STRESS-FG: Diagnosing Visual Grounding in Zero-Shot Fine-Grained Recognition

Ishtiaque Ahmed Khan[1] **and A. Lynn Abbott**[1,2]

[1]Dept. of Computer Science, [2]Bradley Dept. of Electrical and Computer Engineering
Virginia Tech, Blacksburg, Virginia, USA
{ishtiaqueahmedk, abbott}@vt.edu

Abstract

Zero-shot fine-grained recognition is a key test of visual grounding in vision-language models (VLMs), but current evaluations may conflate attribute grounding with reliance on class-label knowledge. We introduce STRESS-FG, a diagnostic benchmark spanning five fine-grained domains that systematically varies the relationship between class names and visual descriptions. The benchmark includes label-only, description-only, cue-conflict, and synthetic-label settings, as well as negated-attribute, multi-image, reference-conditioned, and part-based grounding tasks. Evaluating up to thirteen VLMs shows that label-only prompts are often surprisingly strong, indicating substantial familiarity with class-label priors. Synthetic labels can facilitate stronger attribute grounding evaluation, while familiar but incorrect class labels can suppress grounding capabilities under cue conflict. We further find that negation, multi-image matching, and reference-conditioned selection remain challenging, and part-based analyses show only partial reliance on the correct localized evidence. Finally, our binary question format reveals that models often struggle to reject fine-grained distractors, a weakness hidden by traditional Multiple Choice Question evaluations. These results suggest that standard fine-grained benchmarks may overestimate grounded visual understanding and should more explicitly separate label priors from attribute-based recognition. Code and dataset are available at: https://github.com/Ishtiaque-khan97/Stress-FG.

1 Introduction

Vision–language models (VLMs) are increasingly evaluated on fine-grained recognition tasks to distinguish visually similar categories using subtle attribute differences. VLMs are highly sensitive to lexical cues and may rely on language-driven shortcuts rather than visual evidence (Parisot et al., 2023; Roth et al., 2023; Fu et al., 2025). Multimodal models do not align class names and attribute descriptions equally (He et al., 2025). However, fine-grained VLM benchmarks often rely on class names for evaluations (Geigle et al., 2024; Yu et al., 2026; Atabuzzaman et al., 2025b; Kim and Ji, 2024; Ghosh et al., 2026), allowing models to leverage knowledge priors associated with these labels. Thus, benchmark accuracies may overestimate a model's ability to ground fine-grained attributes in images.

To investigate this issue, we introduce STRESS-FG, a controlled diagnostic benchmark suite for evaluating fine-grained recognition in VLMs. Starting from five established fine-grained datasets (Wah et al., 2011; Maji et al., 2013; Bossard et al., 2014; Khosla et al., 2011; Krause et al., 2013), we construct multiple-choice task variants that systematically manipulate the relationship between class names and visual descriptions. These variants isolate recognition from descriptions alone, recognition from class names alone, robustness under name-description conflict, and sensitivity to label familiarity through synthetic class names. We further include negated-attribute tasks, multi-image and reference-conditioned selection tasks, and a part-based grounding setting that tests whether models rely on the correct localized visual evidence when answering fine-grained attribute questions.

Across thirteen recent VLMs, STRESS-FG reveals that current zero-shot fine-grained evaluations can conflate label familiarity with grounded recognition. Class-label information often acts as a shortcut, while controlled perturbations such as synthetic names and cue-conflict settings expose weaknesses that remain hidden in standard benchmark formats. We also find that negation and multi-image reasoning remain challenging even when models perform strongly on conventional single-image recognition tasks. In addition, part-based experiments show only partial reliance on the rele-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 47614–47640
October 24-29, 2026 ©2026 Association for Computational Linguistics

vant localized evidence, with many models still benefiting from global context or shortcut cues.

These results suggest that existing benchmarks may reward different shortcut behaviors and overestimate visual grounding ability. We argue that future evaluations should explicitly separate label-based or other contextual cue shortcuts from grounding, and include improved robustness settings. Our main contributions are as follows:

- We introduce **STRESS-FG**, a controlled benchmark suite for diagnosing the role of class-label priors in fine-grained evaluation.
- We construct diagnostic task variants that isolate description-based recognition, label-only recognition, cue conflict, and label familiarity, and extend the benchmark to negation, multi-image selection, reference-conditioned composed-image retrieval, and part-based evidence grounding.
- We provide a large-scale empirical study using up to thirteen VLMs and five domains, showing that class-name shortcuts are widespread and can distort conclusions about grounded recognition, while controlled settings also expose weaknesses in negation, multi-image reasoning, and localized evidence use.
- We offer concrete recommendations for future evaluation, including reporting description-only settings and using conflict-based, synthetic-name, and localized-evidence variants along with binary q&a as robustness checks.

2 Benchmark Overview

2.1 Task Design

We construct a suite of controlled multiple-choice tasks that systematically manipulate the relationship between:

- **Nominal information** (N): the class name.
- **Visual description** (D): discriminative natural-language attributes.

Starting from a MCQA dataset based on five established fine-grained datasets, each original instance consists of an image I with ground-truth class c. For every example, there are four answer options $\mathcal{O} = \{o_1, o_2, o_3, o_4\}$, exactly one of which is correct under the task semantics. By selectively preserving, removing, or contradicting N and D, we generate counterfactual task variants probing distinct reasoning behaviors.

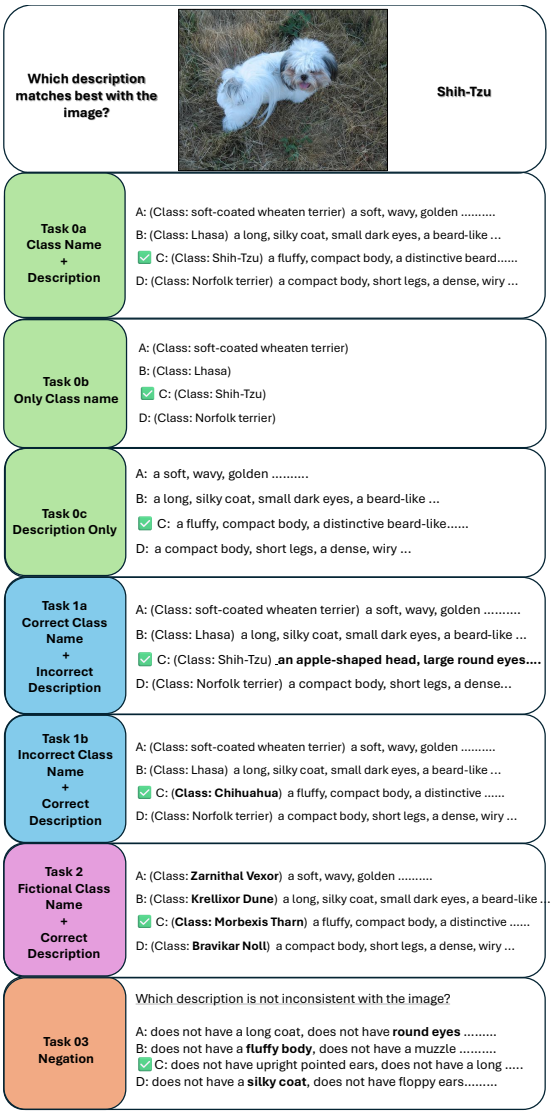


Figure 1: Single-image task examples. The descriptions begin with "This image shows a dog which has ...."

Collectively, the task suite decomposes fine-grained recognition into the following dimensions: (i) recognition from attributes, (ii) recognition from class labels, (iii) robustness under name–description conflict, (iv) sensitivity to label familiarity, (v) logical and compositional reasoning. and (vi) part-based evidence grounding.

2.2 Single-Image Recognition Tasks

Unless otherwise specified, the model receives a single image I and four textual answer options. Examples of the tasks are shown in figure 1.

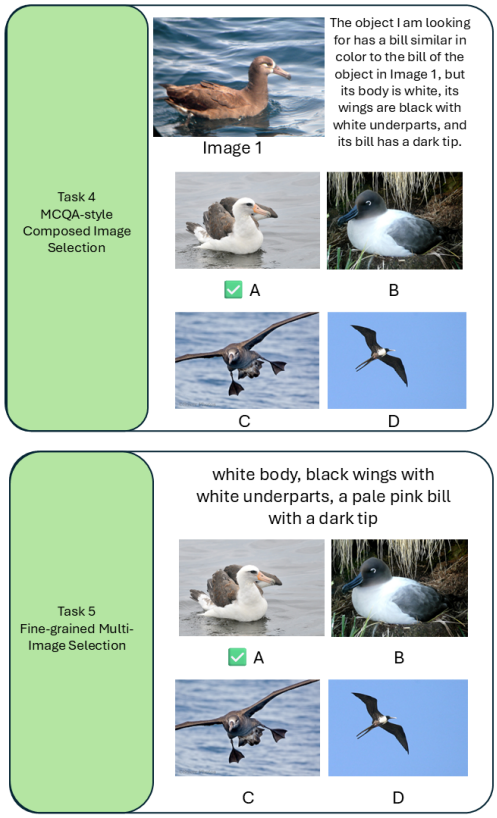


Figure 2: Examples of multi-image task categories.

Task 0a: Consistent nominal–visual recognition (N+D). Each option contains a mutually consistent class name and description, $o_i = (N_i, D_i)$ with $N_i \leftrightarrow D_i$. This is the standard fine-grained recognition setting where nominal and descriptive information agree.

Task 0b: Nominal-only recognition (N). Each option contains only the class name, $o_i = N_i$. This approach isolates recognition based on category labels alone, without descriptive support.

Task 0c: Description-only recognition (D). Each option contains only an attribute description, $o_i = D_i$. This task removes class-name cues and tests recognition from attributes alone.

Task 1a: Name-correct conflict. For the target class c, the correct option is constructed as $o_c = (N_c, D_j)$ where $j \neq c$. Here the class name is correct but the description belongs to another class.

Task 1b: Description-correct conflict. The correct description is paired with an incorrect class name, $o_c = (N_j, D_c)$ where $j \neq c$. Here the description matches the image, but the class name is incorrect. This tests whether grounded attribute evidence can override a misleading label.

The cue-conflict settings are diagnostic tasks with one intended correct answer under the prompt semantics. In the description-correct conflict setting, the intended answer is the option whose description matches the image despite the incorrect class name. In the name-correct conflict setting, the task analogously probes reliance on the class-name cue. In each of these tasks, all the other options are distractors with both incorrect class names and incorrect descriptions. The models are instructed to ignore the incorrect parts (e.g. class labels in incorrect class tasks).

Task 2: Synthetic nominal grounding. We replace real class names with synthetic identifiers:

$$o_i = (\tilde{N}_i, D_i).$$

Because the synthetic names have no real semantic associations, this setting removes label priors while preserving the prompt structure.

Task 3: Negated-attributes. Each option describes attributes that the object *does not* possess. The model must select the option whose negated statements are consistent with the image. We evaluate this task in three input settings:

- **Image-only**: the model sees the image and the negated answer options.
- **Text-only**: the image is replaced by a positive textual description of the target class.
- **Image+description**: the model receives both the image and the positive class description.

These variants help separate failures of visual grounding from linguistic negation reasoning.

2.3 Multi-Image and Reference-Conditioned Tasks

We further extend the benchmark to settings that require handling multiple candidate images. Examples of multi-image tasks are shown in figure 2.

Task 4: Compositional target identification. The model is given: (i) a reference image I_{ref}, (ii) a modification caption describing how the target differs from and resembles the reference, and (iii) four candidate target images $\{I_1, I_2, I_3, I_4\}$. The goal is to select the image consistent with the modification. This task requires integrating visual evidence from the reference image with comparative language, and therefore probes compositional fine-grained reasoning beyond direct description matching.

Task 5: Multi-image MCQA. Given a textual description D and a set of candidate images $\{I_1, I_2, I_3, I_4\}$, the model selects the image that best matches the description. Compared with single-image recognition, this setting is more demanding because the model must discriminate among several visually similar candidates. We run this in two formats mirroring tasks 0a and 0c: (i) the prompt contains both the name and the description (like task 0a) (ii) the prompt contains the description only (similar to task 0c).

2.4 Part-Based Evidence Grounding

This section evaluates whether vision–language models rely on localized body-part evidence when answering fine-grained attribute questions. We construct two attribute-centric questions using images from the CUB dataset (Wah et al., 2011), which provides annotated locations for multiple bird body parts (e.g., crown, throat, beak). Examples are provided in the appendix, Fig. 7.

Attribute verification. The model verifies whether a specific attribute holds for a queried body part (e.g., "Is the color of this bird's crown orange?") with binary answer choices.

Multiple-choice attribute classification. The model identifies the attribute value for a queried body part (e.g., "What color is this bird's throat?") from several candidate options.

Each question is evaluated under several controlled image conditions: (i) **full image**, where the entire image is shown; (ii) **part-only crop**, where only the queried body part is visible; (iii) **target-part masked**, where the body part required to answer the question is masked; and (iv) **irrelevant-part masked**, where an unrelated region is masked while the queried part remains visible. These conditions test whether predictions rely on relevant localized evidence or instead on global context or dataset priors. This task complements the earlier experiments by checking whether models base attribute predictions on the correct localized evidence or on global appearance or class-level correlations.

2.5 Data Generation

We construct our dataset by adapting mainly the "hard" split of description-only questions from (Atabuzzaman et al., 2025a). The underlying images and candidate objects remain unchanged, while option texts are modified according to the task type.

For the composed image selection task, we use GPT-OSS 120B (OpenAI et al., 2025) to generate modification captions describing attribute differences between reference and target objects. The same model is used to generate negated attribute descriptions for the negation task. For other variants, we apply systematic transformations to the option texts. Fictional class names are synthetic labels that do not correspond to real categories but are formatted similarly to natural class names. We manually inspected 600 randomly sampled questions and found that 97.7% of generated descriptions were correct and consistent with the intended task design. The prompts are provided in the appendix.

3 Results and Observations

We evaluate up to thirteen VLMs spanning different families and scales: DeepSeekVL2-tiny (Wu et al., 2024), DeepSeek-VL2-Small (Wu et al., 2024), Gemma3-12B (Team et al., 2025), InternVL-2.5-8B (Chen et al., 2025), Llama Scout 17B (Meta AI, 2025), LLaVA-OneVision-1.5-8B-Instruct (LLaVA-OV) (An et al., 2025), Phi-4 (Abouelenin et al., 2025), Qwen 2.5VL 7B (Team, 2025), GPT4.1-mini (OpenAI, 2025), GPT-5-mini (Singh et al., 2026), Gemini-2.5-flash (Comanici et al., 2025), Qwen-3.5-4B (Qwen Team, 2026a), Qwen-3.6-27B (Qwen Team, 2026b), Qwen3.6-35B-A3B (Qwen Team, 2026c). Due to resource limitations, we were not able to run every model for every task. For tasks 0 to 3, the GPT and Gemini models were evaluated on a fixed 80% subset of image instances. The evaluation covered all classes, questions, and domains; the images were selected before inference, and they were in the common full set used to evaluate the other models. Additionally, image instances were kept fixed when analyzing across task-variant comparisons within each model and the same 80% subset was used for both model families. More details are provide in appendix G.1.

3.1 Single Image Task Results (Task 0–Task 3)

Accuracy results for the single image tasks are shown as a heatmap in Fig. 3. For the tables and results, we used multiple images per question as each question and description can be used with images corresponding to the associated classes.

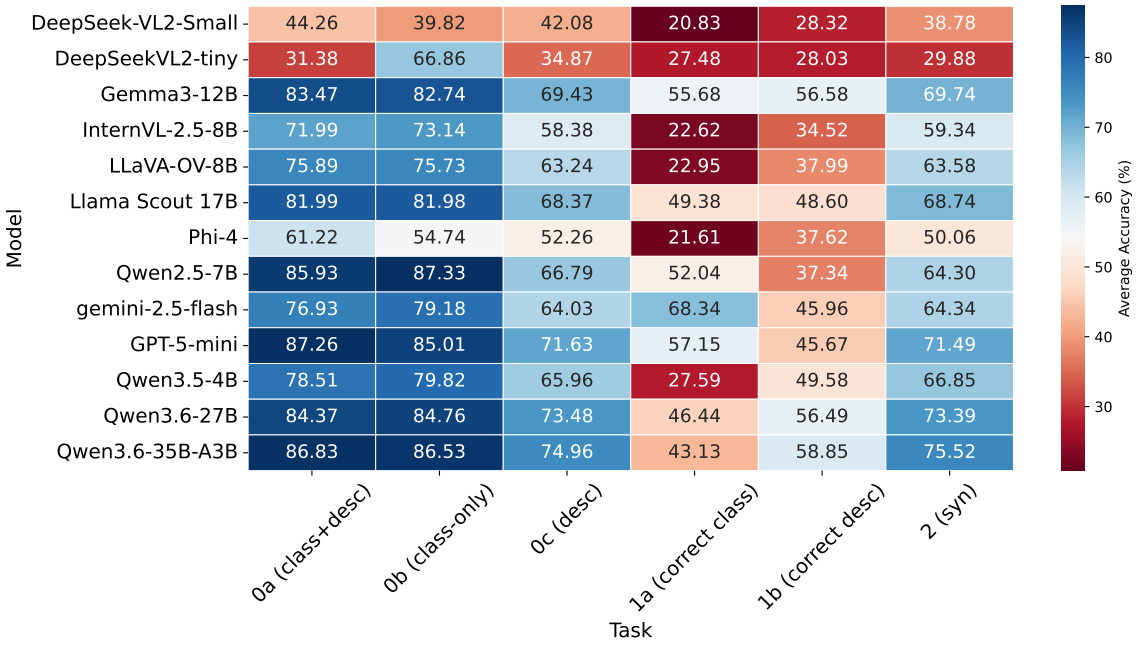


Figure 3: Heatmap of the accuracy (%) of the 13 models on single-image tasks averaged over the five datasets. Including the class names in the options improves performance significantly as tasks 0a and 0b are quite higher than 0c. Including an incorrect class name with the correct description misleads the model (Task 1b) while synthetic fictional names remove the influence (Task 2).

3.1.1 Observation 1: Class-name labels can create shortcuts in zero-shot fine-grained classification

The class-only baseline (Task 0b) is unexpectedly strong, indicating that many instances can be solved by matching images to label strings without grounding the attribute descriptions. Previous research (Atabuzzaman et al., 2025a) reported performance gains from including class names, but their analysis used different answer choices for the with/without name settings, introducing confounding factors. Moreover, they did not examine whether the gains reflected improved attribute grounding or reliance on class-name priors. Our results show that models achieve similarly high accuracies even without descriptions, suggesting that label priors play a substantial role. This raises the question of how much descriptions contribute when both class names and attributes are provided. We therefore recommend that zero-shot fine-grained benchmarks report results both with and without class names and descriptions.

3.1.2 Observation 2: Synthetic labels reduce class-label prior shortcuts

A possible explanation is that class-labels alter the prompt format. To test this, we introduce Task 2, where class names are replaced with fictional labels. Performance is similar to the description-only setting, indicating that the gains from real class name additions are not due to prompt structure. Instead, models largely ignore unfamiliar labels and rely on attribute descriptions. These results indicate that fictional class names provide a useful setting for evaluating attribute-based recognition.

3.1.3 Observation 3: Familiar class labels can override visual evidence

The Task 1a–Task 1b pair reveals how models resolve conflicts between class labels and attribute descriptions. Many models degrade more under Task 1a (correct class, incorrect description) than under Task 1b (incorrect class, correct description), suggesting that predictions often follow attribute descriptions rather than nominal cues under conflict. However, behavior varies across models, indicating different sensitivities to lexical information.

Performance on Task 1b is often lower than on Task 2, where fictional labels remove real-world priors. This gap suggests that familiar class names can suppress visual grounding even when models possess the capability to perform the task. This aligns with (Fu et al., 2025) showing that VLMs may under-utilize visual representations, but our

results indicate a more specific phenomenon.

Models perform quite well on medium-difficulty examples despite mismatched labels or descriptions, performance drops more noticeably on harder instances (Appendix H, tables 10 to 19). For example, Qwen2.5-7B achieves 90%+ accuracy for the Cars dataset (Krause et al., 2013) in the conflict tasks (1a, 1b) for the medium difficulty split but drops to 62.24% and 36.84% in the hard split for the same tasks.

3.1.4 Observation 4: Smaller models may have difficulty parsing complex attribute descriptions

A striking disparity was observed in DeepSeekVL2-tiny. For example, on the CUB dataset, DeepSeekVL2-tiny achieves 72.95% accuracy on Task 0b (class name only). However, simply adding the correct visual description (Task 0a) causes its performance to collapse to 34.35%. This suggests that DeepSeekVL2-tiny has difficulty handling the fine-grained descriptions and struggles with the long prompt.

3.1.5 Observation 5: Negation faces a visual bottleneck

Task 3 substantially reduces accuracy across many models, indicating that reasoning about negated attributes is harder than matching affirmative attribute descriptions. We evaluate three input settings: image-only, text-only, and image+text (Fig. 4).

Across models, the text-only and image+text settings are very close, and performance in both settings is generally high. This suggests that once the relevant semantic information is available in text, most models can handle the linguistic structure of negation reasonably well. Adding the image does not appear to provide a large additional benefit over the text-only condition. In contrast, performance drops more clearly in the image-only setting, indicating that the harder part of the task is not simply understanding negation, but grounding the relevant attributes directly in visual evidence.

Task 3 mainly tests two coupled abilities: recognizing the object's positive attributes from the image, and reasoning over negated textual descriptions. The text-only condition isolates the second ability, while the image-only condition requires both. The image+text condition provides the positive attributes explicitly, so its closeness to text-only suggests that the bottleneck is not multimodal verification once the positive description is available. The weaker image-only results suggest that the remaining gap is less about lexical negation alone and more about aligning negated attribute reasoning with visual evidence.

DeepSeek-VL2-Small performs better in the text-only condition than in the image+text condition, while DeepSeekVL2-tiny remains low across all three settings. This suggests that these models do not benefit from the added visual context in the same way as stronger models, and that visual input may introduce noise when vision–language alignment is weak.

We compare three equivalent prompt forms: a double-negation query (*NOT inconsistent*), a single-negation query (*does not contradict*), and a positive formulation (*consistent*). The first two yield similar results ($+1.15\%$ difference on average), while the positive formulation reduces accuracy by 11.04%. The effect varies across models and datasets and is strongest in the text-only condition, suggesting models' sensitivity to the alignment between the prompt and the negated answer options.

3.2 Multi-Image Task Observations

These tasks evaluate fine-grained understanding under settings that require selecting among multiple visually similar candidate images. The tasks include both standard matching scenarios and a reference-conditioned composed image selection task (Task 4). The results are shown in Fig. 5.

3.2.1 Task 5: Fine-grained multi-image recognition

As shown in Fig. 5, performance is generally higher on standard single-image matching tasks than on the corresponding multi-image variants. While models perform well on image-to-text recognition, accuracy drops when the same decision must be made among multiple candidate images. This suggests that grounding textual attributes becomes more difficult when several visually similar images are present.

3.2.2 Task 4: MCQ based composed image selection using reference image and modification text

Task 4 generally produces results lower than the multi-image recognition tasks. This indicates that reference-conditioned composed selection is more difficult than direct matching. In Task 4, the model must interpret the reference image, understand

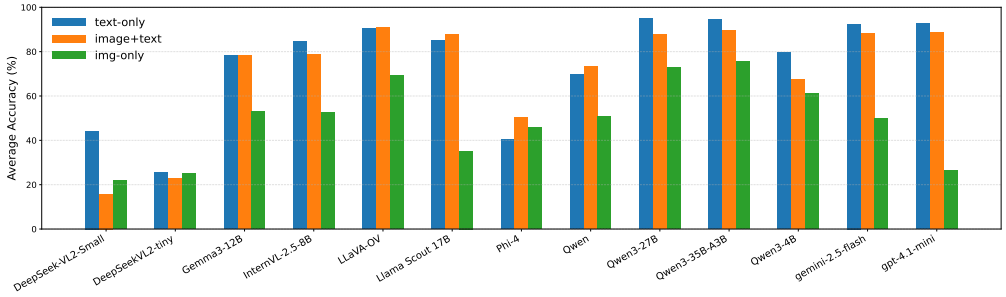


Figure 4: Accuracy (%) of 13 models on the negation task averaged over the five datasets. We show model performance under three input settings: text-only (blue), image+text (orange), and image-only (green). The models are generally able to handle the lexical complexity of negation as shown by the high performance in the text-only setting and the image+text setting. However, they struggle when given only the image.

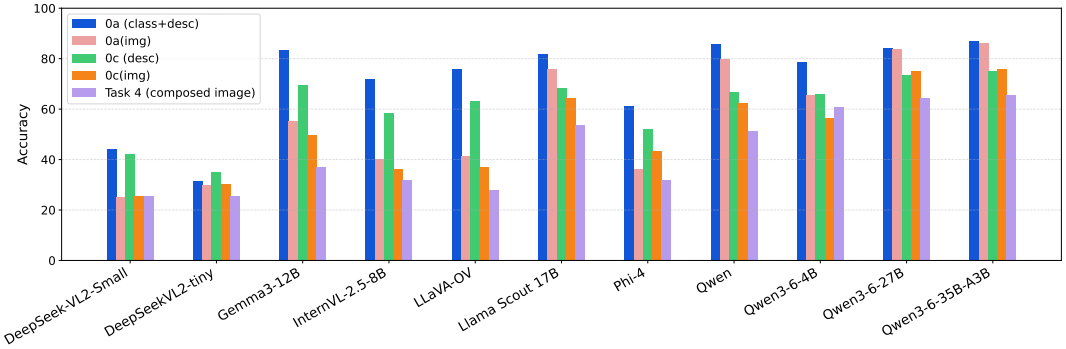


Figure 5: Accuracy (%) of eleven models on single-image tasks (0a, 0c), their multi-image variants, and Task 4. Results are averaged over the five datasets. The models generally struggle more with the multi-image variants.

the modification caption, infer the intended target concept, and then identify the correct image among distractors. The results indicate that reference-conditioned selection is generally more difficult than direct matching tasks and strong performance on standard fine-grained matching does not consistently translate to strong Task 4 performance.

3.3 Task 6: Part-Based Evidence Grounding

Results for Task 6 are shown in Table 1. We omit the results for the irrelevant-masked version here as they are usually very close to the full-image versions. In one evaluation variant, we also allow an abstention option, *Cannot determine from the image*. Overall, models perform better on $q1$ than on the corresponding $q2$ questions. $q1$ serves as a cleaner grounding probe, often showing a sharper contrast between target-part masking and irrelevant-region masking, whereas $q2$ is a harder fine-grained classification task. In the standard setting without abstention, both tasks show performance drops when the queried part is masked, though these drops are often moderate, suggesting that models still rely partly on global context, species-level priors, or correlated attributes.

Allowing abstention reveals substantial calibration differences. Several models shift away from forced predictions on masked inputs, indicating that some earlier masked performance came from contextual guessing. However, some models overuse abstention even when relevant evidence is available, reducing performance on non-masked conditions.

We tested whether models can identify bird species from the Task 6 part-based image variants. Accuracy remains relatively high when the queried part is masked but drops sharply for part-only crops, suggesting that models can rely on global appearance, unmasked regions, and species-level priors. We also find that not all body parts contribute equally to classification. This helps explain why models may answer some part-based attribute questions correctly even when the queried part is masked: the remaining image can still contain residual species-level cues. Details are given in appendix.

Overall, the results suggest that current VLMs exhibit *partial* part-based reasoning: they respond

	q1						q2					
	Without Z			With Z (abstention)			Without Z			With Z (abstention)		
Model	F	C	M	F	C	M	F	C	M	F	C	M
DeepSeek-VL2-Small	35.98	75.30	26.81	30.15	15.46	25.36	34.19	40.85	28.32	26.37	12.63	15.23
DeepSeek-VL2-Tiny	64.94	65.67	63.37	67.71	67.77	64.93	50.40	52.06	44.29	45.21	45.60	38.92
Gemma3-12B	66.81	67.09	52.52	30.71	34.39	3.06	57.58	56.56	48.86	45.73	45.93	6.09
LLaVA-OV-8B	73.73	72.06	68.62	57.30	40.61	1.16	63.65	59.71	57.92	56.00	46.00	2.74
Phi-4	67.99	65.81	64.77	67.18	65.15	57.04	52.43	49.47	46.63	49.19	43.95	27.73
Qwen3-27B	75.97	75.14	52.24	76.20	70.34	4.66	68.17	62.70	59.05	66.67	60.51	2.26
Qwen3-35B-A3B	76.61	72.28	66.84	76.70	72.74	6.99	68.14	63.55	59.48	67.03	63.37	3.14
GPT-5-mini	78.42	72.33	66.00	78.33	71.50	26.42	67.25	58.83	54.92	67.00	59.00	26.58

Table 1: Part-based evidence grounding accuracy (%) on CUB for q1 (attribute verification) and q2 (multiple-choice attribute classification). Here Z = "Cannot Determine." F = full image, C = part crop, and M = target-part masked. Models can reliably answer questions about body part attributes even when the query body parts are masked out but some models choose to abstain if given the option to do so. (Full table is given in the appendix). Due to resource limitations, a smaller subset of the samples were evaluated for the GPT-5-mini model.

to the queried local evidence but can still leverage shortcut cues. Performance on cropped inputs is often close to, and sometimes better than, that on full images, suggesting the relevant evidence is frequently localized.

4 Binary QA ablations

We design a binary question-answering task by converting each four-way MCQ into four independent yes/no questions. Each answer option is presented separately, and the model is asked whether it is correct. We find that models often correctly answer "yes" for the true option, but frequently fail to reject visually similar distractors. Performance can drop by over 38% and so, we suggest that future benchmarks should report both MCQ and binary verification results. Detailed results are given in the appendix D.

5 Related Work

VLMs often struggle with detailed visual discrimination (Kim and Ji, 2024; Yu et al., 2026; Conti et al., 2023). Using synonyms of class names can significantly improve performance, suggesting reliance on lexical cues rather than visual grounding (Parashar et al., 2024). In (Atabuzzaman et al., 2025b), recognition is formulated as an iterative MCQA task using both class names and descriptive attributes. Other fine-grained VLM benchmarks also include class-name labels in the answer choices (Geigle et al., 2024), which may allow models to exploit class-name priors. Visual descriptions have been used to improve zero-shot recognition and interpretability in CLIP-style models (Menon and Vondrick, 2022). Label-based and description-based recognition may behave differently as there is an alignment gap between them in multimodal models (He et al., 2025). However, modern generative VLMs are often instruction-tuned and trained for general multimodal reasoning, making them distinct from CLIP-style classification. It is therefore important to examine how these models handle descriptions and class names.

A growing body of work examines shortcut behaviors and biases in multimodal benchmarks. CLIP-style models exhibit sensitivity to class-name tokens, prompt structures, and lexical cues (Baron et al., 2024; Parisot et al., 2023; Li et al., 2025a). (Roth et al., 2023) show that even nonsensical descriptors can yield comparable zero-shot classification performance with CLIP, suggesting reliance on language priors. Recent analyses suggest that VLMs may under-utilize visual evidence, relying instead on language-driven heuristics (Fu et al., 2025) or default to familiar world priors under counterfactual visual evidence (Vo et al., 2026). While these studies reveal shortcut behavior, they do not directly isolate how class-name priors affect zero-shot fine-grained recognition with natural images and attribute descriptions. Many existing evaluation settings utilize category labels that may allow models to rely on familiar label tokens rather than grounding fine-grained visual attributes. Thus, high performance on such benchmarks may conflate label familiarity with grounded attribute understanding. Understanding how this behavior manifests in modern generative VLMs remains an open question.

Prior work shows that VLM performance degrades significantly under negation-based reasoning (Zhang et al., 2025; Alhamoud et al., 2025; Singh et al., 2024). However, these studies

focus primarily on coarse-grained visual reasoning and do not evaluate fine-grained recognition with multiple negated attribute options. Multi-image fine-grained benchmarking is also limited, as most multi-image works evaluate coarse-grained reasoning (Wang et al., 2025a; Tian et al., 2025; Wang et al., 2025b). Research by (Liu et al., 2024) includes a fine-grained multi-image task, but it probes image-to-image matching rather than text-image alignment. Similarly, composed image retrieval datasets are often restricted to narrow domains (Li et al., 2025b).

Shortcut reliance can also arise from visual context. Work by (Xu et al., 2025) shows that CLIP-style VLMs can assign high confidence to categories using background visual information even when foreground objects are removed. We have, however, found that modern generative VLMs collapse and resort to random answering when faced with fully masked out object images. In contrast, our part-based masking task studies a more subtle setting. We ask models fine-grained attribute questions about specific object parts and tests whether they abstain when the queried evidence is missing or infer answers from relevant global context and shortcut cues.

Motivated by these issues, we introduce a controlled benchmark that separates nominal class labels from visual attribute descriptions. We construct task variants that remove, replace, or conflict class names and descriptions and extend evaluation to negated attributes, multi-image reasoning, reference-conditioned selection, and part-based evidence grounding. This framework enables a clearer analysis of how VLMs balance label priors against grounded visual evidence.

6 Conclusion

We introduced STRESS-FG, a controlled benchmark for testing whether VLMs truly ground fine-grained attributes rather than relying on class-name priors. We find that standard zero-shot evaluations can overstate visual grounding: label-only prompts are often highly competitive, synthetic labels can reveal stronger attribute-based recognition, and familiar labels can interfere with grounded decisions under cue conflict. We also show that negation, multi-image discrimination, and reference-conditioned selection remain difficult for current models. Finally, part-based experiments indicate only partial reliance on the correct localized evidence, with many models still using global context or shortcut cues. We suggest that future fine-grained evaluation should separate label-based and description-based recognition and include conflict-based, synthetic-label, and localized-evidence settings as robustness checks.

7 Limitations

STRESS-FG is designed as a controlled diagnostic benchmark, but it has several limitations. First, it covers five fine-grained domains and a limited number of VLMs, so the results may not fully generalize to other domains, or non-MCQA evaluation settings. We tried to mitigate MCQ confounds by keeping the answer option position distribution relatively uniform but, there could be still some effects of MCQ confounds. Second, the benchmark mainly evaluates recognition through single-turn multiple-choice questions, which isolates specific behaviors but may not fully reflect how these behaviors appear in more open-ended settings. Third, we focus on evaluation rather than intervention, and therefore do not study methods for reducing shortcut reliance or analyze internal mechanisms such as attention patterns. Fourth, we diagnose whether familiar class labels act as shortcut cues in existing evaluation formats but do not determine the origin of these class-label prior shortcuts. This can originate from semantic knowledge, pre-training exposure, dataset contamination etc. Finally, the part-based grounding analysis is limited to CUB, where body-part annotations are available, so this component does not yet test localized grounding across all domains. We hope that our findings will motivate further work in this direction. These limitations should be considered when interpreting the benchmark as a diagnostic tool rather than a complete measure of fine-grained visual understanding.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, and 55 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via Mixture-of-LoRAs. *Preprint*, arXiv:2503.01743.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip H.S. Torr, Yoon Kim, and

Marzyeh Ghassemi. 2025. Vision-language models do not understand negation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29612–29622.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, and 4 others. 2025. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *Preprint*, arXiv:2509.23661.

Md. Atabuzzaman, Ali Asgarov, and Chris Thomas. 2025a. Benchmarking and mitigating MCQA selection bias of large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33548–33562, Suzhou, China.

Md. Atabuzzaman, Andrew Zhang, and Chris Thomas. 2025b. Zero-shot fine-grained image classification using large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23569–23582, Suzhou, China.

Ethan Baron, Idan Tankel, Peter Tu, and Guy Ben-Yosef. 2024. Real classification by description: Extending CLIP's limits of part attributes recognition. *Preprint*, arXiv:2412.13947.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101- Mining discriminative components with random forests. In *European Conference on Computer vision*, pages 446–461. Springer.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. 2023. Vocabulary-free image classification. *Advances in Neural Information Processing Systems*, 36:30662–30680.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025. Hidden in plain sight: VLMs overlook their visual representations. In *Second Conference on Language Modeling*.

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

Dhruba Ghosh, Yuhui Zhang, and Ludwig Schmidt. 2026. Understanding the fine-grained knowledge capabilities of vision-language models. *Preprint*, arXiv:2602.17871.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. 2025. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pages 89927–89944.

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

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

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

Ao Li, Zongfang Liu, Xinhua Li, Jinghui Zhang, Pengwei Wang, and Hu Wang. 2025a. Modeling variants of prompts for vision-language models. *Preprint*, arXiv:2503.08229.

Zixu Li, Zhiheng Fu, Yupeng Hu, Zhiwei Chen, Haokun Wen, and Liqiang Nie. 2025b. FineCIR: Explicit parsing of fine-grained modification semantics for composed image retrieval. *Preprint*, arXiv:2503.21309.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and Weiming Hu. 2024. MIBench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22417–22428, Miami, Florida, USA. Association for Computational Linguistics.

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

Sachit Menon and Carl Vondrick. 2022. Visual classification via description from large language models. *Preprint*, arXiv:2210.07183.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2026-08-29.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI. 2025. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2026-08-29.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. 2024. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12988–12997.

Sarah Parisot, Yongxin Yang, and Steven McDonagh. 2023. Learning to name classes for vision and language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23477–23486.

Qwen Team. 2026a. Qwen3.5: Towards native multimodal agents.

Qwen Team. 2026b. Qwen3.6-27B: Flagship-level coding in a 27B dense model.

Qwen Team. 2026c. Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. 2023. Waffling around for performance: Visual classification with random words and broad concepts. In *2023 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15700–15711.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. OpenAI GPT-5 System Card. *Preprint*, arXiv:2601.03267.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. 2024. Learn "No" to Say "Yes" Better: Improving Vision-Language Models via Negations. *Preprint*, arXiv:2403.20312.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2025. Qwen2.5-VL.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. 2025. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10599–10609.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. 2026. Vision language models are biased. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The Caltech-UCSD Birds-200-2011 dataset.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2025a. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pages 62624–62650.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. 2025b. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3221–3241.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. *Preprint*, arXiv:2412.10302.

Zhuo Xu, Xiang Xiang, and Yifan Liang. 2025. Overcoming shortcut problem in VLM for robust out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15402–15412.

Hongtao Yu, Yuxin Peng, Serge Belongie, and Xiu-Shen Wei. 2026. Benchmarking large vision-language models on fine-grained image tasks: A comprehensive evaluation. In *International Conference on Learning Representations*, volume 2026, pages 44677–44700.

Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. 2025. NegVQA: Can vision language models understand negation? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3707–3716, Vienna, Austria. Association for Computational Linguistics.

A Appendix

This appendix contains:

- B. Dataset Generation Prompts
- C. Dataset Statistics
- D Binary Verification Ablation.
- E Class-only recognition under part perturbations.
- F Summary of Task variants
- G. Model Sizes, Licensing and Computational Setup
- H. Full Tables with detailed results

B Dataset Generation Prompts

B.1 Prompt for Generating the Negation Dataset

```
You are given a single multiple-choice visual
    question answering item in JSON format. It
    contains:
"question"
"options": an object with keys "A""D" and text
    descriptions
"correct_answer": a key "A", "B", "C", or "D"
Example Input:
   "question": "Which description matches best
       with the image?",
   "options": {
     "A": "This image shows a Least Auklet
          object which has a small black and
          white body, a black head and back, a
          white belly, a black bill, a white eye-
          ring, black wings with white spots,
          and black legs and feet.",
     "B": "This image shows a Rhinoceros Auklet
          object which has a black or dark grey
          body, a large orange bill, a white
          face with a black stripe behind the
          eye, and is often seen near water,
          catching fish.",
     "C": "This image shows a Crested Auklet
          object which has a small, dark body, a
           bright orange bill, distinctive white
           facial markings, and a prominent
          crest of feathers on its head.",
     "D": "This image shows a Parakeet Auklet
          object which has a small, stocky body,
           black head and back, white underparts,
           an orange bill, and a distinctive
          white eyebrow stripe, distinguishing
          it from similar seabirds."
   },
   "correct_answer": "C",
Your task is to rewrite all answer options into
    the SAME negated style:
 This image shows a car, which does not have
     and then listing some visual-only
     attributes.
Follow these rules for Creating Negated Answer
    Choices (All Options in Negated Form)
Applies to all answer choices (AD):
1. Remove all class names from every answer
    choice.
2. Rewrite every answer choice to begin with:
    This image shows a car which does not have
For the correct answer choice:
3. Only include visual attributes that the
    object in the image truly does not have (
    true negatives).
4. These true-negative attributes should be
    taken from:
* Attributes mentioned in the original incorrect
     answer choices, or
* Clear visual opposites of the objects real
    features.
5. You may change the colors of these attributes
     when negating them (e.g., does not have a
    green bill), as long as the statement
    remains factually true.
6. Do not negate any feature that the object
    actually has (no false negatives).
7. Include around 3 to 5 true-negative
    attributes in the correct answer.
For each incorrect answer choice:
8. Identify the visual attributes that the
    object actually has from the original
    correct answer choice.
9. In each incorrect answer choice, negate any
    two of these true attributes (creating false
     negatives). Example: if the object has a
    tentacle, say does not have a tentacle.
10. Add two more incorrect visual attributes if
    needed, but they must remain false with
    respect to the correct answer choice. You
    may get these features from the original
    incorrect answer choice.
11. You may randomly change the colors of these
    additional visual attributes while negating
    them (e.g., does not have a blue fin instead
     of its real orange fin). Do not change the
    colors of the negated real attributes from
    Step 09. Random colors may only be applied
    to the extra fake attributes you add, not to
     the negated real ones.
Additional formatting guidelines:
* Only use visual attributes (color, shape, size,
     markings, body parts).
* Do not include behavior, habitat, or non-
    visual information.
* All answer choices must follow the same
    negated style.
* Ensure that two consecutive incorrect choices
    do not have more than one common false
    negatives added in step 9.
Example Negated Version Output:
   "question": "Which description matches best
       with the image?",
   "options": {
     "A": "This image shows a car which does not
          have a small dark body, a bright
         orange bill, a black head and back, a
```

```
          red belly, a green bill, a white eye-
          ring, black wings with white spots,
          and black legs and feet.",
      "B": "This image shows a car which does not
           have a bright orange bill,
          distinctive white facial markings, a
          black or dark grey body, a large green
           bill, a white face with a black
          stripe behind the eye.",
      "C": "This image shows a car which does not
           have black wings with white spots, a
          bright green bill, distinctive yellow
          facial markings, a distinctive white
          eyebrow stripe and a white eye-ring.",
      "D": "This image shows a car which does not
           have distinctive white facial
          markings, a prominent crest of
          feathers on its head, orange head and
          back, white underparts, an orange bill,
           and a distinctive white eyebrow
          stripe."
    },
    "correct_answer": "C",

Now apply this to this MCQ Entry:

{input_json}
```

B.2 Prompt for Generating the Task 4 dataset

```
You are given a multiple-choice question about
     object images. Each option is a natural-
     language caption describing an object. One
     option (the correct answer) is the reference
      object. One of the other options is the
     target object. The remaining options are
     distractors.

Your task is to generate a short Final
     modification caption that describes the
     target object in terms of its visual
     similarities and differences relative to the
      reference object. This caption will be used
      for a composed image retrieval task, where
     a model must use BOTH the reference image
     and the modification caption to find the
     target.

Follow these instructions carefully:

1. Inputs

   * You will receive:

     * A set of options (e.g., A, B, C, D), each
          a caption describing an object.
     * A Reference Class and a field `
          correct_answer` indicating which option
           is the reference object.
     * A separate indication of which option is
          the target object.
   * Assume the reference object corresponds to
       Image 1.
   * The other options (besides the target) are
       distractor objects.

2. Extract only explicit visual attributes

   * For the reference object caption, list all
       explicitly described visual attributes,
       such as:

     * Body color and pattern
     * Head and neck color and pattern
     * Bill (beak) color and shape
     * Wing color, pattern, shape, and length
     * Tail shape or special markings
     * Distinctive visual features (e.g., throat
          pouch, neck band)
     * Visible environment only if explicitly
          described (e.g., resting on grassy
          areas)
   * For the target object caption, list all
       explicitly described visual attributes of
        the same kinds.
   * VERY IMPORTANT: Do NOT invent or add any
       attribute that is not explicitly
       mentioned in the captions. Do not infer
       new colors, shapes, patterns, body parts,
        background, or behaviors beyond what the
        text directly states.

3. Similarities between reference and target (
     relational only)

   * Identify which visual attributes are
       semantically similar between the
       reference and target objects, based
       solely on the explicitly described
       attributes in both captions
   * If there are no similarities between them,
       you can skip this Similarities part and
       go to the next step.
   * When you describe similarities, use purely
       relational language that points back to
       the reference object WITHOUT restating
       the concrete attribute values.
   * Prefer formulations like:

     * a body color similar to that of the bird
          in Image 1
     * wings similar in shape and length to those
           of the object in Image 1
     * a bill similar in color to the bill of the
           bird in Image 1
   * Avoid formulations that explicitly repeat
       the shared attribute values, such as:

     * both objects have long, narrow wings
     * both objects have dark brown bodies
   * The goal is that the reader must look at
       the reference bird to know what the
       shared attribute actually looks like.

4. Differences between reference and target (
     absolute descriptions)

   * Identify all visual attributes where the
       target bird differs from the reference
       bird.
   * Describe these differences using absolute
       descriptions of the targets attributes,
       for example:

     * its body is white
     * it has a dark gray body
     * it has a hooked bill
     * it has a forked tail
```

```
    * it has a red throat pouch
  * These differences should be based ONLY on
      what is explicitly written in the target
      caption (and, when relevant, by contrast
      with what is written in the reference
      caption).
  * Do NOT add new attributes that are not
      mentioned in the text.
  * List no more than three of the most major
      of these differences in the final caption.


5. Check against distractor captions
  * Compare your current similarity-and-
      difference description of the target
      against the captions of the other
      distractor objects.
  * If your description could also match one of
       the distractor objects (because it
      shares many of the same attributes), then
       your caption may not be discriminative
      enough.
  * In that case, look for additional
      attributes in the target caption that:

    * Are explicitly mentioned for the target,
        AND
    * Help distinguish it from the overlapping
        distractor(s).
  * Add one or more of these distinguishing
      attributes to your caption, still
      following the rules above:

    * Similarities must remain relational to the
         object in Image 1.
    * Differences must be absolute descriptions
        of the targets explicitly mentioned
        attributes.

6. Limited and careful use of negation

  * You may use negation (does not have) ONLY
      if:

    * The negated feature is explicitly
        mentioned in a distractor caption, AND
    * Negation is necessary to distinguish the
        target from that object.
  * Do NOT introduce negation about attributes
      that are not mentioned in any caption.
  * Do NOT rely on negation as the main
      distinguishing factor if a more direct
      positive attribute difference is
      available.

7. Output formatting

  * Write 1--3 sentences of fluent natural
      language.
  * Refer to the reference object as the object
       in Image 1.
  * Refer to the target object as the object I
      am looking for.
  * Do NOT mention option letters (A, B, C, D).
  * Do NOT mention any class or species names.
  * Do NOT mention the words reference object,
      target object, distractor, or any
      internal reasoning.

You can output two sections:
[optional] Reasoning: A step-by-step analysis
    where you explain your reasoning.
Final caption: The final, clean 13 sentence
    final caption, with no bullet points, no
    explanations, and no visible reasoning steps
    .

8. Summary of strict constraints

  * Do NOT invent any new features not
      explicitly present in the captions.
  * For similarities: describe them ONLY as
      relations to the object in Image 1 (e.g.,
       similar to the bird in Image 1) and
      avoid repeating the exact attribute
      values.
  * For differences: describe the targets
      attributes with explicit values, but only
       those present in the text.
  * Ensure the caption is sufficiently specific
       to distinguish the target from the
      distractor birds, while still requiring
      the reference image to interpret the
      relational similarities.

Here is an example:

Example 1:

Original:
"
"mcq_id": "Black_footed_Albatross",
"difficulty": "Hard",
"options": {
A: "This image shows a Laysan Albatross bird
    which has a white body, black wings with
    white underparts, a pale pink bill with a
    dark tip, and is often seen soaring
    gracefully over oceans or resting on grassy
    areas.",
B: "This image shows a Sooty Albatross bird
    which has a dark gray body, a slender hooked
     bill, long narrow wings, and a subtle white
     neck band, distinguishing it from similar
    seabirds.",
C: "This image shows a Black-footed Albatross
    bird which has a predominantly dark brown
    body, a pale head and neck, a long pinkish
    bill, and long, narrow wings suited for
    soaring over oceans.",
D: "This image shows a Frigatebird bird which
    has a sleek black body, long wings, a forked
     tail, and a distinctive red throat pouch
    inflated during mating displays,
    distinguishing it from other seabirds."
},
"Reference_Class": "Black-footed Albatross"
"Target_Class": "Sooty Albatross"

Output:
{
**Reasoning**
**Final caption**
"The object I am looking for has a body of
    similar color as the body of the bird in
    Image 1, and similar wing shape and length
    as the bird in image 1. However, it has a
    hooked bill and a subtle white neck band."
}
```

```
Now apply this to this MCQ Entry:
{input_json}
```

C Dataset Statistics

We adapt 1374 medium- and hard-difficulty questions from the original dataset (Atabuzzaman et al., 2025a). These questions form the base set from which we derive task-specific variants. For Tasks 0–2, we systematically modify the correct answer choices while preserving the overall structure of the questions, resulting in the same number of questions (1374) for each task. For the newer models, we only did inference on the hard splits.

For the negation task (Task 3), we generate one additional question per original entry by negating the attribute described in the question while keeping the answer classes unchanged. This results in 1374 negation-task questions.

For the composed image selection task (Task 4), we generate three derived questions for each original entry. In each case, the original ground-truth class serves as the reference class, and one of the incorrect answer options is selected as the target class. Since the original questions contain four answer options (one correct and three distractors), this produces three derived questions per original entry, resulting in 4,122 Task 4 questions in total. Task 5 reuses the original base questions but with images as answer options instead of text and therefore contains 1374 entries.

Each of the tasks described above can be evaluated using different sample images corresponding to the associated classes.

For the part-based evidence grounding task (Task 6), we construct two types of questions: (q1) attribute verification and (q2) multiple-choice attribute classification. These are generated at the image level rather than the question level. In total, we create 12,210 q1 entries and 6,105 q2 entries, where each entry corresponds to a single image and a specific attribute-based query. We additionally apply cropping and masking to each entry to create the corresponding evaluation variants.

D Binary Verification Ablation

Most experiments in our benchmark use a four-way multiple-choice format, where the model only needs to identify the best option among a small candidate set. This format may introduce additional complexity by requiring models to jointly compare four answer options. It may also hide a different failure mode: a model can select the correct option while still assigning high plausibility to visually similar distractors. To test these effects, we construct a binary verification version of the Task 0b and Task 0c questions. Each original four-way MCQ instance is converted into four independent yes/no questions, one for each answer option. For each option, the model is shown the image and a single candidate answer, and is asked whether that answer is correct.

We evaluate the resulting binary task using three increasingly permissive criteria. *All-four* accuracy requires the model to answer all four binary questions correctly, i.e., "yes" for the correct option and "no" for all three distractors. *Two-way* accuracy requires the model to answer the correct option and the first distractor correctly. *One-way* accuracy only checks whether the model answers "yes" for the correct option. The gap between these metrics measures whether a model can recognize the correct answer while also rejecting plausible incorrect alternatives.

The results in Table 2 reveal a strong asymmetry between accepting correct answers and rejecting distractors. For most models, one-way accuracy is very high, often exceeding 85%, showing that the models frequently answer "yes" when presented with the correct option. However, all-four accuracy is much lower. For example, DeepSeek-VL2-Tiny obtains nearly perfect one-way accuracy on Task 0c, but its all-four accuracy is 0.00%, indicating that it almost always accepts the correct option but fails to reject distractors. A similar pattern appears for LLaVA-OV, where Task 0c one-way accuracy is 95.84%, but all-four accuracy drops to 7.60%.

This suggests that the MCQ formulation can overstate fine-grained discrimination. In a four-way MCQ, selecting the correct option does not require the model to explicitly mark every distractor as incorrect. The binary formulation exposes that models often treat multiple options as plausible, especially when distractors are fine-grained and visually close to the target. Thus, MCQ accuracy can hide a tendency toward affirmative verification, where the model recognizes the correct class or description but does not robustly reject incorrect alternatives.

Overall, this ablation shows that multiple-choice evaluation should be complemented with binary verification. The MCQ format is useful for measuring relative preference among options, but

Model	Source Task	One-way	Two-way	All-four	MCQ	$\Delta \downarrow$
DeepSeek-VL2-Small	0b (class-only)	27.07	19.91	12.63	39.82	-27.19
DeepSeek-VL2-Small	0c (desc-only)	30.98	12.17	2.82	42.08	-39.26
DeepSeek-VL2-Tiny	0b (class-only)	98.05	31.41	8.94	66.86	-57.92
DeepSeek-VL2-Tiny	0c (desc-only)	99.83	0.41	0.00	34.87	-34.87
LLaVA-OV-8B	0b (class-only)	98.49	47.71	23.90	75.73	-51.83
LLaVA-OV-8B	0c (desc-only)	95.84	24.80	7.60	63.24	-55.64
Qwen-3-5-4B	0b (class-only)	91.56	69.46	47.02	79.82	-32.80
Qwen-3-5-4B	0c (desc-only)	87.31	40.35	17.03	65.96	-48.93
Qwen-3-6-27B	0b (class-only)	90.01	72.78	53.74	84.76	-31.02
Qwen-3-6-27B	0c (desc-only)	85.65	53.83	28.50	73.48	-44.98

Table 2: Binary verification accuracy (%) averaged over the five fine-grained domains. Each four-way MCQ is converted into four yes/no questions. All-four requires the model to answer all four binary questions correctly. Two-way requires the correct option and first distractor to be answered correctly; One-way only checks whether the model answers 'yes' for the correct option. The models significantly struggle to answer all four questions correctly.

it does not fully test whether models can distinguish correct fine-grained evidence from plausible distractors. Reporting both MCQ and binary verification results gives a more complete picture of grounded recognition and helps reveal affirmative-response biases that are otherwise hidden by forced-choice evaluation.

D.1 False Acceptance rate scaling with distractor similarity

In our binary ablation, we found that the models significantly struggle to reject distractors. A question is whether this is due to the visual complexity or due to a bias towards answering "yes".

Now we have expanded the experiment by also using the Medium difficulty data split for two models, where the distractor classes are much less similar to the correct classes. For each four-way MCQ, we treat each incorrect option as a separate distractor-level binary decision. We define the false acceptance rate (FAR) as the fraction of distractor options for which the model incorrectly answers "yes." Thus, each MCQ contributes three distractor-level observations.

We measured whether false acceptance increases with target-distractor similarity. For each distractor, the cosine similarity between the CLIP text embedding of the correct attribute description and the CLIP text embedding of the distractor's visual-attribute description was used. We used this as a proxy for the inter-category similarities. We then binned distractors by this similarity score and computed FAR within each bin. The results are shown in Fig. 6.

We found that false acceptance generally

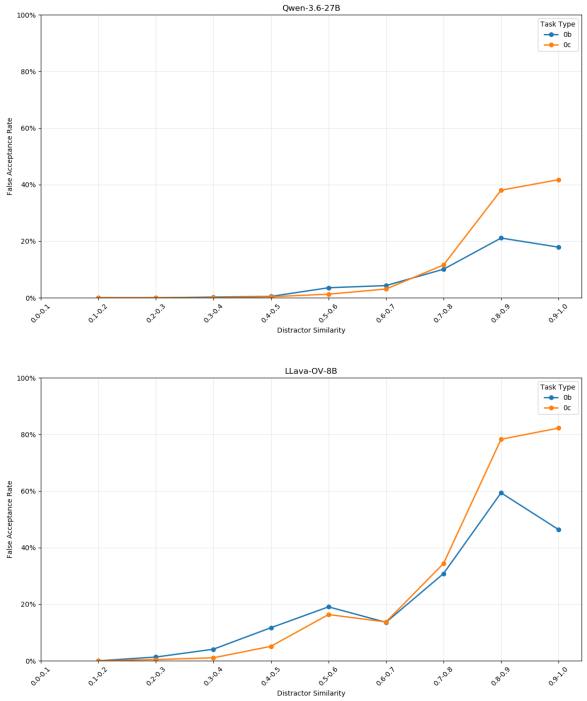


Figure 6: False Acceptance Rate vs. Distractor Similarity for Qwen-3.6-27B and LLava-OV-8B models. As the distractors become more similar to the target class, the models struggle to reject them as incorrect.

increases as distractors become more similar to the correct class. As a brief example, for LLaVA-OV on task 0b (class-name only), distractors with category-similarity around 0.4–0.5 have a FAR of 11.7%, while distractors with similarity above 0.9 have a FAR of 46.37%. For Qwen3.6-27B, the corresponding FAR values are 0.4% and 17.89%. For task 0c (description only), the FAR also increases with similarity. For LLava-OV, it is 5% and 82%. For Qwen27B, it is 0.3% and 41.75%.

Model	Full	Irrelevant Masked	Masked	Cropped
Qwen-3-6-35B-A3B	78.58	76.21	66.79	57.63
Qwen-3-5-4B	63.83	63.17	53.92	33.79
LLaVA-OV-8B	60.83	59.33	54.12	41.92
DeepSeek-VL2-Tiny	50.92	49.58	43.38	43.21
Qwen-3-6-27B	77.08	74.25	63.88	53.83
DeepSeek-VL2-Small	39.29	36.33	31.08	33.54

Table 3: CUB bird class-name recognition accuracy (%) across image perturbations. Models can reliably identify the bird species even when certain body parts are masked out. However, they struggle to identify bird names from only cropped out body parts.

These results indicate that the binary-ablation failures are not merely due to a generic bias toward answering "yes." Instead, models are especially likely to incorrectly accept distractors (false acceptances) that are highly similar to the target. The models are able to successfully reject visually dissimilar distractors.

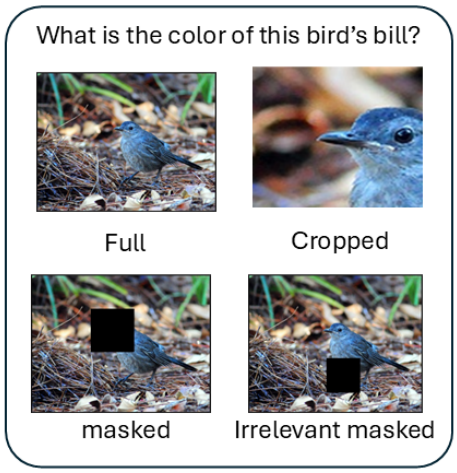


Figure 7: Image conditions for Task 6 (Part-Based Evidence Grounding), where relevant object parts are preserved or masked to evaluate visual evidence grounding.

E Class-only recognition under part perturbations

To further analyze whether the image transformations used in Task 6 preserve enough information for category-level recognition, we conduct an additional class-only ablation. Instead of asking the original part-attribute questions, we reuse the Task 0b (class-only) format MCQ questions and ask the model to identify the bird species from four class-name options. We evaluate this class-only prompt on the same image variants used in Task 6: the full image, the queried-part crop, the target-part-masked image, and the irrelevant-part-masked image (Fig. 7).

This ablation is complementary to the original Task 6 evaluation. While Task 6 tests whether models use localized evidence to answer attribute questions, the class-only ablation tests whether the modified images still support species-level recognition.

The results are shown in table 3. The models maintain high accuracy even after masking the queried part, suggesting that they can still rely on global appearance, unmasked body regions, or class-level priors. These results indicate that the modified Task 6 images preserve different amounts of species-level information. Conversely, accuracy drops sharply on part-only crops as cropped part-only images remove much of the global structure needed for class recognition.

Interestingly, there is a very small gap between full-image and irrelevant-masked image performance. Due to the way we created our dataset, the "tail" parts are usually chosen as the irrelevant region to mask. This suggests that tail regions do not contribute substantially to category recognition. We found that models score on average 61.37% on tail-masked images whereas this is 52.41% on crown-masked images. Conversely, models score on average 29.94% on tail-only crop images whereas this is 46.46% on crown-crop images. Hence, not all parts contribute equally to classification.

Overall, this ablation supports the conclusion that models can often exploit global context and species-level cues even when the queried local evidence is partially removed. Therefore, performance on masked Task 6 images should not be interpreted solely as localized part understanding; it may also reflect residual class-level information in the unmasked image.

Task	Name (N)	Desc (D)	Consistent?	Tests
0a (N+D)	Correct	Correct	Yes	Standard recognition
0b (N-only)	Correct	–	Yes	Name prior reliance
0c (D-only)	–	Correct	Yes	Visual grounding
1a (Name Lure)	Correct	Incorrect	No	Name bias
1b (Visual Lure)	Incorrect	Correct	No	Visual bias
2 (Fictional)	Synthetic	Correct	Yes	Name knowledge removal
3 (Negation)	–	Negated attributes	Yes	Negation reasoning
4 (Composed Img)	Reference + modified	–	Yes	Compositional grounding
5 (Multi-Img)	Correct	Correct	Yes	Multi-image grounding
6 (Part-based)	–	Attribute evidence	Yes	Evidence grounding

Table 4: Summary of task variants. N denotes nominal information (class name) and D denotes visual description.

F Summary of Task variants

We have multiple types of tasks in this benchmark. Each task has its own design and aim. Here we summarize how the tasks are formed and what they analyze in table 4.

G Model Sizes, Licensing and Computational Setup

We used different types of GPUs: A100, L40s, A30 and V100. As there were multiple open weight models, five datasets and different task families, it is very difficult to estimate the to total number of GPU hours.

We evaluate multiple open and closed VLMs spanning different families and scales. We list their parameter sizes here:

- DeepSeekVL2-tiny (Wu et al., 2024): DeepSeek-VL2-Tiny has a total parameter size of 3.37 billion (3.37B), but it only activates 1.0 billion (1.0B) parameters per token thanks to its Mixture-of-Experts (MoE) architecture. It is licensed under the DeepSeek license agreement.
- DeepSeek-VL2-Small (Wu et al., 2024): DeepSeek-VL2-Small has a total of 16 billion (16B) parameters, out of which 2.8 billion (2.8B) parameters are activated during inference using Mixture of Experts. It is licensed under the DeepSeek license agreement.
- Gemma3-12B (Team et al., 2025): This model has 12B parameters. It is licensed under the Gemma 3 license.
- InternVL-2.5-8B (Chen et al., 2025)
- Llama Scout 17B (Meta AI, 2025)" This model has 17B active parameters. It is licensed under the Llama 4 Community License Agreement.
- LLaVA-OneVision-1.5-8B-Instruct (LLaVA-OV) (An et al., 2025): This model has 8B parameters. It is licensed under the Apache 2.0 license.
- Phi-4 (Abouelenin et al., 2025): Phi-4 has 5.6B parameters. It is licensed under the MIT License.
- Qwen 2.5VL 7B (Team, 2025): This model has 7B parameters. It is licensed under the Apache 2.0 license.
- GPT-4.1-mini (OpenAI, 2025): The size is unknown as the model is closed-source.
- GPT-5-mini (Singh et al., 2026): The size is unknown as the model is closed-source.
- Gemini-2.5-flash (Comanici et al., 2025): The size is unknown as the model is closed-source.
- Qwen-3.5-4B (Qwen Team, 2026a): This model has 4B parametes. It is licensed under the Apache 2.0 license.
- Qwen-3.6-27B (Qwen Team, 2026b): This model has 27B parametes. It is licensed under the Apache 2.0 license.
- Qwen3.6-35B-A3B (Qwen Team, 2026c): The Qwen3.6-35B-A3B model features 35 billion total parameters but only 3 billion active parameters per token. It is licensed under the Apache 2.0 license.

G.1 Smaller subset for GPT and Gemini

As mentioned in section 3.1, the GPT and Gemini models were evaluated on the same fixed 80% subset of image instances. The images were selected before inference, and they were in the common full set used to evaluate the other models. Additionally, image instances were kept fixed when

Model	0c(desc.)	1a	1b	2 (F)	0a	0b
Qwen3.5-4B	52.86	33.57	36.79	51.43	76.07	79.29
Qwen3.6-35B-A3B	61.43	65.36	54.64	62.14	85.36	87.14
LLaVA-OV-8B	52.50	33.57	40.00	53.93	77.86	80.36

Table 5: FGVC-Aircraft benchmark — Hard accuracy (%) for 80% subset.

Model	0c(desc.)	1a	1b	2 (F)	0a	0b
Qwen3.5-4B	63.54	34.17	32.92	63.33	78.33	78.96
Qwen3.6-35B-A3B	76.67	41.67	50.42	76.46	87.08	85.00
LLaVA-OV-8B	60.00	38.33	17.92	61.46	75.21	69.58

Table 6: Stanford Dogs benchmark — Hard accuracy (%) for 80% subset.

Model	0c(desc.)	1a	1b	2 (F)	0a	0b
Qwen3.5-4B	62.88	40.05	44.90	65.05	82.14	84.95
Qwen3.6-35B-A3B	75.13	45.79	58.42	77.04	89.41	90.05
LLaVA-OV-8B	58.80	26.79	37.76	61.73	77.55	79.97

Table 7: Stanford Cars benchmark — Hard accuracy (%) for 80% subset.

Model	0c(desc.)	1a	1b	2 (F)	0a	0b
Qwen3.5-4B	84.90	21.53	84.16	85.40	87.13	91.09
Qwen3.6-35B-A3B	88.61	41.09	84.16	89.36	91.09	90.59
LLaVA-OV-8B	80.45	10.89	66.34	79.21	83.66	86.14

Table 8: Food-101 benchmark — Hard accuracy (%) for 80% subset.

Model	0c(desc.)	1a	1b	2 (F)	0a	0b
Qwen3.5-4B	65.75	8.62	46.75	67.25	72.00	65.75
Qwen3.6-35B-A3B	73.88	23.88	44.38	73.62	80.88	76.12
LLaVA-OV-8B	62.88	4.50	26.25	62.50	66.50	60.75

Table 9: CUB benchmark — Hard accuracy (%) for 80% subset.

analyzing across task-variant comparisons within each model.

We believe this subset is representative of the full-set images. We analyzed the performance of three open-weight models (LLaVA-OV-8B, Qwen3.5-4B, and Qwen3.6-35B-A3B) on the exact same 80% subset. These results are given in tables 5 to 9. Across five datasets and six task variants (task 0-2), the mean absolute differences from the full-set accuracies were 0.49%–0.70%, with a maximum difference of 2%. We also compared the trends for 0a>0c, 0b>0c and 2>1b. This led to 45 comparison checks and the trends are the same as the full evaluation set for 44/45 checks. The remaining comparison changed from a 0.60% advantage to an exact tie, with no directional reversal. Similarly, Task 2 remained close to Task 0c with mean absolute gap decreased from 1.22% to 1.14%.

H Full Tables

H.1 Single and Multi-Image Task Tables

Here, we present the full results for all the domains and models for the tasks 0-2 and the multi-image tasks. Additionally, for some of the models, we had run inference on the medium tasks as well. Single image results are in tables 10 to table 19, and the multi-image results are in tables 20 to 24.

Model	n	0c(desc.)	1a	1b	2 (F)	0a	0b
DeepSeek-VL2-Small	350	29.43	23.43	33.14	37.14	38.57	37.14
		[24.90, 34.41]	[19.30, 28.14]	[28.41, 38.23]	[32.24, 42.32]	[33.62, 43.77]	[32.24, 42.32]
DeepSeekVL2-tiny	350	28.57	18.57	18.86	20.00	20.00	68.29
		[24.09, 33.52]	[14.85, 22.98]	[15.11, 23.29]	[16.15, 24.51]	[16.15, 24.51]	[63.24, 72.94]
Gemma3-12B	350	86.29	74.57	76.29	84.57	91.43	92.00
		[82.29, 89.50]	[69.76, 78.85]	[71.56, 80.45]	[80.41, 87.98]	[88.03, 93.93]	[88.68, 94.41]
InternVL-2.5-8B	350	70.29	60.86	68.00	66.86	88.57	94.00
		[65.30, 74.84]	[55.66, 65.83]	[62.94, 72.67]	[61.77, 71.59]	[84.81, 91.49]	[91.00, 96.04]
LLaVA-OV-8B	350	78.57	68.57	67.71	72.57	91.14	92.57
		[73.97, 82.55]	[63.53, 73.21]	[62.64, 72.39]	[67.67, 76.98]	[87.70, 93.69]	[89.34, 94.88]
Llama Scout 17B	350	80.00	88.00	70.00	80.29	91.43	93.14
		[75.49, 83.85]	[84.18, 91.00]	[65.00, 74.56]	[75.80, 84.12]	[88.03, 93.93]	[90.00, 95.35]
Phi-4	350	70.29	39.43	49.14	55.14	64.86	55.43
		[65.30, 74.84]	[34.45, 44.64]	[43.94, 54.36]	[49.90, 60.27]	[59.72, 69.68]	[50.19, 60.55]
Qwen2.5-7B	350	81.71	76.29	48.29	77.43	89.71	94.00
		[77.32, 85.41]	[71.56, 80.45]	[43.10, 53.52]	[72.77, 81.50]	[86.08, 92.47]	[91.00, 96.04]

Table 10: FGVC Aircraft — Medium accuracy (%) with two-sided 95% Wilson score confidence intervals. Each result is reported as accuracy [lower bound, upper bound]. The evaluation size n is the number of questions multiplied by the number of evaluated images per question for the corresponding model.

Model	n	0c(desc.)	1a	1b	2 (F)	0a	0b
DeepSeek-VL2-Small	350	26.57	19.71	29.71	33.71	28.57	32.57
		[22.21, 31.43]	[15.88, 24.20]	[25.16, 34.70]	[28.96, 38.82]	[24.09, 33.52]	[27.87, 37.65]
DeepSeekVL2-tiny	350	32.00	32.86	33.14	32.86	32.86	60.29
		[27.33, 37.06]	[28.15, 37.94]	[28.41, 38.23]	[28.15, 37.94]	[28.15, 37.94]	[55.08, 65.28]
Gemma3-12B	350	57.14	60.57	53.43	60.86	84.00	82.57
		[51.91, 62.22]	[55.36, 65.55]	[48.20, 58.59]	[55.66, 65.83]	[79.79, 87.47]	[78.25, 86.19]
InternVL-2.5-8B	350	37.43	21.14	27.14	40.57	64.57	73.14
		[32.52, 42.61]	[17.19, 25.72]	[22.75, 32.03]	[35.56, 45.79]	[59.43, 69.40]	[68.26, 77.51]
LLaVA-OV-8B	350	52.86	34.00	42.00	52.86	77.43	80.29
		[47.63, 58.03]	[29.23, 39.11]	[36.94, 47.23]	[47.63, 58.03]	[72.77, 81.50]	[75.80, 84.12]
Llama Scout 17B	350	51.71	52.29	38.00	50.57	78.29	82.57
		[46.48, 56.90]	[47.06, 57.47]	[33.07, 43.19]	[45.35, 55.77]	[73.68, 82.29]	[78.25, 86.19]
Phi-4	350	41.43	33.14	43.43	43.14	64.29	61.43
		[36.39, 46.66]	[28.41, 38.23]	[38.34, 48.67]	[38.05, 48.38]	[59.14, 69.13]	[56.23, 66.38]
Qwen2.5-7B	350	51.71	63.43	31.14	47.71	85.43	90.57
		[46.48, 56.90]	[58.26, 68.30]	[26.52, 36.17]	[42.53, 52.94]	[81.35, 88.74]	[87.05, 93.21]
gemini-2.5-flash	280	53.57	75.36	44.29	51.79	77.14	81.79
		[47.72, 59.32]	[69.99, 80.04]	[38.59, 50.15]	[45.95, 57.58]	[71.87, 81.67]	[76.85, 85.87]
GPT-5-mini	280	56.07	56.79	43.21	58.57	86.07	87.50
		[50.21, 61.76]	[50.93, 62.46]	[37.54, 49.07]	[52.72, 64.19]	[81.52, 89.64]	[83.11, 90.87]
Qwen3.5-4B	350	53.71	32.29	37.71	51.71	76.00	80.29
		[48.47, 58.87]	[27.61, 37.36]	[32.79, 42.90]	[46.48, 56.90]	[71.26, 80.18]	[75.80, 84.12]
Qwen3.6-27B	350	55.14	44.57	53.71	57.71	80.57	86.29
		[49.90, 60.27]	[39.45, 49.81]	[48.47, 58.87]	[52.48, 62.77]	[76.10, 84.37]	[82.29, 89.50]
Qwen3.6-35B-A3B	350	60.57	64.29	55.43	60.86	85.71	87.71
		[55.36, 65.55]	[59.14, 69.13]	[50.19, 60.55]	[55.66, 65.83]	[81.66, 88.99]	[83.86, 90.75]

Table 11: FGVC Aircraft — Hard accuracy (%) with two-sided 95% Wilson score confidence intervals. Each result is reported as accuracy [lower bound, upper bound].

Model	n	0c(desc.)	1a	1b	2 (F)	0a	0b
DeepSeek-VL2-Small	600	74.50	46.83	47.83	64.50	70.67	57.00
		[70.86, 77.82]	[42.87, 50.83]	[43.86, 51.83]	[60.59, 68.23]	[66.90, 74.17]	[53.01, 60.90]
DeepSeekVL2-tiny	600	37.17	31.83	32.83	33.50	33.00	83.17
		[33.40, 41.11]	[28.23, 35.66]	[29.19, 36.69]	[29.84, 37.37]	[29.36, 36.86]	[79.97, 85.95]
Gemma3-12B	600	97.33	82.67	83.67	95.50	95.00	87.50
		[95.71, 98.35]	[79.44, 85.49]	[80.50, 86.41]	[93.53, 96.89]	[92.95, 96.48]	[84.61, 89.91]
InternVL-2.5-8B	600	85.83	74.67	62.50	84.00	90.17	81.50
		[82.81, 88.39]	[71.04, 77.99]	[58.56, 66.28]	[80.85, 86.72]	[87.53, 92.30]	[78.20, 84.40]
LLaVA-OV-8B	600	91.50	69.33	48.33	89.50	91.50	82.17
		[89.00, 93.48]	[65.53, 72.89]	[44.35, 52.33]	[86.79, 91.71]	[89.00, 93.48]	[78.91, 85.03]
Llama Scout 17B	600	94.17	86.83	68.00	95.33	97.17	96.00
		[92.00, 95.78]	[83.89, 89.30]	[64.16, 71.61]	[93.33, 96.75]	[95.51, 98.23]	[94.12, 97.30]
Phi-4	600	86.50	39.33	61.67	80.33	81.83	65.00
		[83.53, 89.00]	[35.50, 43.29]	[57.72, 65.47]	[76.96, 83.31]	[78.55, 84.71]	[61.10, 68.71]
Qwen2.5-7B	600	92.33	88.67	54.33	90.33	96.67	95.00
		[89.92, 94.20]	[85.88, 90.96]	[50.33, 58.28]	[87.70, 92.44]	[94.91, 97.83]	[92.95, 96.48]

Table 12: Stanford Dogs — Medium accuracy (%) with two-sided 95% Wilson score confidence intervals. Each result is reported as accuracy [lower bound, upper bound].

Model	n	0c(desc.)	1a	1b	2 (F)	0a	0b
DeepSeek-VL2-Small	600	45.17	21.17	22.00	38.83	45.50	43.67
		[41.23, 49.17]	[18.09, 24.62]	[18.87, 25.49]	[35.01, 42.79]	[41.56, 49.50]	[39.75, 47.67]
DeepSeekVL2-tiny	600	32.00	28.50	29.00	29.50	30.50	68.83
		[28.39, 35.84]	[25.03, 32.24]	[25.51, 32.76]	[25.99, 33.27]	[26.95, 34.30]	[65.01, 72.41]
Gemma3-12B	600	67.33	58.00	47.17	67.67	80.00	79.17
		[63.48, 70.96]	[54.01, 61.89]	[43.21, 51.17]	[63.83, 71.29]	[76.61, 83.01]	[75.74, 82.23]
InternVL-2.5-8B	600	59.17	47.33	27.00	58.33	76.50	74.83
		[55.19, 63.03]	[43.36, 51.33]	[23.60, 30.69]	[54.34, 62.21]	[72.95, 79.72]	[71.21, 78.14]
LLaVA-OV-8B	600	60.50	38.33	18.83	62.00	75.33	70.50
		[56.53, 64.33]	[34.53, 42.28]	[15.90, 22.15]	[58.05, 65.80]	[71.73, 78.61]	[66.73, 74.01]
Llama Scout 17B	600	69.67	55.67	32.67	71.00	82.33	79.67
		[65.88, 73.21]	[51.67, 59.60]	[29.04, 36.52]	[67.24, 74.49]	[79.08, 85.17]	[76.27, 82.70]
Phi-4	600	47.50	22.83	26.00	46.33	59.50	52.33
		[43.53, 51.50]	[19.65, 26.36]	[22.65, 29.65]	[42.38, 50.33]	[55.52, 63.36]	[48.33, 56.30]
Qwen2.5-7B	600	64.83	64.50	23.83	61.83	86.33	86.67
		[60.93, 68.55]	[60.59, 68.23]	[20.59, 27.40]	[57.88, 65.63]	[83.35, 88.85]	[83.72, 89.16]
gemini-2.5-flash	480	65.21	63.54	37.92	67.92	77.71	76.67
		[60.84, 69.34]	[59.14, 67.72]	[33.69, 42.34]	[63.62, 71.94]	[73.77, 81.21]	[72.68, 80.23]
GPT-5-mini	480	71.88	59.17	32.29	71.25	85.62	83.12
		[67.70, 75.72]	[54.72, 63.48]	[28.26, 36.60]	[67.04, 75.12]	[82.20, 88.48]	[79.51, 86.21]
Qwen3.5-4B	600	62.17	34.00	32.83	63.33	77.00	78.50
		[58.22, 65.96]	[30.32, 37.88]	[29.19, 36.69]	[59.40, 67.09]	[73.47, 80.19]	[75.04, 81.60]
Qwen3.6-27B	600	74.50	47.67	42.50	72.33	83.00	84.00
		[70.86, 77.82]	[43.70, 51.67]	[38.60, 46.49]	[68.62, 75.76]	[79.79, 85.79]	[80.85, 86.72]
Qwen3.6-35B-A3B	600	77.50	40.83	50.67	77.33	87.17	86.33
		[73.99, 80.66]	[36.97, 44.81]	[46.68, 54.65]	[73.81, 80.50]	[84.26, 89.61]	[83.35, 88.85]

Table 13: Stanford Dogs — Hard accuracy (%) for single-image tasks with two-sided 95% Wilson score confidence intervals. Each result is reported as accuracy [lower bound, upper bound].

Model	n	0c(desc.)	1a	1b	2 (F)	0a	0b
DeepSeek-VL2-Small	980	81.12	60.20	58.98	75.10	81.53	79.59
		[78.55, 83.45]	[57.10, 63.22]	[55.87, 62.02]	[72.30, 77.71]	[78.98, 83.83]	[76.95, 82.00]
DeepSeekVL2-tiny	980	44.18	20.71	24.08	31.84	36.84	96.12
		[41.10, 47.31]	[18.29, 23.36]	[21.51, 26.85]	[29.00, 34.82]	[33.88, 39.91]	[94.72, 97.16]
Gemma3-12B	980	98.67	95.61	95.31	97.04	99.80	100.00
		[97.74, 99.22]	[94.14, 96.72]	[93.80, 96.47]	[95.78, 97.93]	[99.27, 99.95]	[99.61, 100.00]
InternVL-2.5-8B	980	96.84	87.35	86.53	95.71	99.08	98.78
		[95.55, 97.77]	[85.12, 89.29]	[84.25, 88.53]	[94.25, 96.81]	[98.26, 99.51]	[97.88, 99.30]
LLaVA-OV-8B	980	97.04	65.41	93.06	97.14	99.80	99.90
		[95.78, 97.93]	[62.38, 68.32]	[91.29, 94.49]	[95.90, 98.01]	[99.27, 99.95]	[99.43, 99.98]
Llama Scout 17B	980	97.96	94.08	89.49	98.78	99.80	99.90
		[96.87, 98.68]	[92.42, 95.39]	[87.41, 91.26]	[97.88, 99.30]	[99.27, 99.95]	[99.43, 99.98]
Phi-4	980	94.29	66.63	72.65	88.67	95.20	95.31
		[92.66, 95.58]	[63.62, 69.51]	[69.77, 75.35]	[86.53, 90.51]	[93.68, 96.37]	[93.80, 96.47]
Qwen2.5-7B	980	99.69	98.78	93.98	98.98	99.80	99.80
		[99.10, 99.89]	[97.88, 99.30]	[92.31, 95.30]	[98.13, 99.45]	[99.27, 99.95]	[99.27, 99.95]

Table 14: Stanford Cars — Medium accuracy (%) for single-image tasks with two-sided 95% Wilson score confidence intervals. Each result is reported as accuracy [lower bound, upper bound].

Model	n	0c(desc.)	1a	1b	2 (F)	0a	0b
DeepSeek-VL2-Small	980	36.02	26.12	25.92	35.61	47.86	50.61
		[33.07, 39.07]	[23.47, 28.96]	[23.27, 28.75]	[32.67, 38.66]	[44.75, 50.99]	[47.48, 53.73]
DeepSeekVL2-tiny	980	31.84	24.90	24.39	27.24	29.08	66.02
		[29.00, 34.82]	[22.29, 27.70]	[21.80, 27.18]	[24.55, 30.11]	[26.32, 32.00]	[63.00, 68.92]
Gemma3-12B	980	72.55	66.53	59.69	71.53	88.78	89.08
		[69.67, 75.25]	[63.52, 69.41]	[56.59, 62.72]	[68.62, 74.27]	[86.65, 90.61]	[86.97, 90.88]
InternVL-2.5-8B	980	56.12	30.00	22.55	56.73	71.02	72.65
		[53.00, 59.20]	[27.21, 32.94]	[20.04, 25.27]	[53.61, 59.80]	[68.10, 73.77]	[69.77, 75.35]
LLaVA-OV-8B	980	58.27	26.94	37.76	60.92	76.43	80.10
		[55.16, 61.32]	[24.26, 29.80]	[34.78, 40.84]	[57.83, 63.93]	[73.67, 78.98]	[77.48, 82.48]
Llama Scout 17B	980	67.96	48.78	48.47	69.29	85.41	86.94
		[64.97, 70.81]	[45.66, 51.91]	[45.35, 51.60]	[66.33, 72.10]	[83.06, 87.48]	[84.69, 88.91]
Phi-4	980	50.20	25.92	22.04	43.67	60.00	56.53
		[47.07, 53.32]	[23.27, 28.75]	[19.56, 24.74]	[40.60, 46.79]	[56.90, 63.02]	[53.41, 59.60]
Qwen2.5-7B	980	68.37	62.24	36.84	65.41	89.08	91.12
		[65.39, 71.21]	[59.16, 65.22]	[33.88, 39.91]	[62.38, 68.32]	[86.97, 90.88]	[89.17, 92.74]
gemini-2.5-flash	784	68.62	69.52	42.60	67.73	79.21	79.72
		[65.29, 71.77]	[66.21, 72.64]	[39.18, 46.09]	[64.38, 70.91]	[76.23, 81.90]	[76.76, 82.39]
GPT-5-mini	784	74.49	62.76	42.09	70.66	89.03	87.63
		[71.32, 77.42]	[59.32, 66.07]	[38.68, 45.58]	[67.38, 73.74]	[86.65, 91.03]	[85.14, 89.75]
Qwen3.5-4B	980	64.39	39.18	45.41	65.10	81.53	85.00
		[61.34, 67.33]	[36.17, 42.27]	[42.32, 48.54]	[62.06, 68.02]	[78.98, 83.83]	[82.63, 87.10]
Qwen3.6-27B	980	73.98	44.39	57.04	74.39	85.41	86.43
		[71.14, 76.63]	[41.31, 47.52]	[53.92, 60.11]	[71.57, 77.02]	[83.06, 87.48]	[84.14, 88.43]
Qwen3.6-35B-A3B	980	74.90	46.22	59.59	76.33	88.67	90.31
		[72.09, 77.51]	[43.12, 49.35]	[56.49, 62.62]	[73.57, 78.89]	[86.53, 90.51]	[88.30, 92.01]

Table 15: Stanford Cars — Hard accuracy (%) for single-image tasks with two-sided 95% Wilson score confidence intervals. Each result is reported as accuracy [lower bound, upper bound]. The evaluation size n is the number of questions multiplied by the number of evaluated images per question for the corresponding model.

Model	n	0c(desc.)	1a	1b	2 (F)	0a	0b
DeepSeek-VL2-Small	505	90.30	34.65	66.93	83.96	78.61	54.65
		[87.41, 92.59]	[30.63, 38.90]	[62.71, 70.89]	[80.50, 86.90]	[74.82, 81.96]	[50.29, 58.94]
DeepSeekVL2-tiny	505	63.76	24.75	25.94	38.02	50.10	96.83
		[59.48, 67.83]	[21.19, 28.70]	[22.31, 29.93]	[33.89, 42.33]	[45.75, 54.44]	[94.91, 98.04]
Gemma3-12B	505	97.62	93.47	98.81	96.04	98.42	99.01
		[95.89, 98.63]	[90.97, 95.31]	[97.43, 99.45]	[93.96, 97.42]	[96.91, 99.20]	[97.70, 99.58]
InternVL-2.5-8B	505	98.02	75.05	98.22	98.02	98.61	98.61
		[96.39, 98.92]	[71.10, 78.63]	[96.65, 99.06]	[96.39, 98.92]	[97.16, 99.32]	[97.16, 99.32]
LLaVA-OV-8B	505	98.81	57.82	96.24	98.22	97.82	98.61
		[97.43, 99.45]	[53.47, 62.05]	[94.20, 97.58]	[96.65, 99.06]	[96.14, 98.78]	[97.16, 99.32]
Llama Scout 17B	505	98.02	96.44	96.83	97.62	98.22	98.61
		[96.39, 98.92]	[94.44, 97.74]	[94.91, 98.04]	[95.89, 98.63]	[96.65, 99.06]	[97.16, 99.32]
Phi-4	505	96.44	59.80	96.83	95.64	94.65	92.08
		[94.44, 97.74]	[55.47, 63.99]	[94.91, 98.04]	[93.49, 97.10]	[92.33, 96.30]	[89.39, 94.13]
Qwen2.5-7B	505	98.61	89.70	97.62	98.22	98.81	98.42
		[97.16, 99.32]	[86.74, 92.06]	[95.89, 98.63]	[96.65, 99.06]	[97.43, 99.45]	[96.91, 99.20]

Table 16: Food-101 style benchmark — Medium accuracy (%) for single-image tasks with two-sided 95% Wilson score confidence intervals. Each result is reported as accuracy [lower bound, upper bound]. The evaluation size is $n = 505$ for every model–task combination.

Model	n	0c(desc.)	1a	1b	2 (F)	0a	0b
DeepSeek-VL2-Small	505	64.36	16.63	41.19	53.07	54.46	36.24
		[60.09, 68.41]	[13.64, 20.13]	[36.98, 45.53]	[48.71, 57.38]	[50.10, 58.75]	[32.17, 40.52]
DeepSeekVL2-tiny	505	48.32	25.35	27.92	32.48	35.64	85.74
		[43.99, 52.67]	[21.75, 29.32]	[24.19, 31.99]	[28.54, 36.68]	[31.59, 39.91]	[82.42, 88.52]
Gemma3-12B	505	88.51	68.12	87.13	86.93	91.29	92.48
		[85.43, 91.01]	[63.93, 72.03]	[83.93, 89.77]	[83.71, 89.59]	[88.51, 93.45]	[89.85, 94.47]
InternVL-2.5-8B	505	80.20	7.52	70.89	81.19	84.16	83.76
		[76.50, 83.44]	[5.53, 10.15]	[66.78, 74.68]	[77.55, 84.36]	[80.72, 87.09]	[80.29, 86.72]
LLaVA-OV-8B	505	81.58	11.09	65.74	79.80	83.76	86.34
		[77.96, 84.72]	[8.64, 14.13]	[61.50, 69.75]	[76.08, 83.07]	[80.29, 86.72]	[83.07, 89.06]
Llama Scout 17B	505	87.13	65.74	81.58	86.93	89.11	90.10
		[83.93, 89.77]	[61.50, 69.75]	[77.96, 84.72]	[83.71, 89.59]	[86.09, 91.54]	[87.18, 92.41]
Phi-4	505	74.46	12.28	77.03	72.48	76.83	69.11
		[70.48, 78.07]	[9.70, 15.43]	[73.17, 80.49]	[68.43, 76.19]	[72.96, 80.30]	[64.95, 72.98]
Qwen2.5-7B	505	86.34	45.94	72.67	84.16	90.69	90.30
		[83.07, 89.06]	[41.64, 50.30]	[68.62, 76.37]	[80.72, 87.09]	[87.84, 92.93]	[87.41, 92.59]
gemini-2.5-flash	404	75.74	71.53	74.50	76.49	78.22	80.20
		[71.33, 79.66]	[66.94, 75.71]	[70.03, 78.51]	[72.12, 80.36]	[73.94, 81.97]	[76.04, 83.79]
GPT-5-mini	404	92.82	73.51	88.37	93.07	93.32	92.82
		[89.88, 94.95]	[69.00, 77.58]	[84.88, 91.14]	[90.17, 95.16]	[90.45, 95.37]	[89.88, 94.95]
Qwen3.5-4B	505	84.55	23.37	84.95	86.93	87.52	89.70
		[81.14, 87.44]	[19.89, 27.25]	[81.57, 87.80]	[83.71, 89.59]	[84.35, 90.12]	[86.74, 92.06]
Qwen3.6-27B	505	89.70	61.19	81.78	88.51	91.68	90.69
		[86.74, 92.06]	[56.87, 65.34]	[78.18, 84.90]	[85.43, 91.01]	[88.95, 93.79]	[87.84, 92.93]
Qwen3.6-35B-A3B	505	88.32	40.20	83.76	89.50	91.29	91.49
		[85.23, 90.84]	[36.01, 44.53]	[80.29, 86.72]	[86.52, 91.88]	[88.51, 93.45]	[88.73, 93.62]

Table 17: Food-101 style benchmark — Hard accuracy (%) for single-image tasks with two-sided 95% Wilson score confidence intervals. Each result is reported as accuracy [lower bound, upper bound]. The evaluation size is $n = 505$ for models evaluated with 5 images per question and $n = 404$ for models evaluated with 4 images per question.

Model	n	0c(desc.)	1a	1b	2 (F)	0a	0b
DeepSeek-VL2-Small	1000	83.60	56.50	61.10	78.40	77.80	67.70
		[81.18, 85.77]	[53.41, 59.54]	[58.04, 64.07]	[75.74, 80.84]	[75.12, 80.27]	[64.74, 70.53]
DeepSeekVL2-tiny	1000	51.60	30.00	30.10	35.50	39.90	92.50
		[48.50, 54.69]	[27.24, 32.91]	[27.34, 33.01]	[32.59, 38.52]	[36.91, 42.97]	[90.70, 93.97]
Gemma3-12B	1000	98.60	85.40	95.70	97.60	98.80	99.10
		[97.66, 99.16]	[83.08, 87.45]	[94.26, 96.79]	[96.45, 98.38]	[97.91, 99.31]	[98.30, 99.53]
InternVL-2.5-8B	1000	98.60	66.20	96.00	98.40	99.20	98.40
		[97.66, 99.16]	[63.21, 69.06]	[94.60, 97.05]	[97.42, 99.01]	[98.43, 99.59]	[97.42, 99.01]
LLaVA-OV-8B	1000	98.90	32.50	93.50	99.30	99.60	98.90
		[98.04, 99.38]	[29.67, 35.47]	[91.80, 94.87]	[98.56, 99.66]	[98.98, 99.84]	[98.04, 99.38]
Llama Scout 17B	1000	98.20	86.10	96.80	98.30	99.10	99.40
		[97.17, 98.86]	[83.82, 88.11]	[95.52, 97.72]	[97.29, 98.94]	[98.30, 99.53]	[98.70, 99.72]
Phi-4	1000	92.30	42.50	83.50	89.90	92.40	78.80
		[90.48, 93.80]	[39.47, 45.59]	[81.07, 85.67]	[87.88, 91.62]	[90.59, 93.89]	[76.16, 81.22]
Qwen2.5-7B	1000	99.00	91.90	94.40	98.70	99.50	99.50
		[98.17, 99.46]	[90.04, 93.44]	[92.80, 95.66]	[97.79, 99.24]	[98.83, 99.79]	[98.83, 99.79]

Table 18: CUB (Birds) — Medium accuracy (%) for single-image tasks with two-sided 95% Wilson score confidence intervals. Each result is reported as accuracy [lower bound, upper bound].

Model	n	0c(desc.)	1a	1b	2 (F)	0a	0b
DeepSeek-VL2-Small	1000	38.30	20.50	22.80	32.70	44.90	36.00
		[35.34, 41.35]	[18.11, 23.11]	[20.31, 25.50]	[29.86, 35.67]	[41.84, 48.00]	[33.08, 39.02]
DeepSeekVL2-tiny	1000	30.20	25.80	25.70	27.30	28.80	53.40
		[27.43, 33.12]	[23.18, 28.60]	[23.09, 28.50]	[24.63, 30.14]	[26.08, 31.68]	[50.30, 56.47]
Gemma3-12B	1000	61.60	25.20	35.50	61.70	73.30	70.40
		[58.55, 64.56]	[22.61, 27.98]	[32.59, 38.52]	[58.65, 64.66]	[70.47, 75.95]	[67.50, 73.15]
InternVL-2.5-8B	1000	59.00	7.10	25.00	59.90	63.70	61.30
		[55.92, 62.01]	[5.67, 8.86]	[22.42, 27.78]	[56.83, 62.89]	[60.67, 66.62]	[58.24, 64.27]
LLaVA-OV-8B	1000	63.00	4.40	25.60	62.30	66.50	61.40
		[59.96, 65.94]	[3.29, 5.86]	[22.99, 28.39]	[59.25, 65.25]	[63.52, 69.36]	[58.34, 64.37]
Llama Scout 17B	1000	65.40	24.40	42.30	65.90	74.80	70.60
		[62.40, 68.28]	[21.84, 27.16]	[39.27, 45.39]	[62.91, 68.77]	[72.02, 77.39]	[67.70, 73.34]
Phi-4	1000	47.70	13.90	19.60	44.70	45.50	34.30
		[44.62, 50.80]	[11.89, 16.18]	[17.26, 22.17]	[41.64, 47.80]	[42.44, 48.60]	[31.42, 37.30]
Qwen2.5-7B	1000	62.70	24.10	22.20	62.40	78.10	78.00
		[59.66, 65.64]	[21.55, 26.85]	[19.73, 24.88]	[59.36, 65.35]	[75.43, 80.55]	[75.33, 80.46]
gemini-2.5-flash	800	57.00	61.75	30.50	57.75	72.38	77.50
		[53.54, 60.39]	[58.33, 65.05]	[27.41, 33.78]	[54.30, 61.13]	[69.18, 75.37]	[74.48, 80.26]
GPT-5-mini	800	62.88	33.50	22.38	63.88	82.25	74.00
		[59.48, 66.16]	[30.31, 36.84]	[19.63, 25.40]	[60.49, 67.14]	[79.45, 84.74]	[70.85, 76.92]
Qwen3.5-4B	1000	65.00	9.10	47.00	67.20	70.50	65.60
		[61.99, 67.89]	[7.47, 11.04]	[43.92, 50.10]	[64.23, 70.04]	[67.60, 73.24]	[62.60, 68.48]
Qwen3.6-27B	1000	74.10	34.40	47.40	74.00	81.20	76.40
		[71.30, 76.72]	[31.52, 37.40]	[44.32, 50.50]	[71.19, 76.62]	[78.66, 83.50]	[73.67, 78.93]
Qwen3.6-35B-A3B	1000	73.50	24.10	44.80	73.60	81.30	76.80
		[70.68, 76.14]	[21.55, 26.85]	[41.74, 47.90]	[70.78, 76.24]	[78.77, 83.60]	[74.08, 79.31]

Table 19: CUB (Birds) — Hard accuracy (%) for single-image tasks with two-sided 95% Wilson score confidence intervals. Each result is reported as accuracy [lower bound, upper bound]. The evaluation size n is the number of questions multiplied by the number of evaluated images per question for the corresponding model.

Model	Medium					Hard				
	0c	0c(img)	0a	0a(img)	Task 4	0c	0c(img)	0a	0a(img)	Task 4
DeepSeek-VL2-Small	29.43	14.29	38.57	16.57	27.05	26.57	26.00	28.57	24.57	24.57
DeepSeekVL2-tiny	28.57	20.00	20.00	24.57	25.71	32.00	33.14	32.86	34.86	27.62
Gemma3-12B	86.29	62.57	91.43	64.29	40.10	57.14	53.14	84.00	62.86	31.14
InternVL-2.5-8B	70.29	44.86	88.57	46.00	35.62	37.43	28.29	64.57	35.71	24.76
LLaVA-OV-8B	78.57	28.29	91.14	28.00	26.67	52.86	42.29	77.43	46.29	26.10
Llama Scout 17B	80.00	70.86	91.43	86.00	65.24	51.71	52.00	78.29	69.71	44.10
Phi-4	70.29	44.57	64.86	30.86	33.05	41.43	35.14	64.29	35.14	26.95
Qwen2.5-7B	81.71	76.29	89.71	95.43	63.05	51.71	62.57	85.43	88.00	36.10
Qwen3-4B	–	–	–	–	–	53.71	21.71	76.00	28.86	50.48
Qwen3-27B	–	–	–	–	–	55.14	63.14	80.57	77.71	55.81
Qwen3-35B-A3B	–	–	–	–	–	60.57	66.86	85.71	82.86	53.71

Table 20: FGVC Aircraft — accuracy (%) with single-image and multi-image results (0c, 0a only), plus Task 4.

Model	Medium					Hard				
	0c	0c(img)	0a	0a(img)	Task 4	0c	0c(img)	0a	0a(img)	Task 4
DeepSeek-VL2-Small	81.12	22.86	81.53	18.27	36.29	36.02	23.98	47.86	22.35	25.99
DeepSeekVL2-tiny	44.18	20.92	36.84	26.33	25.75	31.84	25.20	29.08	27.45	25.34
Gemma3-12B	98.67	81.94	99.80	76.63	44.52	72.55	52.35	88.78	60.00	38.91
InternVL-2.5-8B	96.84	48.78	99.08	50.82	46.77	56.12	30.31	71.02	33.06	32.07
LLaVA-OV-8B	97.04	41.84	99.80	45.92	30.17	58.27	34.59	76.43	42.86	27.72
Llama Scout 17B	97.96	96.73	99.80	99.29	60.88	67.96	66.63	85.41	78.88	35.99
Phi-4	94.29	72.55	95.20	45.61	46.46	50.20	41.94	60.00	33.78	31.02
Qwen2.5-7B	99.69	96.63	99.80	99.29	69.52	68.37	64.08	89.08	81.73	54.39
Qwen3-4B	–	–	–	–	–	64.39	52.45	81.53	72.96	59.42
Qwen3-27B	–	–	–	–	–	73.98	78.98	85.41	87.65	61.26
Qwen3-35B-A3B	–	–	–	–	–	74.90	78.98	88.67	88.06	66.19

Table 21: Stanford Cars — accuracy (%) with single-image and multi-image results (0c, 0a only), plus Task 4.

Model	Medium					Hard				
	0c	0c(img)	0a	0a(img)	Task 4	0c	0c(img)	0a	0a(img)	Task 4
DeepSeek-VL2-Small	74.50	30.50	70.67	31.67	29.22	45.17	26.67	45.50	27.50	19.89
DeepSeekVL2-tiny	37.17	36.17	33.00	38.50	25.22	32.00	29.17	30.50	27.50	25.28
Gemma3-12B	97.33	59.33	95.00	64.50	38.89	67.33	44.67	80.00	48.67	35.89
InternVL-2.5-8B	85.83	61.33	90.17	69.67	37.17	59.17	44.67	76.50	48.83	29.89
LLaVA-OV-8B	91.50	43.67	91.50	60.17	32.33	60.50	38.33	75.33	43.17	27.50
Llama Scout 17B	94.17	87.50	97.17	94.50	82.94	69.67	60.67	82.33	74.50	60.06
Phi-4	86.50	50.33	81.83	36.83	38.11	47.50	41.33	59.50	33.83	30.33
Qwen2.5-7B	92.33	78.33	96.67	95.83	69.94	64.83	49.17	86.33	73.83	50.61
Qwen3-4B	–	–	–	–	–	62.17	64.33	77.00	73.00	62.11
Qwen3-27B	–	–	–	–	–	74.50	73.17	83.00	82.00	63.39
Qwen3-35B-A3B	–	–	–	–	–	77.50	71.83	87.17	83.33	66.28

Table 22: Stanford Dogs — accuracy (%) with single-image and multi-image results (0c, 0a only), plus Task 4.

Model	Medium					Hard				
	0c	0c(img)	0a	0a(img)	task04	0c	0c(img)	0a	0a(img)	task04
DeepSeek-VL2-Small	90.30	25.74	78.61	29.11	48.05	64.36	28.12	54.46	27.72	32.34
DeepSeekVL2-tiny	63.76	43.56	50.10	32.08	25.41	48.32	36.24	35.64	29.70	25.15
Gemma3-12B	97.62	73.07	98.42	74.06	46.40	88.51	56.83	91.29	59.41	44.16
InternVL-2.5-8B	98.02	52.87	98.61	51.49	42.24	80.20	42.97	84.16	45.74	39.08
LLaVA-OV-8B	98.81	55.64	97.82	62.38	33.93	81.58	41.58	83.76	43.96	28.91
Llama Scout 17B	98.02	98.81	98.22	99.60	84.82	87.13	82.77	89.11	86.53	71.49
Phi-4	96.44	79.80	94.65	68.71	50.56	74.46	58.22	76.83	46.14	39.87
Qwen2.5-7B	98.61	98.61	98.81	98.81	81.39	86.34	80.79	90.69	86.34	64.16
Qwen3-4B	–	–	–	–	–	84.55	81.58	87.52	84.55	74.65
Qwen3-27B	–	–	–	–	–	89.70	90.50	91.68	92.48	78.09
Qwen3-35B-A3B	–	–	–	–	–	88.32	90.50	91.29	92.87	79.60

Table 23: Food-101 style benchmark — accuracy (%) for Medium and Hard splits. Results include description-based prompts, image-augmented prompts, and Task04 evaluation.

Model	Medium					Hard				
	0c	0c(img)	0a	0a(img)	task04	0c	0c(img)	0a	0a(img)	task04
DeepSeek-VL2-Small	83.60	29.60	77.80	30.30	37.13	38.30	23.80	44.90	22.90	25.30
DeepSeekVL2-tiny	51.60	38.20	39.90	32.90	28.40	30.20	27.50	28.80	29.50	24.40
Gemma3-12B	98.60	82.40	98.80	80.80	52.63	61.60	41.90	73.30	45.00	35.00
InternVL-2.5-8B	98.60	74.50	99.20	77.70	42.70	59.00	35.90	63.70	38.40	33.33
LLaVA-OV-8B	98.90	65.10	99.60	63.20	32.80	63.00	28.10	66.50	29.70	29.33
Llama Scout 17B	98.20	97.70	99.10	99.40	78.27	65.40	59.00	74.80	70.70	56.33
Phi-4	92.30	87.70	92.40	65.40	48.33	47.70	40.50	45.50	32.90	31.87
Qwen2.5-7B	99.00	96.30	99.50	99.30	71.30	62.70	54.80	78.10	68.70	52.13
Qwen3-4B	–	–	–	–	–	65.00	61.80	70.50	68.70	58.03
Qwen3-27B	–	–	–	–	–	74.10	69.80	81.20	79.60	62.93
Qwen3-35B-A3B	–	–	–	–	–	73.50	71.60	81.30	83.60	61.77

Table 24: CUB (Birds) benchmark — accuracy (%) for Medium and Hard splits using description prompts, image-augmented prompts, and Task04 evaluation.

H.2 Full tables for part-based results

In Table 1, we presented the results for multiple models on the part-based evidence grounding task. We had omitted the results for the irrelevant masked task for brevity. Here, we present the full results in table 25.

Without Z								
	q1				q2			
Model	F	C	M	I	F	C	M	I
DeepSeek-VL2-Small	35.98	75.30	26.81	31.22	34.19	40.85	28.32	35.77
DeepSeek-VL2-Tiny	64.94	65.67	63.37	65.07	50.40	52.06	44.29	49.98
Gemma3-12B	66.81	67.09	52.52	64.18	57.58	56.56	48.86	57.71
InternVL-2.5-8B	65.35	66.89	57.35	65.14	56.07	55.94	51.40	56.22
LLaVA-OV-8B	73.73	72.06	68.62	74.25	63.65	59.71	57.92	63.90
Phi-4	67.99	65.81	64.77	67.17	52.43	49.47	46.63	51.04
Qwen2.5-7B	69.39	68.70	56.38	68.27	54.71	52.17	49.03	54.43
GPT-5-mini	78.42	72.33	66.00	76.58	67.25	58.83	54.92	64.42
Qwen3.6-27B	75.97	75.14	52.24	75.80	68.17	62.70	59.05	67.57
Qwen3.6-35B-A3B	76.61	72.28	66.84	76.18	68.14	63.55	59.48	67.91
Qwen3.5-4B	71.79	65.66	62.54	71.41	60.98	51.09	53.99	60.39
With Z								
	q1				q2			
Model	F	C	M	I	F	C	M	I
DeepSeek-VL2-Small	30.15	15.46	25.36	32.98	26.37	12.63	15.23	25.41
DeepSeek-VL2-Tiny	67.71	67.77	64.93	67.36	45.21	45.60	38.92	47.83
Gemma3-12B	30.71	34.39	3.06	26.86	45.73	45.93	6.09	40.13
InternVL-2.5-8B	68.03	65.32	34.68	63.82	45.59	49.42	20.56	43.16
LLaVA-OV-8B	57.30	40.61	1.16	35.61	56.00	46.00	2.74	44.65
Phi-4	67.18	65.15	57.04	67.67	49.19	43.95	27.73	46.99
Qwen2.5-7B	68.63	66.36	30.20	65.71	50.22	45.09	5.63	47.80
GPT-5-mini	78.33	71.50	26.42	76.67	67.00	59.00	26.58	64.42
Qwen3.6-27B	76.20	70.34	4.66	74.79	66.67	60.51	2.26	63.10
Qwen3.6-35B-A3B	76.70	72.74	6.99	75.45	67.03	63.37	3.14	65.39
Qwen3.5-4B	71.53	60.02	34.24	69.28	59.93	43.03	6.50	53.55

Table 25: Part-based attribute reasoning accuracy (%) on CUB for q1 (attribute verification) and q2 (multiple-choice attribute classification). F = full image, C = part crop, M = target-part masked, and I = irrelevant-part masked. Results are shown without and with the unanswerable option (Z). Due to resource limitations, the GPT model was evaluated on a smaller subset.