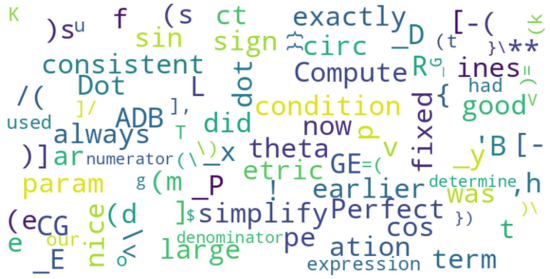
CaRe-R1: High-Entropy Tokens Drive Effective RLVR for Captioning and Reasoning in MRLMs

Anonymous EMNLP submission

Abstract

Multimodal reasoning large language models (MRLMs) have shown strong capability in graphical reasoning tasks, yet their reinforcement learning post-training remains limited by coarse supervision and suboptimal optimization. Our empirical analysis reveals that model failures mainly stem from two sources: visual perception errors and reasoning errors. However, final-answer rewards alone cannot provide sufficient and precise supervision for intermediate visual representations. Moreover, compared with general MLLMs, MRLMs typically generate longer reasoning sequences, in which most tokens are low-entropy continuations, while only a small subset of high-entropy tokens is critical to key reasoning transitions. Motivated by these observations, we propose CaRe-R1, a unified reinforcement learning post-training framework for multimodal reasoning. Specifically, CaRe-R1 improves sample efficiency through difficulty-aware data curation, concentrates optimization on decision-critical reasoning steps through high-entropy token optimization, and jointly rewards intermediate captions via caption faithfulness and caption sufficiency. Experimental results show that, on Qwen3-VL-4B, CaRe-R1 improves the average accuracy across all datasets by 8 percentage points while reducing token consumption by more than 50%.

(a) Frequent tokens with the highest average entropy



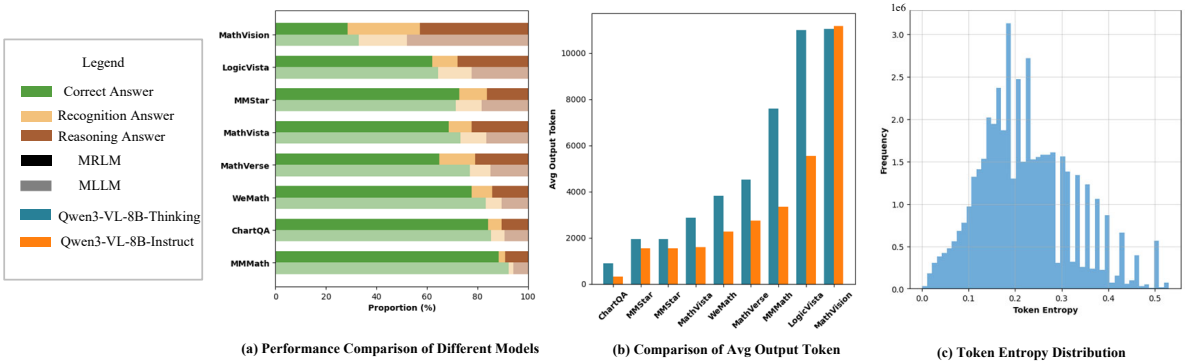
(b) Frequent tokens with the lowest average entropy

Figure 1: (a) and (b) present word clouds of the frequently occurring tokens with the highest and lowest average entropy. High-entropy tokens typically appear at key branching points in the reasoning process, whereas low-entropy tokens are more often used to continue and execute the established reasoning path.

1 Introduction

In recent years, multimodal large language models (MLLMs) (OpenAI et al., 2024) have shown rapid progress on tasks such as visual question answering, chart understanding(Masry et al., 2022), and mathematical visual reasoning(Lu et al., 2024). In particular, multimodal reasoning large language models (MRLMs) (Team, 2025b) equipped with explicit chain-of-thought generation have further advanced multimodal intelligence from simply understanding visual content to reasoning and decision-making based on visual evidence. However, in complex reasoning scenarios, these models still suffer from several major limitations, including visual perception errors, reasoning failures, and excessive reasoning cost, all of which jointly constrain their performance, efficiency, and generalization ability.

To better understand the bottlenecks of current MRLMs, we first conduct a set of motivating analyses, as shown in Figure 2. The results reveal three



(a) Performance Comparison of Different Models (b) Comparison of Avg Output Token (c) Token Entropy Distribution

Figure 2: (a) Analysis of error sources in multimodal large language models and multimodal reasoning large models. (b) Comparison of average token consumption across datasets. (c) Statistical distribution of token entropy in model responses over all datasets.

key observations. First, model failures mainly arise from two sources: visual perception errors and reasoning errors, suggesting that supervision based solely on the final answer is insufficient and that intermediate visual representations should be explicitly modeled. Second, MRLMs typically generate substantially more output tokens which motivates us to focus on MRLMs rather than genera MLLMs. Third, model outputs contain a large proportion of low-entropy tokens, while only a small subset of high-entropy tokens appears to play a critical role in reasoning decisions. Moreover, as illustrated in Figure 1, high-entropy tokens tend to appear at key reasoning transitions and branching points, whereas low-entropy tokens are more often associated with language continuation. This suggests that not all generated tokens have equal learning value, and uniformly optimizing all tokens may dilute training focus (Wang et al., 2025d).

Beyond token-level optimization, existing RL training methods also tend to treat all training samples as having similar learning value. In practice, however, samples differ substantially in difficulty and learnability: overly easy samples are quickly exhausted, while overly hard ones often fail to provide stable training signals. By contrast, medium-difficulty samples near the model's capability boundary are more likely to provide sustained learning value. Together with our earlier observations on visual perception errors, reasoning errors, and the unequal importance of generated tokens, this suggests that effective multimodal RL post-training should be reconsidered from three complementary aspects: which training samples should be selected, which tokens should be emphasized, and how intermediate visual representations should be supervised.

Motivated by these observations, we propose CaRe-R1, a unified RL post-training framework for MRLMs. CaRe-R1 improves multimodal reasoning training from three levels: sample-level data selection, token-level optimization, and intermediate-representation supervision. Specifically, we partition training samples based on repeated-sampling consistency and focus RL training on medium-difficulty samples with higher learning value; we introduce high-entropy token prioritization to concentrate optimization on decision-critical reasoning steps; and we further define two complementary reward signals, caption faithfulness and caption sufficiency, to explicitly supervise intermediate captions from the perspectives of visual fidelity and task sufficiency.

To summarize, our contributions are as follows:

- We present CaRe-R1, a unified RL post-training framework for MRLMs that integrates difficulty-aware data curation, high-entropy token optimization, and intermediate caption quality supervision into a coherent training pipeline.

- We develop two effective strategies for improving training efficiency and optimization quality, namely difficulty-aware data curation at the sample level and high-entropy token optimization at the token level. Exper-

imental results show that these designs provide more informative training signals and improve both reasoning performance and training efficiency.

- We establish a dual-perspective reward framework for intermediate captions by jointly modeling caption faithfulness and caption sufficiency. This design provides explicit supervision for intermediate visual representations and leads to stronger downstream reasoning performance.

2 Related work

Recent progress in multimodal reinforcement learning has mainly advanced along two directions: optimization foundations for stable post-training and visual faithfulness in reasoning.

2.1 MRLMs and RL with Verifiable Rewards

Recent work on multimodal reinforcement learning has largely focused on building stable optimization foundations through initialization, supervision, reward design, and training stabilization. Vision-R1 (Huang et al., 2025b) shows that directly applying reinforcement learning to multimodal reasoning rarely yields stable reasoning behaviors. Advancing Multimodal Reasoning via Reinforcement Learning with Cold Start (Wei et al., 2025) and ReVisual-R1 (Chen et al., 2025a) further emphasize that structured supervised initialization is often essential for effective RL convergence. URSA (Luo et al., 2025) complements this line by showing that large-scale automatically constructed multimodal mathematical chain-of-thought and process-supervision data provide a strong basis for subsequent reasoning training.

Another important direction studies reward design and optimization stability in RLVR-based multimodal training. R1-ShareVL (Yao et al., 2025) improves training signal density and stability through trajectory and advantage sharing. Mixed-R1 (Xu et al., 2025b) unifies diverse multimodal tasks under a single post-training reward framework. SophiaVL-R1 (Fan et al., 2025a) extends reward modeling from outcomes to the reasoning process itself, while using confidence weighting and annealing to mitigate reward hacking.

2.2 Visual Faithfulness

Another related line of work concerns visual faithfulness, whether multimodal reasoning genuinely relies on visual evidence rather than merely producing plausible textual justifications. GRIT (Fan et al., 2025b) improves reasoning auditability by allowing the model to alternately generate text and bounding-box coordinates, thereby explicitly grounding reasoning in image regions. At a finer granularity, v1 (Chung et al., 2026) introduces a point-and-copy mechanism that enables the model to revisit critical visual patches during reasoning. Qwen Look Again (Chu et al., 2025) further shows that attention to visual tokens is progressively diluted as reasoning chains become longer.

3 Main Framework

We propose CaRe-R1, a reinforcement learning framework for improving captioning and reasoning in MRLMs, consisting of three modules: (i) the training pipeline, (ii) the caption reward design, and (iii) the forking token optimization. The details of these modules are presented in this section, as illustrated in Figure 3.

3.1 Preliminary

3.1.1 Policy Optimization

Group Relative Policy Optimization (GRPO)(Shao et al., 2024) is a variant of Proximal Policy Optimization (PPO)(Schulman et al., 2017), designed to improve the reasoning ability of large language models on complex tasks such as mathematical and scientific reasoning. Starting from a pretrained MLLM, GRPO initializes a trainable policy model π_θ and an old policy $\pi_{\theta_{\text{old}}}$. Given an image-text input (I, T), the old policy samples a group of responses $\{O_1, O_2, \ldots, O_G\}$ and a group-based reward function assigns the corresponding rewards $\{R_1, R_2, \ldots, R_G\}$. The advantage of each response is then estimated relative to the group as

$$\hat{A}_i = \frac{R_i - \text{mean}\left(\{R_i\}_{i=1}^G\right)}{\text{std}\left(\{R_i\}_{i=1}^G\right)}. \quad (1)$$

In this work, we adopt Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO) (Yu et al., 2025) as the optimization algorithm. Unlike PPO, which requires training a separate

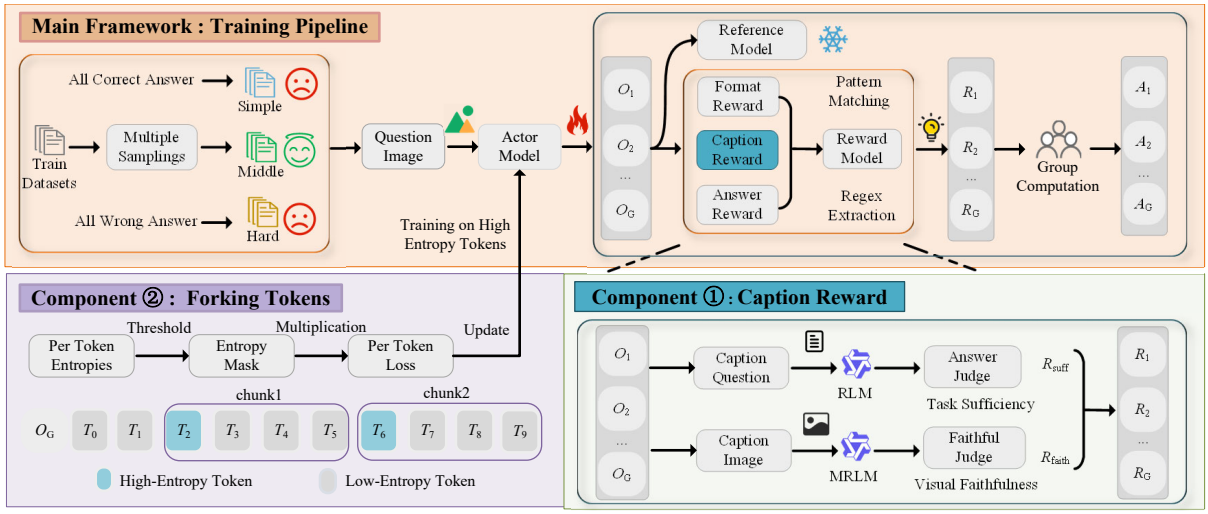


Figure 3: Overall framework of the proposed training pipeline, with illustrations of the caption reward and forking token optimization component.

critic model, DAPO estimates the baseline from a group of sampled rollouts, making it more suitable for critic-free policy optimization. Building on GRPO, DAPO removes the KL penalty, introduces a clip-higher mechanism, incorporates dynamic sampling, applies a token-level policy gradient loss, and adopts overlong reward shaping, applies a token-level policy gradient loss, and adopts overlong reward shaping, leading to the following maximization objective, where $\mathcal{J}_{\text{DAPO}}(\theta)$

$$E_{(I,T)\sim p_{\mathcal{D}},\, o\sim\pi_{\theta_{\text{old}}}(\cdot|I,T)}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\frac{\pi_\theta(o_i \mid I,T)}{\pi_{\theta_{\text{old}}}(o_i \mid I,T)}\hat{A}_i,\right.\right.$$
$$\left.\left.\text{clip}\left(\frac{\pi_\theta(o_i \mid I,T)}{\pi_{\theta_{\text{old}}}(o_i \mid I,T)},\, 1-\epsilon,\, 1+\epsilon\right)\hat{A}_i\right)\right]. \tag{2}$$

3.1.2 Train Dataset Process

In the appendix A.4, we provide a detailed description and analysis of the issues of sparse rewards and advantage vanishing.

Before training, we first preprocess the original training set through a consistency-based procedure using a strong reference model with repeated sampling. Specifically, for each training sample, the reference model is used to generate multiple responses, which are then matched against the ground-truth answer to assess the sample's overall behavior across repeated trials. If a sample is answered correctly in all repeated samplings, it is categorized as simple; if it is answered incorrectly in all repeated samplings, it is categorized as hard; and if both correct and incorrect responses appear across repeated samplings, it is categorized as medium.This partition does not represent the sample's absolute intrinsic difficulty, but rather characterizes its stable solvability under the repeated-sampling condition of the reference model. Based on this procedure, the original training set is further divided into three subsets, namely simple, medium, and hard, while the full dataset is retained for subsequent comparative experiments to examine how different difficulty distributions affect training dynamics and final performance.

3.2 Reward Model

In the reinforcement learning process, the reward function serves as a supervisory signal to adjust and optimize the behavior of the model. Therefore, we design three types of rewards: **format reward**, **answer reward**, and **caption reward**. The following section introduces the details of the reward function.

3.2.1 Format Reward

The format reward is designed to enforce the generation order required by the MRLM, rather than merely checking tag completeness. The target format is defined as *<think><caption><answer>*. Accordingly, the format reward is implemented via pattern matching to encourage outputs that strictly

conform to this predefined order.

$$R_{\text{format}} = \begin{cases} 1.0, & \text{it matches the structure,} \\ 0.0, & \text{otherwise.} \end{cases} \quad (3)$$

3.2.2 Answer Reward

The answer reward is designed to assess the correctness of the final output answer. Specifically, we first use regex extraction to extract the answer content enclosed within the <*answer*> tag. Accordingly, the answer reward is assigned based on the RLM judgment result.

$$R_{\text{answer}} = \begin{cases} 1.0, & \text{answer is correct,} \\ 0.0, & \text{otherwise.} \end{cases} \quad (4)$$

3.2.3 Caption Reward

The caption reward is designed to evaluate the quality of intermediate captions from two complementary perspectives, called *Visual Faithfulness* and *Task Sufficiency*.

Visual Faithfulness measures whether the generated caption is consistent with the input image, focusing on the correctness of object recognition, spatial relations, quantities, and structural details. Task Sufficiency measures whether the caption contains sufficient task-relevant information to support solving the problem. Specifically, Visual Faithfulness is assessed by an MRLM based on the image-caption pair, while Task Sufficiency is evaluated by a stronger external RLM based on whether the problem can be correctly solved using only the problem and the caption.

These two reward signals are introduced because final answer correctness alone cannot fully reflect caption quality: a caption may be visually inaccurate yet still lead to a correct final answer, or it may be visually faithful but still omit critical information required for reasoning. Accordingly, we define two caption-level rewards as follows:

$$R_{\text{faith}} = \begin{cases} 1.0, & \text{the caption is faithful,} \\ 0.0, & \text{otherwise,} \end{cases} \quad (5)$$

$$R_{\text{suff}} = \begin{cases} 1.0, & \text{the caption is task-sufficient,} \\ 0.0, & \text{otherwise.} \end{cases} \quad (6)$$

$$R_{\text{caption}} = \lambda R_{\text{faith}} + (1-\lambda) R_{\text{suff}}, \quad \lambda \in [0,1] \quad (7)$$

Under this unified formulation, different caption reward designs can be instantiated by different choices of λ. The final reward consists of the following components:

$$R_{\text{G}} = R_{\text{format}} + R_{\text{answer}} + R_{\text{caption}} \quad (8)$$

3.3 Forking Tokens Train

3.3.1 Token Entropy Calculation

The token-level generation entropy for token t is defined as H_t.

$$H_t := -\sum_{j=1}^{V} p_{t,j} \log p_{t,j},$$
$$\text{where } (p_{t,1}, \cdots, p_{t,V}) = \boldsymbol{p}_t = \pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{q}, \boldsymbol{o}_{<t}) = \text{Softmax}\left(\frac{\boldsymbol{z}_t}{T}\right). \quad (9)$$

Here, $\pi_{\boldsymbol{\theta}}$ denotes an LLM parameterized by $\boldsymbol{\theta}$, $\boldsymbol{q}$ is the input query, and $\boldsymbol{o}_{<t} = (o_1, o_2, \ldots, o_{t-1})$ denotes the previously generated tokens. V is the vocabulary size, $\boldsymbol{z}_t \in R^V$ represents the pre-softmax logits at step t, $\boldsymbol{p}_t \in R^V$ is the corresponding token distribution over the vocabulary, and $T \in R$ is the decoding temperature.

3.3.2 Threshold Setting

The threshold is used to identify tokens that deserve greater optimization emphasis. Since uncertainty is not uniform across the reasoning process, high-entropy tokens are more likely to correspond to critical decision points, while low-entropy tokens usually represent more deterministic continuations. Based on per token entropies, we construct an entropy mask and apply it to the per token loss, so that training is driven primarily by high-entropy tokens. This allows parameter updates to focus on key reasoning positions instead of treating all generated tokens equally.

4 Experiments

Our experiments investigate the following key questions:

Q1: How does Care-R1 perform on comprehensive multimodal reasoning benchmarks compared with existing MRLMs?

Models	Size	MMStar		ChartQA		WeMath		MMMath		LogicVista		MathVerse		MathVision		MathVista	
		ACC	Token	ACC	Token	ACC	Token	ACC	Token	ACC	Token	ACC	Token	ACC	Token	ACC	Token
Multimodal Large Language Models based on Qwen2.5-VL series																	
Visionary-R1	3B	35.4	174	68.1	120	10.6	456	4.9	728	19.8	215	10.0	494	24.6	419	52.4	202
Qwen2.5-VL	7B	60.3	138	85.6	90	31.7	341	47.8	573	40.4	278	45.7	363	15.4	390	75.6	192
AdaTooler-V	7B	62.3	181	86.8	134	64.1	377	49.2	588	40.8	313	58.5	386	16.1	422	73.0	226
SophiaVL-R1	7B	59.0	166	85.6	83	33.5	365	46.6	592	36.6	292	46.2	348	17.1	422	73.5	188
VL-Rethinker	7B	65.8	164	88.8	114	68.9	334	60.8	642	44.6	267	67.1	351	22.3	425	80.5	195
VL-RLCS	7B	36.1	369	87.9	292	8.4	587	57.8	1006	25.4	508	28.8	647	14.1	664	80.9	417
Revisual-R1	7B	63.6	1588	87.0	707	65.9	2788	59.5	564	51.5	6693	68.2	3387	36.8	7121	80.5	2020
Vision-R1	7B	64.1	1952	82.2	604	77.1	5755	52.5	7191	46.2	5959	56.8	3474	22.6	4709	73.7	1642
+CaRe-R1	7B	64.3	1799	85.6	463	81.7	5456	55.2	6807	50.8	5510	58.5	3002	25.3	4073	74.4	1101
Multimodal Reasoning Language Models																	
Qwen3-VL-T	2B	63.4	3251	72.2	1552	70.1	6971	79.4	1288	52.2	18093	62.1	7510	24.3	15315	70.5	4333
+CaRe-R1	2B	61.5	869	86.1	900	77.2	4662	80.0	8618	53.0	12911	70.0	5845	27.9	12101	82.2	2819
Qwen3-VL-T	4B	71.0	1499	85.1	1491	82.9	4210	84.0	4196	63.8	13625	74.5	5635	31.9	12664	80.6	3751
+CaRe-R1	4B	75.6	1005	88.0	561	85.9	1806	85.0	3529	69.1	4531	80.6	2079	39.8	3630	81.1	1328
Qwen3-VL-T	8B	71.4	1950	85.4	911	83.2	3822	85.7	7609	64.2	11002	76.8	4533	32.8	11052	84.3	2884
+CaRe-R1	8B	71.9	1015	88.2	926	83.3	2859	87.2	5068	67.2	8165	77.9	4074	34.1	8673	87.8	1997
InternVL3_5	8B	70.7	2148	90.0	739	78.4	3897	77.5	7287	57.3	8231	71.2	4725	29.6	12004	83.6	2554
+CaRe-R1	8B	80.6	1467	92.2	735	88.5	3592	87.8	5598	80.4	4355	86.4	4595	49.0	6749	84.9	2444
Step3	10B	73.9	1463	90.4	463	64.4	3678	90.7	7076	64.5	4579	71.3	4856	38.4	10892	96.7	2592

Table 1: Performance of the Qwen2.5 series and reasoning models across different test datasets. The table reports both accuracy (ACC) and the average token consumption (Token). The results highlighted in blue denote models with the same parameter scale further trained using our method CaRe-R1.

Q2: How do the proposed difficulty-aware data curation, high-entropy token optimization, and caption reward design contribute to the overall performance of Care-R1?

Q3: Do the proposed designs effectively improve training signal quality and intermediate caption quality, thereby leading to better multimodal reasoning performance?

4.1 Datasets

We conduct experiments on a diverse set of datasets, including MMStar(Chen et al., 2024), MMMath(Sun et al., 2024), WeMath(Qiao et al., 2025), ChartQA(Masry et al., 2022), LogicVista(Xiao et al., 2024), MathVerse(Zhang et al., 2024) , MathVision (Wang et al., 2024), and MathVista(Lu et al., 2024). We use Vision-R1-RL as the training dataset and We report the Pass@1 accuracy using greedy decoding. We provide comprehensive experimental configurations and additional implementation details in the appendix B.

4.2 Main Results (RQ1)

As shown in table 1, it provides a clear positive answer to Q1. MRLMs trained with Care-R1 achieve consistently stronger performance across comprehensive multimodal reasoning benchmarks. It obtain higher accuracy on most test sets while using fewer reasoning tokens, indicating that the gains come not only from better final performance but also from more efficient inference. It improves both effectiveness and efficiency, rather than simply trading longer reasoning for higher accuracy.

These improvements are not limited to a single model family or a specific parameter scale. Care-R1 shows stable benefits across models of different sizes and across different types of multimodal reasoning models, demonstrating strong cross-model generalization. This suggests that Care-R1 is not specialized for only one narrow task type; It consistently enhances visual understanding, task-relevant information extraction, and reasoning capability.

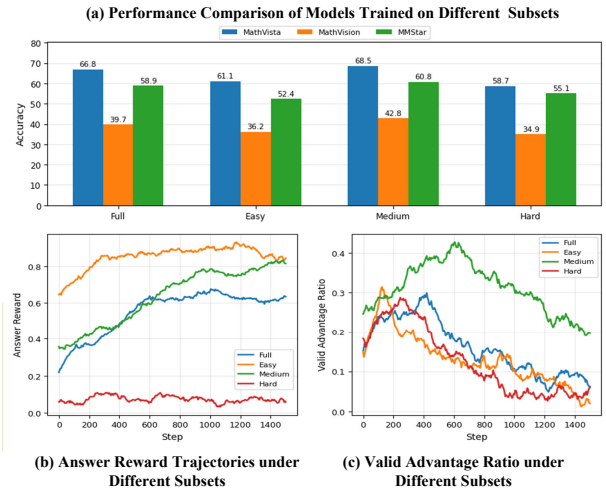


Figure 4: Training dynamics and downstream performance of Qwen3-VL-4B-Thinking.

Dataset	Metric	λ = 1	λ = 0	λ = 1/2
		R_{faith}	R_{suff}	$1/2R_{\text{faith}} + 1/2R_{\text{suff}}$
MMStar	FR (%)	84.7	80.8	83.5
	SR (%)	76.9	83.6	81.9
	ACC (%)	67.8	68.5	**78.8**
	Tokens	1423	1626	1523
MathVision	FR (%)	82.9	79.6	81.8
	SR (%)	74.8	81.7	80.4
	ACC (%)	31.2	32.0	**35.1**
	Tokens	8151	9469	8460
MathVista	FR (%)	83.8	79.9	82.6
	SR (%)	75.5	82.4	80.8
	ACC (%)	79.6	80.4	**81.5**
	Tokens	1417	2434	1426

Table 2: Illustrative comparison of different caption reward designs on representative datasets using Qwen3-4B-VL-Thinking with training. FR, SR, and ACC denote Faithfulness Rate, Sufficiency Rate, and Accuracy, respectively.

4.3 Ablation Study (RQ2)

4.3.1 Data Curation

Figure 4 illustrates the role of difficulty-aware data curation from three perspectives: final performance, training dynamics, and mechanism analysis. (a) compares downstream performance when training on the Full, Easy, Medium, and Hard subsets, and shows that the Medium subset consistently achieves the best overall results across multiple benchmarks, indicating that medium-difficulty samples are more beneficial for generalization. Figure (b) presents the answer reward trajectories under different subsets. Although the Easy subset attains higher rewards more quickly, its final performance is still inferior to Medium, whereas the Medium subset exhibits more sustained reward improvement, suggesting more stable training gains. (c) further provides a mechanism-level explanation: the Medium subset maintains a higher Valid Advantage Ratio throughout training, indicating that it can preserve effective advantage signals for a longer period, while Easy and Hard more easily degenerate into reward-homogeneous groups, leading to rapid signal decay.

4.3.2 Analysis of Caption Reward

Table 2 shows that different caption rewards affect different aspects of intermediate caption quality. When λ=1, the Faithfulness Rate is usually the highest, indicating that the Faithfulness reward is more effective at improving the visual consistency of captions. When λ=0, the Sufficiency Rate tends to be higher, suggesting that the Sufficiency reward is more effective at enhancing the task-supporting information contained in the captions. However, optimizing either dimension alone does not lead to the best final performance. These results indicate that Visual Faithfulness and Task Sufficiency capture two complementary aspects of caption quality, namely fidelity and effectiveness, and that their joint modeling is more beneficial for improving downstream reasoning performance.

4.3.3 High Entropy Token

As shown in Figure 5, different threshold settings markedly affect training dynamics: large thresholds make optimization overly diffuse, very small thresholds over-concentrate the signal, and a moderate threshold such as threshold=0.2 best maintains focus on critical reasoning positions while preserving sufficient training information.

Table 3 shows the same trend: threshold=0.2 achieves the best results on all datasets, with an average accuracy of 75.61, outperforming q=1.0 by 6.86 points. Although threshold=0.5 and threshold=0.1 both improve over the baseline, neither matches threshold=0.2. This suggests that the value of high-entropy thresholding lies not in selecting fewer tokens, but in balancing critical-token

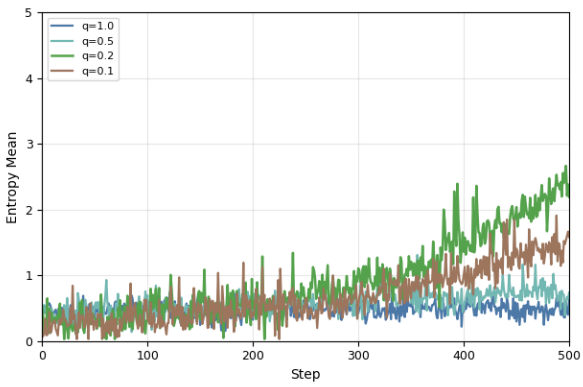


Figure 5: Entropy mean over training steps for different top-entropy threshold settings on Qwen3-4B-VL-Thinking with training.

Dataset	Accuracy (%) under Threshold				Gain
	1.0	0.5	0.2	0.1	(0.2-1.0)
Qwen3-4B-VL-Thinking					
MMStar	71.2	73.8	**75.6**	74.1	+4.4
WeMath	81.7	80.5	**85.9**	81.0	+4.2
MMMath	79.2	80.5	**85.0**	81.0	+5.8
ChartQA	71.4	77.3	**88.0**	78.8	+16.6
LogicVista	64.7	65.4	**69.1**	65.3	+4.4
MathVerse	76.2	75.7	**80.6**	79.9	+4.4
MathVision	33.4	35.5	**39.8**	31.4	+6.4
MathVista	72.2	72.1	**81.1**	75.2	+8.9
Average	68.75	70.10	**75.61**	70.84	+6.86

Table 3: Ablation results (%) under different top-entropy thresholds on the Qwen3-4B-VL-Thinking model. The best result is highlighted in bold.

emphasis with sufficient training signal.

4.4 Comparative analysis (RQ3)

The results in Tables 4 and 6 show that the proposed designs effectively improve both training signal quality and intermediate caption guidance, and that these improvements translate into stronger multimodal reasoning performance. As shown in Table 6, vanilla DAPO with all-token optimization does not consistently improve the base models and even causes average performance degradation, suggesting that unrestricted policy-gradient updates introduce noisy and diffuse training signals. In contrast, restricting optimization to the top 20% high-entropy tokens yields consistent accuracy gains across all benchmarks for both Qwen3-4B-VL-Thinking and InternVL3.5-8B, while also reducing token consumption.

Table 4 further shows that the full CaRe-R1 framework achieves the best accuracy, demonstrating that the proposed components are jointly effective. Among them, removing R_{answer} causes the largest drop in performance, confirming that answer supervision remains the primary optimization signal, while removing R_{caption} also consistently hurts performance, showing that intermediate caption supervision provides useful visual-semantic guidance for downstream reasoning. In addition, removing the threshold leads to clear accuracy degradation and substantially higher token consumption.

Method	MMStar		MathVision		MathVista	
	ACC	Tokens	ACC	Tokens	ACC	Tokens
Qwen3-4B-VL-Thinking						
w/o R_{caption}	70.8	1698	36.8	8150	78.2	2287
w/o R_{format}	72.9	2113	35.7	9231	76.2	3479
w/o R_{answer}	62.1	3471	31.8	5738	72.9	3215
w/o threshold	71.2	1409	33.4	10836	72.2	2541
CaRe-R1	75.6	1005	39.8	3630	81.1	1328

Table 4: Ablation study of the rewards and top-entropy threshold on datasets for Qwen3-4B-VL-Thinking.

5 Conclusion

This paper studies RL post-training for multimodal reasoning models from three complementary perspectives: difficulty-aware data curation, high-entropy token optimization, and intermediate caption quality modeling. Results show that each design improves multimodal reasoning effectiveness and efficiency, while their combination further enhances the overall performance of CaRe-R1.

Overall, our findings highlight that effective RL post-training for multimodal reasoning depends not only on the optimization algorithm, but also on data selection, training-signal allocation, and intermediate-representation supervision. By jointly improving these three aspects, CaRe-R1 achieves stronger performance, higher efficiency, and better generalization.

Limitations

We acknowledge several limitations in our study.

First, the multimodal datasets used in our experiments are primarily in English and Chinese, and the model's performance in other languages has not been investigated.

Second, our discussion of the image captioning task focuses mainly on chart-based images, without considering other image types.

In addition, we do not explore the use of external tools for image processing or interpretation.

Our study takes DAPO as the primary optimization framework. Although many recently popular PO-based algorithms, such as CISPO(Khatri et al., 2025), GSPO(Zheng et al., 2025), and SAPO(Gao et al., 2025), may lead to different training dynamics and downstream performance, we restrict our experiments to DAPO in order to better control experimental variables and ensure a clearer analysis of our proposed method. Exploring whether other PO algorithms can bring additional gains will be an important direction for future work.

Potential Risks

The experiments are mainly based on English and Chinese multimodal datasets, the generalizability of the proposed method to other languages and cultural contexts remains unclear.

In addition, our analysis of the image captioning task focuses primarily on chart-based images, so the applicability of the method to natural images, medical images, or complex real-world scenes still needs further validation.

The model does not rely on external tools for image processing or result verification during reasoning and captioning, which may lead to insufficient visual understanding or inaccurate interpretations in challenging cases.

These factors may further affect the robustness, fairness, and practical applicability of the method in broader multimodal scenarios.

References

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. 2025a. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning.

Yang Chen, Yufan Shen, Wenxuan Huang, Sheng Zhou, Qunshu Lin, Xinyu Cai, Zhi Yu, Jiajun Bu, Botian Shi, and Yu Qiao. 2025b. Learning only with images: Visual reinforcement learning with reasoning, rendering, and visual feedback.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. 2025. Qwen look again: Guiding vision-language reasoning models to re-attention visual information.

Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. 2026. v1: Learning to point visual tokens for multimodal grounded reasoning.

DeepSeek-AI, Daya Guo, and Yang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv e-prints*, page arXiv:2501.12948.

Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. 2025a. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. 2025b. Grit: Teaching mllms to think with images.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. 2025. Soft adaptive policy optimization.

Haowen Gao, Zhenyu Zhang, Liang Pang, Fangda Guo, Hongjian Dou, Guannan Lv, Shaoguo Liu, Tingting Gao, Huawei Shen, and Xueqi Cheng. 2026. Diva-grpo: Enhancing multimodal reasoning through difficulty-adaptive variant advantage.

Qihan Huang, Weilong Dai, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jingyuan Chen, Chang Yao, and Jie Song. 2025a. Boosting mllm reasoning with text-debiased hint-grpo.

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

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S.

Dhillon, David Brandfonbrener, and Rishabh Agarwal. 2025. The art of scaling reinforcement learning compute for llms.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. 2025. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter

Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Miao Pan, Wangjie Gan, Jintao Chen, Wenqi Zhang, Bing Sun, Jianwei Yin, and Xuhong Zhang. 2026. Ground what you see: Hallucination-resistant mllms via caption feedback, diversity-aware sampling, and conflict regularization.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Omar Sharif, Eftekhar Hossain, and Patrick Ng. 2026. From sight to insight: Improving visual reasoning capabilities of multimodal models via reinforcement learning.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024. MM-MATH: Advancing multimodal math evaluation with process evaluation and fine-grained classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1358–1375.

Qwen Team. 2025a. Qwen2.5-vl.

Qwen Team. 2025b. Qwen3 technical report.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, Chaofan Tao, Yangfan He, Mi Zhang, and Shen Yan. 2025. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning.

Chaoyang Wang, Kaituo Feng, Dongyang Chen, Zhongyu Wang, Zhixun Li, Sicheng Gao, Meng Meng, Xu Zhou, Manyuan Zhang, Yuzhang Shang, and Xiangyu Yue. 2025a. Adatooler-v: Adaptive tool-use for images and videos.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025b. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning.

Jiahao Wang, Weiye Xu, Aijun Yang, Wengang Zhou, Lewei Lu, Houqiang Li, Xiaohua Wang, and Jinguo Zhu. 2025c. Enhancing the outcome reward-based rl training of mllms with self-consistency sampling.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, volume 37.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025d. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025e. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. 2025. Advancing multimodal reasoning via reinforcement learning with cold start.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. 2025. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

Donglai Xu, Hongzheng Yang, Yuzhi Zhao, Pingping Zhang, Jinpeng Chen, Wenao Ma, Zhijian Hou, Mengyang Wu, Xiaolei Li, Senkang Hu, Ziyi Guan, Jason Chun Lok Li, and Lai Man Po. 2025a. From exploration to exploitation: A two-stage entropy rlvr approach for noise-tolerant mllm training.

Shilin Xu, Yanwei Li, Rui Yang, Tao Zhang, Yueyi Sun, Wei Chow, Linfeng Li, Hang Song, Qi Xu, Yunhai Tong, Xiangtai Li, and Hao Fei. 2025b. Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models.

Qi Yang, Bolin Ni, Shiming Xiang, Han Hu, Houwen Peng, and Jie Jiang. 2025. R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, et al. 2025. R1-sharevl: Incentivizing reasoning capabilities of multimodal large language models via share-grpo. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. Dapo: An open-source llm reinforcement learning system at scale.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision (ECCV)*.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. 2025. Thyme: Think beyond images.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. Swift:a scalable lightweight infrastructure for fine-tuning.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2026. Deepeyes: Incentivizing "thinking with images" via reinforcement learning.

A Related Work

In recent years, research on multimodal large language model has gradually shifted from early-stage visual question answering and general cross-modal understanding toward the development of multi-step reasoning capabilities for complex tasks. Inspired by pure-text reasoning paradigms such as DeepSeek-R1 (DeepSeek-AI et al., 2025), an increasing number of studies have introduced reinforcement learning into multimodal settings to elicit slow-thinking capabilities in tasks such as mathematical reasoning, chart understanding, spatial relation inference, and visual question answering. Existing work in this area can be broadly organized into four directions: reinforcement learning foundations for multimodal reasoning, visual faithfulness, adaptive and agentic multimodal reasoning, Reward and Advantage Issues.

A.1 Reinforcement Learning Foundations

One major line of work focuses on establishing a stable optimization foundation for multimodal reinforcement learning, especially through cold-start initialization, curriculum-style supervision, reward design, and stabilization strategies. Vision-R1 (Huang et al., 2025b) shows that when reinforcement learning is directly applied to multimodal reasoning, stable and sophisticated reasoning patterns rarely emerge naturally. Advancing Multimodal Reasoning via Reinforcement Learning with Cold Start (Wei et al., 2025) and ReVisual-R1 (Chen et al., 2025a) similarly argue that structured supervised initialization is not merely helpful but often essential for subsequent reinforcement learning to converge effectively. URSA (Luo et al., 2025) further demonstrates that automatically constructed large-scale multimodal mathematical chain-of-thought data and process-supervision data provide a strong basis for unlocking multimodal reasoning capabilities.

Beyond initialization, another key issue concerns reward design and optimization stability in RLVR-based multimodal training. R1-ShareVL (Yao et al., 2025) addresses reward sparsity and advantage vanishing by sharing trajectories and advantage estimation across problem variants, thereby improving both the density and stability of training signals. Mixed-R1 (Xu et al., 2025b) proposes a unified reward perspective that integrates multiple-choice questions, chart understanding, grounding tasks, and open-ended generation into a single post-training framework. SophiaVL-R1 (Fan et al., 2025a) extends reward modeling to the reasoning process itself by combining thinking rewards with outcome rewards, while introducing confidence weighting and annealing mechanisms to reduce reward hacking. Self-Consistency Sampling (Wang et al., 2025c) and two-stage entropy optimization methods (Xu et al., 2025a) further improve multimodal reinforcement learning from the perspectives of trajectory consistency, noise robustness, and more stable single-rollout optimization.

A.2 Visual Faithfulness

A second line of work focuses on visual faithfulness, namely whether multimodal reasoning chains genuinely rely on visual evidence rather than merely producing plausible textual reasoning. Several studies attempt to make visual evidence usage more explicit during reasoning. GRIT (Fan et al., 2025b) improves the auditability of reasoning by allowing the model to alternately generate natural language and bounding-box coordinates, thereby explicitly specifying the image regions on which its reasoning is based. At a finer granularity, v1 (Chung et al., 2026) introduces a point-and-copy mechanism that enables the model to revisit critical visual patches during reasoning instead of relying solely on language-side continuation once image encoding has been completed. From the perspective of attention degradation, Qwen Look Again (Chu et al., 2025) shows that as reasoning chains become longer, attention to visual tokens is progressively diluted.

Closely related to visual faithfulness is the problem of shortcut learning and hallucination. Visionary-R1 (Xia et al., 2025) finds that when training depends only on question-answer pairs and outcome rewards, models may exploit textual shortcuts from easier examples instead of acquiring genuine visual understanding. Ground What You See (Pan et al., 2026) further analyzes hallucination from the perspective of training dynamics, arguing that its causes lie not only in insufficient visual anchoring during reasoning but also in inadequate exploration diversity and inter-sample update conflicts. From Sight to Insight (Sharif et al., 2026) identifies insufficient visual perception, rather than

purely algorithmic weakness, as a core source of failure in visual reasoning, and promotes actual use of visual information during reinforcement learning through multiple reward functions.

A.3 Adaptive and Agentic Reasoning

A third line of work extends multimodal reasoning beyond static chain-of-thought generation toward more adaptive and agentic forms of reasoning. One subdirection emphasizes reflection and controllable reasoning behavior. VL-Rethinker (Wang et al., 2025b) transforms reflection from a naturally emerging phenomenon into an explicitly trainable behavior. SRPO (Wan et al., 2025) further proposes reflection-aware reinforcement learning, where reflection-oriented data construction and reflection reward design encourage reflections that are not only present but also concise and targeted. R-4B (Yang et al., 2025) goes one step further by treating the decision of whether to enter a thinking mode itself as a learnable policy, thereby seeking a dynamic balance between reasoning cost and performance.

Another emerging subdirection investigates tool-augmented and executable multimodal reasoning. DeepEyes (Zheng et al., 2026) attempts to elicit interleaved multimodal reasoning capabilities through end-to-end reinforcement learning without relying on cold-start supervised fine-tuning, while treating grounding ability itself as an internal tool. Thyme (Zhang et al., 2025) moves beyond simply re-inspecting the original image by allowing the model to autonomously generate and execute code for operations such as cropping, rotation, and enhancement, and then use the processed results to support subsequent reasoning. The RRVF framework in Learning Only with Images (Chen et al., 2025b) establishes a closed loop of reasoning, rendering, and visual feedback, in which intermediate outputs are generated directly from raw images and visual consistency, rather than language annotation, serves as the training signal. AdaTooler-V (Wang et al., 2025a) further elevates the decision of whether tools are needed into a learnable policy, aiming to avoid ineffective or excessive tool usage. These studies suggest that multimodal reasoning is gradually evolving from chain-like reasoning over static inputs toward an executable, verifiable, and interactive agentic paradigm.

A.4 Reward and Advantage Issues

In multimodal GRPO-style reinforcement learning, training degeneration typically does not manifest as the conventional layer-wise vanishing gradient problem found in deep neural networks. Instead, it more often appears as sparse rewards, advantage vanishing, and gradient stagnation. VL-Rethinker(Wang et al., 2025b) points out that when all rollouts within the same problem group receive identical rewards, the relative advantage collapses to zero, resulting in ineffective gradient updates. R1-ShareVL(Yao et al., 2025) further observes that this issue in multimodal settings often co-occurs with sparse rewards, and addresses it by enlarging the problem space, sharing reasoning trajectories, and employing hierarchical advantage estimation to increase reward diversity. ReVisual-R1(Chen et al., 2025a) characterizes this phenomenon as gradient stagnation, and alleviates training stagnation through its PAD mechanism, which filters out zero-advantage samples and reweights informative trajectories. From the perspective of low data utilization, Hint-GRPO(Huang et al., 2025a) analyzes the case where all responses within the same problem group are incorrect, leading to zero advantage and thus invalid samples, and improves the proportion of effective samples through adaptive hinting. DIVA-GRPO(Gao et al., 2026) further argues that both overly easy and overly difficult problems can lead to high reward homogeneity within groups, thereby causing sparse rewards and advantage vanishing, and proposes difficulty-adaptive variant augmentation to preserve group-level reward variance. In addition, works such as SophiaVL-R1 improve training stability from the perspectives of advantage shaping, process-supervised rewards, and trust-weighted thinking rewards, respectively. Overall, a shared understanding emerging from existing studies is that the central challenge in multimodal RL is not simply to increase the number of training steps, but to continuously preserve effective relative learning signals throughout training.

Although GRPO has shown strong effectiveness on reasoning-oriented tasks, its application to multimodal large language models still faces two major challenges, namely the sparse reward issue and the advantage vanishing issue.

The sparse reward issue mainly arises from the still limited reasoning capability of current

MLLMs. For a given multimodal problem, only a small number of reasoning trajectories can finally obtain positive rewards, while most sampled responses receive uniformly low or zero feedback. Such sparse supervision makes exploration inefficient and often leads to unstable optimization.

Another closely related problem is advantage vanishing in group-relative policy optimization. In GRPO-style training, the learning signal depends not only on the absolute reward itself, but also on whether different rollouts within the same group exhibit reward differences. When an MLLM generates highly homogeneous responses for the same question and these responses obtain identical rewards, the group-relative advantages collapse to zero, making policy updates ineffective.

B Implementation Details

In this appendix, we provide supplementary experimental details and supporting information omitted from the main text. Specifically, we cover dataset statistics and partitioning, prompt templates and parameter configurations, and the experimental environment along with hyperparameter settings. These materials are intended to enhance reproducibility and give readers deeper insight into the key steps of our model training and evaluation.

B.1 Dataset Statistics

MMStar (Chen et al., 2024) is a high-quality benchmark designed to evaluate the general capabilities of large vision-language models. The benchmark contains 1,500 carefully curated samples, covering 6 core capability domains and 18 fine-grained evaluation dimensions, with relatively balanced distributions across categories.

MMMath (Sun et al., 2024) is a benchmark specifically constructed for evaluating multimodal mathematical reasoning. It contains 5,929 open-ended middle-school-level math problems with visual context. MMMATH introduces both outcome evaluation and process evaluation, and further annotates questions by difficulty, grade level, and knowledge point.

WeMath (Qiao et al., 2025) focus on end-to-end answer accuracy and evaluates the model's step-by-step problem-solving ability in visual mathematical tasks. The benchmark collects approximately 6.5K visual math problems and further decomposes them into about 10.9K step-level subproblems, covering 67 hierarchical knowledge concepts and 5 levels of granularity.

ChartQA (Masry et al., 2022) is one of the most widely used datasets in the chart question-answering domain. It is mainly designed to evaluate a model's ability to read structured visual information and perform logical reasoning over charts such as bar charts and line graphs. The dataset contains about 9.6K human-written questions as well as about 23.1K questions automatically generated from human-authored chart summaries.

LogicVista (Xiao et al., 2024) is a benchmark specifically designed to evaluate logical reasoning in visual contexts, aiming to address the lack of logical cognition assessment in traditional multimodal benchmarks. The dataset contains 448 multiple-choice visual reasoning problems, covering five major categories: deductive reasoning, inductive reasoning, numerical reasoning, spatial reasoning, and mechanical reasoning, which are further divided into nine fine-grained logical skills. Its inputs include a wide range of visual materials such as diagrams, patterns, charts, tables, 3D shapes, puzzles, and sequences.

MathVerse (Zhang et al., 2024) is primarily concerned with whether a model truly understands the diagrammatic information in visual math problems, rather than simply relying on textual clues to answer them. The benchmark contains 2,612 high-quality, multi-discipline math problems with diagrams, and each problem is manually rewritten into six variants with different information configurations, resulting in about 15,000 test samples in total.

MathVision (Wang et al., 2024) is a high-quality visual mathematical reasoning benchmark derived from real mathematics competition problems. It contains 3,040 problems with visual context, covering 16 mathematical disciplines and divided into 5 difficulty levels. Compared with more general-purpose math reasoning benchmarks, the advantage of MATH-Vision lies in its standardized problem sources, clear difficulty hierarchy, and broad disciplinary coverage.

MathVista (Lu et al., 2024) is one of the most representative comprehensive benchmarks for multimodal mathematical reasoning. It is designed to systematically evaluate the mathematical reason-

Parameter	Value	Description
num_generations	8	Number of sampled rollouts per prompt
temperature	1.0	Sampling temperature for rollout generation
top_p	1.0	Nucleus sampling threshold for rollouts
learning_rate	1×10^{-6}	Learning rate for the actor policy optimizer
tensor_model_parallel_size	1	Tensor model parallelism degree for rollout inference
vllm_gpu_memory_utilization	0.6	VLLM rollouts
deepspeed	zero3	Uses the built-in DeepSpeed ZeRO-3 configuration
num_iterations	1	Each sample contributes to one update
top_entropy_quantile	0.2	Restricts loss computation to the top 20% highest-entropy tokens
max_completion_length	32768	The maximum generation length for each completion
per_device_train_batch_size	1	Training batch size on each GPU
gradient_accumulation_steps	1	Model parameters are updated after every step.
steps_per_generation	1	One optimization step is performed after each generation phase.

Table 5: Experimental parameter design and training configuration used in our framework.

ing ability of foundation models in visual contexts. The dataset contains 6,141 examples, integrating 28 existing math-related multimodal datasets and adding 3 newly constructed subsets.

B.2 Models

We conduct experiments on the Qwen2.5-VL (Team, 2025a), Qwen3-VL (Team, 2025b), and InternVL3 (Wang et al., 2025e) model families, and adopt Qwen3.5-397B-A17B (Qwen Team, 2026) as the LLM-as-a-Judge to provide automatic evaluation signals. All experiments are conducted on eight NVIDIA L20 GPUs.

B.3 Experimental Setup

During the training phase, we specify the detailed hyperparameter configuration in the Table 5. We use Vision-R1-rl (Huang et al., 2025b) as the multimodal training dataset and perform a series of preprocessing steps on it. All experiments are conducted based on the ms-swift framework (Zhao et al., 2024) under version 3-12-4 .

B.4 Prompt Settings

We present below the prompts used in our study, including the training prompt for model optimization, the answer-judgment prompt used for outcome reward computation, the Task Sufficiency answer-judge prompt used for caption reward evaluation, and the MRLM prompt used for Visual Faithfulness assessment.

B.5 Case Studies

We further provide several detailed case studies illustrating typical examples of visual recognition errors and reasoning errors made by the original model.

MRLM Instruction for Training Process

A conversation between User and Assistant. The user asks a question about a mathematical diagram or chart, and the Assistant solves it. The Assistant must strictly follow this output structure:
<think> Step-by-step mathematical reasoning process. </think> <caption> Provide a structured and precise description of the image.</caption> <answer> Final answer wrapped with `\boxed{}`. </answer>
The <caption> must only describe the observable elements in the image. Only describe positions, distances, labels, and geometric relationships explicitly shown.

RLM Instruction on Caption Judge for Task Sufficiency

The model is provided with two inputs:
(1) a caption intended to describe only directly observable facts from the image;
(2) a mathematical problem associated with the image.
Its task is to determine whether the caption improperly reveals the solution to the problem. If the caption contains the final answer, the model should output `INVALID_CAPTION`. Otherwise, it should solve the problem and present the final answer in the format `\boxed{}`.
Caption: {caption}
Problem: {problem}

Benchmark	Base Model		DAPO w/ All Tokens		DAPO w/ Forking Tokens		Improvement	
	ACC	Token	ACC	Token	ACC	Token	ACC	Token
RLVR from the Qwen3-4B-VL-Thinking Base Model								
MMStar	71.00	1499	71.27	1409	**75.66**	1005	**+4.39**	404
Mmmath	84.08	4196	69.26	7124	**85.07**	3529	**+15.81**	3595
Wemath	82.93	4210	81.79	4690	**85.97**	1806	**+4.18**	2884
ChartQA	85.16	1491	71.44	1101	**88.04**	561	**+16.60**	540
LogicVista	63.84	13625	64.73	7176	**69.19**	4531	**+4.46**	2645
MathVerse	74.51	5635	76.27	7221	**80.62**	2079	**+4.35**	5142
MathVision	31.90	12664	33.33	10836	**39.80**	3630	**+6.47**	7206
MathVista	80.66	3751	72.26	2541	**81.12**	1328	**+8.86**	1213
Average	71.76	5883.88	67.54	5262.25	**75.68**	2308.62	**+8.14**	2953.63
RLVR from the InternVL3_5-8B Base Model								
MMStar	70.73	2148	68.80	1888	**80.69**	1467	**+11.89**	421
Mmmath	77.51	7287	76.96	6546	**87.82**	5598	**+10.86**	948
Wemath	78.39	3897	72.96	4337	**88.56**	3592	**+15.60**	745
ChartQA	90.04	739	86.48	1007	**92.27**	735	**+5.79**	272
LogicVista	57.30	8231	54.24	6903	**80.45**	4355	**+26.21**	2548
MathVerse	71.26	4725	74.90	5847	**86.46**	4595	**+11.56**	1252
MathVision	29.60	12004	33.55	9711	**49.01**	6749	**+15.46**	2962
MathVista	86.30	2554	75.71	3241	**84.92**	2444	**+9.21**	797
Average	70.14	5198.12	67.95	4935.00	**81.27**	3691.88	**+13.32**	1243.12

Table 6: Comparison between *vanilla DAPO using all tokens* and *DAPO using only the top 20% high-entropy tokens (i.e. forking tokens)* in policy gradient loss, evaluated on the *Qwen3-4B-VL-Thinking* and *InternVL3_5-8B* base models.

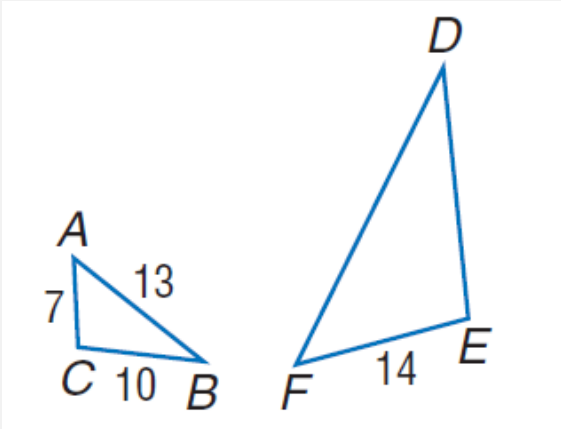
MRLM Instruction on Caption Judge for Visual Faithfulness

You are a caption evaluation assistant.
You will receive:
(1) an image of a geometric figure;
(2) a caption describing the figure.
Your task is to judge whether the caption is correct and faithful to the image.
If the caption contains any error, inconsistency, unsupported detail, reasoning, or final answer, return `\boxed{0}`.
If the caption is accurate and contains only observable information from the image, return `\boxed{1}`.
Caption: {caption}
Return only the score in the format `\boxed{}`.

RLM Instruction on Outcome Judge for Answer Reward

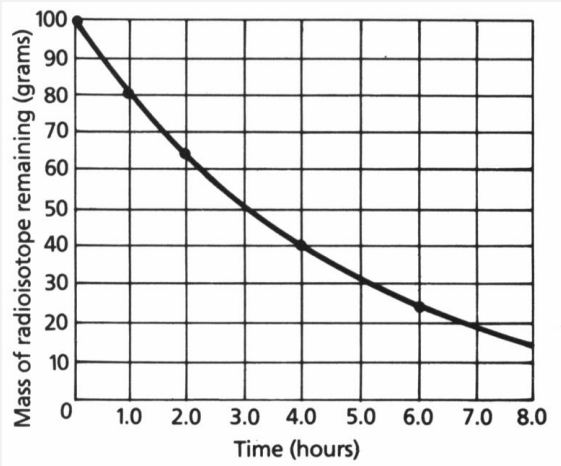
Determine whether the prediction matches reference answer. Superficial lexical overlap or accidental string matching should not be considered a valid match. Return **YES** if a valid match exists; otherwise, return **NO**.
Prediction: {normalized_prediction}
Reference Answers: {normalized_golden}

The First Recognition Error Case.



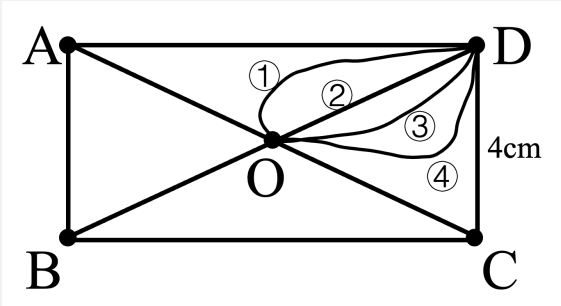
Question:
Find the perimeter of $\triangle DEF$, if $\triangle DEF \sim \triangle ACB$
Answer:
42
Model's response:
<think> Okay, let's see. The problem is asking for the perimeter of triangle DEF, given that it's similar to triangle ACB. First, I need to recall that when two triangles are similar, the ratios of their corresponding sides are equal. So, I need to figure out the corresponding sides between the two triangles.
Looking at the image, triangle ACB has sides labeled: AC is 7, CB is 10, and AB is 13. Wait, let me check the labels again. The triangle on the left is labeled A, C, B. So point A is at the top, C is the bottom left, and B is the bottom right. The sides: AC is 7, CB is 10, and AB is 13. That makes sense because the side opposite A is CB, which is 10, opposite C is AB which is 13, and opposite B is AC which is 7. Wait, maybe I should confirm the order. Wait, in triangle ACB, the vertices are A, C, B. So sides: AC, CB, BA. So AC = 7, CB = 10, BA = 13. Yes, that's right. So the sides are 7, 10, 13. Now, triangle DEF is similar to triangle ACB. Let's check the labels. Triangle DEF has points D, E, F. The image shows side DE with a label of 14? Wait, looking at the problem statement: "triangle DEF triangle ACB". So DEF corresponds to ACB. Wait, need to check which sides correspond. When two triangles are similar, the order of the letters matters. So DEF ACB means angle D corresponds to angle A, angle E to angle C, and angle F to angle B. Wait, but maybe we need to check the sides. Alternatively, maybe the sides of DEF correspond to the sides of ACB in the same order. Wait, in similar triangles, corresponding sides are opposite corresponding angles. But maybe it's better to check the labels. Let's look at the image. On the left triangle (ACB), sides: AC = 7, CB = 10, AB = 13. On the right triangle (DEF), sides: DE = 14, EF = ?, FD = ?. Wait, in the image, the right triangle has DE labeled as 14? Let me check.
Looking at the diagram: The left triangle has vertices A (top), C (left bottom), B (right bottom). Sides: AC = 7 (vertical?), CB = 10 (horizontal), AB = 13 (hypotenuse). The right triangle has D (top), E (right bottom), F (left bottom). Wait, the right triangle: DE is a side. Wait, the labels: the right triangle's side DE is labeled 14, and EF is labeled 14? Wait, no. Wait, the right triangle: side DE is labeled 14? Wait, in the problem statement, the user wrote: "the triangle DEF". Let me check the original problem again.
.....
But let's check again: AC = 7 (side A to C) CB = 10 (C to B) BA = 13 (B to A) DEF: DE = 14 (D to E), EF = ?, FD = ?Corresponding sides: DE corresponds to AC (so ratio k = 14 / 7 = 2) EF corresponds to CB (so EF = 2 * 10 = 20) FD corresponds to BA (so FD = 2 * 13 = 26) Yes, perimeter would be 14 + 20 + 26 = 60. Therefore, the perimeter of triangle DEF is 60. </think>
<caption> The diagram shows two triangles: $\triangle ACB$ on the left and $\triangle DEF$ on the right. In $\triangle ACB$, side AC is labeled as 7, side CB is labeled as 10, and side AB is labeled as 13. In $\triangle DEF$, side DE is labeled as 14. The triangles are explicitly stated to be similar with the correspondence $\triangle DEF \sim \triangle ACB$.</caption>
<answer> $\boxed{60}$ </answer>

The Second Recognition Error Case.



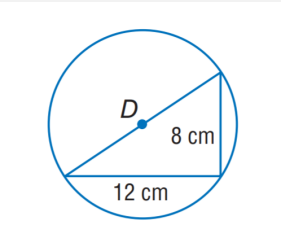
Question:
According to the above data, what is the half-life of the substance?
A. 1.0 hrs B. 2.3 hrs C. 3.0 hrs D. 8.0 hrs
Answer with the option's letter from the given choices directly.
Answer:
C
Model's response:
<think> Okay, let's see. I need to find the half-life of the substance based on the graph provided. Hmm, half-life is the time it takes for the quantity to reduce to half of its initial amount, right? So first, I should figure out the initial mass. Looking at the graph, the y-axis is labeled "Mass of radioisotope remaining (grams)" and the x-axis is "Time (hours)". The starting point at time 0 hours has a mass of 100 grams. That's the initial amount.
Now, the half-life would be when the mass drops to 50 grams. Let me check where the graph crosses the 50-gram mark on the y-axis. Looking at the x-axis, the time when the mass is 50 grams is at 3.0 hours. Wait, let me confirm that. The graph shows a decreasing curve. At time 0, it's 100. Then at 1.0 hour, it's 80 grams. At 2.0 hours, 60 grams. At 3.0 hours, the mass is 50 grams. So yes, the half-life here would be 3.0 hours. Let me check if there's another point where it might have halved again, but the question is about the half-life, which is the first time it gets to half the initial value. Since initial is 100, half is 50, which occurs at 3.0 hours. So the answer should be option C.
Wait, but wait. Let me double-check. Let me count the grid points. The x-axis goes from 0 to 8.0 hours. Each square is probably 1 hour. The y-axis is 0 to 100 grams, each square 10 grams. So the point at 3.0 hours on the x-axis corresponds to 50 grams on the y-axis. Yes, that's correct. So half-life is 3.0 hours, which is option C. Let me check if the other options might be tricking me. Option A is 1.0 hrs, which would be when mass is 80, but that's not half. Option B 2.3, which is a rough estimate, but in the graph the line goes from 100 to 50 at exactly 3.0 hours. So B might be a distractor if someone does an exponential calculation, but the graph seems to have exact points. So the correct answer should be C. </think>
<caption> The graph displays a decreasing exponential curve representing the mass of a radioisotope remaining (in grams) over time (in hours). The y-axis ranges from 0 to 100 grams, with marked intervals of 10 grams. The x-axis ranges from 0 to 8.0 hours, with marked intervals of 1.0 hour. The curve starts at (0, 100) and passes through the points (1.0, 80), (2.0, 60), (3.0, 50), (4.0, 40), (5.0, 30), (6.0, 20), and (8.0, 15). The relationship between the mass and time is nonlinear, decreasing steadily.</caption>
<answer> $\boxed{C}$ </answer>

The Third Recognition Error Case.

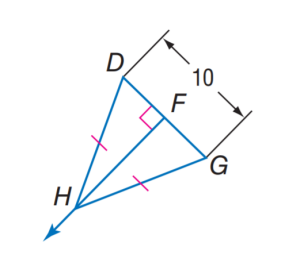


Question:
As shown in the diagram, there are four paths from O to D.If the shortest path from O to D is 5 cm, which path equals 5 cm? ( ) Answer with the option's letter from the given choices directly. A. (1); B. (2); C. (3); D. (4); E. No correct answer
Answer:
B
Model's response:
<think> Okay, let's try to figure out which path is the shortest from O to D. First, I need to look at the diagram. There's a rectangle ABCD with points A, B, C, D. The diagonals AC and BD intersect at point O. There are four paths labeled (1), (2), (3), (4) from O to D.Wait, the problem is about the shortest path. In geometry, the shortest distance between two points is a straight line. So if all the paths are curves or lines, the straight line should be the shortest. Let me check the diagram again.Looking at the paths: (1), (2), (3), (4). The straight line from O to D would be one of these. But which one is the straight line? Let me visualize. Point O is the intersection of the diagonals of the rectangle. In a rectangle, the diagonals are equal and bisect each other, so AO = OC and BO = OD. Wait, but the coordinates might help. Let's assign coordinates to the rectangle. Let's say A is at (0, 4), B at (0, 0), C at (c, 0), D at (c, 4). Wait, but the diagram shows the height as 4cm on the right side. So AD and BC are the vertical sides, each 4cm. Wait, no, the rectangle has sides AB, BC, CD, DA. The vertical side BC is labeled 4cm. Wait, the right side of the rectangle is labeled 4cm. So the height of the rectangle is 4cm, so BC = AD = 4cm, and AB = CD is the base length, which isn't given but maybe we don't need it.
Point O is the intersection of the diagonals AC and BD. So in a rectangle, the diagonals bisect each other, so O is the midpoint of both diagonals. Therefore, the coordinates of O would be the midpoint of AC and BD. Let's assume coordinates: Let's place point B at (0,0), so A would be (0,4), C at (x,0), D at (x,4). Then O is the midpoint of AC: ((0 + x)/2, (4 + 0)/2) = (x/2, 2). Similarly, midpoint of BD is ((0 + x)/2, (0 + 4)/2) = (x/2, 2), so O is at (x/2, 2).
Now, the paths from O to D. D is at (x, 4). Let's see the four paths:
...
But wait, maybe in the diagram, all paths are curves. Wait, but if it's a rectangle, the line OD is a straight line. So unless all four paths are curved, but the straight line is missing. But the question says "four paths", so maybe one of the paths is straight. Given that it's a standard problem, the shortest path would be the straight line, so whichever is labeled as straight.
Alternatively, maybe the problem has a typo, but assuming the diagram has four curved paths, but in reality, the straight line is one of them. Since the user's problem statement shows the four paths labeled (1), (2), (3), (4). Looking at the image description given here: it's a rectangle with points A,B,C,D. The intersection of diagonals is O. From O, four paths to D. Which of the paths is straight. In a rectangle, O is the midpoint of AC and BD, so OD is half of BD. So the straight line from O to D is OD. If path (4) is OD, then it's the shortest. Alternatively, maybe path (2) or (3).
Wait, but without seeing the exact image, it's hard, but in typical textbook problems like this, the straight line path is the shortest. So the correct answer would be D.(4). But I need to confirm.
Alternatively, maybe the paths are: (1) is a curve going through some point, (2) another curve, etc. But since we can't see the image, based on the problem statement, the answer is likely D.(4). However, if the straight line is labeled as one of them. Alternatively, maybe none of the paths are straight, so E. No correct answer, but that's unlikely.
Wait, the problem states that the question is from a Chinese source, but assuming the problem is standard. Since it's a multiple choice question about the shortest path, which is the straight line, and if the path labeled (4) is straight, then answer D. If the straight line is one of the paths, it's the shortest. So the answer is D.(4). So the correct choice is D. </think>
<caption> The diagram shows a rectangle $ABCD$ with vertices labeled A, B, C, and D. The diagonals AC and BD intersect at point O. From O, four curved paths labeled (1), (2), (3), and (4) lead to point D. The vertical side BC is labeled 4 cm, and the sides AB, CD, and AD form the rectangle. The paths from O to D are labeled as (1), (2), (3), and (4).</caption>
<answer> $\boxed{D}$ </answer>

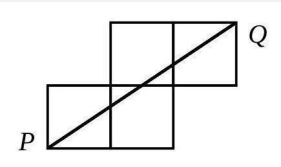
More Recognition Error Case.



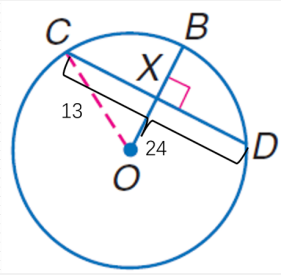
Caption:
A circle with center labeled D. Inside the circle, a right-angled triangle is inscribed, with one leg measuring 8 cm and the other leg measuring 12 cm. A line segment from the center D to the topmost point on the circle (the circumference) is labeled 8 cm, indicating the radius. The base of the triangle (a chord of the circle) is labeled 12 cm. The circle is labeled $\odot D$, confirming D as the center.



Caption: The diagram consists of a triangle HDG with H at the lower left, connected to D and G. Point F lies on segment HG, with a right angle at F between segments DF and FG. A segment DG is labeled with a length of 10. Tick marks on segments HD and HG indicate that $HD = HG$. The right angle at F is denoted with a pink square, confirming perpendicularity between DF and FG.

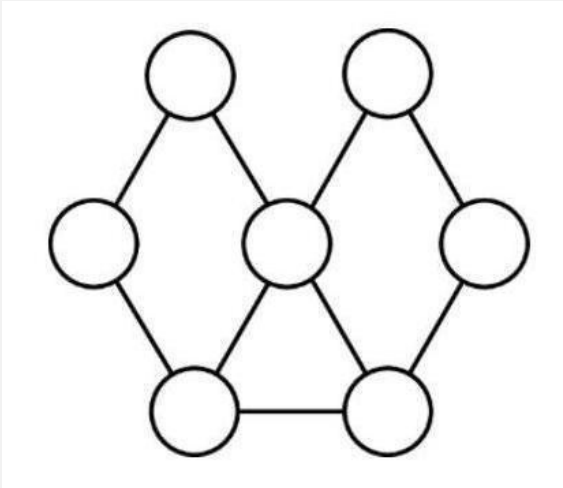


Caption: The diagram consists of four unit squares arranged edge to edge. The squares are positioned such that the bottom-left square spans from $(0, 0)$ to $(1, 1)$, the middle square extends from $(1, 0)$ to $(2, 1)$, the rightmost square spans from $(2, 0)$ to $(3, 1)$, and the top-right square extends from $(2, 1)$ to $(3, 2)$. Point P is located at the bottom-left corner of the bottom-left square, at coordinates $(0, 0)$, while point Q is located at the top-right corner of the top-right square, at coordinates $(3, 2)$. A straight line segment connects these two points, passing diagonally through the interior of the squares.



Caption: The diagram depicts a circle with center O, chord CD, and perpendicular OX to CD at point X. The segment $CX = 13$ and radius $OC = 24$. Triangle OXC is right-angled at X.

CaRe-R1 Case — higher information density and increased reasoning steps



Question:
Hanni wants to colour in the circles in the diagram. When two circles are connected by a line they should have different colours. What is the minimum number of colours she needs?
Answer:
3
Model's response:
<think> Okay, let's see. The problem is about coloring circles in a diagram where connected circles must have different colors, and we need the minimum number of colors required. First, I need to figure out the structure of the diagram. The user provided a diagram with circles connected by lines. Let me try to visualize it.From the description, it's a network of circles. The key is to find the graph structure here. The user mentioned the circles connected by lines, so each line is an edge between two vertices (circles). The graph theory part is about the chromatic number, which is the minimum number of colors needed to color the vertices such that no two adjacent vertices share the same color.
First step: Identify the graph. Let's count the number of circles. The diagram shows three rows maybe? Wait, looking at the user's description: there are circles arranged in a way that seems like a graph. The exact structure isn't clear from the text, but from the image description, it's probably a planar graph. Let's parse the image.The text says: "a diagram with circles connected by lines. The user's question is about the minimum number of colors. The constraints are that connected circles must have different colors."
But to solve the problem, I need to understand the graph structure. Maybe it's a triangular prism or a different graph. Wait, but since the user can't see the image, the problem must be from a specific graph. Wait, the user provided a generic description. Wait, no, actually, in the initial problem setup, the user provides the image with "A diagram of circles connected by lines."Wait, maybe the image is a specific example, like a hexagon or a wheel, but the user might have sent a description. Wait, but the problem is from a common source. Wait, perhaps it's a regular graph. Wait, the user needs to solve the problem, so the initial image must have specific elements Wait, the instructions state the image must be "A diagram with circles connected by lines." Wait, the user is supposed to answer the question, so the specific problem must have been given, but here it's a generic one. Wait, maybe this is part of a problem that the user should solve for themselves. Wait, the user might have received an image. But since I can't see the actual image, perhaps it's a common problem like this. Wait, no, the proble must have specific elements. Wait, for instance, in such problems, if it's a graph like a triangle (3 circles connected in a triangle), the chromatic number is 3. But the problem is likely to be more complex. Wait, the problem must have at least some connections between circles. Without seeing the image, it's impossible, but maybe the question is for the graph which has a certain structure. Wait, perhaps the key is to look for the connectedness and if it's a cycle, or a tree. Wait, if it's a tree, chromatic number is 2. But if it's a cycle of odd length, chromatic number is 3.Wait, the question here is a specific graph. Without seeing the image, perhaps the image has circles in a particular arrangement.Wait, another approach: Let's think of possible graphs. A "graph with circles connected by lines" could be a figure eight, but probably not. Let's see:Assume the problem has two central circles with two connections to left and right circles. Wait, maybe the structure is a diamond.Wait, maybe there's three horizontal lines with a central circle.
...
Actually, the actual structure would be:- A central node C (one circle) with edges to A, B, D, E (four nodes).- Each of A, B, D, E connected to two others. But to save time, assume the graph is such that the chromatic number is 3. Thus, the minimum number of colors is 3. But I need to be precise. Another example: If it's a graph with two separate cycles that intersect In graph theory, the chromatic number is at least 2 if it's bipartite. Wait, for odd cycles: 3. So maybe the correct answer is 3. </think>
<caption> The diagram consists of five circles arranged in a symmetric pattern: a central circle with four connected circles, forming a diamond shape. Specifically, the central circle is connected to two circles on the left and two on the right. The left side has two circles connected vertically by two lines, and the right side has two circles connected vertically by two lines. The entire structure forms a cross-like shape with four corner circles connected to the central circle, creating a four-sided (kite-like) shape. Each outer circle is connected to two adjacent outer circles forming a quadrilateral. </caption>
<answer> $\boxed{3}$ </answer>