

# Benchmarking and Mitigating MCQA Selection Bias of Large Vision-Language Models

Md. Atabuzzaman    Ali Asgarov\*    Chris Thomas

Department of Computer Science

Virginia Tech

{atabuzzaman, aliasgarov, christhomas}@vt.edu

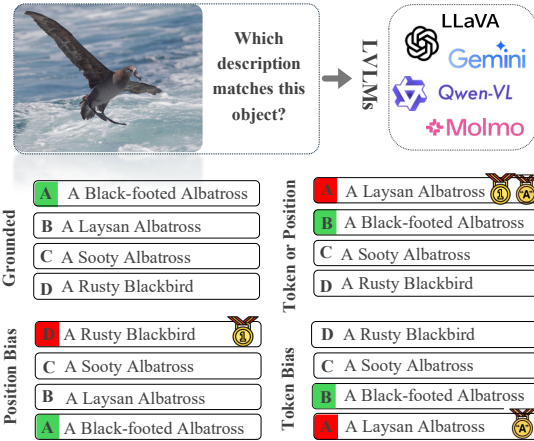
## Abstract

Large Vision-Language Models (LVLMs) have achieved strong performance on vision-language tasks, particularly Visual Question Answering (VQA). While prior work has explored unimodal biases in VQA, the problem of selection bias in Multiple-Choice Question Answering (MCQA), where models may favor specific option tokens (e.g., "A") or positions, remains underexplored. In this paper, we investigate both the presence and nature of selection bias in LVLMs through fine-grained MCQA benchmarks spanning easy, medium, and hard difficulty levels, defined by the semantic similarity of the options. We further propose an inference-time logit-level debiasing method that estimates an ensemble bias vector from general and contextual prompts and applies confidence-adaptive corrections to the model's output. Our method mitigates bias without re-training and is compatible with frozen LVLMs. Extensive experiments across several state-of-the-art models reveal consistent selection biases that intensify with task difficulty, and show that our mitigation approach significantly reduces bias while improving accuracy in challenging settings. This work offers new insights into the limitations of LVLMs in MCQA and presents a practical approach to improve their robustness in fine-grained visual reasoning. Datasets and code are available at: <https://github.com/Atabuzzaman/Selection-Bias-of-LVLMs>

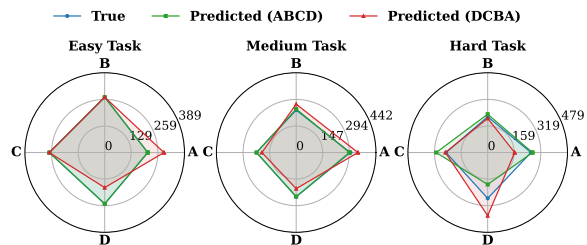
## 1 Introduction

Large Vision-Language Models (LVLMs) (Liu et al., 2024a,b; Li et al., 2023; Dai et al., 2023; Deitke et al., 2024; Team, 2025; Chen et al., 2024c; OpenAI, 2024; Chen et al., 2025) have achieved impressive performance across a wide range of multimodal tasks, including visual question answering (VQA), image captioning, and visual reasoning.

\*The proposed bias mitigation method was fully created and implemented by this author.



(a) Illustration of selection bias (positional and token identity) in LVLm predictions (red) for visual multiple-choice question answering. Answer preferences change with option order and token labels. Correct answers are in green.




(b) As task difficulty increases, Qwen2.5-VL-3B-Instruct exhibits stronger selection biases favoring option ID "A" in easy tasks (**token bias**) and first-position options "A" and "D" in "ABCD" and "DCBA" orderings, respectively, in hard tasks (**positional bias**).

Figure 1: (a) Top: Visual illustration of selection bias in LVLMs. (b) Bottom: Amplification of token and positional biases in the Qwen2.5-VL-3B-Instruct model across increasing task difficulty.

These models exhibit strong zero-shot generalization, attributed to pretraining on large-scale vision-language corpora and subsequent instruction tuning. However, despite their overall success, recent studies have revealed that LVLMs, like their text-only counterparts, are prone to various forms of bias that compromise the fairness, interpretability, and robustness of their outputs (Adila et al., 2024).

One such underexplored phenomenon is selection bias in MCQA. Unlike open-ended VQA for-

**Yellow-headed Blackbird**



Which description matches best with the image?

Easy

**B.** This image shows a **Yellow-headed Blackbird** bird which has a black body, bright yellow head...

Medium

**A.** This image shows a **Yellow-headed Blackbird** bird which has a black body, bright yellow head...

Hard

**D.** This image shows a **Yellow-headed Blackbird** bird which has a black body, bright yellow head...








Figure 2: Examples from our fine-grained visual multiple-choice question answering benchmark for the "Yellow-headed Blackbird" class across three difficulty levels: easy, medium, and hard. Each example presents a multiple-choice question requiring the model to match an image with the most appropriate textual description. The easy task includes distractors (incorrect options) from different domains (e.g., vehicles, food), making the correct choice easily distinguishable. The medium task increases difficulty by using distractors from the same domain (i.e., birds) with less similar visual characteristics. The hard task presents the most visually similar bird species (e.g., blackbirds with subtle distinctions), demanding fine-grained reasoning. This structured difficulty progression enables systematic evaluation of LVLMs' reasoning capabilities and their susceptibility to selection biases, especially when class names are explicitly included or excluded in the options.

mats, MCQA requires models to select one option among predefined choices (e.g., A/B/C/D), introducing the possibility of preference for certain option positions or tokens. Similar forms of bias, such as position bias and token prior bias, have been documented in large language models (LLMs) (Pezeshkpour and Hruschka, 2024; Zheng et al., 2023), but their manifestation in LVLMs remains relatively unexamined. Our preliminary findings suggest that LVLMs exhibit consistent preferences for specific options (e.g., choosing "A" or "D" disproportionately), especially in scenarios where answer candidates are semantically or visually similar (Figure 1). This can result in unstable or unreliable predictions that are influenced more by formatting than content (Adila et al., 2024; Zong et al., 2024).

In this paper, we investigate the presence and nature of MCQA selection bias in LVLMs. We identify multiple sources of bias, including option position bias and token-level prior heuristics that models rely on instead of grounded visual reasoning. To study these phenomena in depth, we introduce new benchmark datasets designed specifically to evaluate LVLMs' MCQA selection behavior. Our dataset is constructed from fine-grained visual classification tasks and includes three levels of difficulty (Easy, Medium, and Hard) based on the semantic similarity between the correct option

and distractor (or incorrect) options (Figure 2). We also incorporate variations with and without class names in the options to test the LVLm's reliance on surface-level cues and prior domain knowledge.

To mitigate the selection bias, we propose an inference-time logit correction mechanism that adjusts the model's output distribution over MCQ options based on an empirically estimated bias vector. Our method does not require retraining and is fully compatible with frozen pretrained LVLMs. It constructs an ensemble bias vector from both general prompts (capturing structural biases) and contextual prompts (reflecting task-specific tendencies), and adaptively corrects the model's logits at inference based on prediction confidence. This approach counteracts option-token and positional biases while preserving the model's ability to reason over visual and semantic content.

Through extensive experiments on several state-of-the-art (SOTA) LVLMs, we show that our fine-grained MCQA benchmarks reveal consistent and intensifying selection biases, particularly when visual evidence is inconclusive and options are fine-grained. We further demonstrate that our logit-based debiasing method improves model accuracy in challenging settings, enhances answer consistency under option reordering, and reduces reliance on spurious token and position priors.

Our main contributions are as follows:

- We propose benchmark datasets to study the selection bias of LVLMs across three difficulty level tasks: easy, medium, and hard.
- Our datasets feature options with and without class names, revealing the nature of selection bias at each difficulty level and allowing investigation of LVLM behavior with and without prior domain knowledge.
- We propose an inference-time logit debiasing method that mitigates selection bias by correcting biased option-token distributions using an ensemble bias vector and confidence-adaptive scaling.

## 2 Related Work

Several studies have explored the selection bias of LLMs (Robinson and Wingate, 2023; Zheng et al., 2023, 2024; Pezeshkpour and Hruschka, 2024; Xue et al., 2024; Shi et al., 2024; Balepur et al., 2025; Wang et al., 2025). Robinson and Wingate (2023) showed that LLMs behave differently when prompted with option IDs compared to cloze-style prompts without IDs. They also evaluated the effect of varying the position of option IDs on model performance. Zheng et al. (2023) & Wang et al. (2024a) found that GPT-4 tends to favor the first-presented answers, potentially leading to unfair evaluation outcomes. Xue et al. (2024) argued that selection bias arises from the model’s inability to effectively associate option IDs with the corresponding option text. Pezeshkpour and Hruschka (2024) observed that LLMs are sensitive to changes in option order in MCQs and attributed this to positional bias and uncertainty. Li et al. (2024) highlighted selection bias in knowledge-intensive scenarios where long-form generation (LFG) is required. Zheng et al. (2024) introduced PriDe and demonstrated that removing option IDs shifts the main source of bias to the model’s prior token bias, which can be mitigated through targeted debiasing. Yang et al. (2024) addressed bias by removing neurons responsible for biased behavior. Zhou et al. (2024) proposed UniBias, an inference-only approach that identifies and eliminates biased feed-forward network (FFN) vectors and attention heads. Wei et al. (2024) quantified the effects of token and option order on selection bias and mitigated them through weight and probability calibration. Guda et al. (2025) introduced a majority-voting method that reduces computational overhead while main-

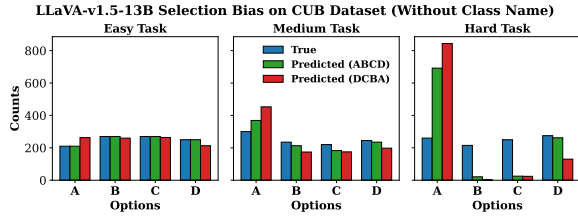
taining effective bias mitigation. Finally, Yang et al. (2025b) proposed a causal debiasing technique that steers key component activations toward unbiased directions, applying stronger interventions to components with higher causal influence.

Recently, researchers have addressed various types of biases in Large Vision-Language Models (LVLMs) using techniques such as data augmentation (Gokhale et al., 2020), model editing (Cheng et al., 2023; Wang et al., 2024b), and post-processing of outputs (Wang et al., 2021; Zhang et al., 2024). Zhang et al. (2024) proposed two strategies to mitigate language prior bias in classification tasks through output probability calibration. Chen et al. (2024a) assessed and mitigated LVLM bias in the VQA task by intervening in both questions and images using causal graphs. Chen et al. (2024b) proposed DCVC, a trainable output calibration network with virtual counterfactual augmentation, to reduce language bias in social intelligence QA. Tan et al. (2024) found that multimodal LLMs favor content at the beginning and end of contexts, and improved inference by strategically placing key elements. Zong et al. (2024) identified permutation vulnerabilities in LVLM-based MCQs such as position bias and weak option-content links and introduced mitigation techniques like majority voting, confidence voting, and context calibration. Adila et al. (2024) proposed inference-type activation steering to reduce selection bias in LVLMs. However, effectively mitigating MCQA selection bias using pretrained LVLMs while preserving model capabilities and performance remains largely unexplored. In particular, how such biases evolve across tasks of increasing difficulty, from easy to medium to hard, has not been systematically studied or addressed.

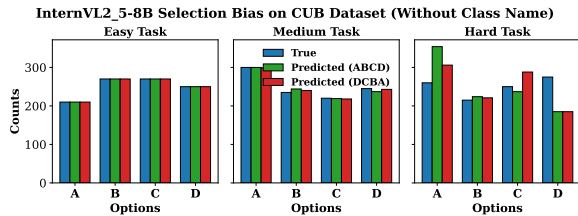
## 3 Dataset Curation

To evaluate the selection bias of LVLMs in the MCQA setting, we curate benchmark datasets that systematically test models under varying degrees of semantic similarity and domain familiarity. Our goal is to assess how effectively these models can select correct descriptive class text when presented with highly similar distractors (incorrect options), and to what extent their predictions are influenced by spurious correlations or prior knowledge.

Each MCQ in our dataset includes one correct class description and three incorrect class descriptions (distractors) as options, selected based on



(a) The LLaVA-v1.5-13B model shows balanced selection in easy tasks but develops strong token bias in hard tasks, with dramatic preference for option ID "A" (reaching nearly 3x the true frequency) when difficulty increases.



(b) The InternVL2\_5-8B model demonstrates balanced behavior across difficulty levels, with predictions closely matching the true distribution in easy and medium tasks. In hard tasks, it shows moderate token bias toward "A" but maintains better distribution consistency between ABCD and DCBA orderings.

Figure 3: Selection bias comparison across two LVLMs on the CUB dataset under the "without class name" setting, organized by increasing task difficulty (easy, medium, hard). Each plot shows distributions for ground truth (True) and model predictions under two option orderings: standard (ABCD) and reversed (DCBA). The comparison reveals how position and token biases emerge and intensify with task difficulty, with varying patterns across architectures.

their cosine similarity to the correct description using CLIP’s text encoder (Radford et al., 2021). By controlling this similarity, we categorize each question into one of three difficulty levels:

**Easy:** Distractors are drawn from unrelated or semantically dissimilar classes. The differences between the correct option and the distractors are clear, even without domain-specific knowledge.

**Medium:** Distractors are from the same domain but moderately different from the target class. These options require more reasoning to eliminate.

**Hard:** Distractors are highly similar to the target class in terms of textual and semantic content, making them challenging to differentiate. These examples often require fine-grained understanding and detailed visual-textual alignment.

To further analyze how LVLm bias behavior changes when models can or cannot rely on domain knowledge, we create two versions of each MCQ:

**With Class Name:** Class names are explicitly mentioned in the options. This version tests whether the model can leverage prior domain knowledge to reduce selection bias and improve prediction accuracy.

**Without Class Name:** Class names are removed from the options, forcing the model to rely on fine-grained visual grounding and descriptive reasoning rather than recalling known labels, potentially increasing susceptibility to selection bias.

To ensure diverse coverage across semantic categories, we construct our benchmark using six fine-grained classification datasets: CUB-200-2011 (Wah et al., 2011) (200 bird species), Stanford Dogs (Khosla et al., 2011) (120 dog breeds), FGVC Aircraft (Maji et al., 2013) (70/102 aircraft variants), Stanford Cars (Krause et al., 2013) (196 car models), Food-101 (Bossard et al., 2014) (101 food categories; referred to as Food-101 or Food throughout this paper), and iNaturalist-2021 (Di Cecco et al., 2021) (9962/10,000 species-level classes). Class descriptions are collected from Atabuzzaman et al. (2025); Kim and Ji (2024).

Using these class descriptions, for each image-class pair, we construct three difficulty-specific MCQs (easy, medium, hard) with two variants each, one with the class name included and one without. This results in six unique MCQs per image. For example, a dataset with 200 (e.g., CUB) classes would yield  $200 \times 3 \times 2 = 1,200$  MCQs. **In total, our dataset contains 63,894 MCQs across 10,649 diverse classes.**

To validate our difficulty categorization, we compute the average and standard deviation of cosine similarity scores between correct and incorrect options across all difficulty levels. As expected, easy tasks show low similarity, medium tasks show moderate similarity, and hard tasks contain high-similarity incorrect options. Table 5 (Appendix A.1) reports these statistics, confirming that our dataset reliably separates difficulty levels based on semantic similarity.

To enable effective detection of positional bias, we balance the position of the correct answer (A–D) across MCQs. Table 6 (Appendix A.1) shows that all datasets maintain balanced correct answer position distributions across both settings (with and without class names), ensuring that any observed positional preferences reflect model bias rather than dataset imbalance.

## 4 Selection Bias of LVLms

In this section, we investigate the presence and nature of selection bias in LVLms when applied to MCQA tasks. Our analysis focuses on three leading models: LLaVA-v1.5-13B (Liu et al.,

2024b), InternVL2\_5-8B (Chen et al., 2024c), and Qwen2.5-VL-3B-Instruct (Team, 2025). We identify consistent and model-specific selection biases that emerge during visual reasoning, particularly as task difficulty increases.

To disentangle token identity bias from positional bias, we design a comparative evaluation using both standard ("ABCD") and reversed ("DCBA") option orderings. As shown in Figures 1 and 3, all models exhibit relatively balanced option selection in easy tasks, with predicted distributions closely matching the ground-truth answer distribution. However, as difficulty increases to medium and hard levels, distinct and intensified bias patterns emerge. These patterns suggest that models begin to rely more heavily on heuristic behaviors, such as defaulting to specific token IDs—under semantic ambiguity, where fine-grained descriptions make it harder to distinguish between closely related classes.

**Token Bias Intensifies with Task Difficulty in LLaVA-v1.5-13B.** Figure 3a shows that LLaVA-v1.5-13B exhibits strong token bias in hard tasks, with the "A" option being selected nearly three times more often than expected. Crucially, this preference persists regardless of the option’s actual position, indicating that the model defaults to the "A" token ID when uncertain. This behavior suggests the presence of a learned token-based prior that increasingly influences predictions as the model struggles with fine-grained visual-textual alignment. While the bias is negligible in easy tasks, it becomes dominant in hard tasks.

**InternVL2\_5-8B Exhibits More Balanced Token Selection Under Difficulty.** Figure 3b demonstrates that InternVL2\_5-8B maintains more balanced token selection behavior across all levels of difficulty. Even in hard tasks where other models exhibit strong biases, InternVL2\_5-8B’s predictions remain relatively aligned with the true answer distribution. Moreover, the model shows consistent behavior across both standard and reversed option formats, suggesting more robust reasoning and less susceptibility to token or positional artifacts. Nevertheless, InternVL still exhibits slight token ("A") bias amplification under difficulty, indicating that even stronger models fall back on heuristics when tasks become challenging.

**Qwen2.5-VL-3B-Instruct Reveals Complex Interaction Between Token and Positional Bias.** Figure 1b reveals that Qwen2.5-VL-3B-Instruct presents the most intricate bias behavior. The

model exhibits a strong interaction between token identity and option position, particularly in hard tasks. In easy and medium tasks, Qwen2.5-VL-3B-Instruct shows mild bias toward "A" even when it appears last. However, in hard tasks, the bias pattern changes significantly: option "C" in the ABCD format becomes dominant, while option "D" when placed first in the DCBA format receives excessive selection. This suggests a conflation of positional and token-specific preferences. The interaction indicates the presence of multiple, competing biases within the model’s decision-making process. The severe collapse to specific options under fine-grained semantic ambiguity implies heightened sensitivity to both token identity and position when semantic distinctions between options are subtle.

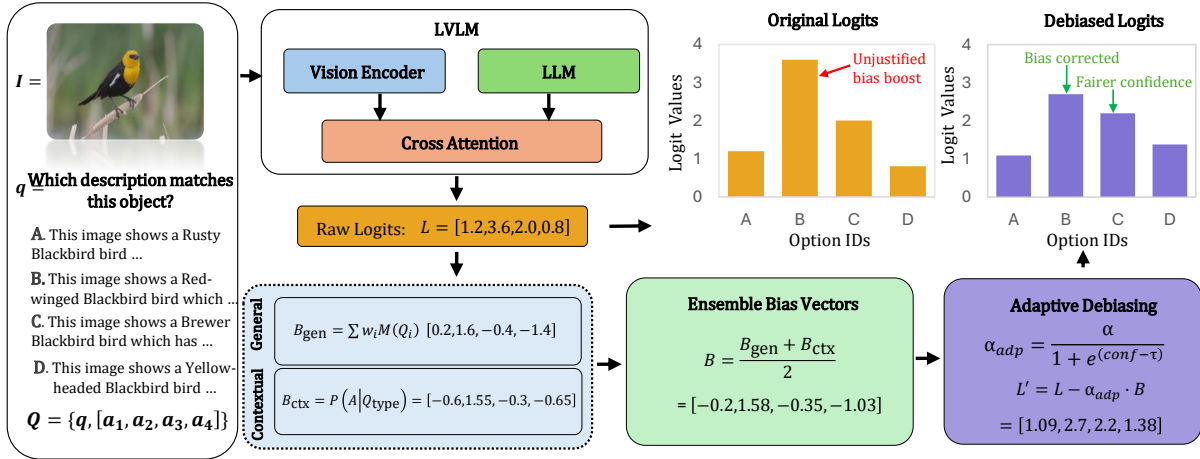
Our findings highlight important limitations in current LVLMs, particularly in fine-grained reasoning under semantic ambiguity. The variation in bias patterns across architectures suggests that each model encodes different shortcuts or priors. Furthermore, the consistent amplification of bias with increasing task difficulty aligns with observations from the LLM literature (Pezeshkpour and Hruschka, 2024), where uncertainty leads models to default to preferred positions. **In our case, LVLMs exhibit both token and positional selection biases, revealing critical challenges in their reasoning reliability when fine-grained visual reasoning is required and MCQ options are semantically similar and fine-grained.**

## 5 Selection Bias Mitigation

To mitigate selection biases (positional and token) in LVLMs during MCQA tasks, we propose a bias mitigation framework comprising two main components: (1) ensemble bias vector estimation and (2) adaptive logit correction at inference time.

**Ensemble Bias Vector Estimation.** LVLMs often exhibit preferences toward certain answer tokens (e.g., "A", "B", "C", "D") due to training artifacts or prompt structures. To capture and correct for these biases, we estimate two types of bias vectors: general bias and contextual bias.

**General Bias Vector.** The general bias vector  $B_{\text{gen}} \in \mathbb{R}^4$  captures systematic biases that arise from model architecture, token position, or prompt format. It is computed by prompting the model with multiple semantically empty templates  $\{Q_1, Q_2, \dots, Q_n\}$ , where each question has randomized answer option orders but lacks meaningful



$I$ : Image,  $Q$ : Question,  $L$ : logits,  $L'$ : Debaised Logits,  $B$ : Bias vector,  $\alpha$ : Debiasing coefficient,  $B_{\text{gen}/\text{ctx}}$ : Position/Question-type

Figure 4: Illustration of our training-free ensemble debiasing framework for LVLM-based MCQA. We estimate general and contextual bias vectors, average them, and apply adaptive logit correction based on model confidence to reduce selection bias and improve prediction accuracy.

content. The model’s output logits for each prompt are converted into probability distributions using softmax, and the results are averaged:

$$B_{\text{gen}} = \frac{1}{n} \sum_{i=1}^n \text{softmax}(f_{\theta}(Q_i)). \quad (1)$$

This estimates structural bias independent of task semantics, content, or reasoning cues.

**Contextual Bias Vector.** The contextual bias vector  $B_{\text{ctx}} \in \mathbb{R}^4$  measures how the model’s predictions are skewed on a representative sample of actual data. Given a small (10%), randomly selected subset of real MCQ examples  $\{(Q_j, A_j)\}_{j=1}^m$ , we take the model’s output logits and average the resulting probability distributions:

$$B_{\text{ctx}} = \frac{1}{m} \sum_{j=1}^m \text{softmax}(f_{\theta}(Q_j)). \quad (2)$$

This captures biases conditioned on realistic visual and textual inputs. Before combining, each bias vector is zero-centered by subtracting its mean to redistribute probabilities without introducing new preference directions.

**Final Ensemble Bias Vector.** The two components are averaged to produce the final ensemble bias vector:

$$B = \frac{B_{\text{gen}} + B_{\text{ctx}}}{2}. \quad (3)$$

**Adaptive Logit Correction.** During inference, the model outputs a logit vector  $L \in \mathbb{R}^4$  for the four answer choices. We debias these logits using the ensemble bias vector:

$$L' = L - \alpha_{\text{adp}} \cdot B, \quad (4)$$

where  $\alpha_{\text{adp}}$  is a confidence-adaptive scaling factor defined as:

$$\alpha_{\text{adp}} = \frac{\alpha}{1 + \exp(\text{conf} - \tau)}. \quad (5)$$

Here,  $\alpha$  is a global scaling hyperparameter (default 1.0), and  $\tau$  is a threshold for model confidence (default 2.0). The confidence is computed as:

$$\text{conf} = \max(L) - \text{mean}(L). \quad (6)$$

This adaptive debiasing applies stronger correction when the model is uncertain (low confidence) and more conservative adjustment when the model is confident, helping prevent over-correction while mitigating selection biases.

## 6 Experiments and Evaluation

In this section, we evaluate the extent of MCQA selection bias in LVLMs and assess the effectiveness of our proposed selection bias mitigation method. We benchmark model performance across different difficulty levels using our curated datasets with Qwen2.5-VL-3B-Instruct and 7B-Instruct (referred to as Qwen2.5-VL-3B or Qwen2.5-VL-7B throughout this paper, respectively), InternVL2\_5-8B, and LLaVA-v1.5-13B. We use 5 images per class for experiments at each difficulty level. For example, in the CUB "without class name" easy category, there are 200 classes, resulting in a total of 1,000 ( $200 \times 5$ ) images for evaluation.

### 6.1 LVLMs’ Performance on MCQA Tasks

Table 1 presents a comprehensive comparison of LVLM performance on our curated fine-grained

| Class Name           | Dataset (Difficulty) | Qwen2.5-VL-7B     | Qwen2.5-VL-3B | InternVL2_5-8B | LLaVA-v1.5-13B |               |
|----------------------|----------------------|-------------------|---------------|----------------|----------------|---------------|
| <b>With</b>          | Aircraft (Easy)      | 100.0 (100.0)     | 100.0 (83.71) | 100.0 (100.0)  | 100.0 (91.43)  |               |
|                      | Aircraft (Medium)    | 99.71 (100.0)     | 98.00 (43.14) | 95.71 (95.43)  | 48.86 (42.29)  |               |
|                      | Aircraft (Hard)      | 71.14 (71.43)     | 65.71 (37.71) | 51.14 (50.86)  | 21.71 (23.71)  |               |
|                      | Cars (Easy)          | 100.0 (100.0)     | 100.0 (81.43) | 100.0 (100.0)  | 100.0 (94.08)  |               |
|                      | Cars (Medium)        | 99.90 (99.59)     | 99.69 (85.41) | 99.59 (99.49)  | 84.69 (68.98)  |               |
|                      | Cars (Hard)          | 82.14 (80.51)     | 76.84 (55.71) | 57.04 (58.57)  | 29.90 (35.51)  |               |
|                      | CUB (Easy)           | 100.0 (99.90)     | 100.0 (93.40) | 100.0 (100.0)  | 99.70 (86.40)  |               |
|                      | CUB (Medium)         | 99.90 (99.70)     | 99.70 (84.90) | 100.0 (99.50)  | 62.70 (51.50)  |               |
|                      | CUB (Hard)           | 76.70 (72.10)     | 69.60 (48.00) | 60.50 (66.00)  | 28.30 (33.70)  |               |
|                      | Dogs (Easy)          | 100.0 (100.0)     | 100.0 (73.17) | 100.0 (100.0)  | 98.50 (68.33)  |               |
|                      | Dogs (Medium)        | 98.50 (98.67)     | 96.83 (67.33) | 96.67 (99.17)  | 45.00 (44.17)  |               |
|                      | Dogs (Hard)          | 76.83 (77.50)     | 71.00 (49.83) | 62.50 (55.83)  | 29.17 (31.67)  |               |
|                      | Food (Easy)          | 100.0 (99.01)     | 100.0 (79.41) | 100.0 (100.0)  | 99.80 (91.29)  |               |
|                      | Food (Medium)        | 98.22 (97.62)     | 98.02 (72.87) | 98.22 (98.61)  | 89.90 (77.62)  |               |
|                      | Food (Hard)          | 89.11 (85.94)     | 83.17 (58.42) | 84.55 (81.39)  | 54.46 (58.42)  |               |
|                      | iNaturalist (Easy)   | 98.88 (95.18)     | 97.99 (76.56) | 98.83 (98.59)  | 68.61 (58.36)  |               |
|                      | iNaturalist (Medium) | 87.95 (85.87)     | 84.11 (60.45) | 84.38 (84.56)  | 45.44 (40.40)  |               |
|                      | iNaturalist (Hard)   | 47.29 (44.87)     | 41.91 (31.88) | 25.06 (40.87)  | 25.46 (27.04)  |               |
|                      | <b>Without</b>       | Aircraft (Easy)   | 100.0 (100.0) | 100.0 (86.29)  | 100.0 (100.0)  | 100.0 (99.43) |
|                      |                      | Aircraft (Medium) | 84.57 (79.43) | 80.57 (42.57)  | 69.43 (75.71)  | 38.57 (33.43) |
|                      |                      | Aircraft (Hard)   | 53.71 (52.86) | 45.14 (33.71)  | 40.29 (45.14)  | 38.86 (36.29) |
|                      |                      | Cars (Easy)       | 100.0 (100.0) | 100.0 (79.90)  | 100.0 (100.0)  | 99.90 (98.98) |
|                      |                      | Cars (Medium)     | 99.49 (98.47) | 99.49 (74.80)  | 95.41 (96.84)  | 84.49 (66.73) |
|                      |                      | Cars (Hard)       | 70.00 (67.86) | 63.37 (44.49)  | 54.69 (56.43)  | 40.41 (35.61) |
|                      |                      | CUB (Easy)        | 100.0 (100.0) | 100.0 (92.60)  | 100.0 (100.0)  | 100.0 (94.70) |
|                      |                      | CUB (Medium)      | 99.40 (99.00) | 99.10 (88.60)  | 98.70 (99.10)  | 91.80 (82.80) |
|                      |                      | CUB (Hard)        | 64.90 (63.30) | 61.80 (46.40)  | 62.20 (61.00)  | 37.50 (34.70) |
| Dogs (Easy)          |                      | 100.0 (100.0)     | 100.0 (70.33) | 100.0 (100.0)  | 99.83 (93.50)  |               |
| Dogs (Medium)        |                      | 93.50 (93.83)     | 93.00 (64.83) | 86.67 (89.83)  | 64.50 (50.33)  |               |
| Dogs (Hard)          |                      | 64.00 (59.83)     | 58.00 (43.50) | 56.83 (55.67)  | 34.67 (33.00)  |               |
| Food (Easy)          |                      | 100.0 (99.80)     | 100.0 (76.24) | 100.0 (100.0)  | 99.80 (88.32)  |               |
| Food (Medium)        |                      | 97.62 (97.03)     | 98.02 (80.99) | 98.02 (97.82)  | 91.09 (78.02)  |               |
| Food (Hard)          |                      | 87.33 (84.95)     | 81.39 (63.56) | 80.59 (84.75)  | 48.91 (50.30)  |               |
| iNaturalist (Easy)   |                      | 95.75 (93.95)     | 94.64 (80.50) | 96.31 (95.69)  | 90.95 (68.79)  |               |
| iNaturalist (Medium) |                      | 83.62 (81.76)     | 82.21 (65.33) | 81.06 (80.21)  | 69.30 (56.87)  |               |
| iNaturalist (Hard)   |                      | 41.84 (40.20)     | 37.54 (29.88) | 37.89 (37.74)  | 25.04 (25.84)  |               |

Table 1: Accuracy (%) comparison of LVLMs across different dataset difficulty levels, with and without class names included in the option descriptions. Values in parentheses correspond to results under the "DCBA" option ordering; all others use the standard "ABCD" format.

MCQA benchmarks, covering six datasets across three difficulty levels, with and without class names included in the option descriptions. We evaluate four prominent LVLm models: Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, InternVL2\_5-8B, and LLaVA-v1.5-13B.

As expected, all models achieve near-perfect accuracy on easy tasks, confirming their ability to solve non-ambiguous MCQs regardless of format, though LLaVA-v1.5-13B shows notably lower performance on the iNaturalist dataset. However, performance diverges substantially on medium and hard tasks, particularly when class names are excluded. Qwen2.5-VL-7B-Instruct consistently outperforms others across all datasets and difficulty

levels, demonstrating robustness under increasingly fine-grained conditions. InternVL2\_5-8B also performs competitively, especially on medium tasks. LLaVA-v1.5-13B, while strong on easy tasks, shows a notable accuracy drop on hard tasks, where the options are semantically fine-grained.

The inclusion of class names generally improves performance across all models, though to varying degrees. Additionally, results in the reversed option format (DCBA), shown in parentheses, indicate that certain models, most notably LLaVA-v1.5-13B, are more sensitive to selection biases. These findings underscore the importance of evaluating both content and structural factors in fine-grained MCQA tasks and highlight the varying degrees of

bias susceptibility across LVLm architectures.

| Class Name | Dataset (Difficulty) | Qwen2.5-VL-7B | LLaVA-v1.5-13B |
|------------|----------------------|---------------|----------------|
| With       | Aircr. (M.)          | –             | 67.71 (+18.86) |
|            | Aircr. (H.)          | 71.71 (+0.57) | 21.71 (+00.00) |
|            | Cars (M.)            | –             | 93.47 (+8.78)  |
|            | Cars (H.)            | 83.67 (+1.53) | 35.31 (+5.41)  |
|            | CUB (M.)             | –             | 69.50 (+6.80)  |
|            | CUB (H.)             | 76.00 (-0.70) | 30.00(+1.70)   |
|            | Dogs (M.)            | –             | 66.33 (+21.33) |
|            | Dogs (H.)            | 78.67 (+1.84) | 34.50 (+5.33)  |
|            | Food (M.)            | –             | 92.67 (+2.77)  |
|            | Food (H.)            | 89.70 (+0.59) | 60.99 (+6.53)  |
|            | Aircr. (M.)          | 84.00 (-0.57) | 48.86 (+10.29) |
|            | Aircr. (H.)          | 54.86 (+1.15) | 42.29 (+3.43)  |
|            | Cars (M.)            | –             | 87.04 (+2.55)  |
|            | Cars (H.)            | 70.82 (+0.82) | 42.24 (+1.83)  |
| W/o        | CUB (M.)             | –             | 93.50 (+1.70)  |
|            | CUB (H.)             | 65.60 (+0.70) | 38.60 (+1.10)  |
|            | Dogs (M.)            | 94.50 (+1.00) | 72.00 (+7.50)  |
|            | Dogs (H.)            | 64.33 (+0.33) | 37.17 (+2.50)  |
|            | Food (M.)            | 97.62 (+0.00) | 95.84 (+4.75)  |
|            | Food (H.)            | 89.11 (+1.78) | 51.88 (+2.97)  |

Table 2: Accuracy (%) of our proposed selection bias mitigation method across multiple datasets and difficulty levels. Results are shown for Qwen2.5-VL-7B-Instruct and LLaVA-v1.5-13B under both class name ("With") and without class name ("W/o") settings. Numbers in parentheses indicate absolute gains over the standard "ABCD" format baseline. "–" denotes settings with near-saturated baseline accuracy ( $\geq 98\%$ ) where further improvement is not meaningful. M. and H. denote medium and hard difficulty, respectively. Aircr. represents FGVC Aircraft dataset.

## 6.2 Bias Mitigation Results

Table 2 reports the performance of our proposed selection bias mitigation method on our curated fine-grained MCQA benchmarks: FGVC Aircraft, Stanford Cars, CUB, Stanford Dogs, and Food-101, under medium (M.) and hard (H.) difficulty levels. Results are shown for Qwen2.5-VL-7B-Instruct and LLaVA-v1.5-13B in both settings: with and without class names included in the option descriptions of the MCQs.

Since most LVLms achieve 100% accuracy on easy tasks in our benchmarks across all MCQA formats (e.g., ABCD, DCBA), we omit those cases from Table 2. However, we verify that our mitigation method does not degrade performance in these cases. For instance, on CUB with class name using LLaVA-v1.5-13B, the model retains 100% accuracy even after applying mitigation.

We focus on Qwen2.5-VL-7B-Instruct and LLaVA-v1.5-13B as representative models due to their contrasting baseline performance: Qwen2.5-VL-7B achieves the highest accuracy overall, while LLaVA-v1.5-13B performs relatively poorly. This contrast enables us to evaluate the robustness and generalizability of our mitigation strategy across both strong and weak model conditions.

Our method consistently improves accuracy in most medium and hard settings, particularly under the "without class name" (W/o) condition, where selection biases are more pronounced. LLaVA-v1.5-13B exhibits substantial gains, such as +21.33% on Dogs (Medium) and +6.53% on Food (Hard). Qwen2.5-VL-7B-Instruct also shows improvements, including +1.15% on Aircraft (Hard) and +1.84% on Dogs (Hard). In medium-difficulty cases where baseline accuracy is already near-saturated ( $\geq 98\%$ ), gains are negligible or omitted. Only a few settings show minor accuracy drops (e.g., CUB (Hard task) with class names on Qwen2.5-VL-7B-Instruct), suggesting model- and dataset-specific variability.

Overall, these results demonstrate that our logit-based mitigation method effectively reduces selection bias without compromising accuracy, and often significantly improves it in challenging scenarios across our curated benchmarks.

| Class Name | Dataset | Model      | ABCD (DCBA)   | 1234 (4321)   |
|------------|---------|------------|---------------|---------------|
| With       | CUB     | LLaVA-13B  | 28.30 (33.70) | 27.20 (29.00) |
|            |         | Qwen2.5-7B | 76.70 (72.10) | 75.60 (66.80) |
|            | Cars    | LLaVA-13B  | 29.90 (35.51) | 25.20 (25.71) |
|            |         | Qwen2.5-7B | 82.14 (80.51) | 82.65 (78.98) |
| W/o        | CUB     | LLaVA-13B  | 37.50 (34.70) | 27.00 (27.10) |
|            |         | Qwen2.5-7B | 64.90 (63.30) | 64.60 (59.90) |
|            | Cars    | LLaVA-13B  | 40.41 (35.61) | 26.12 (26.73) |
|            |         | Qwen2.5-7B | 70.00 (67.86) | 65.92 (62.86) |

Table 3: Accuracy (%) comparison between alphabetic (ABCD/DCBA) and numeric (1234/4321) option identifiers across different models and datasets for hard tasks. Parentheses show results with reversed order. For some datasets, we observe performance degradation with numeric option identifiers.



### 6.3 Alternative Option IDs

We investigate alternative option identifiers to provide crucial insights into whether biases stem from specific token identity versus structural positioning effects. Our experiments with numeric identifiers (1/2/3/4), shown in Table 3, reveal that biases persist regardless of the identifier type. For example, LLaVA-v1.5-13B exhibits a token bias toward ‘A’ in ABCD and DCBA formats, while in 1234 and 4321 formats, it prefers option ‘1’, indicating consistent token-level bias. Qwen2.5-VL-7B-Instruct similarly displays both token and position biases across all formats. Notably, performance drops significantly when using numeric formats (1234/4321). These findings suggest that LVLMs exhibit systematic preferences influenced by both token identity and structural positioning in the prompt format.

### 6.4 Generalizability of the Mitigation Method

We assess the generalizability of our proposed bias mitigation method beyond templated formats to strengthen its robustness. Specifically, we use the Qwen3-32B (Yang et al., 2025a) language model to regenerate questions and options in a more natural MCQ style, given the original question, options, and correct answer. For these experiments, we use the CUB (without class name) and Dogs (with class name) datasets. Unlike our main benchmarks, we do not explicitly control the difficulty level in these regenerated samples. However, the experimental results in Table 4 indicate a comparable level of difficulty. Below is an example of a regenerated MCQ from the CUB (without class name) dataset for the “Black-footed Albatross” class.

#### Multiple Choice Question

**Question:** What color is the majority of the bird’s body in the image?

- A. Dark brown
- B. Yellow
- C. Blue
- D. Red

**Correct Answer:** A

For these experiments, we use the LLaVA-v1.5-13B model and report accuracy in the standard “ABCD” format. As shown in Table 4, our proposed mitigation method effectively reduces selection bias on newly generated benchmarks that

do not follow templated MCQA formats, yielding performance gains of 1.17–5.00%. These results suggest that the mitigation approach addresses underlying model behaviors that persist independently of specific question templates.

| Dataset (Difficulty) | Without Mitigation | With Mitigation |
|----------------------|--------------------|-----------------|
| CUB w/o name (M.)    | 90.40              | 91.70 (+1.30)   |
| CUB w/o name (H.)    | 49.50              | 54.50 (+5.00)   |
| Dogs with name (M.)  | 81.50              | 82.67 (+1.17)   |
| Dogs with name (H.)  | 45.83              | 48.00 (+2.17)   |

Table 4: Accuracy (%) of the LLaVA-v1.5-13B model on beyond-template MCQA tasks, with and without the proposed bias mitigation method. Results show consistent improvements across all datasets, with larger gains on hard (H.) tasks compared to medium (M.) tasks.

We deliberately adopt a templated format (“Which description matches this object?”) to isolate selection bias under controlled conditions. By standardizing the question structure while varying distractor difficulty across diverse domains (birds, dogs, aircraft, cars, food), we can precisely measure token- and position-based biases without introducing confounding variables from question phrasing. Importantly, many real-world applications, such as educational assessments, standardized tests, and evaluation systems, employ structured MCQA formats, making our findings directly applicable.

## 7 Conclusion

We present a systematic analysis of selection bias in Large Vision-Language Models (LVLMs) within the multiple-choice question answering (MCQA) setting. Our findings show that selection bias, driven by both token identity and positional preferences, intensifies with increasing task difficulty, particularly when MCQ options are fine-grained. While stronger models tend to exhibit reduced bias compared to weaker models, none are entirely immune, underscoring a fundamental limitation in current LVLm reasoning capabilities. To address this, we introduce a simple yet effective inference-time logit debiasing method that mitigates these biases without requiring model retraining. Our approach improves accuracy and consistency across varied input configurations and difficulty levels. This work highlights the importance of fine-grained, difficulty-aware benchmarks for revealing nuanced model behaviors and guiding future improvements. In future work, we aim to extend our framework to encompass a broader range of multimodal tasks.

## 8 Limitations

Our work specifically targets selection bias in MCQA, allowing focused investigation of these mechanisms while not addressing other bias forms such as modality imbalance or cultural bias. Our mitigation approach is designed for MCQA tasks rather than open-ended generation, enabling precise logit-level corrections for multiple-choice selection behavior. We curated comprehensive benchmarks within classification-type datasets to systematically analyze how selection biases manifest across diverse domains and difficulty levels. This targeted approach provides foundational insights into selection bias mechanisms in LVLMS. Future work will extend these findings to broader bias types and generative tasks, and will provide more comprehensive analysis of model-specific attention patterns across different architectures..

## 9 Acknowledgments

We acknowledge Advanced Research Computing (ARC) at Virginia Tech for providing the computational resources and technical support that contributed to the results reported in this paper. We thank the reviewers for their constructive feedback, which helped improve this paper.

## References

- Dyah Adila, Shuai Zhang, Boran Han, and Yuyang Wang. 2024. Discovering bias in latent space: an unsupervised debiasing approach. In *Proceedings of the 41st International Conference on Machine Learning*, pages 246–261.
- Md Atabuzzaman, Andrew Zhang, and Chris Thomas. 2025. Zero-shot fine-grained image classification using large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.
- Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above. *arXiv preprint arXiv:2502.14127*.
- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101—mining discriminative components with random forests. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part VI 13*, pages 446–461. Springer.
- Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024a. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469.
- Peng Chen, Xiao-Yu Guo, Yuan-Fang Li, Xiaowang Zhang, and Zhiyong Feng. 2024b. Mitigating language bias of llms in social intelligence understanding with virtual counterfactual calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1300–1310.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. 2023. Can we edit multimodal large language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13877–13888.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 37.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.
- Grace J Di Cecco, Vijay Barve, Michael W Belitz, Brian J Stucky, Robert P Guralnick, and Allen H Hurlbert. 2021. Observing the observers: How participants contribute data to inaturalist and implications for biodiversity science. *BioScience*, 71(11):1179–1188.
- Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892.
- Blessed Guda, Lawrence Francis, Gabriel Zencha Ashungafac, Carlee Joe-Wong, and Moise Busogi. 2025. Tiny: Rethinking selection bias in llms: Quantification and mitigation using efficient majority voting. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

- Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2.
- 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.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of llms? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.
- OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.
- Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. 2024. Order matters: Exploring order sensitivity in multimodal large language models. *arXiv preprint arXiv:2410.16983*.
- Qwen Team. 2025. [Qwen2.5-vl](#).
- Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*.
- Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025. LLMs may perform mcqa by selecting the least incorrect option. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5852–5862.
- Jialu Wang, Yang Liu, and Xin Wang. 2021. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1995–2008.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and 1 others. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.
- Zecheng Wang, Xinye Li, Zhanyue Qin, Chunshan Li, Zhiying Tu, Dianhui Chu, and Dianbo Sui. 2024b. Can we debias multimodal large language models via model editing? In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3219–3228.
- Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5598–5621.
- Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Honglin Han, Meng Zhao, Chengguo Yin, and 1 others. 2024. Strengthened symbol binding makes large language models reliable multiple-choice selectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4344.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang, and so on... 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Nakyeong Yang, Taegwan Kang, Stanley Jungkyu Choi, Honglak Lee, and Kyomin Jung. 2024. Mitigating biases for instruction-following language models via bias neurons elimination. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9061–9073.

Zhen Yang, Ping Jian, and Chengzhi Li. 2025b. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of 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 1902–1917.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024. Debiasing multimodal large language models. *arXiv preprint arXiv:2403.05262*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *ICLR*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. 2024. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. 2024. Fool your (vision and) language model with embarrassingly simple permutations. In *International Conference on Machine Learning*, pages 62892–62913. PMLR.

## A Appendix

This section contains the following topics.

- Dataset Statistics (Appendix A.1)
- Visualization of LVLMS’ Selection Bias (Appendix A.2)

### A.1 Dataset Statistics

To construct each MCQ, we utilize existing fine-grained image classification datasets, including CUB, Stanford Dogs, FGVC Aircraft, Stanford

Cars, Food-101, and iNaturalist. Each image is paired with a curated class description. Class descriptions are collected from Atabuzzaman et al. (2025). For the iNaturalist dataset, we collect class descriptions from Kim and Ji (2024) and use common names of species instead of scientific class names. Upon analysis, we found 38 species with duplicate common names, resulting in 9,962 unique classes.

We analyze the datasets by reporting the average and standard deviation of semantic similarity scores across the three difficulty levels in Table 5. As expected, the Easy set exhibits low average similarity, the Medium set falls in the mid-range, and the Hard set shows high similarity between the target and distractors. These statistics validate our difficulty categorization and provide a quantitative basis for evaluating model performance across different levels of semantic and visual ambiguity.

**Answer Option Distribution.** Table 6 presents the distribution of correct answer positions (A–D) across our benchmark datasets, segmented by dataset (CUB, FGVC Aircraft, Stanford Cars, Stanford Dogs, Food-101, and iNaturalist), difficulty level (Easy, Medium, Hard), and the presence or absence of class names in the answer options. Each row corresponds to a specific configuration, and the columns report how many times each option ID (A, B, C, D) is the correct answer.

To ensure fair evaluation of selection bias, we constructed all datasets to maintain a near-balanced distribution of correct answers across the four option IDs. This prevents any skew that could arise from inherent imbalances in the dataset and helps isolate model behavior due to selection bias rather than label distribution.

Each difficulty level contains both "with class name" and "without class name" variants, allowing us to assess the role of explicit label cues in model predictions. For example, in the CUB (Easy) category, the correct answer is evenly distributed across the four choices (A–D) for both versions. Similar patterns are preserved across other datasets and difficulty levels.

Across all datasets and configurations, the total number of multiple-choice questions is 63,894, with each variant carefully designed to mitigate structural bias in the ground-truth distribution. This careful balancing enables controlled investigation of the positional and token-level biases of LVLMS during multiple-choice question answering.

| Dataset     | Class Names | Difficulty | Domain | Avg. Sim. | Std. Dev. |
|-------------|-------------|------------|--------|-----------|-----------|
| CUB         | Without     | Easy       | Diff.  | 0.2241    | 0.0409    |
|             |             | Medium     | Same   | 0.4259    | 0.0472    |
|             |             | Hard       | Same   | 0.8461    | 0.0573    |
|             | With        | Easy       | Diff.  | 0.1864    | 0.0408    |
|             |             | Medium     | Same   | 0.3696    | 0.0347    |
|             |             | Hard       | Same   | 0.7831    | 0.0707    |
| Dogs        | Without     | Easy       | Diff.  | 0.1736    | 0.0311    |
|             |             | Medium     | Same   | 0.5574    | 0.0557    |
|             |             | Hard       | Same   | 0.8855    | 0.0536    |
|             | With        | Easy       | Diff.  | 0.1382    | 0.0354    |
|             |             | Medium     | Same   | 0.3854    | 0.0573    |
|             |             | Hard       | Same   | 0.7981    | 0.0546    |
| Aircraft    | Without     | Easy       | Diff.  | 0.1478    | 0.0427    |
|             |             | Medium     | Same   | 0.6971    | 0.0509    |
|             |             | Hard       | Same   | 0.9413    | 0.0369    |
|             | With        | Easy       | Diff.  | 0.1623    | 0.0397    |
|             |             | Medium     | Same   | 0.4875    | 0.0531    |
|             |             | Hard       | Same   | 0.8256    | 0.0437    |
| Cars        | Without     | Easy       | Diff.  | 0.1444    | 0.0366    |
|             |             | Medium     | Same   | 0.6458    | 0.0462    |
|             |             | Hard       | Same   | 0.9203    | 0.0356    |
|             | With        | Easy       | Diff.  | 0.1350    | 0.0357    |
|             |             | Medium     | Same   | 0.3571    | 0.0502    |
|             |             | Hard       | Same   | 0.8237    | 0.0732    |
| Food        | Without     | Easy       | Diff.  | 0.1969    | 0.0402    |
|             |             | Medium     | Same   | 0.3228    | 0.0529    |
|             |             | Hard       | Same   | 0.6942    | 0.0615    |
|             | With        | Easy       | Diff.  | 0.1873    | 0.0377    |
|             |             | Medium     | Same   | 0.3856    | 0.0434    |
|             |             | Hard       | Same   | 0.7066    | 0.0532    |
| iNaturalist | Without     | Easy       | Diff.  | 0.1308    | 0.0490    |
|             |             | Medium     | Same   | 0.3438    | 0.1173    |
|             |             | Hard       | Same   | 0.8408    | 0.0708    |
|             | With        | Easy       | Diff.  | 0.0888    | 0.0334    |
|             |             | Medium     | Same   | 0.2844    | 0.1431    |
|             |             | Hard       | Same   | 0.7813    | 0.0818    |

Table 5: Average similarity and standard deviation between the ground truth and distractor options across difficulty levels (Easy, Medium, Hard) on fine-grained datasets, with and without class names. Easy distractors are from different domains; Medium and Hard are from the same domain. Std. is computed among the distractors.

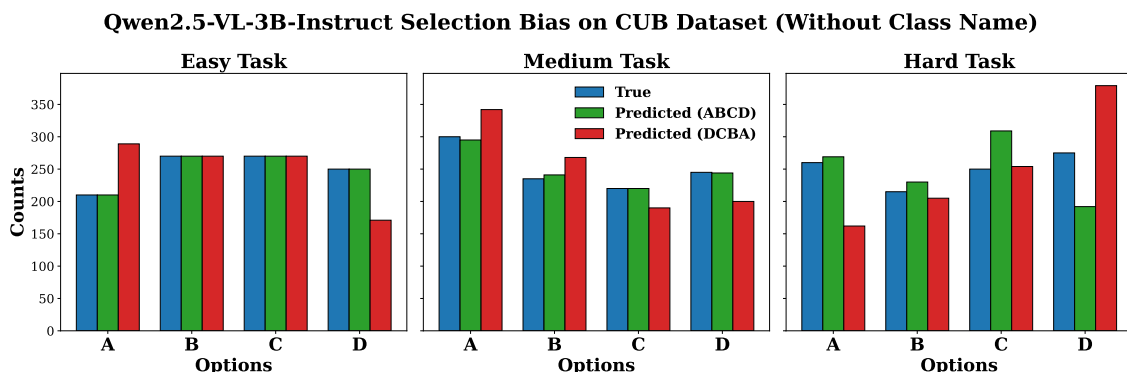
## A.2 Visualization of LVLMS’ Selection Bias

Figure 5 provides a comparative visualization of selection bias exhibited by three state-of-the-art Large Vision-Language Models (LVLMS): Qwen2.5-VL-3B-Instruct, LLaVA-v1.5-13B, and InternVL2\_5-8B on the CUB dataset under the "without class name" setting. The bar plots illustrate how each model’s answer distribution changes across task difficulty levels (Easy, Medium, Hard) and under different option orderings (standard ABCD vs. reversed DCBA). As the task becomes more difficult, Qwen2.5-VL-3B shows increasing bias toward specific token identities and positions, such as a consistent preference for option “A” or the

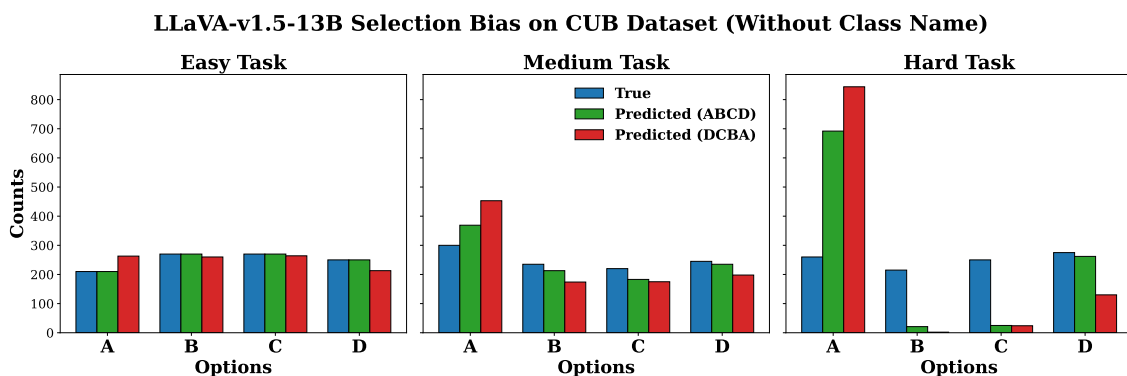
first-position choice. The LLaVA-v1.5-13B model shows balanced selection in easy tasks but develops strong token bias in hard tasks, exhibiting a dramatic preference for option ID “A” (nearly three times the true frequency) as uncertainty increases. In contrast, InternVL2\_5-8B demonstrates more stable behavior across conditions, maintaining a distribution that closely aligns with the ground truth, especially in Medium tasks. These visualizations highlight both shared and model-specific patterns of bias, reinforcing the importance of task difficulty and option formatting when evaluating LVLMS behavior.

| Category               | Class Name | Option IDs |      |      |      | Total |
|------------------------|------------|------------|------|------|------|-------|
|                        |            | A          | B    | C    | D    |       |
| CUB (Easy)             | Without    | 42         | 54   | 54   | 50   | 200   |
|                        | With       | 42         | 54   | 54   | 50   | 200   |
| CUB (Medium+Hard)      | Without    | 112        | 90   | 94   | 104  | 400   |
|                        | With       | 100        | 95   | 103  | 102  | 400   |
| Aircraft (Easy)        | Without    | 21         | 20   | 18   | 11   | 70    |
|                        | With       | 21         | 20   | 18   | 11   | 70    |
| Aircraft (Medium+Hard) | Without    | 36         | 41   | 25   | 38   | 140   |
|                        | With       | 28         | 43   | 37   | 32   | 140   |
| Cars (Easy)            | Without    | 56         | 54   | 38   | 48   | 196   |
|                        | With       | 56         | 54   | 38   | 48   | 196   |
| Cars (Medium+Hard)     | Without    | 85         | 99   | 108  | 100  | 392   |
|                        | With       | 92         | 106  | 97   | 97   | 392   |
| Dogs (Easy)            | Without    | 28         | 25   | 28   | 39   | 120   |
|                        | With       | 28         | 25   | 28   | 39   | 120   |
| Dogs (Medium+Hard)     | Without    | 72         | 54   | 57   | 57   | 240   |
|                        | With       | 60         | 71   | 59   | 50   | 240   |
| Food (Easy)            | Without    | 24         | 26   | 27   | 24   | 101   |
|                        | With       | 24         | 26   | 27   | 24   | 101   |
| Food (Medium+Hard)     | Without    | 53         | 48   | 52   | 49   | 202   |
|                        | With       | 50         | 49   | 49   | 54   | 202   |
| iNaturalist (Easy)     | Without    | 2491       | 2474 | 2439 | 2558 | 9962  |
|                        | With       | 2491       | 2474 | 2439 | 2558 | 9962  |
| iNaturalist (Medium)   | Without    | 2545       | 2505 | 2417 | 2495 | 9962  |
|                        | With       | 2498       | 2533 | 2430 | 2501 | 9962  |
| iNaturalist (Hard)     | Without    | 2469       | 2508 | 2453 | 2532 | 9962  |
|                        | With       | 2469       | 2508 | 2453 | 2532 | 9962  |
| <b>Total = 63894</b>   |            |            |      |      |      |       |

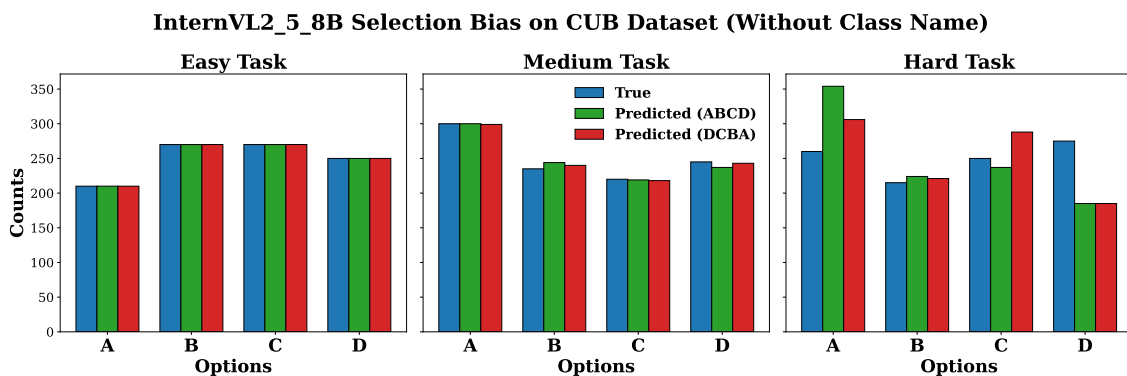
Table 6: Distribution of correct answer options (A–D) across datasets for Easy, Medium, and Hard categories. For brevity, we merge Medium and Hard into a single row for smaller datasets. Each dataset maintains a balanced distribution of answer options across all categories, both with and without the class name.



(a) Qwen2.5-VL-3B-Instruct exhibits stronger selection biases—favoring option ID “A” in Easy tasks (**token bias**) and the first-position options “A” and “D” in “ABCD” and “DCBA” orderings, respectively, in Hard tasks (**positional bias**).



(b) LLaVA-v1.5-13B model shows balanced selection in Easy tasks but develops strong token bias in Hard tasks, with dramatic preference for the option ID "A" (reaching nearly 3x the true frequency) when difficulty increases.



(c) InternVL2\_5-8B model demonstrates the most balanced behavior across difficulty levels, with predictions closely matching the true distribution in Easy and Medium tasks. In Hard tasks, it shows moderate token bias ("A") but maintains better distribution consistency between ABCD and DCBA orderings.

Figure 5: Selection bias comparison across LVLMs on the CUB dataset under the "without class name" setting, organized by increasing task difficulty (Easy, Medium, Hard). Each plot shows distributions for ground truth (True) and model predictions under two option orderings: standard (ABCD) and reversed (DCBA). The comparison reveals how position and token biases emerge and intensify with task difficulty, with varying patterns across architectures.