Revisiting the Data Sampling in Multimodal Post-training from a Difficulty-Distinguish View

Jianyu Qi[1, 2*], **Ding Zou**[2], **Wenrui Yan**[2], **Rui Ma**[2], **Jiaxu Li**[1], **Zhijie Zheng**[1], **Zhiguo Yang**[2], **Rongchang Zhao**[1, †]

[1]School of Computer Science, Central South University, Changsha, Hunan, China
[2]Intelligent System Department, Zhongxing Telecom Equipment(ZTE), Changsha, Hunan, China
qijianyu@csu.edu.cn, zoudinghust@gmail.com, ywraddlk@gmail.com, 214711069@csu.edu.cn, lijiaxu@csu.edu.cn, zhengzhijie@csu.edu.cn, yang.zhiguo@zte.com.cn, zhaorc@csu.edu.cn

Abstract

Recent advances in Multimodal Large Language Models (MLLMs) have spurred significant progress in Chain-of-Thought (CoT) reasoning. Building on the success of Deepseek-R1, researchers extended multimodal reasoning to post-training paradigms based on reinforcement learning (RL), focusing predominantly on mathematical datasets. However, existing post-training paradigms tend to neglect two critical aspects: (1) The lack of quantifiable difficulty metrics capable of strategically screening samples for post-training optimization. (2) Suboptimal post-training paradigms that fail to jointly optimize perception and reasoning capabilities. To address this gap, we propose two novel difficulty-aware sampling strategies: Progressive Image Semantic Masking (PISM) quantifies sample hardness through systematic image degradation, while Cross-Modality Attention Balance (CMAB) assesses cross-modal interaction complexity via attention distribution analysis. Leveraging these metrics, we design a hierarchical training framework that incorporates both GRPO-only and SFT+GRPO hybrid training paradigms, and evaluate them across six benchmark datasets. Experiments demonstrate consistent superiority of GRPO applied to difficulty-stratified samples compared to conventional SFT+GRPO pipelines, indicating that strategic data sampling can obviate the need for supervised fine-tuning while improving model accuracy. Our code will be released at https://github.com/qijianyu277/DifficultySampling.

Introduction

Recent advancement of Multimodal Large Language Models (MLLMs) has witnessed a rapid development in Chain-of-Thought (CoT) reasoning (Wei et al. 2022; Mitra et al. 2024; Kumar et al. 2025), driven largely by innovative post-training techniques that align model behaviors with human reasoning patterns. These advancements are particularly significant as they enable MLLMs to handle complex tasks involving both visual and textual information more effectively. For instance, reinforcement learning frameworks like Group Relative Policy Optimization (GRPO) (Shao et al. 2024) have empowered MLLMs to autonomously discover reasoning paths through reward signals, enhancing their ability to perform intricate reasoning tasks.

Significant efforts have been devoted to enabling CoT reasoning in MLLMs (Zhang et al. 2024; Mitra et al. 2024; Zheng et al. 2023). Previous methods (Xu et al. 2024; Zhang et al. 2023; Guo et al. 2024; Thawakar et al. 2025) construct the datasets manually containing step-level reasoning processes and apply supervised fine-tuning (SFT) to reformat MLLMs' outputs, whose manually designed "MLLs with formatted reasoning outputs" often results in "Pseudo-CoT" reasoning (Huang et al. 2025; Gao et al. 2024; Zheng et al. 2023), lacking essential cognitive processes commonly observed in human thoughts. Then with the success of Deepseek-R1 (Guo et al. 2025), many researchers (Li et al. 2024; Tong et al. 2024; Wang et al. 2022; Luo et al. 2025) have attempted to extend this success to multimodal reasoning, where models process and reason on both visual and textual information. However, mainstream works (Liu et al. 2025a; Meng et al. 2025; Chen et al. 2025) focus on performing RL with multimodal mathematical datasets, which improves the reasoning ability more in terms of textual modality but stresses less cross-modal ability. More recently proposed methods (Ren et al. 2024; Ma et al. 2025; Yu et al. 2025a) focus on perception-enhanced RL training, via incorporating multimodal data (such as detection, grounding, etc.) into Reinforcement Learning with Verifiable Rewards (RLVR) and designing corresponding reward functions or models.

Despite effectiveness, current multimodal post-training methods commonly ignore two crucial questions:

- **How to identify multimodal data of distinct hardness?** Recently proposed RL methods unanimously agreed that proper difficulty means a lot in RL training (Wang et al. 2025b,c; Zhang et al. 2025). But the core problem is that, multimodal data could not be divided into distinguished parts the same way as pure-text data (especically the math or code data), due to their multi-modal feature. It is clear that the text-modality difficulty could not be considered as the sample-difficulty for MLLM, and in many cases text-modality difficulty is not quantizable for MLLM (such as OCR and classify tasks). As a result, a suitable principle or definition for multi-modal hardness is supposed to consider cross-modal features, so as to correctly measure the hardness for MLLMs.

*Work done during internship at ZTE.

†Corresponding author.

- **How to design effective post-training paradigm for MLLM?** After distinguishing the hardness of various multi-modal data, it is necessary to design effective training pipline for MLLM. A straightfoward idea is to follow mainstream method to perform cold-start SFT with the hardest samples then RL with the medium difficulty. However, such a commonly-recognized paradigm is not always sufficient for MLLM post-training, owing to the multi-task features. Actually in post-training of MLLMs, multi-modal data could be roundly divided into two groups, i.e., Visual-Reasoning (Math, Science, Charting, and Puzzle, etc.) and Visual-Perception (Grounding, Counting, and OCR, etc.), it is necessary to design the optimal post-training scheme for each type of dataset.

To address the aforementioned issues, we define a hardness discrimination strategy for multimodal data from both intra-modal and cross-modal perspectives. We then uncover the distributional discrepancies between reasoning and perception data, and propose respective optimal post-training pipelines tailored to different types of datasets, aiming to enhance the reasoning and perception capabilities of multimodal learning models. Our main contributions can be summarized as follows:

- We propose the **Progressive Image Semantic Masking (PISM)**, a random masking strategy at the semantic level of images. By gradually increasing the masking ratio, we observe changes in the model's response state to determine the difficulty of samples.
- We propose the **Cross-Modality Attention Balance (CMAB)**, an attention-based strategy that considers the attention scores of the response tokens generated by the model with respect to the text tokens and image tokens in the original input respectively. The degree of interaction between text and image tokens during response generation is measured by the ratio of their attention scores. Sample difficulty is then classified based on this interaction strength.
- We design and verify the optimal training scheme for different types of datasets, rather than defaulting to supervised fine-tuning and reinforcement fine-tuning.

Related Work

Multimodal Large Language Models Reasoning

The rapid growth of Multimodal Large Language Models (MLLMs) (Chen et al. 2024b; Ma et al. 2025; Su et al. 2025) has endowed them with extensive knowledge and robust multitasking capabilities, enabling their application across complex and diverse domains, particularly for cross-modal tasks such as visual question answering and image captioning. As the requirements for the capabilities of multimodal large models in real-world scenarios gradually increase, the ability of the thinking chain has also been gradually introduced into multimodal large models, such as Llava-COT (Xu et al. 2024), leverage chainof-thought to enhance MLLMs reasoning capabilities. Other similar works employ MCTS-based methods to strengthen their reasoning abilities, such as Mulberry (Yao et al. 2024) which introduces collective knowledge to MCTS to search reasoning paths, thereby enhancing reasoning and reflection capabilities. However, such works which obtain CoT through supervised fine-tuning training often faces issues of weak generalization and high training costs due to the superfacial matching training, which encourages more researchers to design training paradigms that can adapt to the characteristics of large multimodal models to effectively guide complex reasoning processes. Hence more recently proposed MLLMs focus on bringing RL-based reasoning into MLLMs, extend the DeepSeek-R1 training paradigm from large language models to MLLMs, such as VLM-R1 (Shen et al. 2025), Visual-RFT (Liu et al. 2025c), Vision-R1 (Huang et al. 2025), and Perception-R1 (Yu et al. 2025a).

Reinforcement Learning for Post-Training

With the popularity of DeepSeek-R1, several recent studies, including Open-Reasoner-Zero (Hu et al. 2025), SimpleRL-Zoo (Zeng et al. 2025), and Logic-RL (Xie et al. 2025), have explored directly RL post-training with GRPO, without any supplementary supervised fine-tuning stages. Additionally, complementary approaches such as VAPO (Yue et al. 2025), DAPO (Yu et al. 2025b), and Dr.GRPO (Liu et al. 2025b) have sought to refine the GRPO framework by optimizing reward design and enhancing advantage estimation techniques, thus more effectively promoting deeper reasoning behaviors within language models. Meanwhile, recent analyses focus on the effect of data hardness on the RL training stage, such as (Huang et al. 2025) rank samples with human-pre-defined math difficulty, (Xiong et al. 2025) classify sample difficulty through rejection sampling, (Zhang and Zuo 2025) arranges through reward score, and (Wang et al. 2025a) proposes to measure the hardness with the sentence entropy. However, current multimodal post-training methods commonly ignoring the importance of data sampling in multi-modal data, neither ignore the data sampling operation nor filter with text-only principle, such as whether the problem is hard. As a result, image-modal and mutli-interaction signals fail to be modeled, bring a suboptimal sampling strategy and further resulting worse model performance. We hence propose to design a sampling strategy for multi-modal data from both intra-modal and inter-modal perspectives, which could involve cross-modal signals for multimodal post-training.

Methodology

Measurement of the difficulty of multimodal samples remains a fundamental challenge in post-training data curation. We introduce two complementary approaches that capture different aspects of sample complexity: **Progressive Image Semantic Masking (PISM)** focusing on model sensitivity to visual perturbations, and **Cross-Modality Attention Balance (CMAB)** examining the balance of cross-modal interactions during inference.

PISM for sensitivity-based Difficulty Assessment

Intuitively, challenging multimodal samples should exhibit higher sensitivity to visual information loss—when criti-

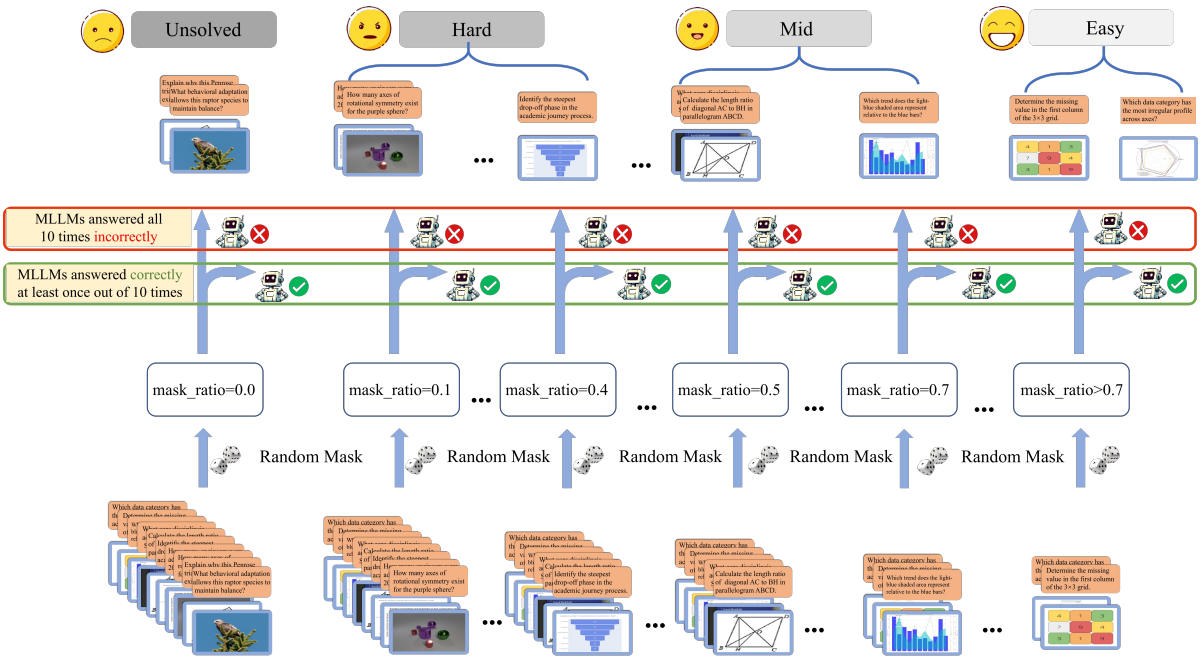

Figure 1: Illustration of the **PISM** (Progressive Image Semantic Masking) method. We progressively mask different portions of the image, from no masking ($mask\_ratio = 0.0$) to heavy masking ($mask\_ratio > 0.7$). Each masked image is created by randomly hiding a certain percentage of pixels. The process simulates varying levels of visual information loss. The model's performance is then evaluated on these masked images to understand how much it relies on visual details for accurate reasoning.

cal visual content is obscured, the model's performance degrades more rapidly compared to easier samples. We operationalize this insight through a systematic masking strategy that gradually removes visual information while monitoring prediction stability.

Mask-based Sensitivity Quantification Given an image-text pair $s = (I, Q)$, we systematically probe the model's visual dependence through controlled perturbation experiments. As shown in Figure 1, we define a series of masking ratios $\Lambda = \{\lambda_i \mid \lambda_i = 0.0, 0.1, 0.2, \ldots, 0.9\}$, spanning from the original unmodified image ($\lambda = 0.0$) to heavily degraded versions where 90% of pixels are occluded.

For each masking level λ_i, we apply the perturbation operation $M(\cdot, \lambda_i)$ that randomly selects and masks the specified proportion of pixels in the original image, yielding $I_{\lambda_i} = M(I, \lambda_i)$. This masking occurs directly in pixel space prior to any feature extraction, thereby simulating realistic scenarios of visual information loss or corruption that might occur in real-world applications.

We then evaluate model performance by feeding each perturbed sample (I_{λ_i}, Q) to the multimodal model $\mathcal{M}$ and obtaining the prediction $A_{\lambda_i} = \mathcal{M}(I_{\lambda_i}, Q)$. The correctness of each prediction is assessed using a binary indicator:

$$\delta_{\lambda_i} = 1[\mathcal{C}(A_{\lambda_i}, A_{\text{gt}})] \tag{1}$$

where $\mathcal{C}$ evaluates whether the predicted answer matches the ground truth A_{gt}.

Given the stochastic nature of random masking, we repeat this evaluation process $K = 10$ times with independent mask realizations for each masking ratio. The robust accuracy estimate is then computed as:

$$P_c(\lambda_i) = \frac{1}{K} \sum_{k=1}^{K} \delta_{\lambda_i}^{(k)} \tag{2}$$

The critical insight lies in identifying the failure threshold λ_s^*—the minimal masking ratio where performance drops below an acceptable level. Formally, we define:

$$\lambda_s^* = \min\{\lambda_i \in \Lambda \mid P_c(\lambda_i) < \tau\} \tag{3}$$

where the threshold τ (we set $\tau = 0.1$)captures the transition point from reliable to unreliable predictions.

Sample Difficulty Classification via PISM The critical masking ratio λ_s^* naturally partitions samples into distinct difficulty categories:

- **Hard Samples:** $\lambda_s^* \leq \lambda_{\text{hard}}$ (we set $\lambda_{\text{hard}} = 0.4$). These samples exhibit fragility to minor visual perturbations, suggesting they require sophisticated visual understanding and tight multimodal coupling.
- **Medium Samples:** $\lambda_{\text{hard}} < \lambda_s^* < \lambda_{\text{easy}}$. These samples show moderate sensitivity to visual masking, indicating that while visual information contributes meaningfully to correct predictions, textual cues alone may still lead to partial or inconsistent performance.
- **Easy Samples:** $\lambda_s^* \geq \lambda_{\text{easy}}$ (we set $\lambda_{\text{easy}} = 0.7$) or undefined (performance remains above τ across all masking levels). Such samples demonstrate robustness to visual

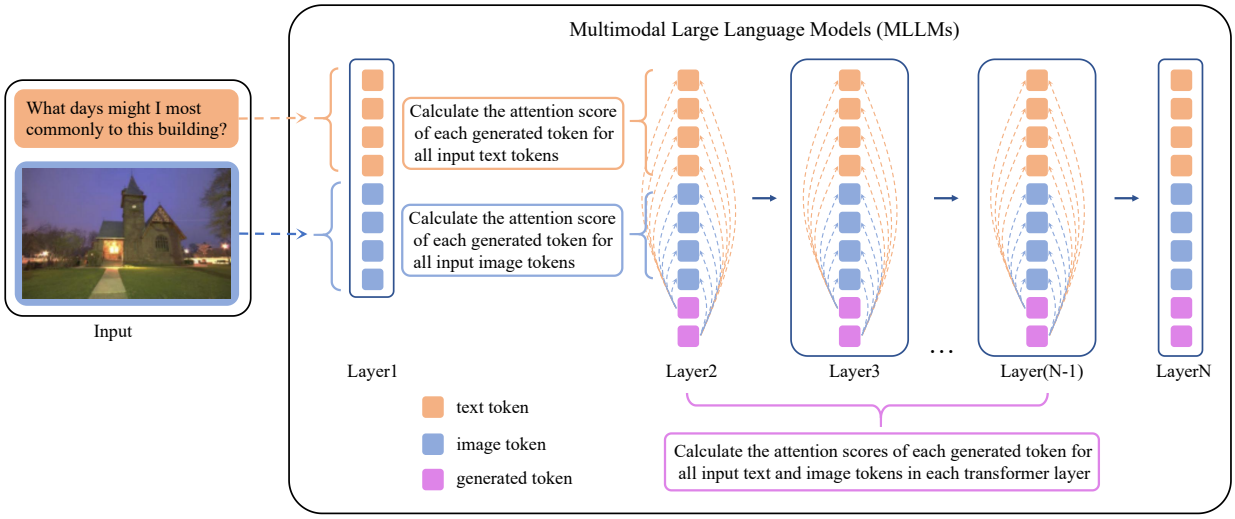

Figure 2: Illustration of the **CMAB** (Cross-Modality Attention Balance) method. For each generated token, we calculate its average attention score over the input text tokens and image tokens across all transformer layers, and then average these scores across all generated tokens. N represents the total number of layers of the transformer.

degradation, indicating that textual information alone largely suffices for correct responses.

- **Unsolved Samples:** Samples with incorrect predictions on the original image ($P_c(0.0) < \tau$) are excluded from difficulty assessment, as their errors reflect model limitations rather than inherent sample complexity.

CMAB for attention-based Difficulty Assessment

While random masking methods based on the semantic layer of images can screen out difficult samples from the perspective of model sensitivity, it may not capture samples that require complex reasoning. We complement our approach by analyzing how the model allocates attention between different modalities during reasoning, which serves as a measure of the cognitive demands imposed by different sample types. A sample is considered difficult when the model allocates balanced attention to both text and images during question answering, as this indicates that information from both modalities is necessary to arrive at the correct answer.

Attention-based Balance Quantification As shown in Figure.2, for input sequences comprising image tokens $\mathbf{X}^{\text{img}} \in \mathbb{R}^{L_{\text{img}} \times d}$ and text tokens $\mathbf{X}^{\text{txt}} \in \mathbb{R}^{L_{\text{txt}} \times d}$, we analyze how the model distributes attention during the generation of each output token y_t in the response sequence $\mathbf{Y} = \{y_1, \ldots, y_T\}$.

During the generation of token y_t at layer l, the cross-attention mechanism produces weights $\mathbf{A}^{(l,t)} \in \mathbb{R}^{1 \times (L_{\text{img}} + L_{\text{txt}})}$ that reveal the model's focus distribution across input tokens. We decompose this attention into modality-specific components by computing the total attention allocated to visual tokens as $S_{\text{img}}^{(l,t)} = \sum_{i=1}^{L_{\text{img}}} A_i^{(l,t)}$ and the attention to textual tokens as $S_{\text{txt}}^{(l,t)} = \sum_{i=L_{\text{img}}+1}^{L_{\text{img}}+L_{\text{txt}}} A_i^{(l,t)}$.

The ratio $\rho^{(l,t)} = S_{\text{img}}^{(l,t)} / S_{\text{txt}}^{(l,t)}$ captures the instantaneous attention balance at layer l for token t. Values significantly above 1 indicate visual dominance, while values below 1 suggest textual preference. To obtain a stable estimate across the model's multiple layers, we compute the geometric mean:

$$\rho_t = \exp\left(\frac{1}{L_{\text{layers}} - 2} \sum_{l=2}^{L_{\text{layer}}-1} \log(\rho^{(l,t)} + \epsilon)\right) \quad (4)$$

where $\epsilon \approx 10^{-8}$ prevents numerical instabilities when attention ratios approach zero. We exclude the first and last transformer layers when computing average attention scores, as they mainly handle input encoding and output decoding, with limited role in high-level semantic or cross-modal reasoning.

The sample-level attention balance emerges as the arithmetic mean across all generated tokens: $\bar{\rho} = \frac{1}{T} \sum_{t=1}^{T} \rho_t$. This metric provides a holistic view of how the model balances visual and textual information throughout the entire response generation process.

Sample Difficulty Classification via CMAB Based on the attention score, we have the following difficulty classification:

- **Easy Samples:** The attention balance ratio $\bar{\rho}$ serves as a window into the cognitive demands imposed by different sample types. For samples that the model answers correctly, extreme values of $\bar{\rho}$ typically indicate straightforward cases where one modality dominates the reasoning process. When $\bar{\rho} < 0.1$, the model relies heavily on textual information, suggesting that the visual content contributes minimally to the final answer. Conversely, when $\bar{\rho} > 1.9$, visual information takes precedence, often indicating questions that can be resolved through direct visual inspection without complex textual reasoning.

- **Medium Samples:** $0.1 \leq \bar{\rho} < 0.4$ or $1.6 < \bar{\rho} \leq 1.9$. These samples are primarily driven by one modality, but still require meaningful support from the other. They demand more than superficial integration, yet fall short of the balanced multimodal reasoning seen in hard samples.
- **Hard Samples:** The most intriguing cases emerge when $\bar{\rho}$ falls within a moderate range, approximately $[0.4, 1.6]$. This balanced attention allocation suggests that successful reasoning requires effective integration of both visual and textual information—neither modality alone provides sufficient context for arriving at the correct answer. Such samples typically involve complex spatial reasoning, visual-textual alignment, or nuanced interpretation that demands sophisticated cross-modal understanding.
- **Usolved Samples:** For samples that the model answers incorrectly, we categorize them as unsolved regardless of their attention patterns. While the attention distribution might provide insights into the model's reasoning process, the primary concern for difficulty assessment is whether the model can successfully leverage the available information to reach the correct conclusion.

The dual-perspective approach captures both the fragility of visual dependencies and the complexity of multimodal reasoning, providing a comprehensive understanding of sample difficulty.

Experiments

Experimental Setup

Our empirical investigation employs the multimodal perception-reasoning benchmark established by (Ma et al. 2025). Experiments were executed on a computing cluster comprising five nodes equipped with NVIDIA A800-SXM4 GPUs (8×80GB memory per node) and two nodes with NVIDIA H20 GPUs (8×96GB memory per node). The implementation leverages PyTorch as the foundational computational framework, where supervised fine-tuning (SFT) was conducted using the LLaMA-Factory framework (Zheng et al. 2024), while GRPO training was implemented through the Swift framework (Zhao et al. 2025) for reward-constrained optimization. All methodologies were systematically evaluated on the Qwen2.5VL-7B foundation model (Bai et al. 2025), ensuring controlled comparisons across experimental conditions.

Evaluation Framework and Benchmarks

Our evaluation encompasses general visual question answering through MMVet (Yu et al. 2023), which test fundamental visual comprehension and reasoning abilities. Mathematical and chart interpretation capabilities are assessed using MathVistar (Lu et al. 2023) and MMMU (Yue et al. 2024), both requiring sophisticated numerical reasoning and visual analysis. For multimodal knowledge integration and complex reasoning, we employ MMStar (Chen et al. 2024a). Document understanding and optical character recognition are evaluated through OCRBench (Liu et al. 2024), while hallucination detection capabilities are measured via HallusionBench (Guan et al. 2024). All benchmarks are evaluated within the OpenCompass framework (Buitrago and Nystrom 2019), using GPT-4o-mini (Hurst et al. 2024) as the unified judge model for consistent and reliable scoring.

Training Setup: GRPO-Only vs. SFT+GRPO

We use **PISM** and **CMAB** to classify samples in the perception and reasoning datasets into difficulty levels, respectively. The resulting sample distributions and associated training strategies are detailed in *Appendix Tables A–D*. Based on these classifications, we compare two post-training paradigms: **(1) GRPO-only**, which applies Group Relative Policy Optimization directly to the full dataset; and **(2) SFT+GRPO**, which first performs supervised fine-tuning (SFT) on a curated subset before applying GRPO. Our goal is to evaluate whether pure reinforcement learning fine-tuning or a hybrid approach yields better performance on perception and reasoning tasks. For the SFT+GRPO paradigm, we further investigate the impact of training order—specifically, how the sequencing of medium- and high-difficulty samples during SFT influences final performance—enabling a more refined training paradigm.

Results and Analysis

The final experimental results are reported in Table 1 and Table 2, which present performance metrics (HBench stands for HallusionBench) for visual perception and visual reasoning datasets, respectively. The **bold** numbers represent the best results, and the * represents the suboptimal results. Here, "mid" denotes medium samples, "rand_m" refers to a random dataset of the same size as the medium sample, and "rand_h" refers to a random dataset of the same size as the hard samples. It is worth noting that due to the large amount of data in the full dataset and unsolved data, their GRPO-only results are only used as a reference and are not included in the comparison of results with other training strategies.

Efficacy of Difficulty-Aware Sampling Strategies The experimental results consistently validate the superiority of our proposed difficulty-aware sampling strategies (PISM and CMAB) across both visual perception and reasoning tasks. As demonstrated in Tables 1–4, models trained on difficulty-stratified samples (mid+hard) using GRPO-only paradigm outperform those trained with conventional SFT+GRPO pipelines, highlighting the critical role of strategic data selection in multimodal post-training. **PISM** exhibits particular strength in tasks requiring robust visual perception. On OCRBench, which evaluates optical character recognition capabilities, the GRPO-only(mid+hard) configuration achieves a score of 77.800 (Table 1), surpassing all SFT+GRPO variants by at least 1.3 points. This performance gain can be attributed to PISM's ability to identify samples where visual information is irreplaceable—by systematically masking image pixels and measuring performance degradation, PISM effectively isolates samples that demand precise visual understanding. The significant drop in performance when using random samples (GRPO-only(random): 77.300) confirms that indiscriminate data inclusion dilutes training efficiency by introducing redundant samples where textual cues suffice. **CMAB**

Training paradigm	MathVista	MMVet	OCRBench	HBench	MMMU	MMStar
GRPO-only(fullset)	53.400	41.697	76.200	67.403	0.440	0.607
GRPO-only(unsolved)	67.200	50.183	78.700	69.295	0.537	0.629
SFT(mid)+GRPO(hard)	67.300	40.596	75.000	68.454	0.507	0.609
SFT(mid)+GRPO(rand_h)	67.700	38.578	74.800	**69.085**	0.504	0.606
SFT(rand_m)+GRPO(hard)	67.500	41.972	75.100	68.349	0.506	0.606
SFT(rand_m)+GRPO(rand_h)	66.600	42.248	74.700	68.980*	0.509	0.609
SFT(hard)+GRPO(mid)	67.300	39.312	74.200	67.613	0.502	0.608
SFT(hard)+GRPO(rand_m)	67.600	39.404	74.800	67.087	0.501	0.610
SFT(rand_h)+GRPO(mid)	67.200	42.661	74.700	68.559	0.504	0.598
SFT(rand_h)+GRPO(rand_m)	67.400	41.560	74.500	68.875	0.501	0.603
GRPO-only(random)	68.200*	**53.257**	77.300*	68.349	0.541*	0.637*
GRPO-only(mid+hard)	**68.300**	48.257*	**77.800**	68.770	**0.547**	**0.639**

Table 1: Comparison of training results using SFT+GRPO and GRPO-only on the visual reasoning dataset through PISM

Training paradigm	MathVista	MMVet	OCRBench	HBench	MMMU	MMStar
GRPO-only(fullset)	70.000	51.147	77.200	68.034	0.557	0.615
GRPO-only(unsolved)	69.200	52.385	77.700	68.875	0.541	0.609
SFT(mid)+GRPO(hard)	67.500	38.119	76.000	66.877	0.496	0.625
SFT(mid)+GRPO(rand_h)	67.400	38.624	75.800	66.141	0.498	**0.628**
SFT(rand_m)+GRPO(hard)	67.400	39.266	76.400	68.454	0.490	0.614
SFT(rand_m)+GRPO(rand_h)	66.500	44.771	75.600	69.085	0.498	0.605
SFT(hard)+GRPO(mid)	67.900*	42.844	76.500	68.033	0.512	0.625
SFT(hard)+GRPO(rand_m)	67.500	41.101	76.000	68.454	0.512	0.610
SFT(rand_h)+GRPO(mid)	65.600	45.459	75.000	68.770*	0.503	0.612
SFT(rand_h)+GRPO(rand_m)	67.700	37.982	75.700	68.665	0.488	0.613
GRPO-only(random)	**68.100**	49.908*	77.300*	68.559	**0.553**	0.627*
GRPO-only(mid+hard)	67.600	**52.477**	**77.600**	**69.716**	0.544*	0.625

Table 2: Comparison of training results using SFT+GRPO and GRPO-only on the visual perception dataset through PISM

Training paradigm	MathVista	MMVet	OCRBench	HBench	MMMU	MMStar
GRPO-only(fullset)	53.400	41.697	76.200	67.403	0.440	0.607
GRPO-only(unsolved)	69.340	54.450	78.300	68.244	0.547	0.636
SFT(mid)+GRPO(hard)	67.200	33.486	74.300	67.298	0.499	0.627
SFT(mid)+GRPO(rand_h)	67.700	33.991	74.300	65.615	0.508	0.621
SFT(rand_m)+GRPO(hard)	67.800	34.541	72.300	67.718	0.502	0.625
SFT(rand_m)+GRPO(rand_h)	67.600	33.394	73.500	64.984	0.498	0.622
SFT(hard)+GRPO(mid)	67.400	34.037	75.200	68.244*	0.501	0.618
SFT(hard)+GRPO(rand_m)	66.900	36.422	74.400	68.034	0.499	0.619
SFT(rand_h)+GRPO(mid)	67.300	34.541	75.300	68.139	0.500	0.617
SFT(rand_h)+GRPO(rand_m)	67.700	34.266	74.200	67.087	0.494	0.619
GRPO-only(random)	68.200*	43.624*	**77.300**	69.085	**0.556**	**0.642**
GRPO-only(mid+hard)	**69.000**	**48.578**	77.100*	**69.085**	0.542*	0.628*

Table 3: Comparison of training results using SFT+GRPO and GRPO-only on the visual reasoning dataset through CMAB

Training paradigm	MathVista	MMVet	OCRBench	HBench	MMMU	MMStar
GRPO-only(fullset)	70.000	51.147	77.200	68.034	0.557	0.615
GRPO-only(unsolved)	68.700	54.541	77.700	69.085	0.536	0.615
SFT(mid)+GRPO(hard)	66.800	41.239	75.100	68.244	0.503	0.627*
SFT(mid)+GRPO(rand_h)	66.500	42.431	75.200	67.823	0.499	0.626
SFT(rand_m)+GRPO(hard)	67.800	36.514	75.100	68.875	0.499	0.625
SFT(rand_m)+GRPO(rand_h)	67.500	42.798	75.000	68.769	0.496	0.623
SFT(hard)+GRPO(mid)	67.400	34.037	75.200	68.244	0.501	0.618
SFT(hard)+GRPO(rand_m)	67.900	48.945	75.900	67.718	0.538	0.609
SFT(rand_h)+GRPO(mid)	68.100*	49.500	76.500	68.454	0.534	0.607
SFT(rand_h)+GRPO(rand_m)	67.600	50.321*	**77.500**	68.980*	0.526	0.610
GRPO-only(random)	67.700	45.550	76.900*	**69.401**	0.545*	0.625
GRPO-only(mid+hard)	**68.300**	**50.367**	76.800	68.244	**0.550**	**0.629**

Table 4: Comparison of training results using SFT+GRPO and GRPO-only on the visual perception dataset through CMAB

shows greater efficacy in complex reasoning tasks that require tight integration of visual and textual information. On MathVista, which assesses mathematical reasoning in visual contexts, CMAB-stratified mid+hard samples yield a GRPO-only score of 69.000 (Table 3), outperforming PISM-based training (68.300) and all SFT+GRPO configurations. This advantage stems from CMAB's unique capability to quantify cross-modal interaction complexity—by analyzing attention distribution between image and text tokens, CMAB identifies samples requiring balanced multimodal processing ($\rho \in [0.4, 1.6]$). The consistent outperformance on MMStar (0.639 vs. 0.625 in SFT+GRPO) further validates that attention balance is a robust indicator of reasoning difficulty. Notably, both strategies demonstrate complementary strengths: PISM excels in perception-heavy tasks (OCRBench, MMVet) by focusing on visual sensitivity, while CMAB dominates reasoning tasks (MathVista, MMMU) through attention analysis. This complementarity confirms that multimodal difficulty assessment requires a dual perspective encompassing both intra-modal sensitivity and inter-modal interaction.

Superiority of GRPO-Only Paradigm Our experiments reveal a counterintuitive finding: GRPO-only training on difficulty-stratified samples consistently outperforms the widely adopted SFT+GRPO pipeline across all evaluated benchmarks. This challenges the prevailing assumption that supervised fine-tuning is a necessary prerequisite for effective reinforcement learning in multimodal systems.

The key advantage of GRPO-only lies in its ability to avoid "Pseudo-CoT" reasoning patterns induced by SFT. As shown in HallusionBench results (Table 2), GRPO-only(mid+hard) achieves a score of 69.716, exceeding SFT+GRPO variants which top out at 68.980. This indicates that SFT's reliance on manually designed reasoning templates may encourage surface pattern matching rather than genuine logical reasoning, increasing hallucination risk. In contrast, GRPO's reward-driven optimization directly reinforces correct reasoning paths without constraining the model to artificial templates.

Another critical observation is the performance gap between difficulty-stratified and full-dataset GRPO training. On MathVista, GRPO-only(mid+hard) using PISM achieves 68.300, substantially outperforming GRPO-only(fullset) at 53.400 (Table 1). This discrepancy highlights the inefficiency of training on unfiltered data, where easy samples and unsolved cases dilute the signal from informative mid+hard samples. The similar trend observed across other benchmarks (MMMU: 0.547 vs. 0.440; MMStar: 0.639 vs. 0.607) confirms that strategic sample selection is more impactful than simply increasing data volume.

These results highlight two key insights: (1) Reinforcement learning, when guided by difficulty-aware sample selection, can effectively learn perception and reasoning without prior supervised fine-tuning. (2) Data quality—particularly the strategic inclusion of medium and hard samples—matters more than quantity in driving multimodal performance. This approach not only simplifies the training pipeline by removing the SFT stage, but also improves model robustness by focusing on samples that truly challenge multimodal reasoning.

Conclusion

In this work, we propose two difficulty-aware sample selection methods—Progressive Image Semantic Masking (PISM) and Cross-Modal Attention Balancing (CMAB)—to enable fine-grained difficulty classification for multimodal data. These metrics assess visual sensitivity and cross-modal alignment, guiding more strategic post-training. Experiments show that GRPO-only training consistently outperforms SFT+GRPO on both visual perception and reasoning tasks, especially on medium and hard samples. This indicates that, with well-chosen samples, reinforcement learning can effectively learn visual and logical reasoning without supervised fine-tuning. The success of the GRPO-only paradigm challenges the assumption that SFT is necessary for stable alignment, suggesting that direct policy optimization on informative samples better preserves reasoning capabilities and avoids overfitting. Our results emphasize intelligent data use over complex training pipelines, pointing to a simpler, more effective path toward multimodal alignment.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Buitrago, P. A.; and Nystrom, N. A. 2019. Open compass: accelerating the adoption of AI in open research. In *Practice and Experience in Advanced Research Computing 2019: Rise of the Machines (learning)*, 1–9.

Chen, L.; Li, J.; Dong, X.; et al. 2024a. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 27056–27087.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024b. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 19–35. Springer.

Chen, M.; Liu, H.; Liang, H.; Huang, H.; Zhang, W.; and He, R. 2025. Unlocking the Potential of Difficulty Prior in RL-based Multimodal Reasoning. *arXiv preprint arXiv:2505.13261*.

Gao, T.; Chen, P.; Zhang, M.; Fu, C.; Shen, Y.; Zhang, Y.; Zhang, S.; Zheng, X.; Sun, X.; Cao, L.; et al. 2024. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9096–9105.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, J.; Zheng, T.; Bai, Y.; Li, B.; Wang, Y.; Zhu, K.; Li, Y.; Neubig, G.; Chen, W.; and Yue, X. 2024. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*.

Hu, J.; Zhang, Y.; Han, Q.; Jiang, D.; Zhang, X.; and Shum, H.-Y. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kumar, K.; Ashraf, T.; Thawakar, O.; Anwer, R. M.; Cholakkal, H.; Shah, M.; Yang, M.-H.; Torr, P. H.; Khan, F. S.; and Khan, S. 2025. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*.

Li, M.; Zhang, Y.; He, S.; Li, Z.; Zhao, H.; Wang, J.; Cheng, N.; and Zhou, T. 2024. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *arXiv preprint arXiv:2402.00530*.

Liu, X.; Ni, J.; Wu, Z.; Du, C.; Dou, L.; Wang, H.; Pang, T.; and Shieh, M. Q. 2025a. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*.

Liu, Y.; Li, Z.; Huang, M.; Yang, B.; Yu, W.; Li, C.; Yin, X.-C.; Liu, C.-L.; Jin, L.; and Bai, X. 2024. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12): 220102.

Liu, Z.; Chen, C.; Li, W.; Qi, P.; Pang, T.; Du, C.; Lee, W. S.; and Lin, M. 2025b. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025c. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Lu, P.; Bansal, H.; Xia, T.; et al. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Luo, R.; Zheng, Z.; Wang, Y.; Ni, X.; Lin, Z.; Jiang, S.; Yu, Y.; Shi, C.; Chu, R.; Zeng, J.; et al. 2025. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*.

Ma, Y.; Du, L.; Shen, X.; Chen, S.; Li, P.; Ren, Q.; Ma, L.; Dai, Y.; Liu, P.; and Yan, J. 2025. One RL to See Them All: Visual Triple Unified Reinforcement Learning. *arXiv preprint arXiv:2505.18129*.

Meng, F.; Du, L.; Liu, Z.; Zhou, Z.; Lu, Q.; Fu, D.; Shi, B.; Wang, W.; He, J.; Zhang, K.; et al. 2025. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*.

Mitra, C.; Huang, B.; Darrell, T.; and Herzig, R. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14420–14431.

Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26374–26383.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; et al. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Su, A.; Wang, H.; Ren, W.; Lin, F.; and Chen, W. 2025. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*.

Thawakar, O.; Dissanayake, D.; More, K.; Thawkar, R.; Heakl, A.; Ahsan, N.; Li, Y.; Zumri, M.; Lahoud, J.; Anwer, R. M.; et al. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*.

Tong, P.; Brown, E.; Wu, P.; Woo, S.; IYER, A. J. V.; Akula, S. C.; Yang, S.; Yang, J.; Middepogu, M.; Wang, Z.; et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37: 87310–87356.

Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; and Wang, L. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Wang, S.; Yu, L.; Gao, C.; Zheng, C.; Liu, S.; Lu, R.; Dang, K.; Chen, X.; Yang, J.; Zhang, Z.; et al. 2025a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Wang, X.; Yang, Z.; Feng, C.; Lu, H.; Li, L.; Lin, C.-C.; Lin, K.; Huang, F.; and Wang, L. 2025b. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Wang, Z.; Cui, G.; Li, Y.-J.; Wan, K.; and Zhao, W. 2025c. Dump: Automated distribution-level curriculum learning for rl-based llm post-training. *arXiv preprint arXiv:2504.09710*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xie, T.; Gao, Z.; Ren, Q.; Luo, H.; Hong, Y.; Dai, B.; Zhou, J.; Qiu, K.; Wu, Z.; and Luo, C. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Xiong, W.; Yao, J.; Xu, Y.; Pang, B.; Wang, L.; Sahoo, D.; Li, J.; Jiang, N.; Zhang, T.; Xiong, C.; et al. 2025. A minimalist approach to llm reasoning: from rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*.

Xu, G.; Jin, P.; Wu, Z.; Li, H.; Song, Y.; Sun, L.; and Yuan, L. 2024. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Yao, H.; Huang, J.; Wu, W.; Zhang, J.; Wang, Y.; Liu, S.; Wang, Y.; Song, Y.; Feng, H.; Shen, L.; et al. 2024. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Yu, E.; Lin, K.; Zhao, L.; Yin, J.; Wei, Y.; Peng, Y.; Wei, H.; Sun, J.; Han, C.; Ge, Z.; et al. 2025a. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Dai, W.; Fan, T.; Liu, G.; Liu, L.; et al. 2025b. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Yu, W.; Yang, Z.; Li, L.; et al. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Yue, X.; Ni, Y.; Zhang, K.; 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.

Yue, Y.; Yuan, Y.; Yu, Q.; Zuo, X.; Zhu, R.; Xu, W.; Chen, J.; Wang, C.; Fan, T.; Du, Z.; et al. 2025. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*.

Zeng, W.; Huang, Y.; Liu, Q.; Liu, W.; He, K.; Ma, Z.; and He, J. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Zhang, J.; and Zuo, C. 2025. Grpo-lead: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. *arXiv preprint arXiv:2504.09696*.

Zhang, R.; Zhang, B.; Li, Y.; Zhang, H.; Sun, Z.; Gan, Z.; Yang, Y.; Pang, R.; and Yang, Y. 2024. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Zhang, Y.-F.; Lu, X.; Hu, X.; Fu, C.; Wen, B.; Zhang, T.; Liu, C.; Jiang, K.; Chen, K.; Tang, K.; et al. 2025. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; et al. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 29733–29735.

Zheng, G.; Yang, B.; Tang, J.; Zhou, H.-Y.; and Yang, S. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 5168–5191.

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; and Ma, Y. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Task Type	Metric	Overall Data	Difficulty Classification			
			Easy	Medium	Hard	Unsolved
Visual Perception	Mask Ratio	20,633	$(0.7, 1)$	$(0.4, 0.7]$	$(0, 0.4]$	0
	Difficulty Level		easy	mid	hard	unsolved
	Data Volume		7,827	4,872	1,454	6,480
Visual Reasoning	Mask Ratio	27,133	$(0.7, 1)$	$(0.4, 0.7]$	$(0, 0.4]$	0
	Difficulty Level		easy	mid	hard	unsolved
	Data Volume		5,048	1,061	1,618	19,406

Table A: Sample distribution and difficulty classification using PISM (Progressive Image Semantic Masking) method.

Task Type	Strategy	Data Subsets and Combinations	
		Combination A	Combination B
Visual Perception	GRPO-only	mid+hard (6.3k), random (6.3k), unsolved (6k), fullset (20.6k)	
	SFT+GRPO-1	mid-4.9k (SFT) + hard-1.4k (GRPO)	hard-1.4k (SFT) + mid-4.9k (GRPO)
	SFT+GRPO-2	random-4.9k (SFT) + hard-1.4k (GRPO)	random-1.4k (SFT) + mid-4.9k (GRPO)
	SFT+GRPO-3	mid-4.9k (SFT) + random-1.4k (GRPO)	hard-1.4k (SFT) + random-4.9k (GRPO)
	SFT+GRPO-4	random-1.4k (SFT) + random-4.9k (GRPO)	random-4.9k (SFT) + random-1.4k (GRPO)
Visual Reasoning	GRPO-only	mid+hard (2.6k), random (2.6k), unsolved (19k), fullset (27k)	
	SFT+GRPO-1	mid-1k (SFT) + hard-1.6k (GRPO)	hard-1.6k (SFT) + mid-1k (GRPO)
	SFT+GRPO-2	random-1k (SFT) + hard-1.6k (GRPO)	random-1.6k (SFT) + mid-1k (GRPO)
	SFT+GRPO-3	mid-1k (SFT) + random-1.6k (GRPO)	hard-1.6k (SFT) + random-1k (GRPO)
	SFT+GRPO-4	random-1k (SFT) + random-1.6k (GRPO)	random-1k (SFT) + random-1k (GRPO)

Table B: Training configurations for PISM-based experiments using GRPO-only and SFT+GRPO strategies.

Task Type	Metric	Overall Data	Difficulty Classification			
			Easy	Medium	Hard	Unsolved
Visual Perception	Range of $\bar{\rho}$	20633	$(0, 0.1) \cup (1.9, +\infty)$	$[0.1, 0.4) \cup (1.6, 1.9]$	$[0.4, 1.6]$	–
	Data Volume		6753	6029	1001	6850
Visual Reasoning	Range of $\bar{\rho}$	27133	$(0, 0.1) \cup (1.9, +\infty)$	$[0.1, 0.4) \cup (1.6, 1.9]$	$[0.4, 1.6]$	–
	Data Volume		2170	3604	2166	19193

Table C: Sample distribution and difficulty classification using the CMAB (Cross-Modality Attention Balance) method.

Task Type	Strategy	Data Subsets and Sample Sizes	
		Combination A	Combination B
Visual Perception	GRPO-only	mid+hard (7k), random (7k), unsolved (6.8k), fullset (20.6k)	
	SFT+GRPO-1	mid-6k (SFT) + hard-1k (GRPO)	hard-1k (SFT) + mid-6k (GRPO)
	SFT+GRPO-2	random-6k (SFT) + hard-1k (GRPO)	random-1k (SFT) + mid-6k (GRPO)
	SFT+GRPO-3	mid-6k (SFT) + random-1k (GRPO)	hard-1k (SFT) + random-6k (GRPO)
	SFT+GRPO-4	random-6k (SFT) + random-1k (GRPO)	random-1k (SFT) + random-6k (GRPO)
Visual Reasoning	GRPO-only	mid+hard (5.7k), random (5.7k), unsolved (19k), fullset (27k)	
	SFT+GRPO-1	mid-3.6k (SFT) + hard-2.1k (GRPO)	hard-2.1k (SFT) + mid-3.6k (GRPO)
	SFT+GRPO-2	random-3.6k (SFT) + hard-2.1k (GRPO)	random-2.1k (SFT) + mid-3.6k (GRPO)
	SFT+GRPO-3	mid-3.6k (SFT) + random-2.1k (GRPO)	hard-2.1k (SFT) + random-3.6k (GRPO)
	SFT+GRPO-4	random-3.6k (SFT) + random-2.1k (GRPO)	random-2.1k (SFT) + random-3.6k (GRPO)

Table D: Training configurations for CMAB-based experiments using GRPO-only and SFT+GRPO strategies.