ReFine-RFT: Improving Reasoning Capability of Multimodal LLMs for Fine-grained Recognition

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-modal large language models (MLLMs) have demonstrated remarkable proficiency in general vision-language tasks but struggle with Fine-grained Visual Recognition (FGVR), which demands distinguishing highly similar subcategories through nuanced visual perception and systematic reasoning. Current methods face critical limitations, including overfitting during fine-tuning, degraded generalization, and unreliable reasoning processes that often lead to plausible yet incorrect rationales. To address these challenges, we propose **ReFine-RFT**, a Reinforcement Fine-Tuning (RFT) framework that enhances MLLMs' reasoning capabilities for FGVR while preserving their general-purpose performance. ReFine-RFT integrates Group Relative Policy Optimization (GRPO) with two types of specially designed reward functions: **enhanced rule-based rewards** for instruction adherence and classification accuracy, and a novel **MLLM-based reasoning reward** evaluated by a teacher MLLM to ensure a rational reasoning process. Additionally, we introduce the ***Reasoning-Answer* evaluation protocol**, which jointly assesses recognition accuracy, reasoning quality, and instruction-following capability. Extensive experiments on six FGVR benchmarks demonstrate the state-of-the-art performance of ReFine-RFT, achieving significant improvements in both accuracy and reasoning fidelity while maintaining data efficiency. Our work bridges the critical gap between MLLMs' reasoning capacity and fine-grained visual understanding, advancing toward trustworthy and expert-level visual recognition systems. The code and model will be publicly released upon publication.

1 Introduction

Multi-modal large language models (MLLMs) have emerged as a prominent paradigm for integrating vision and language, enabling unified interpretation of images and generation of coherent textual responses. While MLLMs excel at general vision-language tasks such as image captioning and visual question answering, their performance on Fine-grained Visual Recognition (FGVR)—which demands distinguishing highly similar subcategories (*e.g.*, bird species or car models with subtle visual differences)—remains suboptimal [14]. Success in FGVR requires not only high-fidelity visual perception to identify nuanced distinctions but also systematic understanding of domain-specific knowledge or comparative attribute analysis. For example, differentiating a *Black-footed Albatross* from a *Laysan Albatross* involves fine-grained cues such as leg color and beak shape, which are often challenging without domain-specific knowledge. Bridging the gap between MLLMs' reasoning capacity and their ability to discriminate fine-grained visual features is thus critical for advancing expert-level visual understanding.

Yet, current MLLMs face key limitations in visual perception tasks, particularly in fine-grained visual recognition (FGVR). First, as shown in Fig. 1, naive fine-tuning on fine-grained datasets often leads

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

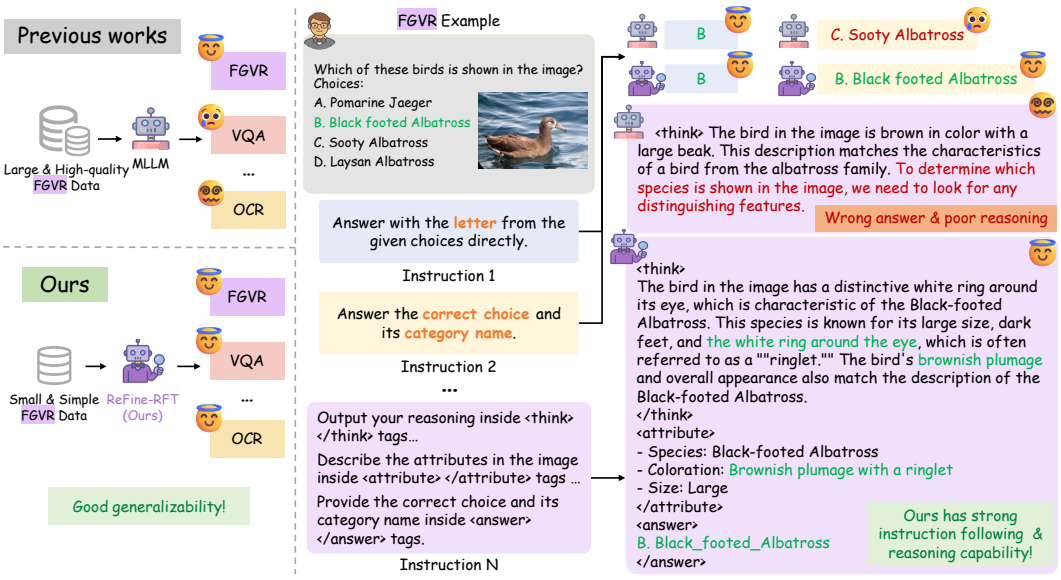

Figure 1: *Left*: Prior works fine-tune MLLMs on large and richly annotated datasets, which (i) significantly reduce cross-task generalization and (ii) yield poor instruction-following and reasoning with diverse instructions. ReFine-RFT is trained on small and simply formatted few-shot datasets, achieving strong reasoning performance and accuracy while maintaining broad generalization.

to overfitting to narrow domains, compromising the model's general-purpose vision-language capabilities. Second, MLLMs struggle to capture subtle visual distinctions such as differences in aircraft turbines or bird wingtips [41], which are essential for FGVR. Third, performance further deteriorates when reasoning is required [25, 39, 20]: producing step-by-step explanations or format-constrained outputs can reduce accuracy, as models may be distracted by irrelevant details or overwhelmed by longer prompts, resulting in plausible but incorrect rationales and misclassification among visually similar categories. Moreover, conventional classification metrics focus solely on final answer accuracy and overlook the quality and generalizability of the reasoning process. While such metrics may reward high nominal performance, they often incentivize superficial heuristics such as pattern matching or statistical guessing, rather than genuine semantic or visual understanding. As a result, models remain brittle under slight variations in question phrasing or when required to produce explicit chain-of-thought (CoT) explanations. This highlights the need for more rigorous evaluation protocols that assess a model's ability to holistically interpret multimodal inputs and reason in a logically grounded manner—an essential step toward building trustworthy AI systems.

To address these challenges, we propose **ReFine-RFT**, a Reinforcement Fine-Tuning (RFT) framework that enhances MLLMs' fine-grained reasoning capabilities in a data-efficient and generalizable manner. To mitigate overfitting to narrow domains, ReFine-RFT adopts Group Relative Policy Optimization (GRPO) [13], enabling robust training with limited supervision. To address MLLMs' difficulty in capturing subtle visual cues, we introduce **enhanced rule-based rewards** with structured reasoning tags: <think>...</think>, <attribute>...</attribute>, and <answer>...</answer>, which guide the model to attend to intra-class distinctions through interpretable outputs. Finally, to improve reasoning without sacrificing recognition accuracy, we propose a **MLLM-based reasoning reward** computed by a teacher MLLM that evaluates the relevance and accuracy of the model's explanation given the image. Teacher MLLMs [15, 48] are employed for their advanced reasoning abilities and consistently superior performance, and this external guidance incentivizes the student MLLM to internalize high-quality reasoning without requiring manually annotated rationales. In addition, we introduce a new evaluation protocol, the ***Reasoning-Answer* setting**, which jointly assesses recognition accuracy, instruction-following, and explanation quality using a powerful MLLM (e.g., GPT [2]) as the evaluator.

In summary, our contributions are as follows:

- We propose **ReFine-RFT**, a novel fine-tuning framework that enhances MLLMs' reasoning capabilities for FGVR while preserving its generalization in a data-efficient manner.

- We design two FGVR rewards: **enhanced rule-based rewards** for instruction-following and classification accuracy, and **MLLM-based reasoning reward** for reasoning performance.
- We extend the FGVR benchmarks with a new evaluation protocol called ***Reasoning-Answer* setting** to evaluate MLLMs on instruction-following, reasoning, and answering.
- Extensive experiments on the FGVR benchmarks demonstrate the superiority of ReFine-RFT, achieving SoTA performance in reasoning quality, recognition accuracy, and instruction-following.

2 Related Works

Fine-grained Visual Recognition in MLLMs. Fine-grained Visual Recognition (FGVR) [42, 45] is a classic problem in computer vision, focused on distinguishing objects at the subcategory level (*e.g.*, differentiating bird species, and car makes). With the advent of large vision-language models, researchers have begun exploring how these models can be adapted for FGVR. Recent methods have shown that incorporating the language modality can be beneficial: for instance, classification performance can be enhanced via a textual description of the given image [33]. Other approaches [41, 8, 24, 4] fine-tune MLLMs on FGVR benchmarks to improve the visual discrimination of MLLMs. For example, Finedefics [14] propose an informative description construction for the FGVR training set and a two-stage fine-tuning to align the visual features with the textual descriptions and category names. However, existing methods either rely on large-scale data or lack logical reasoning and generalization. We propose a data-efficient framework that improves accuracy, enhances reasoning, and preserves strong generalization.

Reasoning Ability of MLLMs. Beyond answering capability, the reasoning of MLLMs has become a focal point of recent research. Investigation of the reasoning capability of MLLMs starts from Chain-of-thought (CoT) prompting, which generates an intervening string of tokens that increases the probability of producing the correct answer by thinking step-by-step [44, 29, 50, 49]. Inspired by DeepSeek-R1-Zero [13], which introduces a rule-based Reinforcement Learning (RL) method that significantly improves reasoning, researchers start to investigate the effects of RL in enhancing the reasoning capability of MLLMs [40, 17, 36, 32]. However, Shaikh et al. [35], Liu et al. [25], Yu et al. [46] reveal that reasoning can also produce harmful outputs that degrade the performance, which is counterintuitive to human expectations. MLLMs can produce convincing explanations that are nevertheless flawed or irrelevant to the true distinctions. To address this, we introduce the *Reasoning-Answer* protocol for evaluating reasoning quality, and an MLLM-based reasoning reward to explicitly enhance it.

Reinforcement Fine-tuning (RFT). Early research primarily focused on RL from Human Feedback (RLHF), which aimed to align model outputs with human preferences [30, 3, 38]. Recent advancements demonstrate that RL can significantly enhance the reasoning capabilities of these models. For instance, DeepSeek-R1 [13] highlights the effectiveness of RL in improving LLMs' reasoning abilities by proposing a Group Relative Policy Optimization (GRPO) framework with rule-based rewards. Follow-up works like Visual-RFT [26], VLM-R1 [36], and LMM-R1 [32] explore GRPO in specific domains such as image classification, visual grounding, or the text-only domain. While prior works have proposed various rule-based rewards (e.g., classification accuracy or IoU [26, 36]), these are primarily task-specific and lack a direct focus on reasoning quality. In contrast, our approach introduces an MLLM-based reasoning reward, offering a level of reasoning supervision that rule-based rewards cannot provide.

3 Methods

In this section, we present an overview of the ReFine-RFT framework. Building on the success of DeepSeek-R1-Zero [13], we adopt Group Relative Policy Optimization (GRPO) as the reinforcement fine-tuning (RFT) strategy and introduce two types of rewards: enhanced rule-based rewards and a novel MLLM-based reasoning reward. The overall architecture of ReFine-RFT is illustrated in Fig. 2.

3.1 Rule-based Rewards

The goal of rule-based rewards is to improve the instruction-following capability and answer accuracy of MLLMs. We introduce two rule-based rewards: format reward and classification reward.

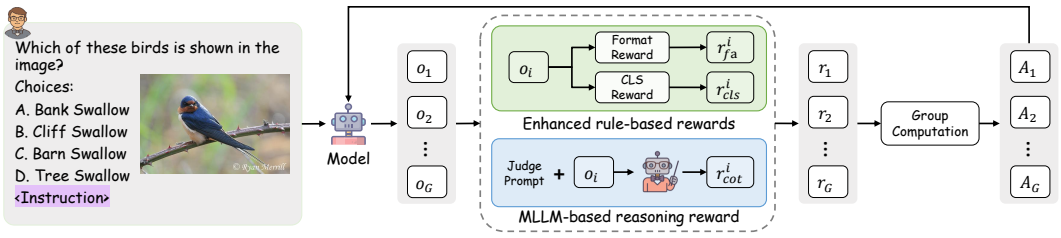

Figure 2: **Overview of ReFine-RFT.** Given a question, the model generates multiple responses, each evaluated with both rule-based rewards (format and classification rewards) and an MLLM-based reasoning reward via a judgment teacher model. Final normalized rewards are used with the policy gradient optimization algorithm to update the model.

Format Reward. Traditional format reward $R_f(o_i)$ ensures strict adherence to a predefined structured response format for the model response o_i, which is widely used in [13, 26, 36]. Unlike traditional methods that enforce basic tag generation (*e.g.*, <think>...</think> and <answer>...</answer>), ReFine-RFT introduces a specialized <attribute>...</attribute> tags (denoted as $R_{fa}(o_i)$) tailored for the FGVR task, explicitly prompting the model to describe discriminative visual attributes critical for fine-grained recognition. The reward is assigned a binary value: 1 if the response strictly follows the sequential structure: <think>...</think> , <attribute>...</attribute> , <answer>...</answer> , and 0 otherwise.

Classification Reward. Based on prior methods Liu et al. [26], Shen et al. [36], we use the classification reward $R_{cls}(a_i, y)$ to quantify the classification accuracy of the model's final prediction within the <answer>...</answer> tags. This metric is computed by verifying alignment between the predicted class a_i in the <answer>...</answer> tags of o_i and the ground-truth label y, yielding a score of 1 for correct predictions and 0 for errors. To enforce structured output compliance, the reward is 0 if the <answer>...</answer> tags are omitted or improperly formatted:

$$R_{cls}(a_i, y) = \begin{cases} 1, & \text{if } \mathrm{a_i} = y, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

3.2 MLLM-based Reasoning Reward

Limitations of Synthetic Reasoning Data. Curating high-quality human-annotated reasoning datasets for specialized tasks is inherently resource-intensive. While conventional methods circumvent this by leveraging synthetic data generated by powerful models (e.g., GPT-4o [2]), such approaches impose an implicit constraint: the fine-tuned model becomes confined to the reasoning patterns and biases present in the pre-generated data. This artificially narrows the model's exploration space, limiting its capacity to discover novel or contextually adaptive reasoning pathways.

MLLM-based reasoning reward. Motivated by the LLM-as-judge paradigm [51], we propose a novel MLLM-based reasoning reward, *e.g.*, $R_{cot}(o_i, x)$. Our framework employs an MLLM as a teacher model (denoted as $\mathcal{F}_{teacher}$) to assess reasoning quality in context. This choice is motivated by the observation that stronger MLLMs, such as bigger or closed-source models, demonstrate more reliable judgment of reasoning quality, especially in visually grounded tasks where nuanced interpretations are required. For a generated response o_i, we concatenate it with the input image x into a judge prompt $\mathcal{J}$ (shown in Fig. 3). $\mathcal{F}_{teacher}$ evaluates two critical dimensions: (1) factual accuracy of visual descriptions based on the image, and (2) helpfulness of these descriptions to the final recognition task. $R_{cot}(o_i, x)$ can be formulated as:

$$R_{cot}(o_i, x) = \mathcal{F}_{teacher}(x \oplus \mathcal{J} \oplus o_i). \tag{2}$$

3.3 Group Relative Policy Optimization (GRPO)

In contrast to RL algorithms such as Proximal Policy Optimization (PPO) [34]— which rely on a critic model to assess policy performance, GRPO eliminates the need for the critic model by directly

Judge Prompt

You are tasked with evaluating the reasoning provided by a model in identifying an object from an image. Based on the image and the reasoning content between the <think> and <attributes> tags, assess whether the reasoning is factually correct, relevant, and helpful for identifying the object. Output only a score between 0 and 100 (number only, no additional text):
- 100 means the reasoning is completely accurate, relevant, and helpful for classification.
- 0 means the reasoning is completely inaccurate or irrelevant.
- Values between 0 and 100 reflect partial correctness or partial relevance.

Image: *<Image>*

Response from the model: *<Response>*

Figure 3: Judge prompt for MLLM-based reasoning reward.

comparing groups of candidate responses. As shown in Fig. 2, 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 rewards $\{r_1, r_2, \ldots, r_G\}$ for o_i based on the reward function $R(q, o_i)$:

$$R(q, o_i) = w_{fa} R_{fa}(o_i) + w_{cls} R_{cls}(a_i, y) + w_{cot} R_{cot}(o_i, x), \tag{3}$$

where w_{fa}, w_{cls}, and w_{cot} are the reward weights for R_{fa}, R_{cls}, and R_{cot}, respectively. GRPO assesses the relative quality by normalizing r_i using the mean and standard deviation of the group reward:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})}, \tag{4}$$

where A_i denotes the advantage of the i-th response. With the group normalization, GRPO encourages the model to sample preferred answers with a higher reward. The model is updated via:

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

where ε and β are the GRPO clipping hyperparameters and the coefficient weight for controlling the Kullback–Leibler (KL) penalty [34], respectively. π_{ref} is the reference model.

4 Extending FGVR Benchmark Evaluation

As shown in Tab. 1, when we change the prompt from *Choice* to *Both*, the performance drops for most of the MLLMs. When we change to the *CoT* prompt, the performance drops even more severely across all the models. This degradation reveals that most of the models neither understand the instruction nor answer correctly, which is a similar phenomenon to [35, 25]. Seemingly correct answers often rely on flawed reasoning, such as logical inconsistencies, hallucinated attributes, or misaligned visual-textual grounding. These issues give a false sense of the model's ability and can lead to failures in real-world applications.

To address this gap, we propose the *Reasoning-Answer* setting (shown in Fig. 4), which jointly evaluates both the final prediction and the quality of the model's reasoning process. Unlike conventional accuracy-based metrics, *Reasoning-Answer* explicitly penalizes models that arrive at correct answers through unreliable or incoherent rationales, thereby promoting interpretable and logically grounded decision-making, which is critical for trustworthy AI in mission-critical applications. *Reasoning-Answer* introduces a complementary metric called the Reasoning Score (*RSC*), which is automatically computed by a powerful teacher MLLM. Given the input image, question, ground-truth answer, and the complete MLLM response, the evaluator produces structured feedback along three dimensions using a single unified prompt. Each score ranges from 0 to 10:

Figure 4: **Evaluation pipeline for *Reasoning-Answer* setting.** Beyond answer accuracy evaluation, we introduce a complementary protocol to assess MLLM performance: Reasoning Score (*RSC*), which evaluates the quality of the response by a powerful model based on three perspectives: accuracy, usefulness, and instruction obedience.

Table 1: **Performance gaps in using different prompts on FGVC-Aircraft [27] and Oxford-IIIT Pets [31].** More results are in the Appendix. [Choice: using prompt *"Answer with the **letter** from the given choices directly"*. Both: using prompt *"Answer the **correct choice** and its **category name**"*. CoT: using the *CoT* prompt in Fig. 4 (See Appendix for complete prompt).]

Model	Aircraft-102			Pets-37		
	Choice	Both	CoT	Choice	Both	CoT
LLaVA-Next-7B (Mistral)	32.5	10.9	19.4	53.7	37.0	47.5
Idefices2-8B	56.2	2.9	2.5	81.3	2.4	9.4
Finedefics-8B	64.2	64.1	0.0	91.8	87.8	0.0
Qwen2-VL-7B-Instruct	71.2	63.7	18.4	91.0	91.1	77.4

- **Answer Accuracy** (RSC_A): Assesses whether the model's answer correctly identifies the target object based on the given image and question.
- **Reasoning Usefulness** (RSC_U): Evaluates whether the explanation meaningfully supports the prediction, focusing on the relevance of visual attributes and their utility in fine-grained recognition.
- **Instruction Obedience** (RSC_O): Measures whether the model follows the instruction (e.g., structured format), accounting for formatting inconsistencies that may otherwise hinder fair evaluation.

This design addresses two critical challenges: the lack of ground-truth reasoning rationales, and inconsistent structured-output generation across models. By submitting the model's full response to the evaluator, we ensure fairness even when outputs deviate from strict formats. The evaluation prompt and implementation details are provided in the Appendix.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate ReFine-RFT on six widely adopted FGVR benchmarks: CUB-200 [42], FGVC-Aircraft [27], Stanford-Cars [18], Stanford Dogs-120 [18], Flowers-102 [28], and Oxford-IIIT Pets [31]. Adhering to established protocols [14, 12], we frame FGVR as a closed-set multiple-choice task with predefined answer candidates, using identical test splits for fair comparison. Unlike Finedefics [14], which employs 38,254 pretraining and 77,051 fine-tuning samples for hybrid open-/closed-set question answering, our approach prioritizes data efficiency. We randomly select 5 images per category across all datasets, yielding a curated training set of 3,775 samples exclusively on closed-set multiple-choice QA.

Table 2: **Comparison of accuracy on six FGVR datasets under the *Answer-only* setting.** Each dataset includes accuracy from *Choice* and *CoT* prompts. [#P: parameters count; CH.: *Choice* prompt; CoT: *CoT* prompt; InternLM XCom.2: InternLM XComposer 2]

Model	#P	Dog-120		Bird-200		Aircraft-102		Flower-102		Pet-37		Car-196		Avg.	
		CH.	CoT	CH.	CoT	CH.	CoT	CH.	CoT	CH.	CoT	CH.	CoT	CH.	CoT
LLaVA 1.5	7B	39.0	–	35.2	–	34.7	–	51.4	–	52.3	–	46.9	–	43.2	–
MobileVLM v2	7B	39.9	–	33.9	–	35.0	–	54.9	–	53.7	–	46.3	–	44.0	–
InstructBLIP	4B	47.1	–	32.2	–	29.2	–	62.3	–	60.0	–	64.6	–	49.2	–
Phi-3-Vision	4B	39.8	–	37.6	–	42.3	–	51.6	–	56.4	–	54.5	–	47.0	–
BLIP2 Flan-T5-XL	4B	46.2	–	33.7	–	32.9	–	64.3	–	65.0	–	67.7	–	51.6	–
InternLM XCom.2	7B	41.5	–	37.4	–	40.5	–	54.3	–	63.2	–	53.9	–	48.5	–
LLaVA-Next	7B	38.9	31.6	34.9	28.9	32.5	19.4	43.9	39.4	53.7	47.5	49.5	29.7	42.2	32.8
Idefics2	8B	58.0	11.6	47.2	14.2	56.2	2.5	72.8	7.2	81.3	9.4	80.3	10.3	66.0	9.2
Finedefics	8B	73.1	0.0	57.3	0.0	64.2	0.0	**89.6**	0.0	**91.8**	0.0	84.7	0.0	76.8	0.0
Qwen2-VL-Instruct	7B	73.9	60.5	65.3	53.5	71.2	18.4	84.8	61.4	91.0	77.4	90.8	52.0	79.5	53.9
GRPO + R_f	7B	73.7	66.1	65.5	60.7	71.2	64.1	84.9	77.2	91.2	84.9	**91.0**	80.5	79.6	72.3
GRPO + R_f + R_{cls}	7B	73.5	65.9	65.6	59.0	71.3	63.0	84.8	78.1	91.0	85.7	90.8	82.1	79.5	72.3
ReFine-RFT (Ours)	7B	**74.4**	**75.5**	**65.8**	**66.1**	**71.8**	**73.6**	85.2	**86.4**	91.5	**91.3**	90.8	**89.8**	**79.9**	**80.5**

Table 3: **Comparison with leading methods on six FGVR datasets in *Reasoning-Answer* setting scaled in [0,100].** [R_A=RSC_A; R_U=RSC_U; R_O=RSC_O; D-120: Dog-120; B-200: Bird-200; A-102: Aircraft-102; F-102: Flowers102; P-37: Pets-37; C-196: Cars-196; Qwen2: Qwen2-VL-Instruct; GRPO: GRPO+R_f+R_{cls}]

Model	D-120			B-200			A-102			F-102			P-37			C-196			Avg.		
	R_A	R_U	R_O	R_A	R_U	R_O	R_A	R_U	R_O	R_A	R_U	R_O	R_A	R_U	R_O	R_A	R_U	R_O	R_A	R_U	R_O
Idefics2	22.4	28.1	35.3	20.3	32.1	40.8	8.4	20.3	18.6	22.6	31.8	30.3	24.1	29.4	27.6	25.8	23.8	29.9	20.6	27.6	30.4
Finedefics	47.4	9.2	14.7	39.1	8.9	10.3	53.3	9.1	8.6	68.7	9.6	13.7	67.1	10.3	15.6	67.1	9.9	13.4	57.1	9.5	12.7
Qwen2	60.8	60.7	79.3	53.6	56.7	75.9	45.6	35.6	47.1	74.0	65.7	76.9	79.9	71.7	85.8	75.8	38.0	52.6	65.0	54.7	69.6
GRPO	61.2	61.0	79.9	54.8	58.8	78.8	52.9	52.0	75.1	73.3	71.8	82.8	81.1	73.3	87.4	81.4	56.0	75.1	67.5	62.2	79.9
Ours	**72.5**	**75.3**	**93.1**	**62.5**	**67.7**	**89.4**	**59.5**	**64.8**	**86.8**	**84.7**	**82.6**	**95.2**	**88.6**	**82.6**	**96.7**	**87.4**	**81.5**	**94.8**	**75.9**	**75.8**	**92.7**

Evaluated MLLMs. We compare ReFine-RFT with recent MLLMs with comparable parameter size including Finedefic [14], Qwen2-VL-Instruct [43], Idefics2 [19], InternLM XComposer 2 [10], BLIP2 Flan-T5-XL [21], Phi-3-Vision [1], InstructBLIP-Flan-T5-XL [9], MobileVLM v2 [7], LLaVA-Next-Mistral [22], and LLaVA 1.5 [23].

Evaluation Metrics. As detailed in Sec. 4, we assess ReFine-RFT under two distinct evaluation regimes: *Answer-only* (aligned with prior works [14, 12]) and our novel *Reasoning-Answer* setting. We report conventional answer accuracy to quantify final prediction correctness (denoted as *Answer-only* setting), and RSC_A, RSC_U, RSC_O for proposed *Reasoning-Answer* setting. In *Answer-only* , we use both the *Choice* and *CoT* prompts from Tab. 1, while in *Reasoning-Answer* , only the *CoT* prompt is used.

Implementation Details. We select Qwen2-VL-7B-Instruct [43] as the base model and use the LoRA technique [16] to fine-tune the model. We trained with 4 NVIDIA A6000 GPUs with 48G of memory. We use InternVL3-38B [6] as the teacher model. We use $\gamma = 64$ and $\alpha = 128$ for LoRA, and a learning rate of $8e-6$ with 64 as the accumulated batch size. We set the number of generations $G = 6$ and $\beta = 0$ (*i.e.*, no KL-divergence) for GRPO hyperparameters. Reward weights are set evenly as $w_{fa} = w_{cls} = w_{cot} = 1/3$. We use GPT-4.1-mini [2] as a judge for *Reasoning-Answer* . All seeds are fixed across the training and evaluation procedures to ensure reproducibility and fairness. More details on the implementation can be found in the Appendix.

5.2 Main Results

Performance on the FGVR Benchmark. As shown in Tab. 2, ReFine-RFT achieves state-of-the-art performance on nearly all FGVR benchmarks in *Answer-only* setting, outperforming both CoT and Choice, and demonstrating superior reasoning ability beyond simple answer generation. Finedefics receives 0 scores in the CoT prompt because its outputs lack the structured reasoning format required

Table 4: Performance comparison between baseline methods and ReFine-RFT on MMStar, MMMU, and TextVQA validation sets.

Model	Size	MMStar $_{Val}$	MMMU $_{Val}$	TextVQA $_{Val}$
Idefics2	8B	49.0	44.9	73.3
Finedefics	8B	40.0 ($\downarrow$ 9.0%)	33.7 ($\downarrow$ 11.2%)	26.8 ($\downarrow$ 46.5%)
Qwen2-VL-Instruct	7B	60.6	53.6	84.5
ReFine-RFT (Ours)	7B	59.9 ($\downarrow$ 0.7%)	54.3 ($\uparrow$ 0.7%)	84.2 ($\downarrow$ 0.3%)

for answer extraction. To the best of our knowledge, this is the first time CoT-based reasoning outperforms simple answer-based methods in FGVR.

We present the RSC performance of the *Reasoning-Answer* setting in Tab. 3. ReFine-RFT significantly outperforms all baseline methods across all metrics, demonstrating strong capabilities in instruction following, reasoning, and fine-grained classification. We acknowledge that the comparison with baselines may not be entirely fair, as our model has been specifically designed to enhance reasoning and instruction-following capabilities, whereas the baselines may not have received such targeted optimization. Nonetheless, compared with GRPO with classic format and classification rewards, ReFine-RFT outperforms in all metrics, demonstrating the effectiveness of the proposed R_{fa} and R_{cot} for FGVR. We encourage readers to consider the effectiveness and robustness of our proposed method: ReFine-RFT maintains a strong generalization performance across diverse tasks and domains, showing minimal performance degradation.

Generalizability on General MLLM Benchmarks. We select some general Visual Question Answering (VQA) benchmarks: MMStar [5], MMMU [47], and TextVQA [37], and use VLMEvalKit [11] to evaluate the generalizability of ReFine-RFT. The experimental results in Tab. 4 demonstrate the robust generalization capability of ReFine-RFT across diverse vision-language benchmarks. While Finedeifcs fine-tunes from Ideifc2 with notable performance drops in general VQA tasks, ReFine-RFT achieves competitive performance, closely matching the base model, despite being explicitly optimized for fine-grained visual recognition. These results underscore that our fine-tuning strategy enhances reasoning fidelity for FGVR without compromising the model's inherent versatility, thereby validating its effectiveness in preserving and transferring knowledge across heterogeneous vision-language tasks.

5.3 Ablation Studies

Effects of Fine-tuning Method. We benchmark ReFine-RFT against Supervised Fine-Tuning (SFT) in the *Answer-only* setting (Tab. 5a), using identical training data. SFT exhibits performance degradation compared to zero-shot, especially using the *CoT* prompt. In contrast, ReFine-RFT achieves superior performance, underscoring its efficacy in harmonizing data-efficient training with enhanced classification performance while mitigating overfitting risks.

Effects of Reward Functions. Tab. 5b highlights the impact of the proposed rewards R_{fa} and R_{cot}. Initial training with the format reward alone ensures structural compliance but lacks answer correctness. Incorporating R_{cls} does not improves classification accuracy in CoT, while replacing R_f with R_{fa} significantly improves by 4.3%. The addition of R_{cot} further enhances results ($+3.9\%$), underscoring its critical role in guiding to the correct answers. This progression demonstrates that R_{fa} and R_{cot} are key to fostering task-aware understanding and promoting attention to semantically meaningful visual features during decision-making. More details are in the Appendix.

Effects of Reward Weights. To investigate the role of each reward component, we perform an ablation on different weight configurations. As shown in Tab. 5c, the performance remains consistent across various weight combinations, with CoT scores ranging narrowly between 80.4 and 80.6. This indicates that our approach is relatively insensitive to the exact setting of reward weights and the three reward functions contribute complementarily to the final performance.

Case Analysis. We visualize an example to demonstrate the reasoning capability of ReFine-RFT compared with other baseline methods in Fig. 5. While baseline methods overlook instructions

Table 5: **Ablation studies of ReFine-RFT in *Answer-only* settings.** We report average performance. [CH.: *Choice* prompt; Zero: zero-shot performance of Qwen2-VL-7B-Instruct; R_{fa}: format reward with attribute tags.]

(a) Effects of FT methods.

Method	CH.	CoT
Zero	79.5	53.9
SFT	78.0	47.6
Ours	79.9	80.5

(b) Effects of reward functions.

R_f	R_{fa}	R_{cls}	R_{cot}	CoT
✓				72.3
✓		✓		72.3
	✓	✓		76.6
	✓	✓	✓	80.5

(c) Effects of reward weights.

w_{fa}	w_{cls}	w_{cot}	CoT
0.33	0.33	0.33	80.5
0.20	0.50	0.30	80.6
0.20	0.30	0.50	80.4

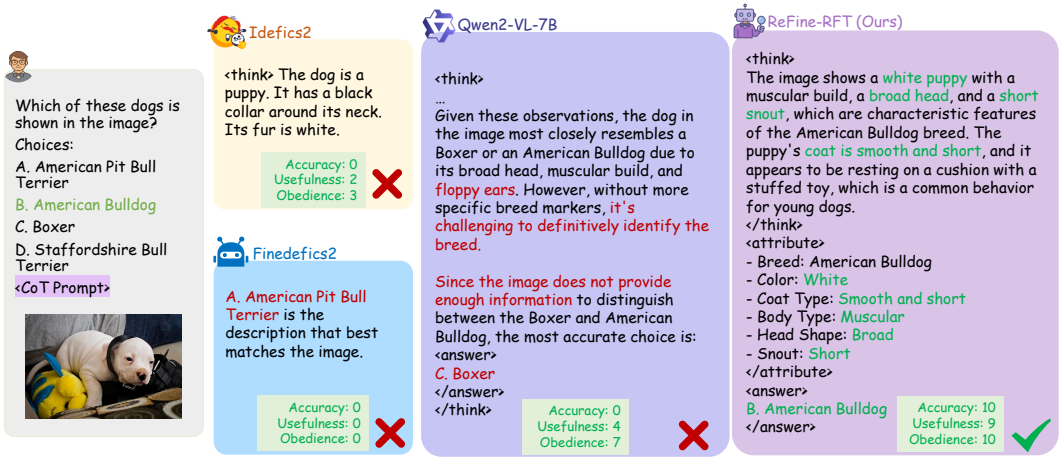

Figure 5: **Responses and *RSC* scores comparison.** ReFine-RFT outperforms all baseline methods w.r.t. reasoning quality, answer accuracy, and instruction-following. The completed responses and GPT feedback are provided in the Appendix.

or yield weak reasoning, ReFine-RFT delivers stronger instruction alignment and more accurate reasoning, demonstrating greater reliability in FGVR. Specifically, ReFine-RFT produces fine-grained, attribute-based explanations grounded in clear visual evidence, such as coat texture, head shape, and snout structure, which directly support and justify the final classification. In contrast, baseline models show notable shortcomings: Qwen2-VL-7B tends to hallucinate attributes not present in the image, Idefics2 often generates irrelevant or nonsensical responses, and Finedefics2 ignores the instruction and misclassifies the breed entirely. These examples underscore the robustness of ReFine-RFT in reasoning and visual grounding. Additional qualitative comparisons and feedback from GPT are provided in the Appendix.

6 Conclusion

In this work, we identified key limitations of existing MLLMs in fine-grained visual recognition, particularly their susceptibility to overfitting, reasoning misalignment, and inadequate attention to discriminative attributes. To address these challenges, we introduce ReFine-RFT, a RFT framework that enhances reasoning fidelity and classification accuracy through a hybrid reward mechanism combining rule-based and MLLM-based incentives. Our method not only achieves state-of-the-art results across six FGVR benchmarks but also ensures robust generalization without requiring large-scale annotated datasets. Experimental results validate the effectiveness of our approach, with significant performance gains over existing methods while maintaining competitive performance on general vision-language tasks. This work underscores the importance of reasoning-answer alignment and external feedback in developing reliable MLLMs for mission-critical applications.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).

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

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

[4] Junwen Chen, Jie Zhu, and Yu Kong. 2023. ATM: Action Temporality Modeling for Video Question Answering. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4886–4895.

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint arXiv:2403.20330* (2024).

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.

[7] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766* (2024).

[8] Chenhang Cui, An Zhang, Yiyang Zhou, Zhaorun Chen, Gelei Deng, Huaxiu Yao, and Tat-Seng Chua. 2024. Fine-Grained Verifiers: Preference Modeling as Next-token Prediction in Vision-Language Alignment. *arXiv preprint arXiv:2410.14148* (2024).

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2024. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2402.03766* (2024).

[10] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420* (2024).

[11] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11198–11201.

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

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

[14] Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. 2025. Analyzing and Boosting the Power of Fine-Grained Visual Recognition for Multi-modal Large Language Models. *arXiv preprint arXiv:2501.15140* (2025).

[15] Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071* (2022).

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

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

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

[19] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246* (2024).

[20] Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848* (2024).

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. PMLR, 19730–19742.

[22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26296–26306.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

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

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

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

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

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

[29] 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. 2021. Show Your Work: Scratchpads for Intermediate Computation with Language Models.

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

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

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

[33] Zhiyuan Ren, Yiyang Su, and Xiaoming Liu. 2023. ChatGPT-powered hierarchical comparisons for image classification. *Advances in neural information processing systems* 36 (2023), 69706–69718.

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

[35] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061* (2022).

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

[37] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8317–8326.

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

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

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

[41] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9568–9578.

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

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

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[45] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. 2021. Fine-grained image analysis with deep learning: A survey. *TPAMI* 44, 12 (2021), 8927–8948.

[46] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. 2025. Perception-R1: Pioneering Perception Policy with Reinforcement Learning. *arXiv preprint arXiv:2504.07954* (2025).

[47] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9556–9567.

[48] Chen Zhang, Chengguang Tang, Dading Chong, Ke Shi, Guohua Tang, Feng Jiang, and Haizhou Li. 2024. TS-Align: A Teacher-Student Collaborative Framework for Scalable Iterative Finetuning of Large Language Models. *arXiv preprint arXiv:2405.20215* (2024).

[49] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. 2023. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research* (2023).

[50] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

[51] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.