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

Jie Zhu[1] Yiyang Su[1] Xiaoming Liu[1,2]
[1]Michigan State University [2]University of North Carolina at Chapel Hill
{zhujie4, suyiyan1}@msu.edu liuxm@cs.unc.edu

Abstract

Multi-modal large language models (MLLMs) exhibit strong general-purpose capabilities, yet still struggle on Fine-Grained Visual Classification (FGVC), a core perception task that requires subtle visual discrimination and is crucial for many real-world applications. A widely adopted strategy for boosting performance on challenging tasks such as math and coding is Chain-of-Thought (CoT) reasoning. However, several prior works have reported that CoT can actually harm performance on visual perception tasks. These studies, though, examine the issue from relatively narrow angles and leave open why CoT degrades perception-heavy performance. We systematically re-examine the role of CoT in FGVC through the lenses of zero-shot evaluation and multiple training paradigms. Across these settings, we uncover a central paradox: the degradation induced by CoT is largely driven by the reasoning length, in which longer textual reasoning consistently lowers classification accuracy. We term this phenomenon the "Cost of Thinking". Building on this finding, we make two key contributions: (1) MRN, a simple and general plug-and-play normalization method for multi-reward optimization that balances heterogeneous reward signals, and (2) ReFine-RFT, a framework that combines ensemble rewards with MRN to constrain reasoning length while providing dense accuracy-oriented feedback. Extensive experiments demonstrate the effectiveness of our findings and the proposed ReFine-RFT, achieving state-of-the-art performance across FGVC benchmarks. Project page: ReFine-RFT.

1. Introduction

Multi-modal large language models (MLLMs) have demonstrated remarkable capabilities in general vision-language understanding, enabling seamless interaction across images and text and driving progress toward versatile, general-purpose AI systems [1, 2, 77]. As these models are increasingly deployed as unified interfaces for perception and reasoning, their ability to handle fine-grained visual understanding becomes particularly critical [14]. Fine-grained Visual Classification (FGVC) requires discriminating among subordinate-level categories that exhibit only subtle visual differences (*e.g.*, car models or plant varieties) and serves as a foundation for more advanced perception-centric tasks such as object-centric visual question answering [10, 14, 68, 70]. For example, a model unable to reliably differentiate between similar-looking pet breeds (*e.g.*, golden retriever *vs.* labrador) is also likely to fail in answering follow-up questions about their behavioral traits or health conditions. Unlike other recognition tasks [9, 17, 47, 73, 74], FGVC demands precise visual perception and sensitivity to subtle cues such as fur texture, fine-grained shape differences, or minor pattern variations. Studying FGVC in the context of MLLMs directly probes their visual grounding and fine-grained feature extraction abilities. This lets us assess whether MLLMs can act as trustworthy assistants in visually demanding domains (*e.g.*, ecology, medical imaging, industrial inspection) [68].

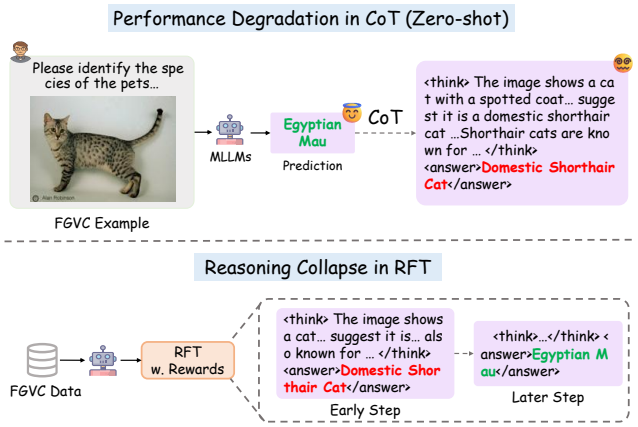

Figure 1. **Performance degradation with CoT and reasoning collapse in RFT.** In zero-shot evaluation (top), MLLMs predict the correct label directly, but adding CoT reasoning leads to a wrong answer. During RFT (bottom), reasoning length steadily shrinks while accuracy improves, indicating a reasoning collapse.

Despite their sophisticated architectures, current MLLMs

exhibit clear limitations on FGVC, often failing to capture the subtle distinctions that define fine-grained categories [10, 14, 68], especially in few-shot and open-ended scenarios where training data is scarce and models must adapt to specialized domains from limited supervision without overfitting. This naturally raises the question of whether the textual description can help compensate for these perceptual weaknesses. A widely held belief in the community is that eliciting Chain-of-Thought (CoT) reasoning improves performance on complex tasks such as math and coding [18, 42, 48, 58, 67, 69], and recent visual-oriented frameworks such as Visual-RFT [32] introduce CoT to steer MLLMs toward enhanced visual perception capabilities, achieving state-of-the-art performance on FGVC. However, several prior works have shown that explicit textual reasoning can paradoxically *reduce* accuracy compared to direct predictions [20, 26, 30, 51]. These studies, however, though informative, generally examine only limited settings: either focusing solely on zero-shot evaluation or comparing CoT and answer-only predictions at a coarse level. This motivates us to systematically re-examine the role of textual reasoning from broader and in-depth evaluation and training perspectives. Specifically, we formulate a key research question:

Is textual reasoning detrimental to fine-grained visual perception, or do current methods simply employ it in a suboptimal way?

To answer this question, we conduct a comprehensive investigation through two aspects: i) zero-shot scenario, and ii) using a Reinforcement Fine-Tuning (RFT) framework to diagnose and manage this trade-off between reasoning and visual accuracy. Our diagnostic experiments reveal several key observations: First, CoT harms zero-shot FGVC performance, shown in Fig. 1 (top) and Tab. 1; second, Reasoning Collapse in RFT, shown in Fig. 1 (bottom) and Fig. 2, where MLLMs gradually learn to suppress verbose reasoning while optimizing for accuracy during RFT. Our in-depth analysis not only corroborates previous observations but further extends them by revealing a central insight: *the length of textual reasoning itself is a critical factor for fine-grained visual perception.* We observe a consistent negative correlation between reasoning length and accuracy: the longer the reasoning content, the worse the visual perception performance. We term these the **"Cost of Thinking"**, which reveals that fine-grained visual perception tasks might benefit from concise rather than elaborate reasoning for MLLMs.

Based on our findings, we introduce **ReFine-RFT**, a novel RFT framework designed to constrain reasoning and improve accuracy. Our framework features two key technical innovations to solve the core challenges of this task. First, to overcome the sparse and semantically naive signal of binary accuracy rewards, we introduce an **ensemble, semantically-aware reward**, which provides a dense and continuous learning signal while explicitly restricting the reasoning length. Second, to optimize multi-objective reward space, we propose **Multi-reward Normalization (MRN)**, a plug-and-play module that stabilizes training by reducing and smoothing the variance of the reward signals. Our contributions are fourfold:

- We empirically characterize the **"Cost of Thinking"** in FGVC, showing that verbose CoT systematically degrades MLLM performance on fine-grained perception tasks.
- We propose **MRN**, a plug-and-play normalization that independently normalizes heterogeneous reward signals in multi-objective settings.
- We introduce **ReFine-RFT**, an RFT framework that integrates MRN with ensemble rewards to primarily optimize accuracy while controlling reasoning length.
- ReFine-RFT achieves state-of-the-art across multiple FGVC benchmarks, validating the effectiveness of our findings and the proposed method.

2. Related Works

Reasoning Ability of MLLMs. Prior works have demonstrated that reasoning could improve performance on complex tasks like math and coding [25, 31, 36, 45, 58, 71, 72]. However, empirical evidence [20, 30, 46, 64] shows that CoT reasoning can introduce spurious explanations and degrade visual perception accuracy. For example, No-Thinking-RFT [26] presents that visual tasks do not need thinking. However, it primarily compares performance between training with CoT and answer-only prompts. We systematically re-evaluate textual reasoning for fine-grained perception across zero-shot and different training regimes, revealing that thinking length is the key to "Cost of Thinking".

Fine-grained Visual Classification in MLLMs. FGVC [22, 28, 33, 35, 37, 54, 59, 61, 76] focuses on subcategory-level recognition that requires capturing subtle visual cues. With the advent of MLLMs, recent works [4, 8, 12, 14, 29, 40, 53] explore prompting and adaptation strategies to improve fine-grained visual discrimination of MLLMs. Visual-RFT [32] further applies RFT with CoT reasoning and achieves additional gains. Building on our findings, we propose ReFine-RFT, which combines ensemble rewards with MRN to explicitly constrain textual reasoning while further enhancing the fine-grained visual perception capability of MLLMs.

Reinforcement Fine-tuning. Reinforcement Learning (RL) originated in control theory for optimal decision-making in dynamic environments [13, 21, 34, 50, 60, 75]. Recent research demonstrates that RL can significantly enhance the reasoning and problem-solving capabilities of LLMs and MLLMs compared with Supervised Fine-tuning (SFT) [3, 11, 18, 42, 49]. DeepSeek-R1 [11] introduces Group Relative Policy Optimization (GRPO), substantially

Model	Aircrafts-102		Flowers-102		Cars-196		Pets-37		Average	
	Answer-only	*CoT*	*Answer-only*	*CoT*	*Answer-only*	*CoT*	*Answer-only*	*CoT*	*Answer-only*	*CoT*
Open-source Non-reasoning Models										
Qwen2-VL-2B	47.5	45.9	55.7	54.8	82.6	56.8	56.4	66.4	60.5	55.9
Qwen2-VL-7B	53.5	42.3	55.8	51.1	83.9	76.5	51.4	61.1	61.2	57.8
Qwen2.5-VL-7B	54.0	41.7	51.1	36.3	73.0	66.9	52.5	62.4	57.7	51.8
InternVL2.5-8B	13.8	11.9	20.1	12.9	33.5	31.9	48.1	50.4	28.9	26.8
InternVL3-8B	14.2	14.5	23.2	10.1	42.2	36.5	37.6	42.8	29.3	25.9
Open-source Reasoning Models										
R1-OneVision-7B-RL	-	42.0	-	59.2	-	49.3	-	68.8	-	54.8

Table 1. **Performance Degradation of CoT on FGVC benchmarks.** We evaluate several open-source MLLMs on four FGVC datasets under two prompting settings: *Answer-only* and *Chain-of-Thought (CoT)*. Results show notable performance degradation when CoT reasoning is applied. For R1-OneVision-7B-RL, *Answer-only* are omitted, as it generates CoT-style outputs even under *Answer-only*.

improving reasoning and generalization. Follow-up works further apply GRPO to other tasks such as visual grounding [5, 16, 32, 38, 43, 52, 62, 66, 78], typically using simple rule-based signals such as accuracy. However, existing methods ignore heterogeneity across reward functions. We propose MRN to balance multi-reward signals, and use an ensemble of rewards to provide denser reward feedback.

3. Cost of Thinking in FGVC

3.1. Experiment Setup

Datasets. We select widely adopted FGVC benchmarks: FGVC-Aircraft [33], Stanford-Cars [22], Flowers-102 [35], and Oxford-Pets [37]. Considering a real-world scenario where fine-grained labeled data might be scarce, we use a 4-shot dataset provided by [32]. We perform FGVC as an open-ended QA task to mimic the real-world application.

Models and Prompts. We evaluate *Answer-only* and *CoT* prompts across several open-source MLLMs: Qwen2/2.5-VL series [2, 57], InternVL series [6, 77], and the reasoning model R1-OneVision [62]. For RFT training, we use Qwen2-VL-2B [57] as the base model. We use the *CoT* prompt from [32], and the following *Answer-only* prompt as an example for Flowers-102:

> *This is an image containing a plant. Please identify the species of the plant based on the image. Only provide the* ***final answer directly****, without any explanation or special formatting.*

Reward Functions for RFT. We follow Visual-RFT [32] and use format reward $R_f(o_i)$ and accuracy reward $R_{cls}(a_i, y)$ to improve the instruction-following capability and answer accuracy. The format reward is a binary signal that enforces strict adherence to the required structured output template, assigning 1 if the model's response o_i correctly follows the sequential `<think>...</think>` and `<answer>...</answer>` tags, and 0 otherwise. The classification reward $R_{cls}(a_i, y)$ measures prediction accuracy based on the class encoded within the `<answer>...</answer>` tags, yielding 1 when the the ground-truth label y is in the predicted label a_i extracted from the answer tags of o_i and 0 otherwise;

To investigate the effects of thinking length, we introduce a thinking length reward $R_{len}(o_i)$ that assigns a binary score based on whether the thinking length lies within a specified range. Let t_i denote the reasoning content extracted from the model output o_i, and let $L_i = |t_i|$ be its character length. Given predefined bounds $(L_{\min}, L_{\max})$, the reward is computed as:

$$R_{len}(o_i) = \begin{cases} 1, & \text{if } L_{\min} \leq L_i \leq L_{\max}, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

This formulation encourages the model to produce reasoning traces whose lengths fall within the desired interval, enabling explicit control over the degree of internal deliberation. When the `<think>` tags are missing or improperly formatted, the reward is set to 0.

3.2. Results & Findings

Performance Degradation in CoT under Zero-Shot. As shown in Tab. 1, incorporating Chain-of-Thought (CoT) prompting consistently leads to accuracy degradation across all FGVC datasets. Non-reasoning models exhibit an average drop of 3–6% when switching from *Answer-only* to *CoT* prompts, while even reasoning-oriented model, R1-OneVision, achieves only moderate accuracy under *CoT* and still generates *CoT* response for *Answer-only*. This indicates that visual reasoning chains often introduce useless or hallucinatory explanations rather than improving decision quality. [20, 30, 51] also reveal similar phenomena that reasoning might be harmful to visual recognition. However, this observation raises a fundamental question: *Is reasoning intrinsically harmful to visual perception tasks, or is the degradation simply a byproduct of zero-shot misalignment*

Figure 2. **Dynamics of reasoning length during RFT across FGVC datasets.** The dark green lines denote the running average of completion lengths throughout RFT FGVC tasks. Across all datasets, the reasoning content length rapidly decreases and stabilizes at a shorter range, suggesting that RFT discourages excessive reasoning generation and promotes concise, decision-focused responses. [Zero-shot: average content length of base model on the evaluation set; Step: the cumulative number of gradient update steps.]

between CoT prompting and model training? We further explore how reasoning evolves under RFT, as models adapt their generation strategy through reward-driven learning.

Reasoning Collapse in RFT. To probe the dynamics of reasoning adaptation, we track the change of completion length throughout RFT. We set up the experiment following Visual-RFT [32] with format reward and classification reward described in Sec. 3.1. As shown in Fig. 2, the average reasoning length exhibits a consistent downward trend across all FGVC datasets. At the early stages of RFT, model outputs are verbose and exploratory, but as training progresses, the content length rapidly declines and stabilizes at a compact range. Notably, the final completion lengths after RFT are shorter than those in the zero-shot setting. We refer to this phenomenon as *reasoning collapse*: an emergent behavior where RFT implicitly discourages long reasoning chains for visual perception tasks, optimizing instead for concise, confident answer prediction. This collapse suggests that the model learns to suppress reasoning steps that do not contribute to reward maximization, thereby aligning its behavior more closely with discriminative objectives. In other words, RFT appears to regularize the reasoning process itself, favoring precision and efficiency over verbosity and exploration, a tendency that aligns with findings from [26].

However, this behavior may also result from reward hacking, since no explicit constraint is imposed on the reasoning process, leading the MLLM to generate only minimal reasoning content. Building upon this observation, we design the subsequent experiments to further verify and quantify the effect for reasoning.

Effects of Thinking Length in RFT. To further examine whether the reasoning length is beneficial or detrimental, we explicitly manipulate the reasoning process by involving the thinking length reward during RFT. We gradually limit the reasoning content length from $[0, 20]$ to $[60, 80]$. As shown in Fig. 3, enforcing longer reasoning sequences leads to a clear decline in classification accuracy across all FGVC datasets. This inverse correlation demonstrates that

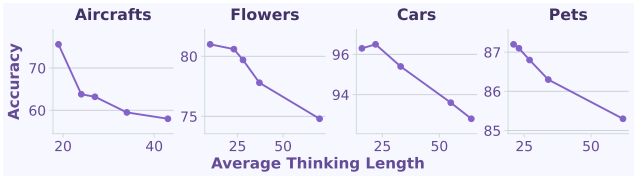

Figure 3. **Impact of reasoning length on FGVC performance.** We analyze the relationship between average reasoning (thinking) length and classification accuracy across FGVC datasets. As the average thinking length increases, performance consistently declines, indicating that excessive reasoning generation introduces noise or distracting the model from key discriminative visual cues.

extended reasoning is not only unhelpful but can actively degrade performance by introducing textural reasoning into the responses. In contrast, shorter reasoning traces yield higher accuracy, indicating that optimal visual performance is achieved with minimal reasoning and precise visual perception and localization. However, the degradation may also stem from low-quality CoT, as reasoning quality is unsupervised during RFT. This motivates analyzing whether higher-quality CoT can close this gap.

Answer-only Surpasses CoT in SFT. To analyze the effects of CoT quality, we use GPT-4o [19] to generate the high-quality CoT data for SFT-CoT training. As shown in Tab. 2, SFT-AO consistently outperforms SFT-CoT, indicating that the degradation is not simply due to the quality of CoT. This finding complements our "Cost of Thinking" analysis from the SFT perspective, showing that excessive reasoning can harm fine-grained visual perception in both training and inference. Taken together with the observations under zero-shot and RFT, these results reveal a key finding:

Finding 1: For fine-grained visual tasks, **thinking length is the key factor**: excessive reasoning hurts performance, and MLLMs benefit more from **concise** responses than from elaborate reasoning.

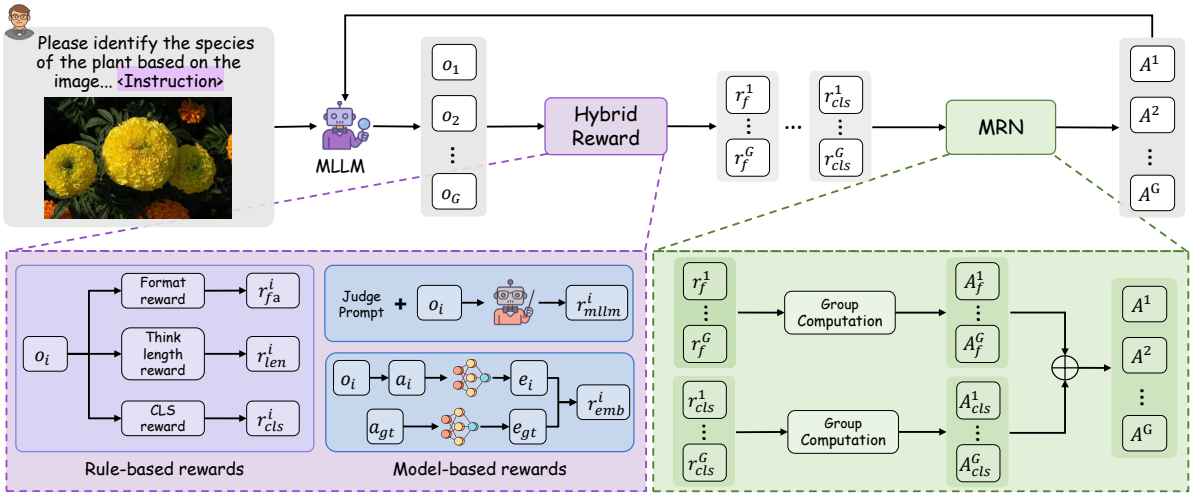

Figure 4. **Overview of ReFine-RFT.** Given a question, the model generates multiple candidate responses, each evaluated using an ensemble reward that combines rule-based rewards and model-based rewards like MLLM-based accuracy reward and embedding similarity reward. The proposed MRN then normalizes the rewards for each function to compute the final advantages used to update the MLLM.

4. Methods

Inspired by our findings, we propose ReFine-RFT, which improves RFT by combining the ensemble reward with a Multi-Reward Normalization scheme. The ensemble reward jointly constrains reasoning length and provide dense accuracy feedback, while Multi-Reward Normalization stabilizes optimization across heterogeneous reward signals. An overview of ReFine-RFT is shown in Fig. 4.

4.1. Ensemble Reward

The ensemble reward is composed of the format, accuracy, and thinking-length rewards defined in Sec. 3.1, together with two complementary rewards: an MLLM-based accuracy reward and an embedding similarity reward, which jointly provide a richer and accuracy-centric feedback.

MLLM-based Accuracy Reward. The classification reward provides only binary supervision through exact string matching between the predicted and ground-truth answers, which fails to capture semantic similarity. For example, predictions such as "*Datura stramonium*" vs. "*thorn apple*" denote the same subcategory but would be penalized under hard matching, and "*Dodge Dakota*" vs. "*2007 Dodge Dakota Club Cab*", which is missing some details. To provide a smoother and more semantically-aware signal, we introduce an MLLM-based Accuracy Reward R_{mllm} that employs an MLLM as a teacher to grade each prediction. Given a predicted answer and the reference label, the MLLM is prompted to output a score from 0 to 10 based on its semantic alignment, which is then normalized to $[0, 1]$. This continuous reward function provides fine-grained feedback by assigning high scores to fully correct answers, intermediate scores to semantically similar ones, and low scores to irrelevant predictions. To mitigate potential scoring biases of the reward model [23, 44], we include few-shot grading examples in the prompt and use this reward together with the embedding similarity reward.

Embedding Similarity Reward. To further provide a smooth and continuous supervision signal, we introduce an embedding similarity reward R_{emb} that measures the semantic closeness between the predicted and ground-truth answers in an embedding space. Given the extracted predicted answer a_i from the `<answer>` tags of the model output and the reference label, both are encoded into text embeddings using a pretrained text embedding model. The cosine similarity between the predicted embedding $\mathbf{e}_i$ and the ground-truth embedding $\mathbf{e}_{gt}$ is used as the reward:

$$R_{emb} = \cos(\mathbf{e}_i, \mathbf{e}_{gt}) \in [0, 1].$$

This continuous reward provides a differentiable measure of semantic alignment, encouraging the model to produce answers that are semantically close to the reference even when lexical forms differ.

4.2. Multi-reward Normalization (MRN)

As shown in Fig. 4, for a given question q and image x, GRPO requires the model to sample G diverse responses $o_1, o_2, \ldots, o_G$ from the current model π_θ and obtains final rewards $r^1, r^2, \ldots, r^G$ for $o_1, o_2, \ldots, o_G$, respectively. GRPO assesses the relative quality by normalizing r^i using the mean and standard deviation of the group reward:

$$A^i = \frac{r^i - \text{mean}(r^1, \ldots, r^G)}{\text{std}(r^1, \ldots, r^G)}, \tag{2}$$

where A^i denotes the advantage of the i-th response. With the group normalization, GRPO encourages the model to

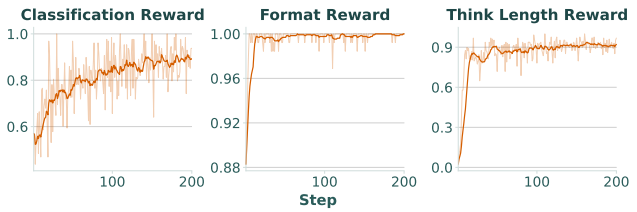

Figure 5. **Differences among rewards during training.** Each reward exhibits distinct convergence speed, value range, and saturation point, reflecting the heterogeneity of different rewards.

sample preferred answers with a higher reward.

In practical training scenarios, multiple reward signals (*e.g.*, format and classification) are often combined to guide optimization. In the original GRPO, these heterogeneous rewards are first aggregated into a single scalar value before performing group normalization. However, in reality, different rewards exhibit varying levels of difficulty and convergence rates. As shown in Fig. 5, the format reward may quickly saturate in early training, thereby dominating the total reward and diluting the influence of other, more informative rewards such as accuracy. To address this issue, we propose Multi-reward Normalization (MRN), which performs group normalization independently for each reward component before aggregation. Specifically, given K reward types $r^i_{(1)}, r^i_{(2)}, \ldots, r^i_{(K)}$ for the i-th response o_i, we compute the normalized advantage for each reward as:

$$A^i_{(k)} = \frac{r^i_{(k)} - \text{mean}(r^1_{(k)}, \ldots, r^G_{(k)})}{\text{std}(r^1_{(k)}, \ldots, r^G_{(k)})}, \tag{3}$$

and then aggregate them to obtain the final advantage:

$$A^i = \sum_{k=1}^{K} A^i_{(k)}. \tag{4}$$

This normalization places all rewards on a comparable scale, leading to a more stable and balanced optimization. The pseudocode is provided in Alg. 1.

5. Experiments

Implementation Details. Qwen2-VL-2B-Instruct [57] is used as the base model. We train with 4 NVIDIA H100 GPUs with 81G of memory. We use Qwen2-VL-7B-Instruct [57] as the reward model for MLLM-based accuracy reward, and E5 [56] as the embedding model for answer embedding similarity reward. To constrain the reasoning length, we set $L_{\min}$=0, $L_{\max}$=10 for ReFine-RFT. We use γ=64 and α=128 for LoRA, and a learning rate of $2e-5$ with 64 as the accumulated batch size. We set the number of generations G=8 and β=0.04 for GRPO. Each experiment trains with 200 step maximum, and all seeds are fixed across the training and evaluation procedures to ensure reproducibility and fairness. More details are in the supplementary.

Algorithm 1 Advantage Normalization with MRN

Input: Number of generations G, number of reward functions M, reward matrix $\boldsymbol{R} \in \mathbb{R}^{G\times M}$, ϵ.
Output: Aggregated advantage $\boldsymbol{A} \in \mathbb{R}^G$.

// 1. Compute mean and std for each reward function
$\boldsymbol{\mu} \leftarrow \text{Mean}(\boldsymbol{R}, \text{axis=0})$ ▷ $\mathbb{R}^M$
$\boldsymbol{\sigma} \leftarrow \text{Std}(\boldsymbol{R}, \text{axis=0})$ ▷ $\mathbb{R}^M$

// 2. Normalize rewards per function (element-wise)
$\boldsymbol{A}_{\text{norm}} \leftarrow (\boldsymbol{R} - \boldsymbol{\mu})/(\boldsymbol{\sigma} + \epsilon)$ ▷ $\mathbb{R}^{G\times M}$

// 3. Aggregate scores into a final advantage
$\boldsymbol{A} \leftarrow \text{Sum}(\boldsymbol{A}_{\text{norm}}, \text{axis=1})$ ▷ $\mathbb{R}^G$

// — Original GRPO Method (for comparison) —
// 1. Aggregates rewards
// $\boldsymbol{R}_{agg} \leftarrow Sum(\boldsymbol{R}, axis=1)$
// 2. Normalizes the aggregated rewards
// $\mu_{agg} \leftarrow Mean(\boldsymbol{R}_{agg})$; $\sigma_{agg} \leftarrow Std(\boldsymbol{R}_{agg})$
// $\boldsymbol{A}_{GRPO} \leftarrow (\boldsymbol{R}_{agg} - \mu_{agg})/(\sigma_{agg} + \epsilon)$
return $\boldsymbol{A}$

5.1. Results of ReFine-RFT

Tab. 2 summarizes the performance across four FGVC datasets, from which several consistent patterns emerge:

Lora Outperforms Fully-FT. We find that LoRA fine-tuning [15] consistently surpasses fully fine-tuning (Fully-FT) under both SFT and RFT settings. This confirms that LoRA not only reduces computational cost but also enables the model to better leverage limited fine-grained visual data.

RFT Provides Stronger Gains. Transitioning from SFT to RFT yields substantial accuracy improvements across all FGVC benchmarks. RFT enables the model to directly optimize accuracy-centric objectives, correcting undesirable generation behaviors and yielding more stable, discriminative predictions. This is especially desirable in few-shot FGVC, where precise category boundaries must be learned from limited supervision.

Superiority of ReFine-RFT. Building upon our findings, our proposed ReFine-RFT further improves performance across all benchmarks. By integrating: (i) an ensemble, semantically-aware reward that provides dense and accuracy-aligned feedback, and (ii) the Multi-Reward Normalization module (MRN) that stabilizes heterogeneous rewards, ReFine-RFT achieves consistent gains over Visual-RFT and other baselines. Importantly, ReFine-RFT with only a 2B backbone and 4-shot training significantly surpasses Finedefics-8B trained on the full FGVC datasets, underscoring the efficiency and scalability of our approach.

Methods	FT Methods	FT Types	Aircrafts-102	Flowers-102	Cars-196	Pets-37	Average
Qwen2-VL-2B [57]	*Zero-shot*	-	45.9	54.8	56.8	66.4	56.0
Finedefics-8B [14]	*SFT*	*Fully-FT*	63.8	89.9	84.7	92.2	82.7
SFT-AO	*SFT*	*Fully-FT*	67.9	58.5	40.5	55.5	55.6
SFT-AO		*Lora*	78.3	74.8	80.0	<u>87.6</u>	80.2
SFT-CoT		*Lora*	73.9	74.4	52.3	87.5	72.0
Visual-RFT [32]	*RFT*	*Fully-FT*	74.8	71.4	95.3	86.1	81.9
Visual-RFT [32]		*Lora*	75.6	74.1	<u>95.7</u>	86.0	82.9
No-Thinking-RFT [26]		*Fully-FT*	-	71.2	-	86.1	-
ReFine-RFT-AO (Ours)		*Lora*	<u>78.7</u>	**81.4**	93.1	<u>87.6</u>	<u>85.2</u>
ReFine-RFT-CoT (Ours)		*Lora*	**79.3** (+3.7%)	81.0 (+6.9%)	**97.1** (+1.4%)	**88.6** (+2.6%)	**86.5** (+3.6%)

Table 2. **Performance comparison on FGVC datasets.** Compared to SFT (with/without CoT data) and Visual-RFT baselines, our ReFine-RFT achieves the best results with consistent gains across datasets (values in parentheses denote relative improvements over the Visual-RFT (*Lora*) [32] baseline). [AO: *Answer-only* prompt; SFT-CoT: SFT with CoT data; **Best** and <u>second best</u> are highlighted.]

Reasoning Control Outweighs Prompt Style. We observe that controlling the reasoning length has a larger impact on performance than the choice of prompt style. First, No-Thinking-RFT uses an *Answer-only*-style prompt, whereas Visual-RFT uses a CoT-style prompt, yet they yield similar performance. This suggests that prompt style is not the primary performance determinant. Second, within our ReFine-RFT, ReFine-RFT-CoT is slightly better than the ReFine-RFT-AO. Together, these observations suggest that performance depends more on reasoning-length control than on whether the prompt elicits CoT. In our view, suppressing reasoning encourages the model to focus on accuracy as the main optimization target, while still allowing it to generate reasoning when reasoning is genuinely beneficial. This further supports our Cost of Thinking analysis and Finding 1 in Sec. 3, confirming that reasoning length is the key factor influencing fine-grained visual perception. We then derive the following conclusion:

Conclusion 1: For fine-grained visual perception, jointly using **multi-perspective, accuracy-centric rewards** and **explicit reasoning length control** leads to stronger visual perception capabilities.

5.2. Ablation Studies

Effects of MRN. We employ format reward and classification reward to investigate the impact of MRN. As shown in Tab. 3, integrating MRN consistently improves performance across all three FGVC datasets, yielding gains of +1.1%/+2.1%/+0.4% under full fine-tuning and +0.7%/+1.5%/+0.6% under LoRA. The improvements are more pronounced with larger training capacity, suggesting that MRN can better leverage additional parameters. Overall, these results confirm that MRN effectively boosts model performance while preserving strong efficacy in parameter-efficient training regimes.

Methods	Aircrafts-102	Flowers-102	Cars-196
Fully fine-tuing			
GRPO [11]	74.0	68.6	94.7
+ MRN (Ours)	75.1	70.7	95.1
Lora (r=64, α=128)			
GRPO [11]	75.6	74.1	95.7
+ MRN (Ours)	76.3	75.6	96.3

Table 3. **Comparison of MRN under two training regimes.** MRN serves as a plug-and-play module and consistently enhances accuracy across datasets under fully fine-tuning and LoRA.

Effects of Ensemble Reward. We ablate the effects of the ensemble reward design in Tab. 5. The results demonstrate that incorporating multiple reward components leads to consistent performance gains. Each reward contributes complementary information, guiding the model toward robust learning objectives. Notably, when all reward functions are jointly combined as the ensemble reward, the model achieves the best overall performance, suggesting that aggregating complementary reward signals provides richer and more stable guidance for optimization than any individual reward. Fig. 6 shows the reward curves during training, validating the effectiveness of ReFine-RFT.

Effects of Trainable Parameters. We analyze the impact of trainable parameters using the format and classification rewards on ReFine-RFT. As shown in Tab. 4, model performance consistently improves with increasing LoRA capacity. In particular, the configuration with r=64, α=128 achieves the highest accuracy, surpassing the Fully-FT baseline across all datasets. In contrast, the smaller setting (r=16, α=32) leads to a notable performance drop. These results indicate that appropriately chosen LoRA rank and scaling factors can outperform fully fine-tuning in few-shot scenarios, providing an efficient and effective approach for model adaptation.

Methods	Aircrafts-102	Flowers-102	Cars-196
Fully-FT	75.1	70.7	95.1
	Lora		
$r = 16, \alpha = 32$	71.3	64.6	94.0
$r = 32, \alpha = 64$	75.4	70.4	95.0
$r = 64, \alpha = 128$	76.3	75.6	96.3

Table 4. **Comparison of trainable parameters.** LoRA with larger ranks (r) and scaling factors (α) progressively improves accuracy, eventually surpassing full fine-tuning across all datasets.

R_f	R_{cls}	R_{len}	R_{mllm}	R_{emb}	Aircrafts-102	Pets-37
✓	✓				76.3	86.8
✓	✓	✓			78.5	87.5
✓	✓	✓	✓		77.5	85.7
✓	✓	✓		✓	79.0	86.3
✓	✓	✓	✓	✓	79.3	88.6

Table 5. **Effects of ensemble reward.** Combining multiple reward functions consistently improves performance, and using all rewards yields the best overall results.

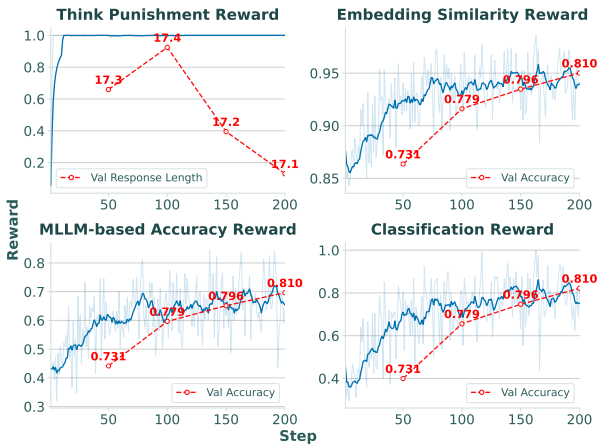

Figure 6. **Reward curves of ReFine-RFT on Flowers-102.** Rewards and validation accuracy consistently increase over training, demonstrating the effectiveness of our reward design.

Reward Distribution Comparison. As shown in Fig. 7, our proposed MRN consistently achieves higher reward values and maintains significantly lower reward standard deviation compared to the baseline GRPO. Throughout training, MRN exhibits a steady improvement in reward, indicating more stable and efficient policy optimization. In contrast, GRPO shows larger standard deviation, reflecting less stable learning behavior. The notably lower reward variance of MRN suggests that it effectively mitigates gradient noise and reduces policy fluctuation, leading to smoother and more reliable reward progression. These observations demonstrate that MRN not only enhances training stability but also enables more consistent reward maximization, thereby improving optimization robustness and efficiency.

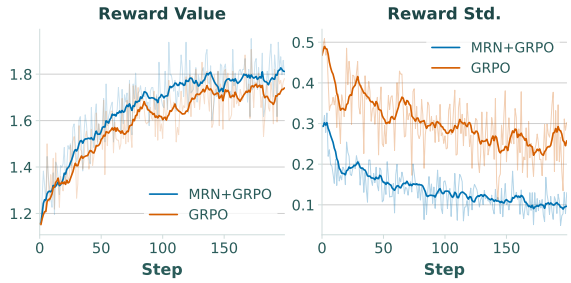

Figure 7. **Training reward and its standard deviation comparison on Aircrafts-102.** MRN + GRPO achieves consistently higher reward values and lower variance throughout training, indicating improved stability and optimization efficiency.

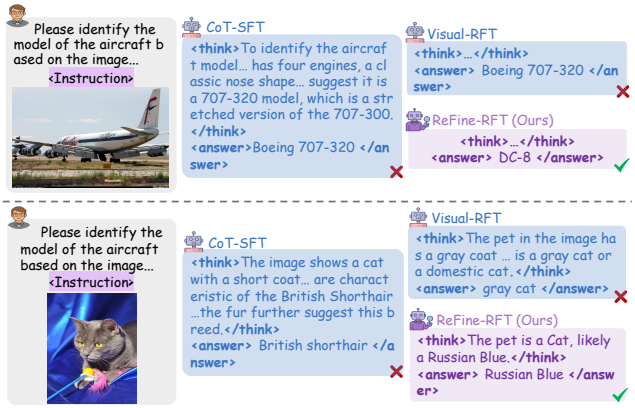

Figure 8. **Comparison of responses.** SFT-CoT and Visual-RFT produce long reasoning with incorrect answers, while ReFine-RFT achieves concise reasoning and higher accuracy. More results and analyses are in the supplementary.

Qualitative Results. As shown in Fig. 8, ReFine-RFT demonstrates a clear advantage in both reasoning efficiency and accuracy, encouraging minimal reasoning steps.

6. Conclusion

We reveal the "Cost of Thinking" in FGVC for MLLMs, showing that excessive textual reasoning degrades fine-grained visual perception performance from both the inference and training perspectives. Our systematic study across zero-shot and multiple fine-tuning regimes indicates that perception-centric tasks benefit more from concise reasoning. Motivated by this, we propose ReFine-RFT, a reasoning-constrained RFT framework that enhances visual perception by integrating ensemble, semantically-aware rewards with a Multi-Reward Normalization (MRN) for optimization. Extensive results demonstrate that ReFine-RFT achieves state-of-the-art performance across FGVC benchmarks, highlighting that effective visual perception emerges from constraint thinking and accuracy-centric reward shaping. Future work will probe the mechanisms behind the Cost of Thinking and extend ReFine-RFT to broader multimodal tasks.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 3

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 2

[4] Junwen Chen, Jie Zhu, and Yu Kong. Atm: Action temporality modeling for video question answering. In *ACM MM*, 2023. 2

[5] Xiaxu Chen, Wei Li, Chunxu Liu, Chi Xie, Xiaoyan Hu, Chengqian Ma, Feng Zhu, and Rui Zhao. On the suitability of reinforcement fine-tuning to visual tasks. In *CVPR*, 2025. 3

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3

[7] LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy, 2023. 1

[8] Chenhang Cui, An Zhang, Yiyang Zhou, Zhaorun Chen, Gelei Deng, Huaxiu Yao, and Tat-Seng Chua. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. *arXiv preprint arXiv:2410.14148*, 2024. 2

[9] Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, et al. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025. 1

[10] 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. 1, 2

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 7

[12] Xiao Guo, Jie Zhu, Anil Jain, and Xiaoming Liu. On the holistic approach for detecting human image forgery. *arXiv preprint arXiv:2601.04715*, 2026. 2

[13] Dong Han, Beni Mulyana, Vladimir Stankovic, and Samuel Cheng. A survey on deep reinforcement learning algorithms for robotic manipulation. *Sensors*, 2023. 2

[14] 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. 1, 2, 7

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 6

[16] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 3

[17] Zhanbo Huang, Dingqiang Ye, Xiaoming Liu, and Yu Kong. Unlocking motion from large vision models with a semantic and kinematic duality for gait recognition. In *CVPR*, 2026. 1

[18] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024. 2

[19] 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. 4, 1

[20] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025. 2, 3

[21] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 1996. 2

[22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013. 2, 3

[23] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *NAACL*, 2025. 5

[24] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. 4

[25] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026. 2

[26] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *NeurIPS*, 2025. 2, 4, 7

[27] Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. When thinking fails: The pitfalls of reasoning for instruction-following in llms. *NeurIPS*, 2025. 4

[28] Feng Liu, Nicholas Chimitt, Lanqing Guo, Jitesh Jain, Aditya Kane, Minchul Kim, Wes Robbins, Yiyang Su, Dingqiang Ye, Xingguang Zhang, et al. Person recognition at altitude and range: Fusion of face, body shape and gait. *arXiv preprint arXiv:2505.04616*, 2025. 2

[29] Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. *arXiv preprint arXiv:2401.13837*, 2024. 2

[30] Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*, 2024. 2, 3

[31] Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, et al. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*, 2026. 2

[32] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *ICCV*, 2025. 2, 3, 4, 7, 1

[33] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2, 3

[34] Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural Networks*, 2022. 2

[35] 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*. IEEE, 2008. 2, 3

[36] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models, 2021. 2

[37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*. IEEE. 2, 3

[38] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025. 3

[39] Akshara Prabhakar, Thomas L Griffiths, and R Thomas McCoy. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. *arXiv preprint arXiv:2407.01687*, 2024. 4

[40] Zhiyuan Ren, Yiyang Su, and Xiaoming Liu. Chatgpt-powered hierarchical comparisons for image classification. *NeurIPS*, 2023. 2

[41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1

[42] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2

[43] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 3

[44] Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. In *EMNLP*, 2023. 5

[45] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *arXiv preprint arXiv:2506.21656*, 2025. 2

[46] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024. 2

[47] Yiyang Su, Yunping Shi, Feng Liu, and Xiaoming Liu. Hamobe: Hierarchical and adaptive mixture of biometric experts for video-based person reid. In *ICCV*, 2025. 1

[48] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025. 2

[49] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 2

[50] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998. 2

[51] 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. 2, 3

[52] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025. 3

[53] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. 2

[54] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. 2011. 2

[55] Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025. 4

[56] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text

embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. 6

[57] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 6, 7, 1

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. 2

[59] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *TPAMI*, 2021. 2

[60] Feng Xu, Guangyao Zhai, Xin Kong, Tingzhong Fu, Daniel FN Gordon, Xueli An, and Benjamin Busam. Stare-vla: Progressive stage-aware reinforcement for fine-tuning vision-language-action models. *arXiv preprint arXiv:2512.05107*, 2025. 2

[61] Haolan Xu, Keli Cheng, Lei Wang, Ning Bi, and Xiaoming Liu. Emotag: Emotion-aware talking head synthesis on gaussian splatting with few-shot personalization. In *CVPR*, 2026. 2

[62] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025. 3

[63] Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. *NeurIPS*, 2022. 4

[64] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025. 2

[65] Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Enhancing vision-language few-shot adaptation with negative learning. In *WACV*. IEEE, 2025. 3

[66] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025. 3

[67] Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *NeurIPS*, 2024. 2

[68] 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? *NeurIPS*, 2024. 1, 2

[69] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 2

[70] Zhiahao Zhang, Yiwei Chen, Weizhan Zhang, Caixia Yan, Qinghua Zheng, Qi Wang, and Wangdu Chen. Tile classification based viewport prediction with multi-modal fusion transformer. In *ACM MM*, 2023. 1

[71] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. Multimodal chain-of-thought reasoning in language models. *TMLR*, 2023. 2

[72] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *ICLR*, 2023. 2

[73] Zhihao Zhang, Abhinav Kumar, Girish Chandar Ganesan, and Xiaoming Liu. Unleashing the power of chain-of-prediction for monocular 3d object detection. In *CVPR*, 2026. 1

[74] Zhihao Zhang, Abhinav Kumar, and Xiaoming Liu. Towards intrinsic-aware monocular 3d object detection. In *CVPR*, 2026. 1

[75] Jie Zhu, Mengsha Hu, Amy Zhang, Ruoming Jin, and Rui Liu. Fairness-sensitive policy-gradient reinforcement learning for reducing bias in robotic assistance. In *IEEE ROMAN*, 2024. 2

[76] Jie Zhu, Yiyang Su, Minchul Kim, Anil Jain, and Xiaoming Liu. A quality-guided mixture of score-fusion experts framework for human recognition. In *ICCV*, 2025. 2

[77] 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. 1, 3

[78] Jie Zhu, Xiao Guo, Yiyang Su, Anil Jain, and Xiaoming Liu. Fusionagent: A multimodal agent with dynamic model selection for human recognition. In *CVPR*, 2026. 3

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 [41], 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 [32]. 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 [57] 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 [19] 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 [32] 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 [32], 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 [32] and No-Thinking-RFT [26].

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.

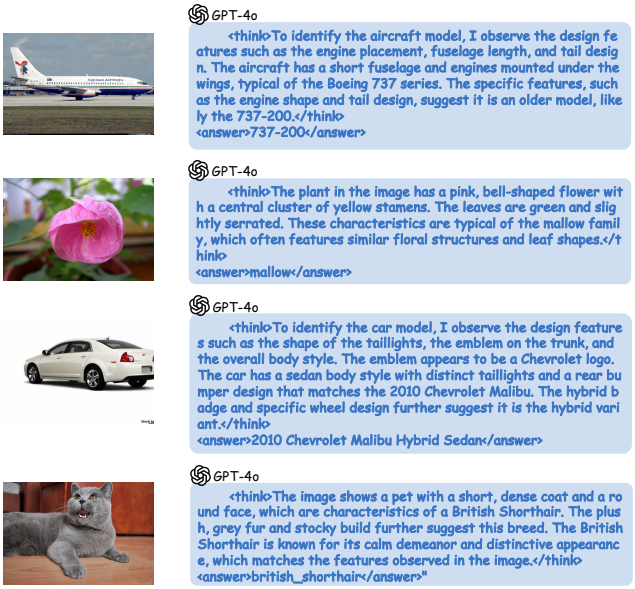

Figure 13. Examples of CoT annotations generated by GPT-4o, featuring long and fine-grained reasoning traces that describe each image in detail.

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
CLIP^{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: [65].

standard discriminative baselines.

4Correlation 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

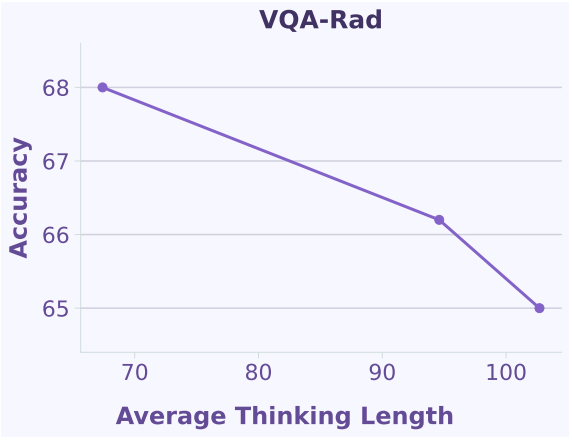

Figure 14. Cost of thinking on VQA-Rad dataset. Performance decreases as the thinking length increases.

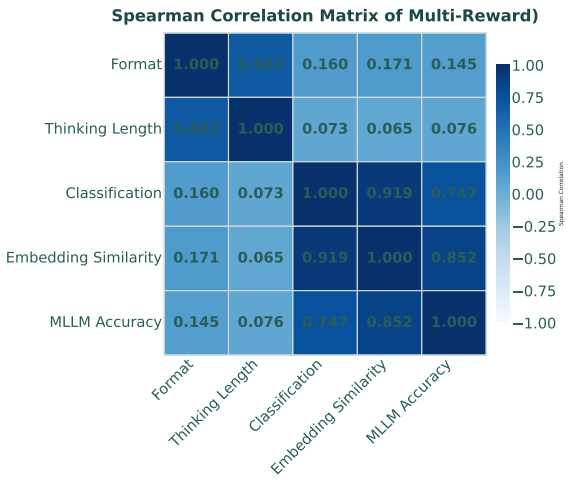

Figure 15. Correlation heatmap of ensemble rewards.

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.

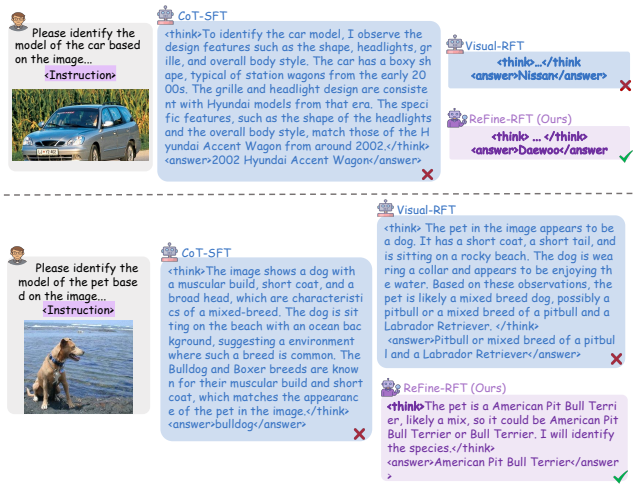

Figure 16. **Additional comparison of responses.**

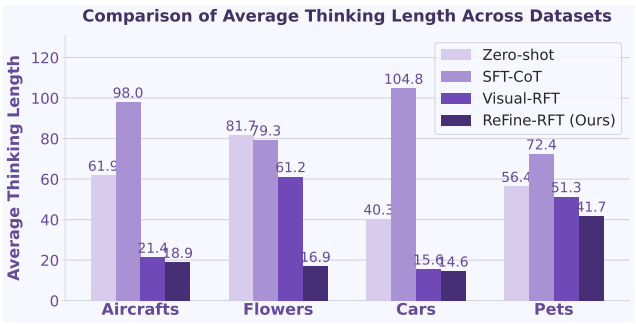

Figure 17. **Comparison of average thinking length across datasets.** We average the number of thinking tokens as the thinking length per dataset. SFT-CoT consistently yields the longest chains of thought, Zero-shot produces medium-length traces, while both Visual-RFT and especially ReFine-RFT generate 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 [27], 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 [39, 55].

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 [24, 63], 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.