Can Textual Reasoning Improve the Performance of MLLMs on Fine-grained Visual Classification?

Supplementary Material

Dataset	#Categories	Train	4-shot (%)	Test
Aircraft–102	100	3 334	400 (12.0%)	3 333
Flower–102	102	1 020	408 (40.0%)	2 463
Pet–37	37	3 680	148 (4.0%)	3 669
Car–196	196	8 144	784 (9.6%)	8 041

Table 6. Statistics of FGVC datasets. The "4-shot" column shows the number of images we used for training.

Answer-only Prompt

This is an image containing an {DATASET}. Please identify the {DATASET} of the {DATASET} based on the image. Only provide the final answer directly, without any explanation or special formatting.

Figure 9. *Answer-only* prompt.

7. GRPO Algorithm

GRPO requires the model to sample G diverse responses $\{o_1, o_2, \ldots, o_G\}$ from the current model π_θ and obtains rewards $\{r_1, r_2, \ldots, r_G\}$ for o_i. GRPO assesses the relative quality by normalizing r_i using the mean and standard deviation of the group reward (details provided in the main paper). With the group normalization, GRPO encourages the model to sample preferred answers with a higher reward. The model is updated via:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} A_i, \text{clip}\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\varepsilon, 1+\varepsilon \right) A_i \right) - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \right] \tag{5}$$

where ε and β are the GRPO clipping hyperparameters and the coefficient weight for controlling the Kullback–Leibler (KL) penalty [42], respectively. π_{ref} is the reference model.

8. Additional Implementation Details

8.1. Datasets

Statistics of Training and Evaluation Set. We use the 4-shot data provided by [33]. The statistics of the training set and evaluation set can be found in Tab. 6.

Prompts. Fig. 9 and Fig. 10 show the prompts for *Answer-only* and *CoT*, respectively, while Fig. 11 provides the prompt for the MLLM-based accuracy reward. The *Answer-only* prompt is used for SFT training, and the *CoT* prompt for both CoT-SFT and RFT training. Placeholders DATASET, PRED, and GT are used to denote the specific dataset name (*e.g.*, plants, aircrafts), the model's predicted label, and the ground truth label, respectively.

Reward Model Implementation. We deploy Qwen2-VL-7B [58] as the reward model for MLLM-based accuracy reward using LMDeploy [7]. To optimize performance, we employ mixed precision and the TurboMind inference backend. LMDeploy provides a flexible framework with OpenAI-compatible APIs, ensuring broad compatibility and facilitating the potential integration of other teacher models in the future.

CoT Data Curation. We employ GPT-4o-2024-08-06 [20] to generate high-quality Chain-of-Thought data. For each sample, we provide the image, question prompt, and corresponding ground truth label, instructing the model to generate reasoning that leads to the correct answer. This ensures the accuracy of the synthesized CoT data. The prompt template shown in Fig. 12 uses SOLUTION as a placeholder for the ground truth label, while Fig. 13 displays representative examples of the generated data.

Additional Training Implementation Details. To ensure reproducibility, all experiments use fixed random seeds. We employ BF16 precision and apply LoRA with a rank $\gamma = 64$ and scaling parameter $\alpha = 128$ to the following modules: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`. Models are trained for a maximum of 200 steps with a completion length capped at 256 tokens.

Performance Validation Details. Performance is evaluated exclusively by answer accuracy. For *CoT*, we follow [33] to extract answers from the `<answer>...</answer>` tag. A prediction is considered correct if a normalized substring match exists in either direction between the extracted answer and the ground truth. For *Answer-only*, responses are evaluated directly, as their output format is inherently comparable to the ground truth.

CoT Prompt

This is an image containing an {DATASET}. Please identify the {DATASET} of the {DATASET} based on the image. Output the thinking process in `<think>...</think>` and final answer in `<answer>...</answer>` tags. The output answer format should be as follows: `<think>...</think> <answer>species name</answer>`. Please strictly follow the format."

Figure 10. *CoT* prompt.

Judge Prompt

You are a scoring assistant. Based on the similarity between the "Predicted Answer" and the "Correct Answer", provide a score from 0 to 10. A score of 10 means a perfect match, and 0 means a complete mismatch. You must output only the numerical score.
—
[Example 1]
Predicted Answer: "2007 Dodge Dakota Club Cab"
Correct Answer: "2007 Dodge Dakota Club Cab"
Score: 10
—
[Example 2]
Predicted Answer: "Boeing 707"
Correct Answer: "707-320"
Score: 6
—
[Example 3]
Predicted Answer: "Nasturtium"
Correct Answer: "watercress"
Score: 0
—
[Your Task]
Predicted Answer: "{PRED}"
Correct Answer: "{GT}"
Score:

Figure 11. Judge prompt for MLLM-based accuracy reward.

9. Additional Experimental Results

Additional results on prompt-type choices for RFT. We further study how the prompt type used during RFT affects performance. Following Visual-RFT [33], we train both the chain-of-thought (*CoT*) and answer-only (*Answer-only*) variants under the same setting. As reported in Tab. 8, the two variants achieve almost identical accuracies on Aircrafts-102 and Cars-196, suggesting that explicitly generating long CoT traces brings little additional benefit beyond an answer-only prompt, which is consistent with our comparison between Visual-RFT [33] and No-Thinking-RFT [27].

Extending to Other FGVC Tasks. Fig. 14 shows that the Cost of Thinking exists in the medical imaging dataset VQA-Rad.

Comparisons against discriminative models. We show the comparison in Tab. 7. Compared with CLIP and its discriminative variants, our method achieves the best performance on Cars and Aircrafts, improving over CLIP^{LP} by a large margin (+10.4% and +19.8%, respectively). This suggests that our method is particularly effective on fine-grained categories with subtle inter-class differences. On Flowers and Pets, however, linear probing on CLIP remains more competitive, indicating that discriminative adaptation is still advantageous on domains with relatively cleaner visual cues or stronger alignment to CLIP pre-training. Overall, these results show the promise of MLLMs on fine-grain visual understanding, and our method is highly competitive with

SFT-CoT Prompt

This is an image containing a pet. Please identify the species of the pet based on the image. Output the thinking process in <think> </think> and final answer in <answer> </answer> tags.The output answer format should be as follows: <think> ... </think> <answer>species name</answer> Please strictly follow the format. The ground truth answer is {SOLUTION}. Limit your response to 100 words.

Figure 12. Judge prompt for MLLM-based accuracy reward.

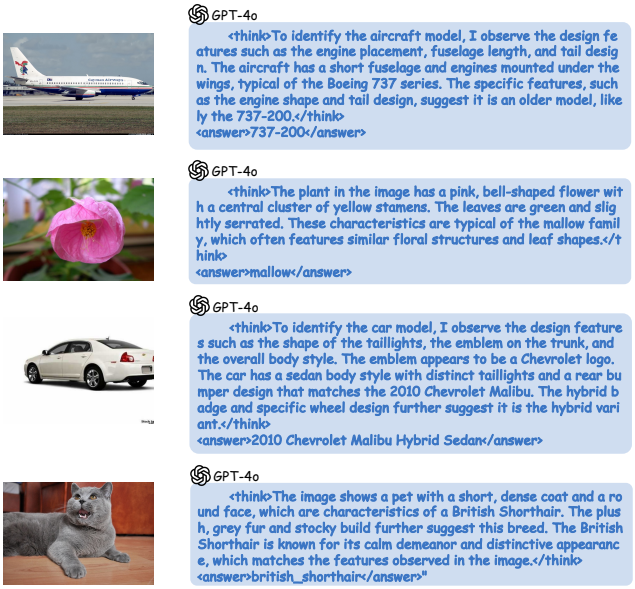

Figure 13. Examples of CoT annotations generated by GPT-4o, featuring long and fine-grained reasoning traces that describe each image in detail.

Methods	Cars	Aircrafts	Flowers	Pets
CLIP (ViT-B/16)*	65.6	27.1	70.4	88.9
CLIP^{LP} (ViT-B/16)*	86.7	59.5	98.1	93.1
$\text{CLIP}^{\text{SimNL}}$ (4-shot)	68.0	29.0	92.0	88.1
Ours	97.1	79.3	81.0	88.6

Table 7. Comparison with discriminative models. *: from CLIP official report. LP: Linear Probing. SimNL: [66].

standard discriminative baselines.

4**Correlation Analysis of Rewards.** As shown in Tab. 5 and Fig. 6, combining all rewards yields the best performance, and R_{cls}, R_{emb}, and R_{mllm} show consistent positive trends. Fig. 15 on the Flowers test set further shows that the rewards are correlated yet distinct. This indicates that these three rewards are aligned in encouraging semantically correct predictions, but they are not redundant and still provide complementary learning signals. By contrast, the format reward and thinking-length reward have much weaker correlations with the task-related rewards, suggesting that they

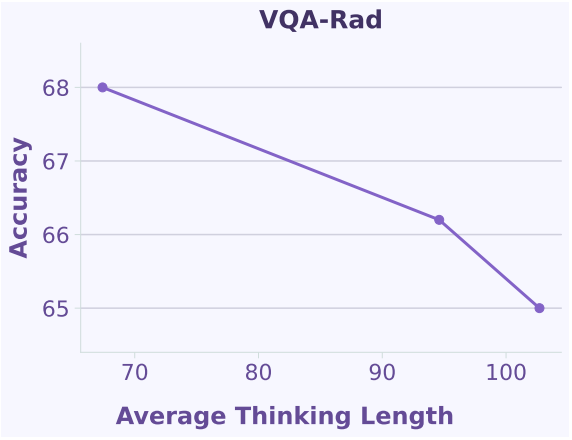

Figure 14. Cost of thinking on VQA-Rad dataset. Performance decreases as the thinking length increases.

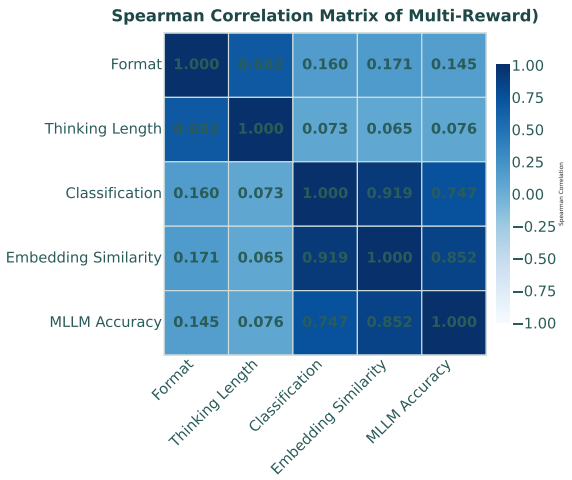

Figure 15. Correlation heatmap of ensemble rewards.

mainly regularize output structure and response behavior rather than directly optimizing classification performance. Interestingly, the format reward is relatively correlated with thinking length, implying that shorter reasoning often co-occurs with better-formatted responses.

Additional Qualitative Results. We provide additional qualitative results in Fig. 16.

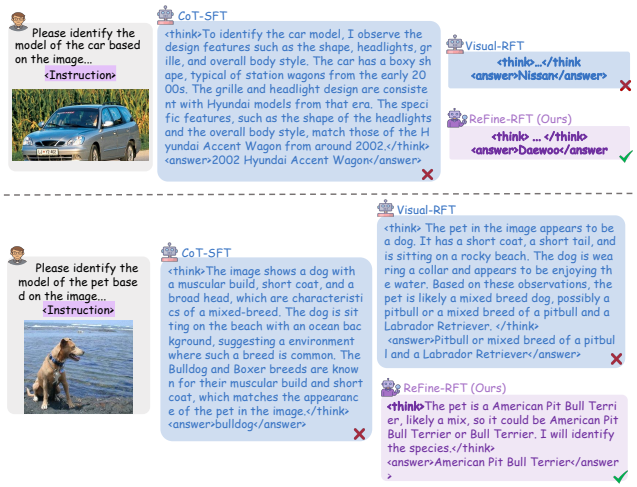

Figure 16. **Additional comparison of responses.**

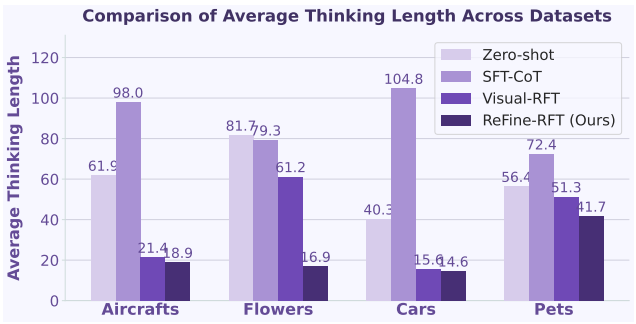

Figure 17. **Comparison of average thinking length across datasets.** We average the number of thinking tokens as the thinking length per dataset. SFT-CoT consistently yields the longest chains of thought, Zero-shot produces medium-length traces, while both Visual-RFT and especially ReFine-RFT generate much more concise reasoning. Our method attains the shortest thinking length on all datasets, indicating that strong performance does not require long reasoning traces.

Thinking Length comparison. Across all four datasets, there is a clear and consistent ordering of thinking lengths: SFT-CoT produces the longest chains of thought, as SFT-CoT data contains long reasoning traces. Zero-shot sits in the middle, while Visual-RFT substantially shortens the reasoning, and ReFine-RFT is the most concise. The gap is especially striking on Cars and Aircrafts, where SFT-CoT more than doubles or even triples the thinking length of ReFine-RFT. Combined with our empirical observation that training with longer thinking lengths actually hurts task performance, this pattern suggests that excessive CoT introduces redundancy and noise rather than useful intermediate supervision. Long SFT-CoT traces likely contain distracted or unhelpful information, which dilutes the gradient signal and encourages the model to mimic verbosity instead of learning the decision-critical answering. Zero-shot, which is not explicitly trained to be verbose, yields somewhat shorter traces and better aligns with test-time behavior, but still carries uncontrolled overthinking. This is possibly because of the pretraining data distribution. In contrast, ReFine-RFT explicitly regularize the model toward concise, high-utility rationales: we focus on accuracy-centric signals that are tied to the final prediction and constrain reasoning tokens. This not only reduces token cost, but empirically correlates with higher accuracy, suggesting that there is an optimal, concise reasoning horizon, and that pushing the model to produce ever-longer CoT drives it into a worse visual perception performance.

10. Potential Reasons of CoT Degradation on Visual Tasks.

We hypothesize that the observed "Cost of Thinking" arises from two interacting effects. First, long textual chains-of-thought may compete with visual processing for the model's finite attention and context budget: as more self-generated tokens accumulate, the transformer increasingly attends to its own linguistic history rather than the image embeddings, amplifying language priors while suppressing subtle visual cues that are crucial for FGVC. A closely related "attention diversion" phenomenon has been documented in instruction-following LLMs, where explicit CoT reduces focus on constraint tokens and significantly harms compliance accuracy [28], and in multimodal reasoning, where reasoning primarily in the language space leads to strong language bias and under-utilization of image features, motivating architectures that explicitly replay or re-ground visual information during reasoning [40, 56].

Second, extending the CoT sequence increases exposure to noisy or unfaithful reasoning: each additional step is an opportunity to introduce hallucinated details, spurious correlations, or incorrect intermediate visual descriptions, which are then propagated and rationalized downstream. Prior analyses of CoT on text-only tasks have shown that explanations are often unfaithful to the model's true decision process and can rationalize biased or incorrect predictions [25, 64], and that error rates grow with the number of implicit reasoning operations, consistent with a "noisy reasoning" view where

Methods	Aircrafts-102	Cars-196
Visual-RFT-AO	75.8	95.8
Visual-RFT-CoT	75.6	95.7

Table 8. **Comparison of prompt types in Visual-RFT.** Using an *Answer-only* prompt (Visual-RFT-AO) attains almost identical accuracy to using an explicit *CoT* prompt (Visual-RFT-CoT), indicating that long reasoning traces are not necessary for effective RFT on these benchmarks.

longer chains accumulate more mistakes. In fine-grained visual classification, where decisions hinge on subtle, localized perceptual evidence, such mis-grounded or noisy chains are particularly detrimental: once the CoT commits to an incorrect local description (e.g., misidentifying a part or texture), subsequent reasoning tends to reinforce that error instead of revisiting the image, making verbose CoT systematically worse than concise, answer-focused predictions.

11. Potential Social Impact

Our work advances the reasoning and fine-grained recognition capabilities of MLLMs, with the potential to significantly benefit real-world applications in domains such as biodiversity monitoring, medical diagnostics, industrial inspection, and scientific research, where expert-level fine-grained categorization is crucial. By enabling MLLMs to generate interpretable reasoning steps in addition to accurate predictions, our method promotes transparency and trustworthiness, critical factors for safe AI deployment in high-stakes environments. We believe this research contributes to the broader goal of making MLLMs more reliable, interpretable, and aligned with human values, while acknowledging the necessity of continuous ethical scrutiny as these systems become increasingly capable.

12. Limitation

While ReFine-RFT achieves strong performance on FGVC, several limitations remain. First, although suppressing excessive thinking length indirectly improves training efficiency, the overall RFT pipeline is still more time-consuming than standard SFT due to the rollout sampling strategy and RFT optimization. Second, our analysis mainly focuses on the impact of thinking length and the comparison between SFT-AO, SFT-CoT, and our RFT variants, whereas the effects of *thinking quality* during the RFT process remain unexplored. Finally, we conduct a detailed study only on fine-grained visual classification (FGVC); extending our framework and analyses to other visual tasks such as object detection, visual grounding, or more open-ended vision–language reasoning is an important direction for future work.