoning collection directly from GPT-5 vs our pipeline. Notice how our reasoning text is more accurate and does not suffer from hallucinations.

## F Qualitative Reasoning Samples

As shown in Fig. 5 in the main paper, we show the reasonings generated from the Qwen model and compare them to our model. We report three cases: 1) where Qwen fails and our model gets it right, 2) where both fail, and 3) where our model fails and Qwen gets it right. For all the figures, we present the generated reasoning from both cases, the input image, and species descriptions, along with the attention visualization to illustrate where the model attends while generating the output, as shown in Fig. 5.

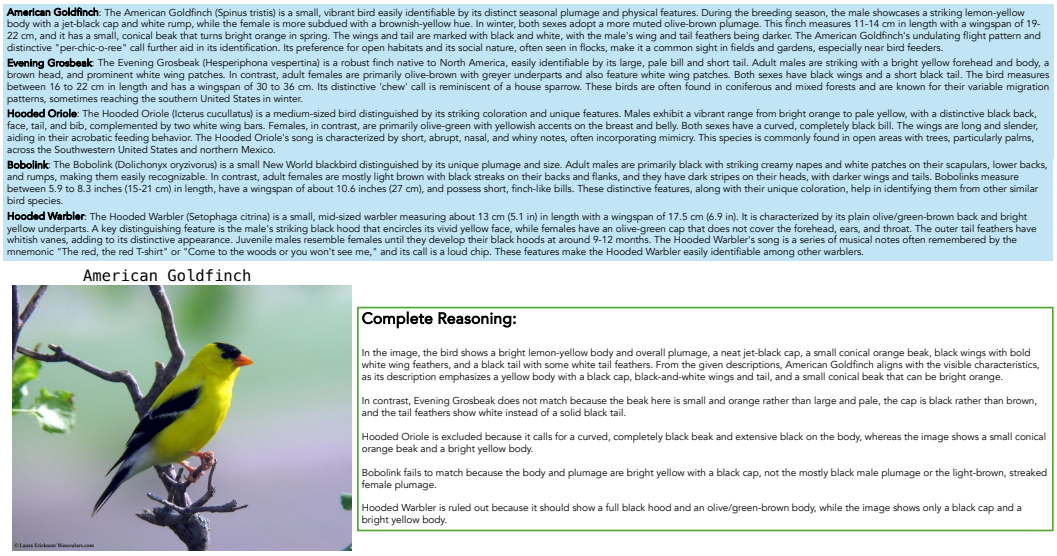



Figure 8: Full reasoning text with both positive and negative reasoning for an image of an American Goldfinch.

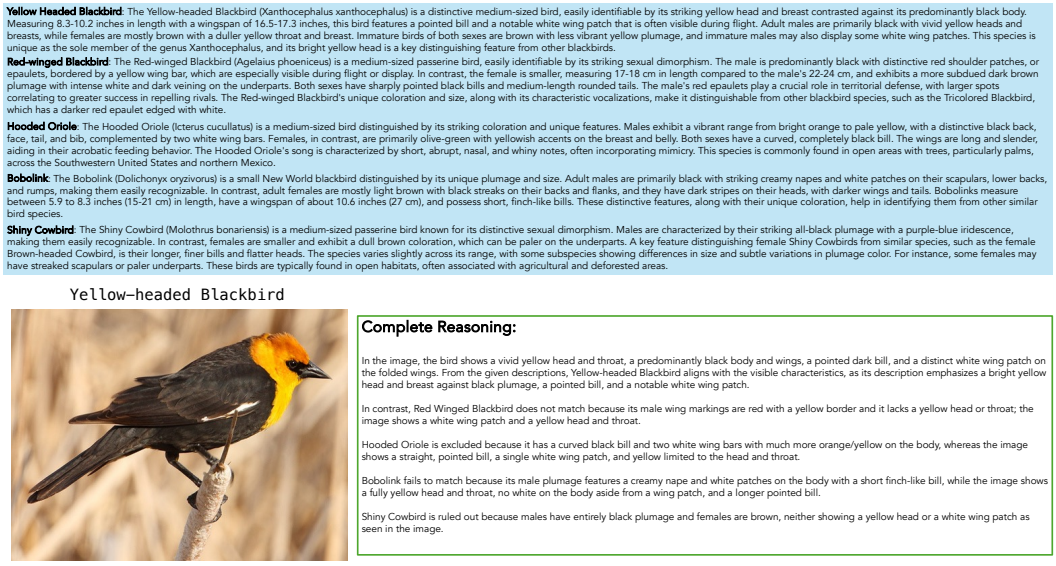



Figure 9: Full reasoning text with both positive and negative reasoning for an image of a Yellow-headed Blackbird.

Figures 10, 11, and 12 show cases where the Qwen model fails and our model produces the correct answer. We can see that:

**Figure 10 (Yellow-headed Blackbird vs. Evening Grosbeak)**:

- The Qwen model incorrectly identifies the bird as the Yellow-headed Blackbird by focusing on the yellow head, black body, and white wing patch.
- Our model correctly identifies the bird as the Evening Grosbeak. Our reasoning correctly notes the bird's short and stout beak (characteristic of a grosbeak, not a warbler), its relatively large size, and its yellow face and chest with brown/black on the body and white wing patches, which specifically aligns with the Evening Grosbeak's description of a bright yellow forehead and body, a brown head, and prominent white wing patches.

**Figure 11 (White Pelican vs. "not found")**:

- The Qwen model incorrectly identifies the bird as the White Pelican by noting its large size, large, flat-topped bill, and predominantly white plumage.
- Our model correctly determines that the species is "none of the above" among the provided options (White Pelican, Caspian Tern, California Gull, Western Gull, Brown Pelican). Our reasoning highlights that the bird in the image has a brownish body and a white head with a dark patch behind the eye, which rules out the White Pelican, which is predominantly bright white. This analysis demonstrates a more accurate observation of the bird's full plumage.

**Figure 12 (Forster's Tern vs. Least Tern)**:

- The Qwen model incorrectly identifies the bird as Forster's Tern, noting the black cap extending down the neck, black-tipped orange bill, and bright orange legs.
- Our model correctly identifies the bird as the Least Tern. Our reasoning, although noting that the black-tipped orange bill and orange legs are consistent with both Forster's Tern and Least Tern, correctly matches the features to the Least Tern, including the black cap, white underparts, pale gray upperparts, and specifically the black line running through the eye to the base of the bill and the small white patch on the forehead, which is a key identifying feature of the Least Tern.

**Figure 13 (LeConte's Sparrow Identification)**:

- The Qwen model focuses on the Lincoln's Sparrow features: the brown cap with a central grey stripe, a grey face with brown cheeks, and a buffy mustache and brown line through the eye. It concludes that the bird's appearance, including the brown and beige body tones, was indicative of the Lincoln's Sparrow's finely streaked light brown breast and olive-brown upperparts.
- Our model returns "none of the above". However, the reasoning demonstrates a strong attempt at fine-grained analysis, suggesting the failure is primarily due to the subtle nature of color variation in this fine-grained recognition task. Our model correctly identified the bird's small, compact body, short tail, and brown–buffy coloration, including the pale supercilium and darker mustache region. It also noted the absence of key features described for LeConte's Sparrow, such as the yellow-orange face and lilac-gray nape—indicating a mismatch. The model then systematically ruled out Savannah, Song, and Lincoln's Sparrows by comparing required traits (e.g., Lincoln's Sparrow's white belly and finely streaked breast) with the visible evidence, ultimately concluding that none of the candidate species matched the image.

In the third reported case, shown in Figure 14 (where an image of a Blue Jay was the target), the Qwen model successfully identified the species as none of the options, while our model failed by misidentifying it as the Florida Jay. The correct bird in the image is a Blue Jay, a species that was not listed in the provided candidates (Florida Jay, Blue Grosbeak, Green Jay, Clark's Nutcracker).

**Figure 14 (Blue Jay Identification)**:

- The Qwen model correctly returned "none of the above," systematically ruling out all provided species. It noted the bird's predominantly blue coloration on the wings and tail but emphasized the absence of the Green Jay's characteristic black bib or mask. It also observed that the beak shape and body structure did not match those of the Florida Jay or Blue Grosbeak, and that the coloration and wing patterns were inconsistent with those of the Blue Grosbeak. By cautiously eliminating each option based on these mismatches, the model concluded that the correct answer was not among the listed species.
- Our model incorrectly identified the bird as a Florida Jay. While it accurately noted the blue coloration, medium size, and robust build, and correctly ruled out the Green Jay, Clark's Nutcracker, and Blue Grosbeak, it made a key mistake when evaluating the Florida Jay. The model recognized that the bird lacked the species' defining whitish forehead and supercilium, but nonetheless concluded that it "could be a Florida Jay" based on its overall

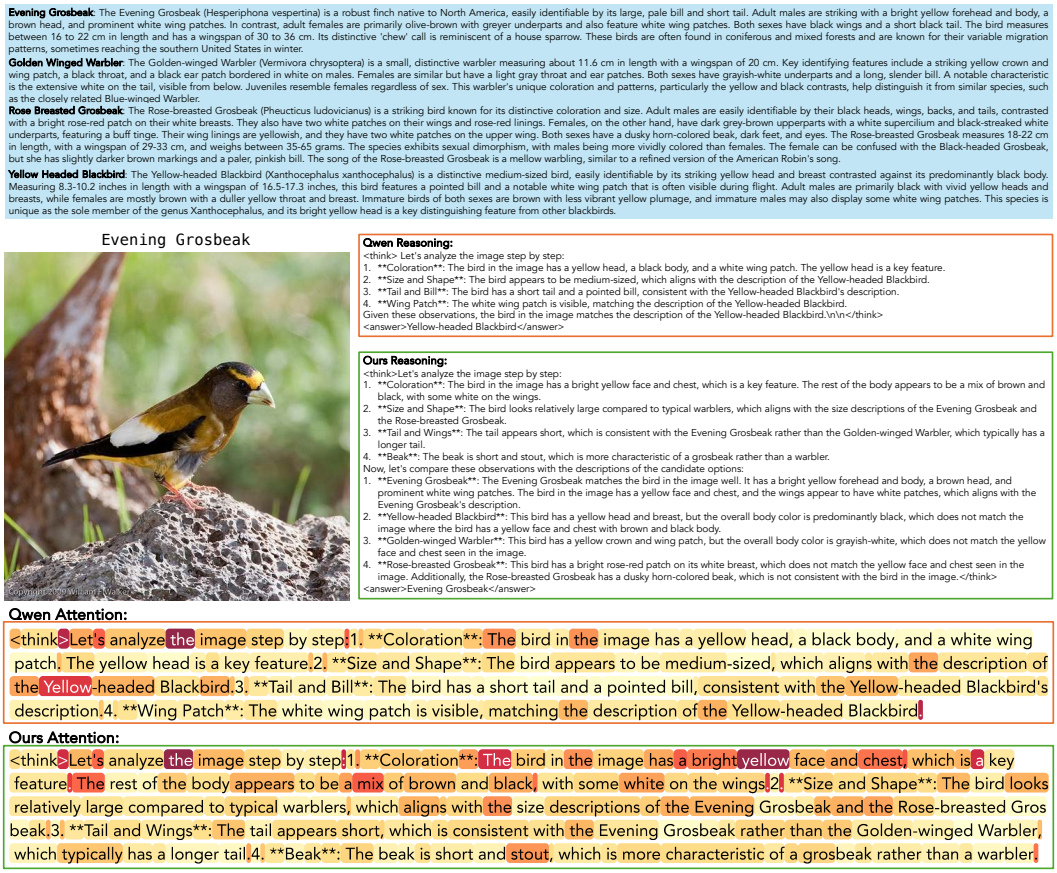



Figure 10: Our model correctly identifies the Evening Grosbeak while Qwen fails by classifying it as the Yellow-headed Blackbird.

blue coloration and a long tail. This reflects a failure to apply discriminative features strictly and a tendency to force a match to the nearest option when the correct species is not present, a common pitfall in fine-grained classification.

**White Pelican**: The American white pelican (Pelecanus erythrorhynchos) is a strikingly large bird, notable for its impressive size and distinctive features. It measures between 50 to 70 inches in length and boasts a wingspan of 95 to 120 inches, making it one of the largest birds in North America. The pelican's plumage is predominantly bright white, with contrasting black primary and secondary feathers visible only during flight. A key identifying feature is its enormous, flat-topped bill, which is vivid orange during the breeding season, along with the bare skin around the eyes and feet. Unique to this species, a keratinous "horn" develops on the upper bill during breeding, which is later shed. Outside the breeding season, the bill and other bare parts become a dull pink-orange. Males and females are similar in appearance, though males are slightly larger. Immature birds display light grey plumage with darker brownish nape and remiges. These pelicans are often seen soaring gracefully, utilizing their large wingspan for efficient flight.

**Caspian Tern**: The Caspian Tern (Hydroprogne caspia) is the world's largest tern, measuring 48-60 cm in length with a wingspan of 127-145 cm. It is distinguished by its robust red-orange bill with a small black tip and black legs. The bird's plumage features a striking white head with a black cap, which remains even in winter, albeit with some white streaking on the forehead. Its body is predominantly white, with pale grey upper wings and back, and underwings that are pale with dark primary feathers. In flight, the Caspian Tern's tail is less forked compared to other terns, and its wingtips are black on the underside. These features, along with its loud heron-like croak, make it identifiable from other similar species.

**California Gull**: The California Gull (Larus californicus) is a medium-sized gull, distinguishable by its brown eyes, yellow legs, and a more rounded head compared to the similar herring gull. It features a mainly white body with a grey back and upper wings. A key identifying trait is its yellow bill, which has a distinctive black ring and a red spot near the tip. The wings have black primaries with white tips. Immature California Gulls have browner plumage than immature ring-billed gulls. They range in length from 46 to 55 cm (18 to 22 inches), with a wingspan of 122 to 137 cm (48 to 54 inches), and weigh between 430 to 1,045 grams (0.948 to 2.304 pounds). These physical characteristics, along with their opportunistic feeding behavior, help distinguish them from other gull species.

**Western Gull**: The Western Gull (Larus occidentalis) is a large gull native to the west coast of North America, distinguishable by its robust size, measuring 55 to 68 cm in length with a wingspan of 130 to 144 cm. It features a striking white head and body, with dark grey upperparts or mantle. The gull's large, bulbous-tipped yellow bill is notable for a red subterminal spot, which is a key feeding stimulus for chicks. Eye color varies, with southern populations typically having pale yellow eyes, while northern populations have darker eyes. The Western Gull's wings are broad and powerful, aiding in its predominantly marine lifestyle. It closely resembles the Slaty-backed Gull but can be differentiated by its darker mantle and less streaked head in nonbreeding plumage. The Western Gull is known for its aggressive territorial behavior and is often found in coastal regions, rarely venturing far inland.

Brown Pelican




**Qwen Reasoning:**

<think>The bird in the image appears to be a pelican, specifically an American white pelican, based on several key features:
1. **Size and Shape**: The bird is quite large, which aligns with the description of the American white pelican being one of the largest birds in North America.
2. **Bill Characteristics**: The bird has a large, flat-topped bill, which is a distinctive feature of pelicans. The bill is not as vividly colored as described for the Caspian Tern, but it is consistent with the pelican's appearance.
3. **Plumage**: The bird's plumage is predominantly white, with darker feathers on the wings, which matches the description of the American white pelican.
4. **Legs and Feet**: The legs appear to be a dull pink-orange, which is consistent with the description of the American white pelican.
Given these observations, the bird in the image most closely matches the description of the American white pelican.</think>
<answer>White Pelican</answer>

**Ours Reasoning:**

<think>Let's analyze the image step by step:
1. **Size and Shape**: The bird in the image appears to be relatively large, with a long neck and a large bill. This is consistent with the description of a pelican.
2. **Coloration**: The bird has a white head with a dark patch behind the eye, a brownish body, and a long, curved bill. The plumage is not entirely white, which rules out the White Pelican as it is predominantly white with black primary and secondary feathers.
3. **Bill Characteristics**: The bill is long and curved, which is characteristic of a pelican. The bill also appears to have a dark band near the base, which could be indicative of a pelican species.
4. **Legs and Feet**: The legs appear to be short and robust, which is typical for pelicans.
Now, let's compare these observations with the candidate options:
1. **Caspian Tern**: The Caspian Tern has a robust red-orange bill with a small black tip and black legs. The bird in the image does not have a red-orange bill; instead, it has a long, curved bill with a dark band near the base, which is more consistent with a pelican.
2. **White Pelican**: The White Pelican is described as having a bright white plumage with a large, flat-topped bill that is vivid orange during the breeding season. The bird in the image does not have a predominantly white plumage; it has a brownish body and a white head with a dark patch behind the eye.
3. **Western Gull**: The Western Gull has a white head and body with dark grey upperparts, a bulbous-tipped yellow bill with a red subterminal spot, and broad wings. The bird in the image does not have a white head or body; it has a brownish body and a white head with a dark patch behind the eye.
4. **California Gull**: The California Gull has a mainly white body with a grey back and upper wings, a yellow bill with a black ring and a red spot near the tip, and black primaries with white tips. The bird in the image does not have a white body or a grey back; it has a brownish body and a white head with a dark patch behind the eye.</think>
<answer>none of the above</answer>

**Qwen Attention:**

<think>The bird in the image appears to be a pelican, specifically an American white pelican, based on several key features:1. **Size and Shape**: The bird is quite large, which aligns with the description of the American white pelican being one of the largest birds in North America.2. **Bill Characteristics**: The bird has a large, flat-topped bill, which is a distinctive feature of pelicans. The bill is not as vividly colored as described for the Caspian Tern, but it is consistent with the pelican's appearance.3. **Plumage**: The bird's plumage is predominantly white, with darker feathers on the wings, which matches the description of the American white pelican.4. **Legs and Feet**: The legs appear to be a dull pink-orange, which is consistent with the description of the American white pelican.

**Ours Attention:**

<think>Let's analyze the image step by step:1. **Size and Shape**: The bird in the image appears to be relatively large, with a long neck and a large bill. This is consistent with the description of a pelican.2. **Coloration**: The bird has a white head with a dark patch behind the eye, a brownish body, and a long, curved bill. The plumage is not entirely white, which rules out the White Pelican as it is predominantly white with black primary and secondary feathers.3. **Bill Characteristics**: The bill is long and curved, which is characteristic of a pelican. The bill also appears to have a dark band near the base, which could be indicative of a pelican species.4. **Legs and Feet**: The legs appear to be short and robust, which is typical for pelicans.

Figure 11: Our model correctly determines the bird is not present in the options, accurately rejecting the White Pelican based on the non-white plumage. Qwen incorrectly identifies it as the White Pelican.

**Least Tern**: The Least Tern (Sternula antillarum) is a small, distinctive tern species, measuring 22-24 cm in length with a wingspan of about 50 cm. It is characterized by its pale gray upperparts and white underparts. A key feature is its black cap and a black line running through the eye to the base of the bill, with a small white patch on the forehead, which becomes more extensive in winter. The bill is yellow with a black tip during the breeding season, turning all blackish in winter. Its legs are yellowish. The wings are mostly pale gray with noticeable black markings on the outermost primaries. In flight, it exhibits fast, jerky wingbeats and a distinctive hunchback appearance, with the bill pointing slightly downward. The Least Tern can be distinguished from similar species by its gray rump and tail, as opposed to the white seen in the Little Tern, and its unique squeaking call.

**Common Tern**: The Common Tern (Sterna hirundo) is a medium-sized seabird distinguished by its slender build, long pointed wings, and deeply forked tail. Breeding adults exhibit pale grey upperparts and very pale grey underparts, with a striking black cap on the head. Their legs are orange-red, and the bill is narrow and pointed, varying in color from mostly red with a black tip to entirely black, depending on the subspecies. The upper wings are pale grey, developing a distinctive dark wedge as the summer progresses. In non-breeding adults, the forehead and underparts turn white, and the bill becomes all black or black with a red base. Juveniles have pale grey upper wings with a dark carpal bar and a ginger forehead that fades to white by autumn. The Common Tern can be differentiated from similar species like the Arctic Tern by its larger head, thicker neck, longer legs, and more triangular wings, as well as its more powerful flight. The Arctic Tern, in contrast, has greyer underparts and a consistently white wing throughout the summer. The Common Tern's tail does not extend beyond the wingtips when standing, unlike the Arctic and Roseate Terns. These features are crucial for identifying the Common Tern in photographs.

**Elegant Tern**: The Elegant Tern (Thalasseus elegans) is a medium-large tern distinguished by its long, slender orange bill and pale grey upperparts with white underparts. It has black legs and, in winter, a white forehead. Juveniles display a scalier pale grey back. A key identifying feature is the black crest that extends from the crown through the eye, creating a small black "smudge" in front of the eye. This differentiates it from the Royal Tern, which has a larger, thicker bill and more white on the forehead in winter, and the Lesser Crested Tern, which has a less shaggy crest. The Elegant Tern measures 15.3-16.5 inches in length, weighs 6.7-11.5 ounces, and has a wingspan of 76-81 cm. Its call is a loud grating noise similar to that of a Sandwich Tern.

**Forsters Tern**: Forster's Tern (Sterna forsteri) is a medium-sized tern distinguished by its slender body, deeply forked long tail, and relatively long legs. In breeding plumage, it features a striking black cap extending down the neck, pale gray wings and back, bright white underparts, a black-tipped orange bill, and bright orange legs. During non-breeding seasons, the crown turns white, and a distinctive black comma-shaped patch covers the eye and ear-covert, with a black bill and dull brownish-red legs. The wings are gray with dark silvery gray primaries. Juveniles resemble non-breeding adults but often have darker primaries. Forster's Tern is similar to the Common Tern but can be distinguished by its black eye mask in winter and a gray center to its white tail, with whiter upper wings lacking the darker primary wedge seen in the Common Tern. Its call is a descending "kerr," and it exhibits aggressive behavior when defending its nest.

Least Tern



**Qwen Reasoning:**

<think>Let's analyze the image step by step:
1. **Body Shape and Size**: The bird in the image appears to be medium-sized with a slender body, long legs, and a deeply forked tail, which aligns with the description of Forster's Tern.
2. **Head and Neck**: The bird has a black cap extending down the neck, which matches the description of Forster's Tern.
3. **Wings and Back**: The wings appear to be pale gray with dark silvery gray primaries, consistent with the description.
4. **Underparts**: The underparts are bright white, matching the description.
5. **Bill and Legs**: The bill is black-tipped orange, and the legs are bright orange, which also matches the description.
6. **Tail**: The tail is deeply forked, which is characteristic of Forster's Tern.
7. **Overall Appearance**: The bird's overall appearance, including the shape of the wings, the forked tail, and the coloration, closely matches the description of Forster's Tern.
Given these observations, the bird in the image most closely resembles Forster's Tern.</think>
<answer>Forster's Tern</answer>

**Ours Reasoning:**

<think>Let's analyze the image step by step:
1. **Body Shape and Size**: The bird appears to be a medium-sized tern with a slender body, long legs, and a deeply forked tail. This matches the description of Forster's Tern, Least Tern, Common Tern, and Elegant Tern.
2. **Head and Crown**: The bird has a black cap on its head, which is characteristic of Forster's Tern, Least Tern, and Common Tern. However, the Elegant Tern also has a black crest that extends from the crown through the eye, creating a small black "smudge" in front of the eye.
3. **Underparts**: The bird has white underparts, which is consistent with Forster's Tern, Least Tern, and Common Tern. The Elegant Tern has a white forehead in winter, but the image does not show the bird in winter plumage.
4. **Wings and Tail**: The wings appear to be pale gray with dark markings on the outermost primaries, which is consistent with the Least Tern and Common Tern. The Elegant Tern has a pale gray upperpart with a white underpart, which does not match the image.
5. **Bill**: The bill is black-tipped orange, which is consistent with Forster's Tern and Least Tern. The Common Tern's bill varies in color from mostly red with a black tip to entirely black, and the Elegant Tern's bill is long and slender with an orange base.
6. **Legs**: The legs are orange, which is consistent with Forster's Tern and Least Tern. The Common Tern's legs are orange-red, and the Elegant Tern's legs are black.
Based on the above analysis, the bird in the image most closely matches the description of the Least Tern. The black cap, white underparts, pale gray upperparts, and orange legs are consistent with the Least Tern. Additionally, the black line running through the eye to the base of the bill and the small white patch on the forehead are also consistent with the Least Tern.
The other options do not fit as well:
1. **Forster's Tern**: While it has a black cap and white underparts, the bill is not described as black-tipped orange, and the legs are not described as orange.
2. **Common Tern**: The bill is described as narrow and pointed, varying in color from mostly red with a black tip to entirely black, which does not match the image.
3. **Elegant Tern**: The bill is described as long, slender, and orange-based, which does not match the image.</think>
<answer>Least Tern</answer>

**Qwen Attention:**

<think>Let's analyze the image step by step:1. **Body Shape and Size**: The bird in the image appears to be medium-sized with a slender body, long legs, and a deeply forked tail, which aligns with the description of Forster's Tern.2. **Coloration**: - **Head and Neck**: The bird has a black cap extending down the neck, which matches the description of Forster's Tern. - **Wings and Back**: The wings appear to be pale gray with dark silvery gray primaries, consistent with the description. - **Underparts**: The underparts are bright white, matching the description. - **Bill and Legs**: The bill is black-tipped orange, and the legs are bright orange, which also matches the description.3. **Tail**: The tail is deeply forked, which is characteristic of Forster's Tern.4. **Overall Appearance**: The bird's overall appearance, including the shape of the wings, the forked tail, and the coloration, closely matches the description of Forster's Tern.

**Ours Attention:**

<think>Let's analyze the image step by step:1. **Body Shape and Size**: The bird appears to be a medium-sized tern with a slender body, long legs, and a deeply forked tail. This matches the description of Forster's Tern, Least Tern, Common Tern, and Elegant Tern.2. **Coloration**: - **Head and Crown**: The bird has a black cap on its head, which is characteristic of Forster's Tern, Least Tern, and Common Tern. However, the Elegant Tern also has a black crest that extends from the crown through the eye, creating a small black "smudge" in front of the eye. - **Underparts**: The bird has white underparts, which is consistent with Forster's Tern, Least Tern, and Common Tern. The Elegant Tern has a white forehead in winter, but the image does not show the bird in winter plumage. - **Wings and Tail**: The wings appear to be pale gray with dark markings on the outermost primaries, which is consistent with the Least Tern and Common Tern. The Elegant Tern has a pale gray upperpart with a white underpart, which does not match the image.3. **Bill and Legs**: - **Bill**: The bill is black-tipped orange, which is consistent with Forster's Tern and Least Tern. The Common Tern's bill varies in color from mostly red with a black tip to entirely black, and the Elegant Tern's bill is long and slender with an orange base. - **Legs**: The legs are orange, which is consistent with Forster's Tern and Least Tern. The Common Tern's legs are orange-red, and the Elegant Tern's legs are black.

Figure 12: Our model correctly identifies the Least Tern based on the black line through the eye and small white forehead patch, features consistent with its orange legs and black-tipped orange bill. Qwen incorrectly identifies it as Forster's Tern.

**Le Conte Sparrow**: LeConte's Sparrow (Ammospiza leconteii) is a small, distinctive sparrow notable for its relatively large head and short, pointed tail. It features a short grey bill and a buffy yellow-orange face with grey cheeks, complemented by a dark brown crown with a white central stripe. The nape is lilac grey with chestnut streaks, and its back is streaked with brown and beige. The bird's underparts are characterized by an off-white belly and a buffy orange-yellow breast and sides, adorned with dark brown streaks. Its legs and feet are a brownish-pink color. Measuring about 4.7 inches in length with a wingspan of 7.1 inches, LeConte's Sparrow is often confused with similar species like Nelson's Sparrow and Grasshopper Sparrow, but its unique color patterns and head markings help in distinguishing it.

**Savannah Sparrow**: The Savannah Sparrow (Passerculus sandwichensis) is a small, variable sparrow distinguished by its dark-streaked brown back and whitish underparts with brown or blackish streaks on the breast and flanks. It features a whitish crown and supercilium stripes, often with a hint of yellow near the beak, and brown cheeks with a white throat. The flight feathers are blackish-brown with light brown or white edges. Its eyes are dark, and the legs and feet are horn-colored, with a dark grey upper bill and a lighter lower bill. The Savannah Sparrow's size ranges from 11 to 17 cm in length, with a wingspan of 18 to 25 cm, and it weighs between 15 to 29 grams. This species is highly variable, with numerous subspecies that differ mainly in plumage darkness, following Gloger's rule, where Alaskan and interior races are paler, and southwestern coastal forms are darker. These sparrows are typically found in open grassland habitats across North America, with migratory populations breeding in Canada and the northern United States and wintering in Mexico and the southern United States.

**Song Sparrow**: The Song Sparrow (Melospiza melodia) is a medium-sized sparrow native to North America, known for its adaptability and variability across numerous subspecies. It typically measures 11 to 18 cm in length with a wingspan of 18 to 25.4 cm. Key identifying features include brown upperparts with dark streaks, a white underside with dark streaking, and a distinctive dark brown spot in the center of the breast. The bird has a brown cap, a long rounded tail, and a gray face marked by a brown streak through each eye. The Song Sparrow's beak is generally slender, with variations in size correlating to different climates. It is often confused with the Lincoln's Sparrow, which has a shorter, grayer tail and a distinct head pattern, and the Savannah Sparrow, which has a forked tail and yellowish facial flecks. The Song Sparrow's diverse plumage and song variations make it a fascinating subject for bird enthusiasts and researchers alike.

**Lincoln Sparrow**: The Lincoln's Sparrow (Melospiza lincolnii) is a small, secretive sparrow native to North America, distinguishable by its finely streaked light brown breast, white belly, and dark-streaked olive-brown upperparts. It features a brown cap with a central grey stripe, grey face with brown cheeks, a buffy mustache, and a brown line through the eye with a narrow eye ring. Its olive-brown wings and narrow tail complement its overall trim appearance. Measuring 5.1-5.9 inches in length with a wingspan of 7.5-8.7 inches, it is smaller and more slender than the similar Song Sparrow, with finer breast streaks. Juveniles resemble juvenile Swamp Sparrows but typically lack a unicolored crown. The Lincoln's Sparrow's subtle yet distinct plumage and its preference for dense, brushy habitats make it a unique find for bird watchers.

Le Conte Sparrow



**Qwen Reasoning:**

<think>Let's analyze the bird in the image step by step:
1. **Size and Shape**: The bird appears to be small and compact, which aligns with the descriptions of all four options. However, the specific details will help us narrow it down.
2. **Head**: The bird has a brown cap with a central grey stripe, which is consistent with the Lincoln Sparrow's description.
3. **Face**: The bird has a grey face with brown cheeks, matching the Lincoln Sparrow's description.
4. **Breast**: The bird has a buffy mustache and a brown line through the eye with a narrow eye ring, which also matches the Lincoln Sparrow's description.
5. **Body**: The bird's body appears to have a mix of brown and beige tones, which could be indicative of the Lincoln Sparrow's olive-brown upperparts and finely streaked light brown breast.
6. **Tail**: The bird's tail is not clearly visible in the image, but the Lincoln Sparrow's description mentions a narrow tail, which might be inferred from the overall shape.
Given the detailed comparison, the bird in the image most closely matches the description of the Lincoln Sparrow.</think>
<answer>Lincoln Sparrow</answer>

**Ours Reasoning:**

<think>Let's analyze the image step by step:
1. **Size and Shape**: The bird appears to be small and compact, with a short, stout body and a relatively short tail. This matches the description of a sparrow.
2. **Coloration**:The bird has a mix of brown and buffy colors on its back and wings. The underparts appear to have a mix of white and brown, with some streaking. There is a noticeable pattern around the head, including a pale area above the eye and a darker area below it.
3. **Head Features**:The head has a pale area above the eye, which could be interpreted as a "supercilium" stripe. There is a darker area below the eye, which could be part of the "mustache" marking.
4. **Legs and Feet**: The legs appear to be a light brownish-pink color, which aligns with the description of Le Conte's Sparrow.
Now, let's compare these observations with the candidate options:
1. **Le Conte Sparrow**: The Le Conte's Sparrow has a distinctive pattern with a buffy yellow-orange face, grey cheeks, and a dark brown crown with a white central stripe. The nape is lilac grey with chestnut streaks, and the back is streaked with brown and beige. The underparts are off-white with dark brown streaks. The bird in the image does not have the bright yellow-orange face or the lilac grey nape; instead, it has a more muted brown and buffy coloration. Additionally, the head pattern does not match the Le Conte's Sparrow's description.
2. **Savannah Sparrow**: The Savannah Sparrow has a whitish crown and supercilium stripes, often with a hint of yellow near the beak, and brown cheeks with a white throat. The flight feathers are blackish-brown with light brown or white edges. The bird in the image does not have a whitish crown or supercilium stripes, nor does it have the blackish-brown flight feathers with light edges. Instead, it has a more uniform brown and buffy coloration.
3. **Song Sparrow**: The Song Sparrow has a brown cap, a long rounded tail, and a gray face marked by a brown streak through each eye. The bird in the image does not have a brown cap or a long rounded tail. Instead, it has a more compact body and a shorter tail.
4. **Lincoln Sparrow**: The Lincoln's Sparrow has a finely streaked light brown breast, a white belly, and dark-streaked olive-brown upperparts. It features a brown cap with a central grey stripe, grey face with brown cheeks, a buffy mustache, and a brown line through the eye with a narrow eye ring. The bird in the image does not have a white belly or a finely streaked light brown breast. Instead, it has a more uniform brown and buffy coloration.</think>
<answer>none of the above</answer>

**Qwen Attention:**

<think>Let's analyze the bird in the image step by step:1. **Size and Shape**: The bird appears to be small and compact, which aligns with the descriptions of all four options. However, the specific details will help us narrow it down.2. **Coloration**: - **Head**: The bird has a brown cap with a central grey stripe, which is consistent with the Lincoln Sparrow's description. - **Face**: The bird has a grey face with brown cheeks, matching the Lincoln Sparrow's description. - **Breast**: The bird has a buffy mustache and a brown line through the eye with a narrow eye ring, which also matches the Lincoln Sparrow's description. - **Body**: The bird's body appears to have a mix of brown and beige tones, which could be indicative of the Lincoln Sparrow's olive-brown upperparts and finely streaked light brown breast.3. **Tail**: The bird's tail is not clearly visible in the image, but the Lincoln Sparrow's description mentions a narrow tail, which might be inferred from the overall shape.

**Ours Attention:**

<think>Let's analyze the image step by step:1. **Size and Shape**: The bird appears to be small and compact, with a short, stout body and a relatively short tail. This matches the description of a sparrow.2. **Coloration**: - The bird has a mix of brown and buffy colors on its back and wings. - The underparts appear to have a mix of white and brown, with some streaking. - There is a noticeable pattern around the head, including a pale area above the eye and a darker area below it.3. **Head Features**: - The head has a pale area above the eye, which could be interpreted as a "supercilium" stripe. - There is a darker area below the eye, which could be part of the "mustache" marking.4. **Legs and Feet**: The legs appear to be a light brownish-pink color, which aligns with the description of Le Conte's Sparrow.

Figure 13: Both models fail to identify the LeConte's Sparrow. Qwen incorrectly selects the Lincoln's Sparrow, while our model returns 'none of the above' due to the image's subtle coloration not matching the descriptive 'buffy yellow-orange face' of the true species.

**Florida Jay**: The Florida scrub jay (Aphelocoma coerulescens) is a medium-sized bird endemic to Florida, measuring 23 to 28 cm in length with a wingspan of 33-36 cm. It is distinguished by its striking blue head and nape, which lack a crest, and a whitish forehead and supercilium. The bird features a blue bib and wings, a long blue tail, and grayish underparts and back. Its strong black bill and black legs and feet are also notable. Juvenile scrub jays can be identified by the brown feathers on their heads, which gradually turn blue as they mature. This species is unique to the scrub habitats of Florida and is known for its tameness and cooperative breeding behavior.

**Blue Grosbeak**: The Blue Grosbeak (Passerina caerulea) is a medium-sized passerine bird notable for its striking plumage and large, deep bill. Males are predominantly deep blue with distinctive black and brown on their wings, along with two prominent brown wing bars. Females, in contrast, are mostly brown with scattered blue feathers on their upperparts and also feature the characteristic double wing bars. This species is larger than the similar Indigo Bunting, with a length ranging from 14 to 19 cm (5.5 to 7.5 in) and a wingspan of 26 to 29 cm (10 to 11 in). The Blue Grosbeak's robust size and unique coloration make it distinguishable in its habitat, which includes open areas with scattered trees and thickets across the southern United States and northern Mexico.",

**Green Jay**: The Green Jay (Cyanocorax) is a striking bird known for its vibrant plumage and distinctive features. It is medium-sized, typically measuring around 25-29 cm in length. The Green Jay's most notable characteristic is its vivid green body, which contrasts with its bright blue crown and nape. The face is marked by a bold black bib and a black line extending from the beak through the eyes, giving it a masked appearance. Its wings and tail are a mix of green and blue, with the tail often showing a yellowish tip. The beak is stout and black, well-suited for its omnivorous diet. These features, along with its lively and social behavior, make the Green Jay easily distinguishable from other bird species in its habitat.

**Clark Nutcracker**: Clark's Nutcracker (Nucifraga columbiana) is a medium-sized bird, measuring about 10.6 to 11.8 inches in length with a wingspan of 24 inches. It is distinguished by its ashy-grey body plumage, contrasted with black and white wings and tail. The central tail feathers are black, while the outer ones are white, creating a striking pattern. Its long, stout, and cone-shaped bill, along with black legs and feet, are key features. This bird is often found in the mountainous conifer forests of western North America. Its unique behavior of caching pine seeds and its remarkable memory for retrieving them are notable traits. The Clark's Nutcracker's varied vocalizations, often described as a khraaaah-khraaaah call, further aid in its identification.

Blue Jay



**Qwen Reasoning:**

<think>Let's analyze the bird in the image step by step:
1. **Coloration**: The bird has a predominantly blue coloration on its wings and tail, with some white on the underside. The head appears to have a mix of blue and white, but there is no distinct black bib or mask like the Green Jay.
2. **Size and Shape**: The bird seems to be medium-sized, consistent with the Florida Jay and the Blue Grosbeak. However, the shape of the beak and the overall body structure do not align perfectly with either of these options.
3. **Behavior and Habitat**: The bird is perched on what looks like a bench, suggesting a more open environment rather than the scrub habitats of the Florida Jay or the mountainous regions of the Clark's Nutcracker.
Given these observations, the bird in the image does not match the descriptions of the Green Jay, Florida Jay, or Clark's Nutcracker. The Blue Grosbeak's description does not fully align with the bird's appearance either, particularly the lack of the distinctive black and brown markings on the wings and the overall coloration.\n\nTherefore, based on the analysis, the bird in the image does not match any of the provided options.</think>
<answer>none of the above</answer>

**Ours Reasoning:**

<think>Let's analyze the bird in the image step by step:
1. **Coloration**: The bird has a predominantly blue coloration on its head, back, and wings, with some white on its belly and tail. The tail appears to have a blue base with a lighter tip.
2. **Size and Shape**: The bird is medium-sized, with a robust build and a relatively short, stout beak. The wings appear broad and rounded, typical of many jay species.
3. **Behavior and Habitat**: The bird is perched on what looks like a wooden structure, possibly a bench, which suggests it might be in an open area with some vegetation, consistent with the habitat of scrub jays.
Now, let's compare these observations with the descriptions of the candidate options:
1. **Green Jay**: The Green Jay is described as having a vibrant green body with a blue crown and nape, a black bib, and a mix of green and blue wings and tail. The bird in the image does not have a green body; instead, it is primarily blue with some white on the belly and tail. Therefore, it does not match the Green Jay.
2. **Florida Jay**: The Florida Jay is described as having a striking blue head and nape, a whitish forehead and supercilium, a blue bib and wings, a long blue tail, and grayish underparts and back. The bird in the image has a blue head and nape, but lacks the whitish forehead and supercilium, and the underparts and back are not clearly visible in the image. However, the overall blue coloration and the long tail suggest it could be a Florida Jay.
3. **Clark Nutcracker**: The Clark Nutcracker is described as having an ashy-grey body, black and white wings and tail, and a long, stout, and cone-shaped bill. The bird in the image does not have an ashy-grey body or black and white wings and tail, so it does not match the Clark Nutcracker.
4. **Blue Grosbeak**: The Blue Grosbeak is described as having a predominantly deep blue male with black and brown on the wings and a large, deep bill. The bird in the image does not have a deep blue coloration or a large, deep bill, so it does not match the Blue Grosbeak.</think>
<answer> Florida Jay </answer>

**Qwen Attention:**

<think>Let's analyze the bird in the image step by step:1. **Coloration**: The bird has a predominantly blue coloration on its wings and tail, with some white on the underside. The head appears to have a mix of blue and white, but there is no distinct black bib or mask like the Green Jay.2. **Size and Shape**: The bird seems to be medium-sized, consistent with the Florida Jay and the Blue Grosbeak. However, the shape of the beak and the overall body structure do not align perfectly with either of these options.3. **Behavior and Habitat**: The bird is perched on what looks like a bench, suggesting a more open environment rather than the scrub habitats of the Florida Jay or the mountainous regions of the Clark's Nutcracker.

**Ours Attention:**

<think>Let's analyze the bird in the image step by step:1. **Coloration**: The bird has a predominantly blue coloration on its head, back, and wings, with some white on its belly and tail. The tail appears to have a blue base with a lighter tip.2. **Size and Shape**: The bird is medium-sized, with a robust build and a relatively short, stout beak. The wings appear broad and rounded, typical of many jay species.3. **Behavior and Habitat**: The bird is perched on what looks like a wooden structure, possibly a bench, which suggests it might be in an open area with some vegetation, consistent with the habitat of scrub jays.

Figure 14: Qwen correctly returns 'none of the above', recognizing the bird (a Blue Jay) lacks the required features of the listed species (e.g., Green Jay's black bib), while our model incorrectly forces a match to the Florida Jay.