FREE-TO-ANSWER: RETHINKING OUTPUT-FORMAT CONSTRAINTS IN REINFORCEMENT FINE-TUNING FOR FINE-GRAINED VISUAL RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual concepts in the real world are naturally organized from coarse semantic groups to increasingly fine-grained sub-categories, whose recognition requires both rich category knowledge and sensitivity to subtle visual differences. Rule-based reinforcement fine-tuning (RFT) has recently improved the fine-grained visual recognition (FGVR) capabilities of Multimodal Large Language Models (MLLMs). To enable automatic reward computation, existing methods commonly require responses to follow rigid output structures, such as separate reasoning and answer tags. However, it remains unclear whether these structures are merely parsing interfaces or whether they also affect recognition behavior. In this work, we systematically investigate how output-format constraints affect FGVR before and during reinforcement fine-tuning. Across five benchmarks, removing mandatory structural tags generally improves recognition accuracy, although the effect varies across models and datasets. Motivated by this observation, we develop Free-to-Answer, which optimizes free-form responses using LLM-based label extraction followed by deterministic verification. Under Open-Ended QA evaluation, Free-to-Answer achieves higher average accuracy than format-constrained RFT across the 1-, 2-, and 4-shot settings, while controlled experiments suggest that the improvement is not solely attributable to the reward design. We further analyze how free-form responses interact with closed-set label grounding. Our results reveal that rigid output structures can affect recognition behavior in visual RFT, rather than serving only as interfaces for reward parsing.

1 INTRODUCTION

Visual concepts in the real world are naturally organized from coarse semantic groups to increasingly fine-grained sub-categories, whose recognition requires both rich category knowledge and sensitivity to subtle visual differences. This task remains challenging for MLLMs because discriminative differences are often subtle, while visual appearances can vary substantially within the same category (Geigle et al., 2024; Zhang et al., 2024; Pang et al., 2025). Although MLLMs possess broad visual and semantic knowledge, they do not always ground this knowledge reliably to fine-grained category labels (Zhang et al., 2024; He et al., 2025).

Recent methods improve FGVR through rule-based reinforcement fine-tuning (RFT) with verifiable rewards (Liu et al., 2025; Kumar et al., 2025; Li et al., 2025b; He et al., 2026). To facilitate automatic reward computation, these methods commonly require responses to follow predefined structures, such as `<think> reasoning </think>` followed by `<answer> answer </answer>`. Such structures make it straightforward to extract the final prediction and evaluate format compliance during optimization. However, the model is consequently optimized not only for task correctness but also for conformity to a prescribed response structure. Prior studies suggest that rigid output structures can affect model performance in reasoning and prediction tasks (Tam et al., 2024; Wang et al., 2025; Molfese et al., 2025; Lawrence et al., 2026; Sterz et al., 2025). As illustrated in Figure 1 (a), we observe a similar phenomenon in FGVR: requiring the model to separate its reasoning and final answer using predefined tags can change its prediction and, in some cases, produce an incorrect result. This raises an underexplored question: *do output-format constraints merely simplify reward computation, or can they also affect fine-grained recognition behavior?*

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

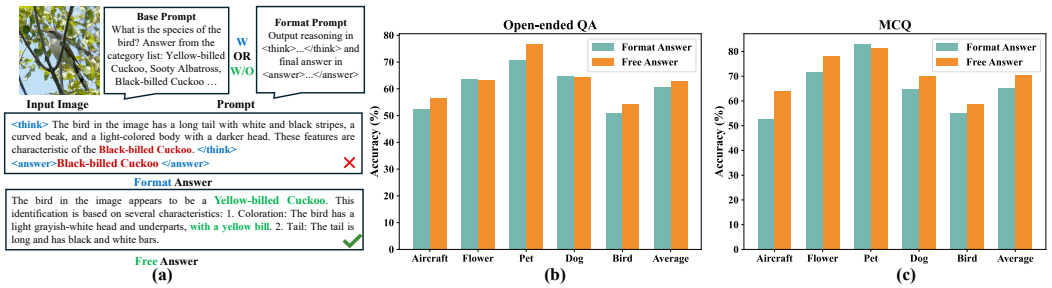


Figure 1: Motivation and pilot study on the effect of output-format constraints on Qwen2.5-VL-7B. (a) A representative example where rigid structural tags lead to an incorrect prediction, while free-form generation produces the correct prediction. (b–c) Zero-shot fine-grained recognition accuracy under format-constrained and free-form prompting across five benchmarks in Open-Ended QA and MCQ settings. Free-form prompting improves the average accuracy in both settings and outperforms format-constrained prompting on most datasets.

We investigate this question both before and during reinforcement fine-tuning. Across five FGVR benchmarks, removing mandatory structural tags generally improves average recognition accuracy, although the magnitude and direction of the effect vary across models and datasets. For example, without any fine-tuning, free-form prompting improves the average Open-Ended QA accuracy of Qwen2.5-VL-7B from 60.53% to 63.08%. These results suggest that output formatting is not always a behaviorally neutral interface and motivate an examination of its role during RFT.

Based on this observation, we propose **Free-to-Answer**, a training-time RFT framework optimized with Group Relative Policy Optimization (GRPO) (Guo et al., 2025). Free-to-Answer removes mandatory reasoning and answer tags, allowing the model to generate responses without predefined structural boundaries. To compute verifiable rewards from these responses, an LLM-based evaluator first maps each response to the closed-set label space, after which deterministic verification checks whether the extracted label is supported by the generated response. Here, "format-free" refers specifically to removing mandatory output structures rather than removing task supervision or label-space constraints. Under Open-Ended QA evaluation, Free-to-Answer improves average accuracy over format-constrained RFT by 1.51, 1.67, and 1.45 percentage points in the 1-, 2-, and 4-shot settings, respectively. Controlled experiments using the same evaluator further suggest that these improvements are not attributable solely to differences in reward design.

Free-to-Answer differs from *Nlg2choice* (Lawrence et al., 2026) in its primary objective and training setting. While *Nlg2choice* keeps the MLLM frozen and maps free-form outputs to candidate labels at inference time, Free-to-Answer introduces free-form generation into policy optimization. For standardized closed-set prediction, we additionally adopt an *nlg2choice*-style two-stage grounding protocol. Rather than treating this inference protocol as a new method, we use it to analyze how free-form responses are mapped to formal category labels and which types of prediction errors can be corrected during this process.

In summary, the main contributions are:

- A systematic empirical investigation showing that rigid output-format constraints generally reduce the average FGVR performance of MLLMs, although their effects vary across models and datasets.
- This work proposes **Free-to-Answer**, a training-time format-free RL framework that couples semantic label extraction with deterministic evidence verification, enabling verifiable optimization of unconstrained responses.
- An analysis of how free-form generation interacts with closed-set label grounding, including the roles of semantic aliases, category granularity, and candidate-set information in correcting predictions.

2 Related Work

2.1 Reinforcement Learning

Reinforcement learning (RL) has become a central paradigm for improving foundation models by reducing reliance on large-scale human annotations. For large language models (LLMs), reinforcement learning from human feedback (RLHF) (Ziegler et al., 2019; Ouyang et al., 2022) has played a critical role in aligning models such as the GPT series (Achiam et al., 2023; Hurst et al., 2024) to better follow instructions and handle complex reasoning and coding tasks. More recently, DeepSeek-R1-Zero (Guo et al., 2025) demonstrates that strong logical and mathematical reasoning capabilities can be elicited through reinforcement learning with verifiable rewards, obviating the need for explicit reasoning demonstrations. Recent studies have extended RL to enhance the reasoning capabilities of MLLMs across diverse vision-centric tasks (Huang et al., 2025; Ma et al., 2025). Early efforts in multimodal RL primarily focused on improving instruction following and mitigating hallucinations, often leveraging PPO-style optimization or human feedback augmented with factual grounding (Schulman et al., 2017; Sun et al., 2024; Yu et al., 2024). Recent progress has further introduced Rule-based Reinforcement Fine-Tuning (RFT) with automatically verifiable rewards into the multimodal domain to enhance perception and inference (Li et al., 2025a; Tan et al., 2025; Li et al., 2025c; Linghu et al., 2026). This paradigm has been successfully applied to a variety of tasks, including image classification (Li et al., 2025a), object detection and grounding (Liu et al., 2025; Ni et al., 2025), and complex visual question answering (VQA) (Tan et al., 2025). For example, Visual-RFT (Liu et al., 2025) incorporates perception-aware rewards derived from spatial metrics such as Intersection over Union (IoU) for detection tasks, while 3D-RFT (Linghu et al., 2026) extends verifiable RL to video-based 3D scene understanding by directly optimizing geometry-related objectives.

2.2 Fine-Grained Visual Recognition with MLLMs

Although MLLMs have achieved strong general visual understanding (An et al., 2025; Bai et al., 2025a;b; Zhu et al., 2025), their performance often drops substantially on FGVR benchmarks (Zhang et al., 2024; He et al., 2025). FINER (Kim & Ji, 2024) suggests that the bottleneck lies less in the absence of category knowledge than in aligning discriminative visual cues with the corresponding fine-grained concepts. Analyses of MLLM projectors further indicate that patch-level visual–semantic alignment can remain weak and coarse (Jiang et al., 2026). To bridge this gap, prior work explores both supervised and training-free strategies. Supervised Fine-Tuning (SFT) methods such as Finedefics (He et al., 2025) enhance part-level alignment and class boundary sensitivity by leveraging attribute descriptions and contrastive objectives. Training-free inference frameworks such as Nlg2choice (Lawrence et al., 2026) decouple open-ended querying from a subsequent constrained decision stage, improving efficiency and robustness for fine-grained retrieval and selection. FineR (Liu et al., 2024) bridges visual observations and category knowledge by converting part-level attributes into textual descriptions for LLM-based category reasoning. More recently, RFT with verifiable rewards has also been explored for FGVR. Visual-RFT (Liu et al., 2025) extends verifiable optimization across multiple visual tasks via GRPO, and CLS-RL (Li et al., 2025b) studies when explicit reasoning tokens are necessary under RL fine-tuning. DiVE-K (Kumar et al., 2025) constructs verifiable rewards by converting the model's top-k candidates into multiple-choice questions, improving generalization among base-to-novel settings, while Fine-R1 (He et al., 2026) combines Chain-of-Thought (CoT) SFT with Triplet Augmented Policy Optimization (TAPO).

3 Method

We begin by examining the effect of structural output constraints through a pilot study and then present Free-to-Answer based on the resulting observation. As shown in Figure 2, Free-to-Answer removes mandatory structural tags during reinforcement fine-tuning, allowing the policy model to respond directly to a natural-language query. To compute rewards from free-form responses, a text-only LLM maps each response to the closed-set label space, followed by deterministic verification against the generated response and the ground-truth label.

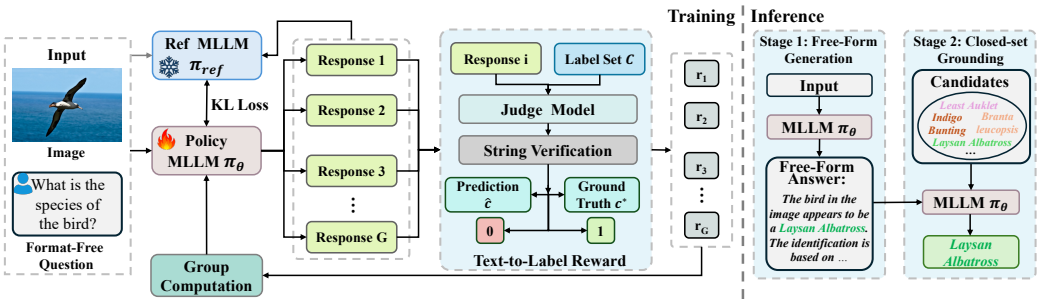


Figure 2: Overview of the proposed format-free reinforcement learning framework. During training, we remove rigid structural tags to allow for unconstrained, free-form reasoning, utilizing an LLM as judge model to optimize the policy via GRPO. During inference, we employ a two-stage decoupled strategy: (1) *Free-form Generation*, where the model produces an open-ended response; and (2) *Closed-set Grounding*, where the same model acts as its own judge to map its previous response to a standardized category from a candidate list.

3.1 The Bottleneck of Structural Constraints

We first examine whether structural tags affect FGVR performance by comparing format-constrained and free-form prompting. As shown in Figure 1 (b-c), enforcing structural tags yields a notable degradation in FGVR performance. We attribute this phenomenon to the disruption of continuous visual reasoning, as the model is forced to allocate its context attention to format compliance rather than visual deduction.

3.2 Problem Formulation

Consider a K-shot FGVR task with a closed-set label space $\mathcal{C}$. Each training instance is represented as (I, q, c^*), where I is an image, q is a natural-language query without mandatory formatting instructions, and $c^* \in \mathcal{C}$ is the ground-truth label. The training set $\mathcal{D}$ contains K labeled examples per category. Given $x = (I, q)$, the objective is to optimize a multimodal policy π_θ to generate a response y from which the correct category can be recovered.

3.3 Format-Free Reinforcement Learning

Let $x = (I, q)$ denote the input tuple. The policy model π_θ generates a free-form response sequence $y = (y_1, y_2, \ldots, y_T)$ via autoregressive decoding:

$$\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t}), \tag{1}$$

where $y_{<t} = (y_1, \ldots, y_{t-1})$ represents the previously generated tokens. Because the output is structurally unconstrained, the inferred category token y_{answer} can organically map to any token position within the response y. A sparse scalar reward $R(x, y, c^*)$ is evaluated exclusively on the terminated sequence. The optimization objective is formulated to maximize the expected trajectory reward:

$$J(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta}[R(x, y, c^*)]. \tag{2}$$

Optimization via Group Relative Policy Optimization (GRPO) During each training iteration, for a given context $x \sim \mathcal{D}$, we sample a group of G diverse responses $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ from the old policy. Each response is evaluated to obtain a scalar reward $r_i = R(x, y_i, c^*)$. We then compute the group-relative advantage $\hat{A}_i$ by standardizing the rewards within the sampled group:

$$\hat{A}_i = \frac{r_i - \mu_r}{\sigma_r + \varepsilon}, \quad \text{where} \quad \mu_r = \frac{1}{G}\sum_{j=1}^{G} r_j \quad \text{and} \quad \sigma_r = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \mu_r)^2}, \tag{3}$$

and ε is a small constant ensuring numerical stability.

The policy is updated by maximizing a PPO-style clipped surrogate objective. To maintain policy stability and prevent the unconstrained generation from drifting away from the natural language distribution of the base model, a token-level Kullback-Leibler (KL) divergence penalty is applied against the frozen pre-trained reference policy π_{ref}. The comprehensive GRPO objective is formulated as:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G}\Big(\min\big(\rho_i(\theta)\hat{A}_i,\ \text{clip}(\rho_i(\theta), 1-\delta, 1+\delta)\hat{A}_i\big) - \beta\,\mathbb{D}_{\text{KL}}(y_i \parallel \pi_{\text{ref}} \mid x)\Big)\right], \quad (4)$$

where $\rho_i(\theta) = \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}$ denotes the sequence-level probability ratio, δ is the clipping margin, and β controls the strength of the KL penalty.

3.4 LLM-Based Text-to-Label Reward

In our format-free setting, we cannot rely on rigid templates to parse predicted labels from responses due to the natural variability in generated text. To address this, we design an LLM-based text-to-label reward function that evaluates response correctness. Specifically, we leverage a pre-trained large language model to map the free-form response y to a predicted label $\hat{c}$, then compare $\hat{c}$ with the ground-truth label c^* to compute the reward. We prompt the language model to output the predicted label from the allowed label set $\mathcal{C}$ based on the response content. If insufficient information is present to determine a specific label, we prompt the model to output "None". To eliminate extraction stochasticity, we apply deterministic decoding via greedy decoding (temperature $T = 0$). This mapping is formulated as:

$$\hat{c} = \text{LLM}_{\text{label}}(y, \mathcal{C}), \quad \hat{c} \in \mathcal{C} \cup \{\text{None}\}. \quad (5)$$

String Verification Process We further apply a deterministic string verification step to mitigate potential hallucinations (Zheng et al., 2023; Yu et al., 2023). Let $\mathbb{I}(\cdot)$ denote the indicator function. The verification strictly asserts that the extracted category syntactically exists within the unstructured generative text y. The verified predicted label $\tilde{c}$ and the ultimate reward R are formalized as follows:

$$\tilde{c} = \begin{cases} \hat{c}, & \text{if } \mathbb{I}(\hat{c} \in y) = 1 \\ \text{None}, & \text{otherwise} \end{cases}, \quad R(x, y, c^*) = \mathbb{I}(\tilde{c} = c^*) \in \{0, 1\}. \quad (6)$$

3.5 Evaluation and Inference

For open-ended question answering tasks, we evaluate the model's response using the same LLM-based text-to-label mapping and string verification process as during training. For real-world applications requiring standardized outputs, we avoid *constrained decoding* or secondary judge LLMs, which can be computationally expensive or resource-intensive. Instead, we employ a simple two-stage prompting strategy following Nlg2choice (Lawrence et al., 2026) to ground free-form responses into the closed-set label space. Specifically, we first prompt the model with the base prompt and input image to generate its unconstrained, open-ended response. Subsequently, we prompt the model again with only the textual input, letting it choose the best-matching label from the candidate list based only on the previously generated answer.

4 Experiments

4.1 Experimental Settings

Datasets We evaluate on five FGVR benchmarks: CUB-200-2011 (Wah et al., 2011), Stanford Dogs (Khosla et al., 2011), Oxford Flowers (Nilsback & Zisserman, 2008), FGVC-Aircraft (Maji et al., 2013), and Oxford-IIIT Pet (Parkhi et al., 2012). For few-shot training, we sample $K \in \{1, 2, 4\}$ labeled images per class and use the original test splits for evaluation. Dataset statistics and the complete sampling protocol are provided in Appendix A.1.

Table 1: Prompt templates used in training and evaluation. The **Base prompt** defines the basic recognition query, while the **Format prompt** and **MCQ prompt** are used in specific settings. The placeholders {type}, {domain}, and {label_list} are dynamically filled based on the dataset (e.g., {type} could be "species", {domain} could be "bird", and {label_list} would be the list of candidate labels).

Type	Template
Base prompt	`What is the {type} of this {domain} in the image?`
Format prompt	`Output the thinking process in <think></think> and final answer in <answer></answer> tags. Please strictly follow the format.`
MCQ prompt	`Answer only from the following list: {label_list}.`

Evaluation and Baselines We report results under both Open-Ended QA and closed-set MCQ settings. We compare against zero-shot inference, answer-only supervised fine-tuning (SFT), and the tag-constrained think-then-answer baseline ViRFT (Liu et al., 2025) . All fine-tuned models are trained with open-ended prompts; for MCQ evaluation, ViRFT receives the candidate-list prompt directly, whereas our method uses the two-stage pipeline described in Section 3.5. Reproduction and verification details are provided in Appendices A.2 and A.3.

Table 2: Performance comparison of fine-grained classification accuracy (%) across five datasets under the Open-Ended QA setting. "Ours" denotes our method. Parenthesized values report absolute percentage-point differences from the indicated reference. The best results are highlighted in bold.

Model	Oxford Flowers	CUB Birds	Oxford Pets	Stanford Dogs	FGVC Aircrafts	Average
Qwen2.5-VL-7B						
Format Answer	**63.54**	51.00	70.65	**64.94**	52.54	60.53
Free Answer	63.26 (-0.28)	**54.16** (+3.16)	**76.75** (+6.10)	64.53 (-0.41)	**56.68** (+4.14)	**63.08** (+2.55)
			1-shot			
SFT	51.24	43.82	70.46	56.61	59.47	56.32
ViRFT (Liu et al., 2025)	62.20	59.91	**79.56**	**72.20**	63.10	67.39
Ours (Δ vs. ViRFT)	**69.79** (+7.59)	**60.94** (+1.03)	78.55 (-1.01)	69.31 (-2.89)	**65.92** (+2.82)	**68.90** (+1.51)
			2-shot			
SFT	58.95	46.30	61.98	66.44	65.64	59.86
ViRFT (Liu et al., 2025)	72.79	64.15	**82.53**	72.49	64.57	71.31
Ours (Δ vs. ViRFT)	**76.53** (+3.74)	**65.30** (+1.15)	81.06 (-1.47)	**74.69** (+2.20)	**67.30** (+2.73)	**72.98** (+1.67)
			4-shot			
SFT	**83.96**	55.28	75.93	74.90	70.45	72.10
ViRFT (Liu et al., 2025)	76.82	69.23	**88.77**	77.06	67.81	75.94
Ours (Δ vs. ViRFT)	80.80 (+3.98)	**69.88** (+0.65)	86.29 (-2.48)	**78.48** (+1.42)	**71.50** (+3.69)	**77.39** (+1.45)

Implementation Details We use Qwen2.5-VL-7B (Bai et al., 2025b) as the base model and train with GRPO in the easyr1 framework (Zheng et al., 2025; Sheng et al., 2024). Qwen3-30B-A3B-Instruct (Yang et al., 2025) serves as the reward judge. Our method and ViRFT use identical training configurations for a fair comparison. Prompt templates are shown in Table 1, with full hyperparameters and deployment details deferred to Appendix A.2.

4.2 Free Answer Enhances Open-Ended Recognition

We first analyze the experimental results under the Open-Ended QA setting (Table 2), which directly evaluates the model's unconstrained visual perception and generation capabilities. In the zero-shot pilot study with Qwen2.5-VL-7B, removing the rigid `<think>` and `<answer>` tags increases the average accuracy from 60.53% to 63.08%, an absolute gain of 2.55 percentage points. The gains are particularly pronounced on Oxford Pets ($+6.10$), FGVC-Aircraft ($+4.14$), and CUB Birds ($+3.16$). This result indicates that format constraints generally impede open-ended fine-grained recognition. The same overall advantage carries over to RL fine-tuning. Our Free-to-Answer method achieves average accuracies of 68.90%, 72.98%, and 77.39% in the 1-shot, 2-shot, and 4-shot settings, respectively, outperforming ViRFT by 1.51, 1.67, and 1.45 percentage points. It also exceeds the corresponding SFT averages by 12.58, 13.12, and 5.29 points. At the dataset level, our method

Table 3: Performance comparison of fine-grained classification accuracy (%) across five datasets under the MCQ setting. "Ours" denotes our method. Parenthesized values report absolute percentage-point differences from the indicated reference. The best results are highlighted in bold. For ViRFT (Liu et al., 2025), we add the candidate answer options as additional input to the trained model during inference.

Model	Oxford Flowers	CUB Birds	Oxford Pets	Stanford Dogs	FGVC Aircrafts	Average
Qwen2.5-VL-7B						
Format Answer	71.62	55.07	**83.02**	64.81	52.51	65.41
Free Answer	**78.03** (+6.41)	**58.66** (+3.59)	81.28 (-1.74)	**70.00** (+5.19)	**63.97** (+11.46)	**70.39** (+4.98)
1-shot						
ViRFT (Liu et al., 2025)	71.21	60.56	**86.59**	72.14	58.03	69.71
Ours (Δ vs. ViRFT)	**82.62** (+11.41)	**65.15** (+4.59)	86.26 (-0.33)	**74.52** (+2.38)	**67.78** (+9.75)	**75.27** (+5.56)
2-shot						
ViRFT (Liu et al., 2025)	76.82	63.12	**88.44**	71.61	58.72	71.74
Ours (Δ vs. ViRFT)	**84.86** (+8.04)	**68.86** (+5.74)	87.68 (-0.76)	**77.51** (+5.90)	**68.56** (+9.84)	**77.49** (+5.75)
4-shot						
ViRFT (Liu et al., 2025)	79.86	65.77	88.77	74.90	64.51	74.76
Ours (Δ vs. ViRFT)	**85.95** (+6.09)	**72.26** (+6.49)	**89.26** (+0.49)	**79.58** (+4.68)	**72.16** (+7.65)	**79.84** (+5.08)

improves over ViRFT on three of five benchmarks in the 1-shot setting and four of five benchmarks in both the 2-shot and 4-shot settings. Notable gains include $+7.59$ points on Oxford Flowers in the 1-shot setting, $+3.74$ points on Oxford Flowers and $+2.20$ points on Stanford Dogs in the 2-shot setting, and $+3.69$ points on FGVC-Aircraft in the 4-shot setting. Although our method remains below ViRFT on Oxford Pets across the few-shot settings and on Stanford Dogs in the 1-shot setting, its consistently higher average accuracy demonstrates the overall effectiveness of removing output-format constraints during RL fine-tuning.

4.3 Two-Stage Closed-Set Inference

While open-ended evaluation accurately reflects a model's intrinsic visual understanding, it inevitably suffers from a "perception-naming gap". Even when the model correctly perceives visual features, it frequently produces semantic aliases or exhibits granularity mismatch. To bridge this gap and adapt unconstrained MLLMs for structured downstream applications, we propose a two-stage decoupled inference pipeline. As detailed in Section 3.5, instead of forcing the MLLM to adhere to a strict classification format while performing visual reasoning, we decouple the recognition process from the final label grounding. Table 3 validates the remarkable effectiveness of this paradigm. For the zero-shot base model, applying decoupled mapping on top of the Free Answer achieves 70.39%, significantly outperforming the single-stage Format Answer + MCQ (65.41%). This confirms that separating visual extraction from strictly formatted classification drastically improves performance. Furthermore, when evaluating our RL-trained models under this decoupled pipeline, we achieve average absolute gains of $+5.56\%$ (1-shot), $+5.75\%$ (2-shot), and $+5.08\%$ (4-shot) over ViRFT (Liu et al., 2025). These results conclusively demonstrate that by interleaving an unconstrained reasoning stage with a structured grounding stage, our method effectively unlocks the potential of format-free MLLMs for fine-grained visual recognition, while also ensuring compatibility with standardized expert systems that require deterministic outputs.

Why does the decoupled paradigm excel? Across the five datasets under 4-shot settings, the MCQ stage yields an overall correction rate of 10.40% for cases that fail open-ended evaluation. Among the corrected cases, 12.37% arise from semantic aliases or lexical variations, 72.12% from granularity mismatch or evidence grounding, and 15.51% from closed-set elimination. These results reveal three primary mechanisms: (1) **Resolving Semantic Aliases and Lexical Variations:** The unconstrained MLLM often generates valid alternative names or morphological variants that fail rigid matching, such as "water lilies" versus "water lily." The MCQ stage maps these variants to the standardized ontology. (2) **Bridging Granularity Mismatch via Evidence Grounding:** The model may identify an object at a different taxonomic tier (e.g., "Boeing 707-328C" instead of "Boeing 707-320"), while its free-form visual description still provides sufficient evidence for the secondary prompt to select the intended class. (3) **Error Mitigation via Closed-Set Elimination:**

Table 4: Performance comparison of fine-grained classification accuracy (%) across five datasets using Qwen3-VL-8B-Instruct as the base model.

Model	Oxford Flowers	CUB Birds	Oxford Pets	Stanford Dogs	FGVC Aircrafts	Average
Qwen3-VL-8B under open-ended QA						
Format Answer	56.39	36.50	59.99	51.05	24.96	45.78
Free Answer	**61.63** (+ 5.24)	**37.26** (+ 0.76)	64.57 (+ 4.58)	**54.00** (+ 2.95)	**35.43** (+ 10.47)	**50.58** (+ 4.80)
Qwen3-VL-8B under MCQ						
Format Answer	74.79	**45.84**	82.39	61.17	35.97	60.03
Free Answer	**82.64** (+ 7.85)	44.27 (- 1.57)	**85.25** (+ 2.86)	**61.41** (+ 0.24)	**38.01** (+ 2.04)	**62.32** (+ 2.29)

When an initially predicted name is absent from the candidate list, the model re-evaluates its generated evidence against valid labels, narrowing the error space and enabling self-correction. Detailed statistics and additional examples are provided in Appendix A.9 and Figure 6.

4.4 Generalization to Additional Models

To further validate the generalizability of our approach, we extend our experiments using Qwen3-VL-8B-Instruct as the base policy model. First, we conduct the same zero-shot evaluation on the 5 datasets under both open-ended and MCQ settings. As detailed in Table 4, the format-free paradigm consistently outperforms the structurally constrained baseline across both evaluation settings, yielding average improvements of +4.80% in Open-Ended QA and +2.29% in MCQ. This demonstrates that the benefits of liberating format constraints are not limited to a specific model architecture, but can be observed across different MLLM variants.

Table 5: Open-ended QA accuracy (%) of Qwen3-VL-8B-Instruct before and after Free-to-Answer training.

Setting	Flowers	Dogs	Pets
Before RFT	61.63	54.00	64.57
After RFT	**68.17**	**58.26**	**73.97**

Table 6: Ablation of trainable components in Free-to-Answer.

Setting	Flowers	Birds	Pets	Dogs	Aircraft	Avg.
Language Only	66.42	57.27	77.21	68.53	61.39	66.16
Full Model	**69.79**	**60.94**	**78.55**	**69.31**	**65.92**	**68.90**

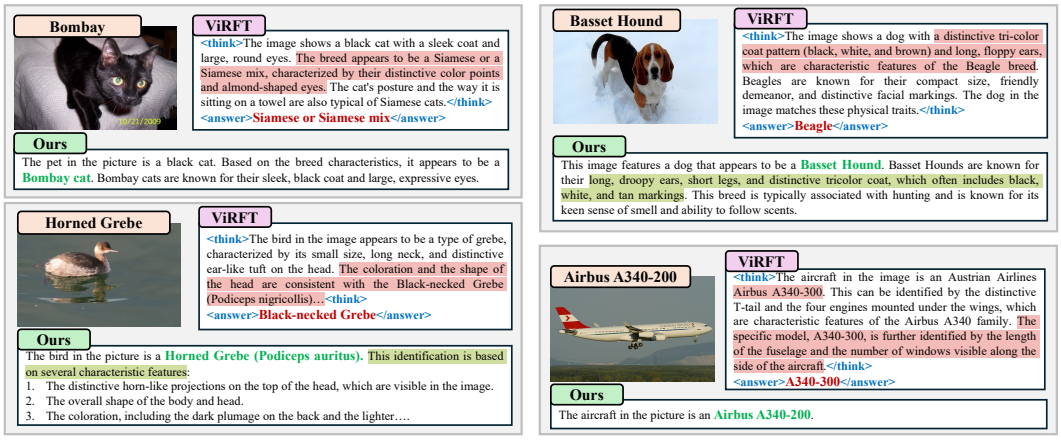


Figure 3: Qualitative examples comparing Free-to-Answer (ours) with ViRFT on fine-grained visual recognition tasks.

Furthermore, we apply our Free-to-Answer training to the Qwen3-VL-8B-Instruct (Yang et al., 2025) model and repeat the evaluation on the Oxford Flowers, Stanford Dogs, and Oxford Pets datasets. The results are shown in Table 5. We observe that the format-free RL training still provides significant improvements over the zero-shot baseline, with an average improvement of +6.73%

Table 7: Controlled comparison of output-format constraints under the 1-shot setting.

Dataset	ViRFT	Controlled Format-RL	Free-to-Answer
Birds	59.91	59.82	**60.94**
Flowers	62.20	62.97	**69.79**

Table 8: Reward ablation on FGVC-Aircraft under the 4-shot setting.

Evaluation	String Matching	LLM Based
Open-Ended QA	68.95	**72.00**
MCQ	70.27	**72.16**

in the Open-Ended QA evaluation. This confirms that the efficacy of the Free-to-Answer framework is not restricted to a specific model and can be effectively applied to enhance the performance of various MLLM architectures in fine-grained visual recognition tasks.

4.5 Visualization

Figure 3 presents qualitative comparisons between our format-free RL approach and ViRFT (Liu et al., 2025). Each row shows an image from a different dataset alongside the predictions from both models. Our method generates concise and natural reasoning, adapting the depth of explanation as needed, without being restricted by rigid formatting. For example, on the Airbus A340-200 image, our model directly answers "The aircraft in the picture is an Airbus A340-200." For the Horned Grebe image, our approach not only correctly identifies the species but also provides a succinct rationale. These examples illustrate that removing strict formatting enables the model to focus its generative capacity on visual discrimination.

4.6 Ablation Studies

Isolating structural format constraints. To separate the effect of structural prompting from reward design, we introduce a controlled Format-RL variant that retains the structured prompt but removes the format reward and uses the same LLM-based verifier as Free-to-Answer. As shown in Table 7, Free-to-Answer performs best on both datasets, supporting the conclusion that relaxing structural output constraints during RL provides gains beyond the verifier itself.

Roles of text and vision components. We ablate the effect of updating the vision encoder by comparing "Language Only" (frozen vision) and "Full Model" RL training on the 1-shot Open-ended QA setting. Table 6 shows that freezing the vision encoder reduces average accuracy from 68.90% to 66.16%. This confirms that adapting both vision and language components is important for fine-grained recognition.

Effect of LLM-based reward. We evaluate the necessity of our proposed LLM-based text-to-label reward mechanism by comparing it against a simple string-matching baseline. The string-matching reward yields a positive signal only if the generated response contains the ground-truth category string. As shown in Table 8, our LLM-based reward outperforms this heuristic approach across both evaluation settings. This performance gap demonstrates that while string matching is computationally cheap, simple keyword inclusion does not guarantee logical correctness. In contrast, the LLM evaluation acts as a semantically aware judge, providing a more robust signal for policy optimization. A human audit of 1,000 responses in Appendix A.4 further shows 99.4% agreement for the complete extraction-and-verification pipeline.

5 Conclusion

This paper investigates how output-format constraints affect MLLMs in fine-grained visual recognition. Motivated by the finding that removing mandatory structural tags generally improves recognition accuracy, we introduce Free-to-Answer, a reinforcement fine-tuning scheme that optimizes free-form responses through LLM-based label extraction and deterministic verification. Experiments show improvements over supervised and format-constrained baselines, while further analysis illustrates how two-stage closed-set grounding can correct errors in free-form predictions.

AI USE STATEMENT

We used OpenAI Codex to assist with interpreting and presenting already-computed experimental results. We additionally used it to draft and edit portions of the manuscript for readability. The research questions, methodology, and experimental protocol were determined by the authors, and this assistance tool was not used to generate or alter datasets, experimental measurements, or model outputs reported as experimental results. All AI-assisted prose, calculations, and source-code changes were reviewed by the authors against the underlying experimental results. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses established, publicly available fine-grained visual recognition benchmarks. We use these datasets only for their intended research purposes and cite their original sources. The study does not involve human subjects, private or sensitive personal data, or other ethical concerns.

REPRODUCIBILITY STATEMENT

The implementation details required to reproduce our results are provided in Sections 3, 3.4, 3.5, and 4, as well as Appendices A.1, A.2, and A.3. In addition, we provide the source code in the supplementary materials.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training. In *arXiv*, 2025.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Hulingxiao He, Zijun Geng, and Yuxin Peng. Fine-r1: Make multi-modal llms excel in fine-grained visual recognition by chain-of-thought reasoning. *arXiv preprint arXiv:2602.07605*, 2026.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

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

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. *Advances in Neural Information Processing Systems*, 38:874–899, 2026.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

Jeonghwan Kim and Heng Ji. 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*, pp. 6187–6207, 2024.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Logan Lawrence, Oindrila Saha, Megan Wei, Chen Sun, Subhransu Maji, and Grant Van Horn. You may speak freely: Improving the fine-grained visual recognition capabilities of multimodal large language models with answer extraction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1428–1437, 2026.

Hengtao Li, Pengxiang Ding, Runze Suo, Yihao Wang, Zirui Ge, Dongyuan Zang, Kexian Yu, Mingyang Sun, Hongyin Zhang, Donglin Wang, et al. Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators. *arXiv preprint arXiv:2510.00406*, 2025a.

Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. To think or not to think: A study of thinking in rule-based visual reinforcement fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025c.

Xiongkun Linghu, Jiangyong Huang, Baoxiong Jia, and Siyuan Huang. 3d-rft: Reinforcement fine-tuning for video-based 3d scene understanding. *arXiv preprint arXiv:2603.04976*, 2026.

Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *International conference on learning representations*, volume 2024, pp. 16942–16973, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

Xinyu Ma, Ziyang Ding, Zhicong Luo, Chi Chen, Zonghao Guo, Derek F Wong, Xiaoyi Feng, and Maosong Sun. Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*, 2025.

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

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18477–18494, 2025.

Minheng Ni, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Kevin Lin, Wangmeng Zuo, and Lijuan Wang. Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning. *arXiv preprint arXiv:2505.19702*, 2025.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Cong Pang, Hongtao Yu, Zixuan Chen, Lewei Lu, and Xin Lou. Towards fine-grained recognition with large visual language models: Benchmark and optimization strategies. *arXiv preprint arXiv:2512.10384*, 2025.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Hannah Sterz, Jonas Pfeiffer, and Ivan Vulić. Dare: Diverse visual question answering with robustness evaluation. *Transactions of the Association for Computational Linguistics*, 13:1121–1145, 2025.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *arXiv preprint arXiv:2408.02442*, 2024.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv e-prints*, pp. arXiv–2503, 2025.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Tech. Report CNS-TR-2011-001*, 2011.

Darren Yow-Bang Wang, Zhengyuan Shen, Soumya Smruti Mishra, Zhichao Xu, Yifei Teng, and Haibo Ding. Slot: Structuring the output of large language models. *arXiv preprint arXiv:2505.04016*, 1(2):3, 2025.

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

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, Yuwen Xiong, and Richong Zhang. Easyr1: An efficient, scalable, multi-modality rl training framework. `https://github.com/hiyouga/EasyR1`, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Appendix

A.1 Dataset Details

Comprehensive statistics for the fine-grained visual recognition (FGVR) benchmarks utilized in this work are detailed in Table 9. For each dataset, we report the total number of fine-grained classes and the size of the original test split used for evaluation. Under the few-shot setting, we randomly sample $K \in \{1, 2, 4\}$ images per class from the training split. Training uses only these images and their corresponding category labels, without additional annotations.

A.2 Training Details

We use Qwen2.5-VL-7B (Bai et al., 2025b) as the base policy model and implement all RL experiments with easyr1 (Zheng et al., 2025; Sheng et al., 2024), an efficient multimodal RL framework based on veRL. We adopt Group Relative Policy Optimization (GRPO), where multiple responses are sampled per prompt and optimized using group-relative rewards and a PPO-style clipped objective. Table 10 lists the complete hyperparameters. All policy-training experiments are conducted on 8 NVIDIA A6000 GPUs (8×48GB).

GRPO Configuration. For each prompt, we sample $G = 8$ responses to form a group and compute relative advantages within the group. We adopt a KL-regularized objective to stabilize training, with a coefficient $\beta = 0.01$.

Optimization. The policy model is optimized using AdamW with a learning rate of 1×10^{-6} and a weight decay of 0.01. The model is trained for 4 epochs.

Supervised Fine-Tuning. Because the datasets do not provide step-by-step reasoning annotations, the SFT baseline is trained on answer-only targets using standard cross-entropy loss. We perform full-model fine-tuning for the same number of optimization steps as the RL methods.

ViRFT Implementation. For a controlled comparison, we reproduce ViRFT (Liu et al., 2025) using the same EasyR1 pipeline and hyperparameters as Free-to-Answer. Following DiVE-K (Kumar et al., 2025), we retain only the `ground_truth in answer` matching rule.

Table 9: Dataset Statistics.

Dataset	Classes	Test Samples
CUB-200-2011	200	5,794
Stanford Dogs	120	8,580
Oxford Flowers	102	2,463
FGVC-Aircraft	100	3,333
Oxford-IIIT Pet	37	3,669

Table 10: Hyperparameters for Free-to-Answer RL training.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	1×10^{-6}
Weight Decay	0.01
GRPO Group Size (G)	8
KL Coefficient (β)	0.01
Rollout Temperature	1.0
Top-p	1.0
Global Batch Size	4
Max Prompt Length	1024
Max Response Length	1024
Training Epochs	4
Maximum Training Steps	240

A.3 Evaluation Pipeline Details

For the **Open-Ended QA** evaluation, we use Qwen3-30B-A3B-Instruct (Yang et al., 2025) as the judge model to extract deterministic labels from unconstrained MLLM responses, followed by the string verification process described in Section 3.4. The judge is deployed with vLLM-Serve (Kwon et al., 2023) on 2 NVIDIA A6000 GPUs. Its prompt explicitly requests one category from the dataset label set, as shown in Figure 4. The placeholders {`type`}, {`domain`}, and {`label_list`} are filled according to the dataset; for CUB-200-2011, they correspond to "species," "bird," and the list of 200 bird species, respectively.

For the **Multiple Choice Question (MCQ)** evaluation, we follow a two-stage inference pipeline: (1) The MLLM generates a free-form visual deduction based solely on the image and base query; (2) Provide the generated response along with the candidate list to the model again to select the best-matching label. The second-stage prompt is designed to encourage the model to directly compare its initial response with the candidate labels and make a final selection based on semantic alignment. The prompt is shown in Figure 5.

```
What is the most likely {type} of the {domain} indicated in the
following response?
Response: {model_response}
If there is no answer, output exactly: None
Allowed labels: {label_list}
Answer using only labels from the following list or ''None'', with no
other text.
```

Figure 4: Prompt template for label extraction in Open-Ended QA evaluation.

```
What is the most likely {type} of the {domain} indicated in the
following response?
Response: {model_response}
Answer only from the following list with no other text: {label_list}
```

Figure 5: Prompt template for the second-stage MCQ evaluation.

A.4 Judge Reliability and Verification

We assess the text-to-label evaluator on 1,000 randomly sampled responses using human annotations as the reference. As summarized in Table 11, the LLM-only extractor achieves 98.7% agreement, whereas the complete extraction-and-verification pipeline reaches 99.4% and reduces false-positive decisions from 11 to 1. String verification is applied only after generation and therefore does not constrain the policy response. Before verification, we normalize superficial formatting differences, such as hyphens and whitespace, so equivalent surface forms are not rejected.

Table 11: Agreement with human annotations on 1,000 sampled responses. FP and FN denote false-positive and false-negative decisions.

Evaluator	**Agreement (%)**	**FP**	**FN**
LLM-only extraction	98.7	11	2
Extraction + verification	**99.4**	**1**	5

The reward signal is also robust to a smaller verifier. On Aircraft 1-shot Open-Ended QA, replacing Qwen3-30B-A3B-Instruct with Qwen2.5-VL-7B as the judge yields 66.73% accuracy, comparable to 65.92% with the default judge. This result suggests that reward verification mainly requires reliable extraction within a fixed label space rather than a large dense model.

A.5 Generalization to Additional Models

We also test whether the zero-shot effect persists at a larger policy scale using Qwen2.5-VL-32B. Table 12 shows that free-form prompting improves four of the five datasets and raises average accuracy from 53.41% to 57.37% (+3.96 percentage points). The only exception is CUB, where the format-constrained prompt is 1.60 points higher.

Table 12: Zero-shot Open-Ended QA accuracy (%) with Qwen2.5-VL-32B.

Prompt	**Flowers**	**CUB**	**Pets**	**Dogs**	**Aircraft**	**Avg.**
Format Answer	61.71	**49.11**	60.07	60.27	35.91	53.41
Free Answer	**64.72**	47.51	**70.37**	**61.26**	**42.99**	**57.37**

A.6 Response Length before and after RL

To investigate the impact of removing structural constraints on the model's generation behavior, we compare the average response length (in number of tokens) before and after 4-shot RL training

for our Free-to-Answer model and the baseline ViRFT model. The results are shown in Table 13. As observed, our Free-to-Answer approach exhibits a consistent but moderate increase in response length across all five datasets after RL training. For instance, on the CUB dataset, the average length increases from 116.61 to 131.11 tokens. In contrast, the ViRFT model, which enforces strict structural constraints, shows inconsistent length dynamics. On datasets like CUB and Stanford Dogs, the response length actually decreases after RL training (from 108.96 to 86.00 on CUB, and 93.49 to 87.34 on Dogs).

Table 13: Average response length (in tokens) before and after 4-shot RL training across five datasets. Δ denotes After RL minus Before RL.

Dataset	Ours			ViRFT		
	Before RL	After RL	Δ	Before RL	After RL	Δ
Flowers	86.99	93.29	+6.30	106.37	112.87	+6.50
Birds	116.61	131.11	+14.50	108.96	86.00	-22.96
Pets	42.96	51.61	+8.65	86.52	94.87	+8.35
Dogs	62.32	74.71	+12.39	93.49	87.34	-6.15
Aircraft	108.34	120.66	+12.32	119.88	121.55	+1.67

A.7 Visual Grounding Audit

To examine whether unconstrained generation encourages unsupported explanations, we manually verify 2,000 correctly classified responses (400 per dataset), assisted by a vision-language model for initial screening. We find inconsistent visual details in only 4 Free-to-Answer responses and 6 ViRFT responses. Together with the moderate response-length changes in Table 13, this audit indicates that the performance gains are not primarily explained by verbosity or ungrounded reasoning.

A.8 Computational Cost Analysis

The training-time judge is Qwen3-30B-A3B-Instruct, a mixture-of-experts model with approximately 3B activated parameters per inference, served on two NVIDIA A6000 GPUs. Its reward computation accounts for only 0.024% of the measured step time; rollout generation, policy updates, and old/reference log-probability computation account for 19.8%, 58.9%, and 21.1%, respectively. Table 14 further shows that the end-to-end training cost of Free-to-Answer is comparable to ViRFT rather than being dominated by the judge.

Table 14: End-to-end training time (minutes) under identical hardware settings.

Dataset	Method	2-shot	4-shot
Birds	ViRFT	124.5	250.7
	Free-to-Answer	102.8	202.2
Dogs	ViRFT	102.0	228.0
	Free-to-Answer	111.0	246.9

The external judge is not used at inference time. For MCQ evaluation, Free-to-Answer only adds a lightweight label-grounding stage. We measure 500 examples per dataset on one NVIDIA A6000 GPU with batch size 1. As shown in Table 15, the additional stage takes about 11 ms per example; total latency is moderately higher than ViRFT on CUB but lower on Aircraft because generation length varies by dataset.

Table 15: Average MCQ inference latency per sample (ms), measured over 500 examples with batch size 1.

Dataset	Stage 1	Stage 2	Ours Total	ViRFT
CUB	2685.60	10.63	2696.23	2229.10
Aircraft	3021.70	11.29	3032.99	3927.20



Figure 6: Additional qualitative examples where the model's response is incorrect under the Open-Ended QA evaluation but correct under the MCQ evaluation.

A.9 Analysis of Errors Corrected by Closed-Set Grounding

We analyze the prediction changes introduced by the second-stage MCQ grounding procedure under the 4-shot setting. A corrected case is an instance that is judged incorrect under Open-Ended QA evaluation but is correctly classified after the free-form response is grounded to the closed-set label space. Across the five datasets, Open-Ended QA produces 5,517 incorrect predictions, of which 574 are corrected by the MCQ stage.

This correction rate measures the proportion of open-ended errors recovered by closed-set grounding and should not be interpreted as the net accuracy improvement. The MCQ stage also changes 90 initially correct predictions into incorrect ones, resulting in a net gain of $574 - 90 = 484$ correct predictions. Table 16 reports the corresponding dataset-level statistics. The overall correction rate

is computed from the aggregated numbers of open-ended errors and corrections, rather than by averaging the five dataset-level rates.

Table 16: Dataset-level prediction transitions introduced by the second-stage MCQ grounding procedure under the 4-shot setting. "Corrections" denotes incorrect open-ended predictions that become correct after MCQ grounding, whereas "Regressions" denotes initially correct predictions that become incorrect. The correction rate is computed over open-ended errors.

Dataset	Open-Ended Errors	Corrections	Regressions	Correction Rate
Aircraft	950	42	20	4.42%
CUB-200	1,745	138	0	7.91%
Dogs	1,846	123	29	6.66%
Flowers	473	165	38	34.88%
Pets	503	106	3	21.07%
Total	**5,517**	**574**	**90**	**10.40%**

To characterize how closed-set grounding recovers open-ended errors, we assign each of the 574 corrected cases to one of three mutually exclusive categories according to its primary correction pattern. As summarized in Table 17, most corrected cases involve granularity mismatch or evidence-based grounding.

Table 17: Distribution of correction patterns among the 574 cases recovered by the second-stage MCQ grounding procedure.

Correction pattern	Count	Share
Semantic aliases or lexical variations	71	12.37%
Granularity mismatch or evidence-based grounding	414	72.13%
Closed-set elimination	89	15.51%
Total	**574**	**100.00%**

Semantic-alias cases arise when the free-form response refers to the target category using an alternative name or lexical variant that is not accepted by open-ended evaluation. Granularity-mismatch and evidence-grounding cases arise when the initial response predicts a related category at a different level of specificity, or when its descriptive evidence enables the second stage to recover the target label. Closed-set elimination cases occur when the initial prediction is incompatible with the candidate set, but comparison with the available labels leads to the correct category. Representative examples are provided in Figure 6.

A.10 Additional Qualitative Analysis

Figure 6 provides a qualitative analysis illustrating how the decoupled two-stage evaluation rectifies Open-Ended QA failures. These examples highlight three primary patterns: (1) **Resolving Semantic Aliases**: The MLLM often generates accurate synonyms or lexical variations (e.g., "Water lilies" vs. "Water lily", or "California poppy" vs. "Californian poppy") that rigid string-matching fails to recognize. The MCQ stage successfully maps these semantically equivalent terms to the dataset's ontology. (2) **Bridging Granularity Mismatches**: The model may output a more specific taxonomic prediction (e.g., "Boeing 707-328C" instead of "Boeing 707-320") that does not exactly match the label. The MCQ stage utilizes the candidate list to accurately associate granular answers with the corresponding class. (3) **Error Mitigation via Closed-Set Elimination**: When the model produces visually similar but incorrect predictions (e.g., "Rhododendron" instead of "Azalea"), the absence of the erroneous term from the candidate list compels the model to re-evaluate its reasoning against the available valid options. Overall, these cases demonstrate that the MCQ stage acts as a robust semantic bridge, ensuring that the model's fine-grained visual reasoning is more accurately reflected in the final classification metrics.