MM-EUREKA: EXPLORING VISUAL AHA MOMENT WITH RULE-BASED LARGE-SCALE REINFORCEMENT LEARNING

Fanqing Meng* **Lingxiao Du*** **Zongkai Liu** **Zhixiang Zhou** **Quanfeng Lu**
Daocheng Fu **Botian Shi** **Wenhai Wang** **Junjun He** **Kaipeng Zhang**
Ping Luo **Yu Qiao** **Qiaosheng Zhang†** **Wenqi Shao†**

Shanghai AI Laboratory **Shanghai Innovation Institute**
Shanghai Jiao Tong University **The University of Hong Kong**

ABSTRACT

We present MM-Eureka, a multimodal reasoning model that successfully extends large-scale rule-based reinforcement learning (RL) to multimodal reasoning. While rule-based RL has shown remarkable success in improving LLMs' reasoning abilities in text domains, its application to multimodal settings has remained challenging. Our work reproduces key characteristics of text-based RL systems like DeepSeek-R1 in the multimodal space, including steady increases in accuracy reward and response length, and the emergence of reflection behaviors. We demonstrate that both instruction-tuned and pre-trained models can develop strong multimodal reasoning capabilities through rule-based RL without supervised fine-tuning, showing superior data efficiency compared to alternative approaches. We open-source our complete pipeline to foster further research in this area. We release all our codes, models, data, etc. at `https://github.com/ModalMinds/MM-EUREKA`

1 INTRODUCTION

Large-scale reinforcement learning (RL) (Sutton et al., 1998) has demonstrated remarkable progress in improving the reasoning ability of Large Language Models (LLMs), particularly in the math and code domains (OpenAI, 2024; DeepSeek-AI et al., 2025). Recent research, such as o1 (OpenAI, 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025), shows that large-scale RL can achieve breakthrough improvements in complex reasoning tasks during post-training phases, sometimes even without supervised fine-tuning (SFT) (Radford et al., 2019). Despite great success in the text domain, many real-world reasoning tasks such as interpreting scientific diagrams and geometrical reasoning can only be effectively solved with the image input. However, transferring large-scale RL techniques that work well for LLMs to multimodal scenarios remains underexplored.

Recent efforts largely fail to reproduce key characteristics of DeepSeek-R1 in multimodal settings, such as stable growth in response length and accuracy reward. For example, R1-V (Chen et al., 2025) shows improvements only in simple counting tasks, but does not reproduce the increase in response length and "aha moments". R1-Multimodal-Journey (Meng & Du, 2025) explores geometric problems, but the response length decreases as the training goes on. In addition, LMM-R1 (Peng et al., 2025) achieves gains in accuracy reward and response length. However, such a success has not been verified in the large-scale training with image-text data. Although Kimi1.5 (Team et al., 2025) has achieved competitive results in multimodal reasoning, it has not open-sourced its model or training data to the community. In this report, we aim to investigate the effectiveness of large-scale RL in multimodal reasoning and open-source our pipeline of reproducing multimodal reasoning.

† Corresponding Authors: shaowenqi@pjlab.org.cn; zhangqiaosheng@pjlab.org.cn
∗ Equal contribution

To this end, we build the multimodal reasoning models MM-Eureka, which successfully reproduces key characteristics of DeepSeek-R1 in the multimodal reasoning scenarios, including steady increases in accuracy reward and response length as training progresses, and the emergence of reflection and backtracking operations (e.g., aha-moments) during training.

Through the journey of developing MM-Eureka, we have several findings. **First,** we observe visual aha moments where the model rechecks the intermediate step by looking for more clues from the image, implying that multimodal reasoning ability can also be cultivated with large-scale RL. **Second,** both instruction-tuned and pre-trained models can obtain improvements in multimodal reasoning ability with minimal training setups (e.g., rule-based reward functions without KL divergence constraints) combined with difficulty-based data selection strategies. **Third,** rule-based RL demonstrates competitive performance and remarkable data efficiency compared to post-training strategies like MPO (Wang et al., 2024b) and SFT across various benchmarks.

Specifically, we present MM-Eureka-8B and MM-Eureka-Zero-38B, which are trained from InternVL2.5-Instruct-8B and InternVL2.5-Pretrained-38B, respectively. We evaluate their performance on 5 representative benchmarks including MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024a), OlympiadBench (He et al., 2024), and our mannualy-collected K12 math test set. MM-Eureka employs 54K image-text data for rule-based RL, achieving an average performance on the above benchmarks that surpasses the model trained with 1M data using MPO (Wang et al., 2024b). The overall benchmark accuracy is also comparable to the model trained with 12M data via CoT SFT (Guo et al., 2024). Moreover, MM-Eureka-Zero directly applies rule-based RL with only 8K image-text math reasoning data, outperforming the instruct model trained with 16.3M data on certain benchmarks such as OlympiadBench and demonstrating comparable performance on MathVerse.

Our goal is to share our implementation experiences and complete open-source pipeline with the community, including data, code, and models. We believe this comprehensive open-source framework would help the community better explore the multimodal reasoning task. The main contributions are sumarized as follows:

- We contribute a multimodal large-scale reinforcement learning framework based on OpenRLHF (Hu et al., 2024), supporting various models including InternVL (Chen et al., 2024b), as well as multiple RL algorithms. Compared to frameworks like R1-V (Chen et al., 2025), our framework demonstrates enhanced scalability, enabling the training of substantially larger models, such as the InternVL2.5-38B.
- We build the multimodal reasoning models MM-Eureka-8B and MM-Eureka-Zero-38B, where MM-Eureka is based on an InternVL2.5-Instruct-8B, and MM-Eureka-Zero is based on InternVL2.5-Pretrained-38B. Both models present visual aha moments, achieving steady increases in accuracy reward and response length.
- Our experiments demonstrate that simple rule-based RL has significant advantages in data efficiency compared to other post-training approaches like MPO and SFT. Despite utilizing only 0.05% of the training data compared to the instruct model, MM-Eureka-Zero-38B demonstrates superior performance with an 8.2% accuracy improvement on the K12 benchmark while maintaining comparable results on the MathVerse evaluation.

2 METHOD

2.1 BASIC SETTINGS

We employ InternVL2.5 (Chen et al., 2024a) as the base model because it offers a wide range of model sizes, making it suitable for scale-up experiments. Thanks to the strong performance of the base model across various tasks, we can systematically study the impact of scaling in RL while ensuring that our models remain competitive in terms of capability. We use large-scale RL on models with different sizes (8B or 38B), instruction-tuned or pretrained models, and models with or without cold starts using distilled data, and analyze the experimental results. Our RL algorithm is similar to DeepSeek-R1 (DeepSeek-AI et al., 2025), using rule-based format rewards $r_{\text{format}} \in \{0, 1\}$ and accuracy rewards $r_{\text{accuracy}} \in \{0, 1\}$ for training. Furthermore, we develop a multimodal input RL framework based on OpenRLHF, which is compatible with commonly used models such as

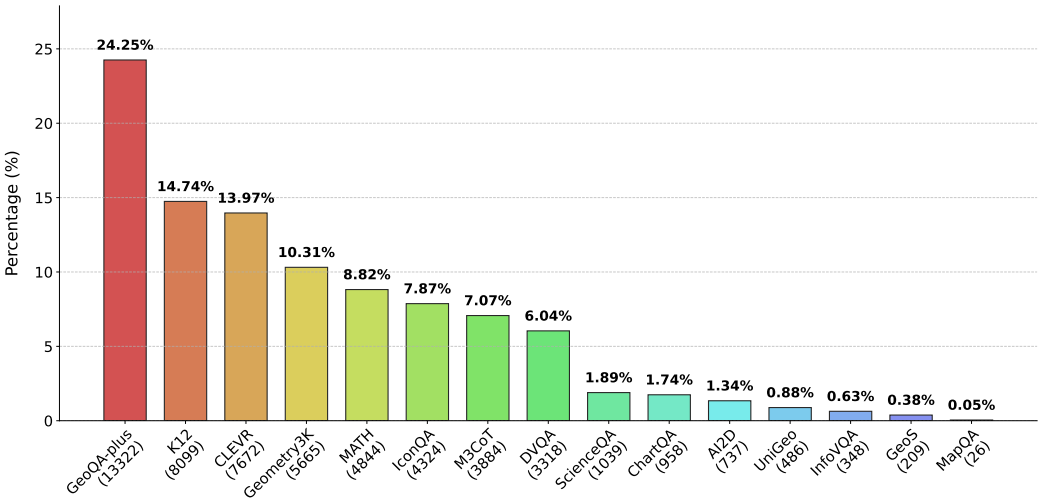


Figure 1: The distribution of the training dataset.

InternVL (Chen et al., 2024b), supporting various model sizes and RL algorithms. In the following sections, we provide detailed settings for our scale-up RL training.

2.2 Dataset

In this section, we introduce our dataset construction and cleaning process. Our dataset primarily consists of open-source data. To enhance the reasoning characteristics of the data, we manually collect visual questions and answers on math at the K-12 level. Note that the chain-of-thought reasoning steps are unnecessarily collected due to the simplicity of the RL algorithm and strong performance of the base model (Ye et al., 2025; Li et al., 2025). To further improve data quality for stable rule-based RL training, we implemented the following cleaning process. In the end, the composition of training dataset is shown in Figure 1.

Data Collection. We collect data with mathematical reasoning properties from multiple open-source datasets, including GeoQA (Chen et al., 2022), ChartQA (Masry et al., 2022), MATH (Hendrycks et al., 2021), etc. We collect them from MMPR (Wang et al., 2024b). We also manually collect K-12 level multimodal mathematical reasoning problems. In the end, we collect 75514 samples, which encompasses several key categories:

- Chart Comprehension: This task requires models to accurately interpret charts and formulate solution strategies based on question key points, including datasets like ChartQA (Masry et al., 2022) and DVQA (Kafle et al., 2018). These data help models understand and reason about real-world scientific visualizations to generate appropriate responses.
- General Scientific Reasoning: This component demands scientific common knowledge combined with reasoning capabilities to derive answers. Datasets such as AI2D (Kembhavi et al., 2016) and ScienceQA (Lu et al., 2024) are included. These materials strengthen the model's scientific knowledge foundation.
- Mathematical Reasoning: This type of data focuses on mathematical reasoning tasks, including geometry and function problems, with datasets like K12 and GeoQA (Chen et al., 2022). These resources enhance the model's ability to perform mathematical reasoning with visual inputs.

Data Filter. Two steps are adopted to remove low-quality data in RL training. First, we exclude problems without clear answers or those difficult to correctly parse with our rule-based reward function, such as proof problems and multi-choice problems. Second, we use InternVL2.5-8B-instruct to generate 8 responses for each problem, estimating the difficulty of the problem based on the accuracy among 8 responses. We remove the problems with estimated accuracy of 0 or 1 to ensure a stable RL training process. After filtering, we have 54931 samples left.

2.3 REWARD FUNCTION

Following DeepSeek-R1, we also adopt the simple rule-based reward function rather than using outcome or process reward models, thereby alleviating reward hacking (Gao et al., 2022). Specifically, we use two types of rewards: accuracy reward and format reward. The former uses math-verify library* to extract the answer from model responses and compare it with the reference one, returning 1 or 0 based on correctness; the latter checks whether the response follows the specified format (`<think>...</think><answer>...</answer>`), also returning 1 or 0 based on compliance.

The final reward is defined as $r = r_{\text{accuracy}} + \lambda r_{\text{format}}$ where r_{accuracy} is the accuracy reward, r_{format} is the format reward, and λ is a scaling coefficient that balances the contribution of the format reward. We find that this simple and sparse reward is sufficient to significantly improve the model's multimodal reasoning ability.

2.4 ADVANTAGE ESTIMATION AND POLICY UPDATE

We employ the REINFORCE Leave-One-Out (RLOO) algorithm (Kool et al., 2019; Ahmadian et al., 2024) in our reinforcement learning training phase. Similar to GRPO used in DeepSeek-R1 (DeepSeek-AI et al., 2025), the RLOO algorithm does not require a critic model, which effectively reduces the training cost. Moreover, it leverages a leave-one-out baseline to reduce the variance in the policy gradient estimates. Specifically, for each query $\mathbf{x}$, the model generates K responses $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \cdots, \mathbf{y}^{(K)}\}$. Each query-response pair $\{\mathbf{x}, \mathbf{y}^{(i)}\}$ receives a score $r^{(i)}$ determined by a rule-based reward function, which comprises both an accuracy reward and a format reward. Consequently, the advantage estimator is computed as follows:

$$A^{(i)} = r^{(i)} - \frac{1}{K-1}\sum_{j \neq i} r^{(j)}, \quad i = 1, \cdots, K.$$

Regarding the actor loss, rather than employing the standard REINFORCE objective, we adopt a PPO-clip loss (Schulman et al., 2017):

$$J_{\text{PPO}}(\theta) = -\mathbb{E}_t\Big[\min\Big(\frac{\pi_\theta(y_t^{(i)}|\mathbf{x}, \mathbf{y}_{<t}^{(i)})}{\pi_{\theta_{\text{old}}}(y_t^{(i)}|\mathbf{x}, \mathbf{y}_{<t}^{(i)})}A^{(i)}, \text{clip}\big(\frac{\pi_\theta(y_t^{(i)}|\mathbf{x}, \mathbf{y}_{<t}^{(i)})}{\pi_{\theta_{\text{old}}}(y_t^{(i)}|\mathbf{x}, \mathbf{y}_{<t}^{(i)})}, 1-\epsilon, 1+\epsilon\big)A^{(i)}\Big)\Big],$$

where ϵ is the clipping parameter, π_θ is the current policy, and $\pi_{\theta_{\text{old}}}$ is the old policy before the update. This design choice follows established implementations in TRL (von Werra et al., 2020) and OpenRLHF (Hu et al., 2024) frameworks, while maintaining theoretical consistency since the REINFORCE gradient can be regarded as a special case of PPO loss (Huang et al., 2022). For the KL divergence loss between the policy π_θ and the reference policy π_{ref}, we follow the same method as GRPO by directly adding it as a regularization term to the loss:

$$J(\theta) = J_{\text{PPO}}(\theta) + \alpha_{\text{KL}}\mathcal{D}_{\text{KL}}(\pi_\theta, \pi_{\text{ref}}),$$

where α_{KL} is the weight parameter. However, it is important to note that we typically set the weight α_{KL} to 0 in practice, as it performs better in the experiments.

2.5 KEY FINDINGS

Data filtering is crucial for stable RL training. We find that the difficulty-based data filtering strategy described in Section 2.2 is crucial for stable RL training in multimodal reasoning scenarios. For comparison, we randomly sampled an equal amount of unfiltered data versus filtered data, and trained using the same parameters. As shown in Figure 2, we find that model training on unfiltered data is extremely unstable, with accuracy reward showing fluctuating trends and response length showing a declining trend. We believe this is because data with all incorrect answers causes unstable rewards, making learning difficult. How to utilize these difficult data requires further exploration.

The Simplest RL training setups are sufficient. We find that the simplest experimental setting, particularly without KL divergence, effectively reproduces DeepSeek-R1's reasoning patterns in multimodal inference scenarios, which is similar to what Open-Reasoner-Zero (Hu et al., 2025) has

*https://github.com/huggingface/Math-Verify

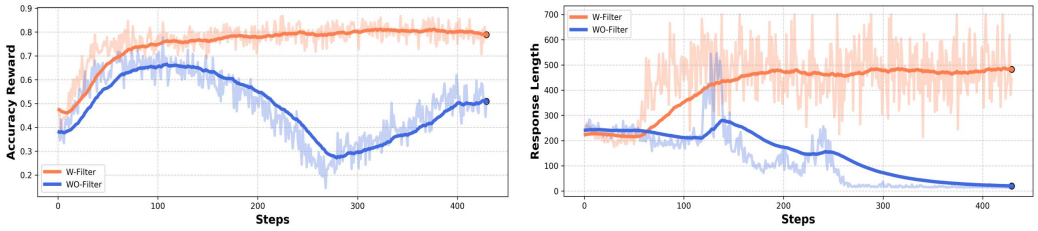


Figure 2: Comparison of training on InternVL2.5-Instruct-8B with and without data filtering. When data filtering is not used, the accuracy reward shows fluctuating trends, while response length exhibits a downward trend.

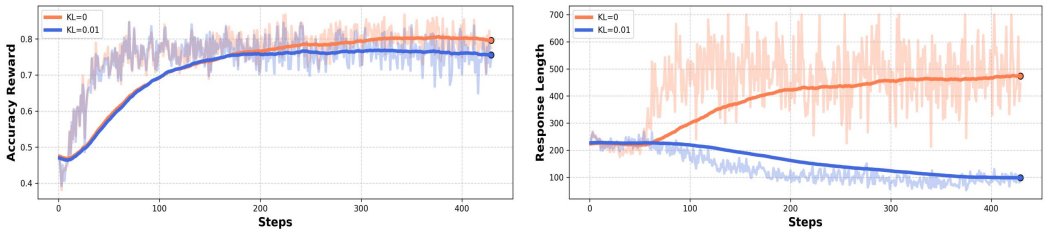


Figure 3: Comparison of RL training on InternVL2.5-Instruct-8B with and without KL divergence. When KL divergence is added, response length consistently shows a downward trend, and the peak accuracy reward is marginally lower than in the case without KL divergence.

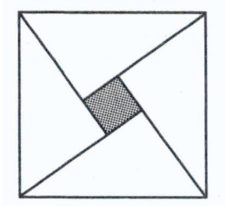
found in language-context reinforcement learning. Although prior work suggests that maintaining KL divergence helps prevent models from excessively focusing on reward maximization at the expense of the degradation of linguistic patterns acquired during pretraining (Ouyang et al., 2022; Xiong et al., 2024), our results demonstrate that when fine-tuning from an instruction-tuned model, omitting KL divergence can not only reduce computational cost during training but offers a stable growth in response length.

As shown in Figure 3, we compare training with initial $\alpha_{\text{KL}} = 0$ versus $\alpha_{\text{KL}} = 0.01$. With $\alpha_{\text{KL}} = 0.01$, response lengths consistently decrease, and the accuracy reward is lower than when $\alpha_{\text{KL}} = 0$. We attribute this to the fact that when initializing from an instruction-tuned model, KL divergence further constrains exploration, preventing the model from engaging in thorough reasoning and resulting in lower accuracy rewards compared to training without KL divergence.

Visual Aha Moment. We find that rule-based RL training in multimodal reasoning scenarios can similarly elicit emergent behaviors such as reflection and verification (e.g., aha moment). Notably, we observe that models engage in re-perception of images in multimodal reasoning contexts. As shown in Figure 4, the model first plans a solution approach, and upon completion, reflects on its work and adopts a new method to resolve the problem. Interestingly, during this process, the model demonstrates behavior resembling image re-perception (e.g., "Upon closer inspection"). Finally, through a verification step, the model arrives at the correct answer. Additionally, even for non-mathematical problems, the model has learned similar thinking patterns. As shown in Figure 5, for visual common sense questions and chart understanding tasks, the model shows extended thinking processes that include reflection and similar operations.

3 Experiments

In this section, we present a comprehensive overview of our model's training process, including detailed training configurations and an in-depth analysis of its performance. Our models, built upon InternVL2.5-Instruct and InternVL2.5-Pretrain, are trained in two sizes: 8B and 38B. We develop MM-EUREKA-8B upon InternVL2.5-Instruct-8B and MM-EUREKA-Zero-38B upon InternVL2.5-Pretrain-38B. We introduce the different prompt settings and training hyperparameters in Sec. 3.1, baselines in Sec. 3.2, and evaluation benchmarks in Sec. 3.3. Subsequently, we analyze



The figure is a 'Zhao Shuang Xian Tu' composed of four congruent right-angled triangles with legs of $$3$$ and $$4$$. Jack randomly throws a needle into the large square region once. The probability that the needle lands in the shaded area is ___

<think>To solve this problem, we need to calculate the area of the large square and the area of the shaded region, then find the probability that the needle lands in the shaded area. **Planning**

First, let's determine the side length of the large square. Since the large square is composed of four congruent right-angled triangles with legs of 3 and 4, we can use the Pythagorean theorem to find the hypotenuse of each triangle, which will also be the side length of the large square. The hypotenuse c can be computed as:

c = sqrt(3^2 + 4^2) = sqrt(9 + 16) = sqrt(25) = 5
Thus, the side length of the large square is 5. The area of the large square is therefore:
Area of large square = 5 * 5 = 25

Next, we need to calculate the area of the shaded region. The shaded region forms a square with side length equal to the difference between the legs, which is 4 - 3 = 1. Therefore, the area of the shaded square is:
Area of shaded region = 1 * 1 = 1

The probability that the needle lands in the shaded area is the ratio of the area of the shaded region to the area of the large square. Thus, the probability P is:
P = (Area of shaded region) / (Area of large square) = 1 / 25

However, upon re-evaluating the problem statement and the figure description, it appears there might have been an oversight. **Reflection**

The correct interpretation should consider how the triangles are arranged and overlap to form the shaded region. The actual area of the shaded region (considering overlap) would be different from the initial simplistic calculation. Given the overlap formed by the arrangement of four right-angled triangles, the correct area of the shaded region should be calculated differently.

Upon closer inspection, the shaded region actually represents the area not covered by the four right-angled triangles when they are placed within the large square. **Perception**

The total area of the four triangles is: Total area of four triangles = 4 * (1/2 * 3 * 4) = 4 * 6 = 24
Thus, the area of the shaded region is: Area of shaded region = (Area of large square) - (Total area of four triangles) = 25 - 24 = 1
Therefore, the probability that the needle lands in the shaded area remains:
P = (Area of shaded region) / (Area of large square) = 1 / 25

However, considering the arrangement and overlap more accurately leads us to realize the shaded region is actually the square remaining after the triangles are removed, which happens to be a square of side length 1. This reaffirms our initial conclusion </think> **Verification**

<answer>boxed(1/25)</answer> **Correct**

Figure 4: An example of visual aha moment. We see that our MM-Eureca reaffirms the answer by re-percepting the image.

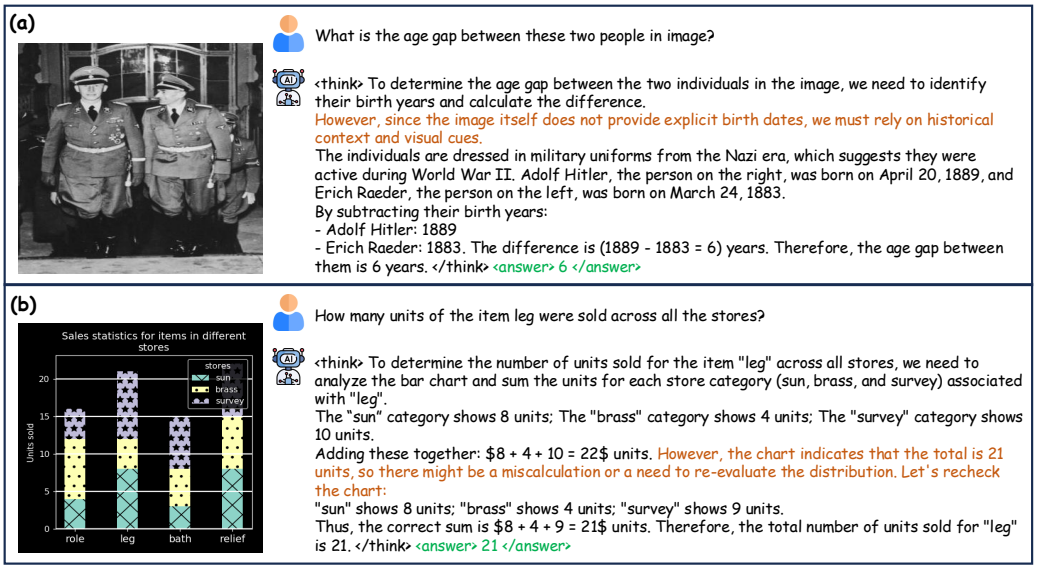


Figure 5: visual perception aha moment. In (a), the model cannot directly find the answer in the image and instead reasons by re-examining the visual cues within the image. In (b), the model detects a discrepancy between its computed result and the extracted information from the image, prompting it to re-read the image and arrive at the correct answer.

the training dynamics. Finally, we evaluate our model on multiple benchmark tests to assess its overall performance in Sec. 3.4 and Sec. 3.5.

3.1 Experiments Setup

For the instruct model and the pretrain model, we adopt different prompt strategies. For the instruct model, we retain the model's built-in system prompt and included format-related information in the user prompt. In contrast, for the base model, we follow the approach of DeepSeek-R1-Zero by providing format information within the system prompt. The different prompt settings for each model are shown in Table 1.

In our reinforcement learning setup, we assign different weights to the format reward for the two models. Since the instruct model already exhibits stronger instruction-following capabilities, we set the format reward coefficient to 0.5. In contrast, as the base model has weaker instruction-following abilities, we assign a higher coefficient of 1.0 to the format reward to encourage stricter adherence to the specified response structure.

Regarding training hyperparameters, we set the rollout batch size to 128 and the training batch size to 64, with each sample generating 8 rollouts. The temperature for model generation is set to 1, and KL divergence is not included in the loss calculation. For the 8B model, we use a learning rate of 3e-7, while for the 38B model, the learning rate is set to 5e-7.

Table 1: Prompt setting for MM-Eureka and MM-Eureka-Zero.

MM-Eureka	**SYSTEM:** {{built-in system prompt}} **USER:** <image> You should first think about the reasoning process in the mind and then provide the user with the answer. Your answer must be in latex format and wrapped in $...$.The reasoning process and answer are enclosed within <think></think>and <answer></answer>tags, respectively, i.e., <think>Since $1+1=2$, so the answer is 2.</think><answer>2</answer>, which means your output should start with <think>and end with </answer>. Question: {{question}}
MM-Eureka-Zero	**SYSTEM:** A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think></think>and <answer></answer>tags, respectively, i.e., <think>reasoning process here </think><answer>answer here </answer> **USER:** <image> Answer the following question: {{question}} Please reason step by step, and put your final answer within \boxed{}.

3.2 Baselines

To evaluate the effectiveness of rule-based RL, we compare several different post-training strategies:

- SFT: We directly use the RL data for SFT training, following the default settings of InternVL2.5 for one epoch
- COT SFT: Since our collected data does not include COT annotations (Wei et al., 2022), we test MAmmoTH-VL-8B (Guo et al., 2024) as a baseline for this post-training strategy, which uses 12M COT SFT data to fine-tune on LLaVA-OneVision-7B (Li et al., 2024) and shows strong performance on multimodal mathematical reasoning.
- MPO: As our collected data lacks negative examples, we use MMPR (Wang et al., 2024b) to conduct MPO training on InternVL2.5 with default settings as a baseline.

It is important to note that while the data used for COT SFT and MPO differs from our rule-based RL data, there is substantial overlap. Both approaches utilize original data from MathV360k (Shi et al., 2024) for construction, including ChartQA (Masry et al., 2022), GeoQA (Chen et al., 2022), and ScienceQA (Lu et al., 2022). These datasets also constitute the majority of our RL data, accounting for 76.5% of the total.

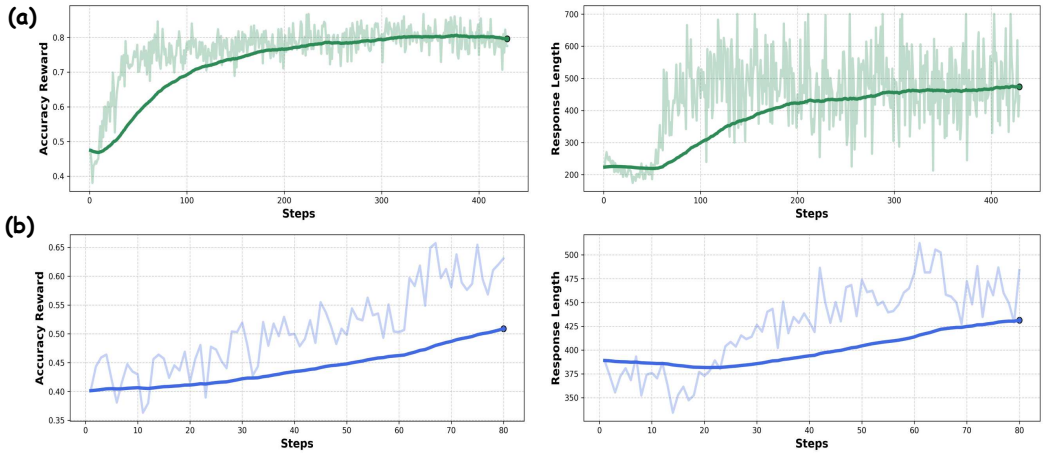


Figure 6: Train Time Scale-up on Accuracy Reward and Response Length of Rule-Based RL. (a) represents the training scenario on InternVL2.5-instruct-8B, while (b) corresponds to the training scenario on InternVL2.5-pretrained-38B. It can be observed that stable improvements in accuracy reward and response length can be achieved regardless of whether the model is based on an instruct model or a pretrained model.

3.3 BENCHMARKS

To assess the performance of our model, we conduct evaluations on multiple benchmark datasets, including MathVista(testmini) (Lu et al., 2024), MathVerse(testmini) (Zhang et al., 2024), MathVision(test) (Wang et al., 2024a), and OlympiadBench(OE_MM_maths_en_COMP) (He et al., 2024). MathVista is one of the most widely used multimodal mathematical benchmarks, offering a diverse set of problems that span general visual question answering, figure question answering, logic, algebra, and geometry. MathVerse, on the other hand, focuses specifically on the model's ability to comprehend images, with tasks categorized into areas such as algebra and geometry. MathVision takes this a step further by emphasizing more abstract visual understanding, testing the model's capacity for recognizing and reasoning beyond conventional mathematical contexts. OlympiadBench presents graduate-level mathematical competition problems, from which we select the multimodal, open-ended English mathematical questions for evaluation.

Beyond these established benchmarks, we also introduce a new **K12 Math** dataset, which consists of 500 fill-in-the-blank mathematics questions spanning from middle to high school levels. Unlike OlympiadBench, which is designed to assess advanced graduate-level mathematical reasoning, this dataset focuses on evaluating the model's ability to solve fundamental mathematical problems commonly encountered in K12 education.

During evaluation, we adopt greedy decoding with a temperature of 0, ensuring deterministic outputs. We do not use beam search, and both top-p and top-k sampling are disabled.

3.4 RL FROM INSTRUCT MODEL

Building upon InternVL-8B-Instruct, we apply rule-based RL and observe a synchronized increase in both accuracy reward and response length as shown in Figure 6 (a). Throughout the training, the model not only learned to allocate more tokens to interpreting the problem but also dedicated more effort to reflecting on and correcting its own mistakes. Both aspects played a crucial role in enhancing the model's reasoning ability and overall performance.

Besides, with only 54K training samples, our model shows improvements across all benchmarks compared to the instruct model. As shown in Table 2, compared to other post-training strategies, it outperforms the MPO training method that uses 1M data samples; and achieves comparable performance on multimodal mathematical reasoning to models trained with 12M COT SFT data. This highlights the simplicity and effectiveness of rule-based RL, as well as its data efficiency.

Table 2: Performance comparison on various multimodal math benchmarks. MM-Eureka-8B trained with 50k data on InternVL2.5-Instruct-8B outperforms the model trained with 1M data using MPO and achieves performance comparable to MAmmoTH-VL-8B after SFT with 12M COT data.

Model	Data Scale	MathVista	MathVerse	MathVision	Olympiad.	K12	Avg.
InternVL2.5-8B-Instruct	-	64.4	39.5	19.7	8.0	24.8	31.3
+ SFT	54k	54.5	28.9	17.7	6.6	20.6	25.5
+ MPO	1M	66.3	33.3	20.8	10.0	19.4	30.0
MAmmoTH-VL-8B	12M	**67.6**	35.0	**24.4**	**10.0**	24.8	32.4
MM-Eureka-8B	54k	67.1	**40.4**	22.2	8.6	**27.0**	**33.0**

Table 3: Performance comparison on different multimodal mathematical benchmarks. MM-Eureca-Zero-38B, trained with only 8k data on InternVL2.5-pretrained-38B, surpasses the instruct model trained with 16.3M data via SFT on the Olympiadbench and K12, while achieving comparable performance on MathVerse.

Model	Data Scale	MathVista	MathVerse	MathVision	Olympiad.	K12	Avg.
InternVL2.5-38B-Pretrained	-	60.9	44.8	27.6	28.7	38.2	40.0
InternVL2.5-38B-Instruct	16.3M	**71.9**	**49.4**	**31.8**	29.3	37.2	43.9
MM-Eureka-Zero-38B	9.3k	64.2	48.9	26.6	**37.3**	**46.4**	**44.7**

3.5 RL From Pretrained Model

In addition to our experiments on the instruct model, we also conduct rule-based RL on InternVL-Pretrain models with 8B and 38B, using only 9.3K K-12 data samples selected through the process described in Section 2.2 due to resource constraints. For the 8B model, its limited model size prevents stable training, as discussed in Section 4.3. Here, we focus on the results of the 38B model. As shown in Figure 6 (b), despite the limited data, the training process exhibited a clear trend. In the early stages, the model experienced a slight decrease in response length, likely as it prioritized learning the response format. As training progressed, both response length and reasoning depth increased synchronously, mirroring the pattern observed in the instruct models' experiments.

Remarkably, as shown in Table 3, despite the small training dataset, RL training on the 38B pretrain model resulted in a significant 8.2% improvement on the K-12 benchmark, along with measurable performance gains across other evaluation benchmarks. These results highlight the efficiency of our RL approach in leveraging minimal data to achieve substantial enhancements in model performance, while also demonstrating the existence of a "Zero moment" (Hu et al., 2025) in the field of multimodal reasoning.

4 Discussion

In this section, we discuss methods that we expected to be effective but failed in our experiments. This does not necessarily mean these approaches lack merit, as several works have demonstrated their effectiveness (Team et al., 2025; Cui et al., 2025). We share our attempts and hope the community can refine these strategies.

4.1 Curriculum Learning

Thanks to the difficulty levels labeled in Section 2.2, we sort the data by difficulty to conduct curriculum learning RL experiments using K12 data. As shown in Figure 7, accuracy reward using curriculum learning fails to achieve stable growth. Despite the intuition that curriculum learning allows models to learn gradually, we find no advantage compared to direct training (e.g., shuffle). We believe this simple curriculum learning setup may cause the model to fixate on simple problems in the early and middle stages, hindering exploration of difficult problems and preventing accuracy improvements in later stages.

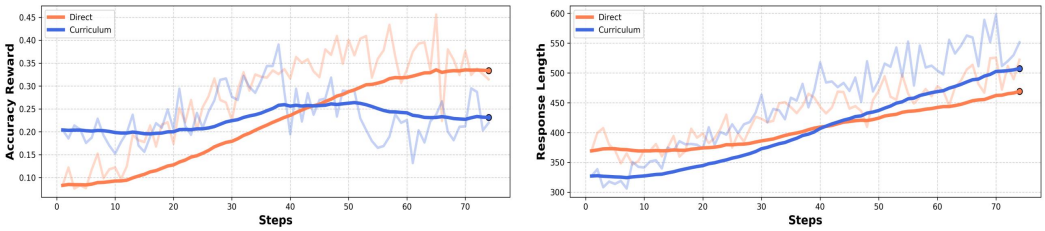


Figure 7: Comparison of training with Curriculum Learning RL and Direct RL. Although both show an increasing trend in response length, the accuracy reward for Curriculum Learning struggles to improve in the later stages.

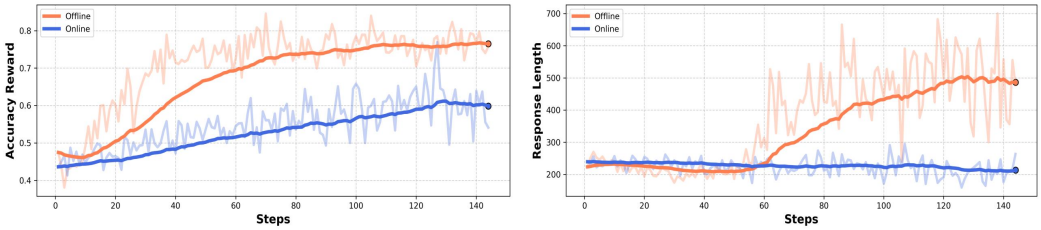


Figure 8: Comparison of training with Offline Data Filter strategy and Online Data Filter strategy. Although both achieve stable increases in accuracy reward, the peak value reached with Online Data Filter is lower. Additionally, the response length with Online Data Filter fails to show an upward trend.

4.2 Online Data Filter

We refer to the pre-difficulty-based data filtering strategy proposed in Section 2.2 as the Offline Data Filter. Although it stabilizes RL training, this approach wastes a portion of the available data. To improve data utilization efficiency, we experiment with an Online Data Filter approach similar to PRIME (Cui et al., 2025). By selecting only prompts that are neither completely correct nor completely incorrect for each training iteration, we enhance training stability; additionally, this online filtering method dynamically selects different data as the model improves. However, as shown in Figure 8, we find that this training approach struggles to achieve significant improvements in accuracy reward or response length. We attribute this limitation to the varying batch sizes used for updating in each training round, which likely leads to gradient instability.

4.3 Model Size

Despite some works successfully reproducing R1-Zero scenarios with small models in pure language settings (Hu et al., 2025; Zeng et al., 2025), we find that in multimodal mathematical reasoning scenarios, small models (e.g., 8B) struggle to maintain stable rule-based RL training compared to larger models (e.g., 38B). As shown in Figure 9, when training InternVL2.5-pretrained-8B using the same data and training strategies as in Section 3.5, its accuracy reward struggles to increase, and the response length also shows a fluctuating trend. However, as shown in the results of Section 3.5, increasing the model size to 38B under the same data conditions enables stable training and consistent performance improvements. Therefore, how to reproduce the R1-Zero moment with small models in the multimodal reasoning domain remains an issue that requires further exploration.

5 Conclusion

We explore methods for adapting DeepSeek R1 to multimodal reasoning scenarios. Through a minimalist RL setup (e.g., RLOO algorithm without KL divergence penalty) combined with difficulty-based data filtering strategies, we successfully achieved stable improvements in accuracy rewards and increased response length on InternVL2.5-8B-Instruct and InternVL2.5-38B-Pretrained mod-

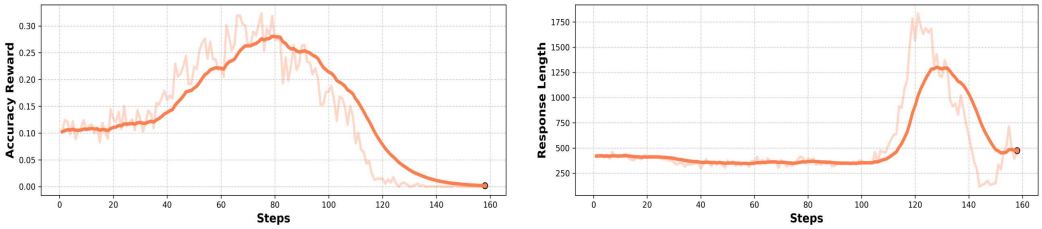


Figure 9: Visualization of rule-based RL training using InternVL-2.5-pretrained-8B directly on 8K K12 data. The results show that both accuracy reward and response length exhibit fluctuating trends, making stable growth difficult to achieve.

els, while also observing visual "aha-moments". We share our complete training pipeline and unsuccessful attempts with the community, aiming to collaboratively advance the field of multimodal reasoning through open-source contributions.

6 ACKNOWLEDGEMENTS

We acknowledge the outstanding open-source contributions from vLLM, LMM-R1 and vLLM. We also extend our gratitude to DeepSeek-R1 and InternVL for their open-source techniques and base models, which have enabled us to further our exploration.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pp. 12248–12267. Association for Computational Linguistics, 2024.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning, 2022. URL https://arxiv.org/abs/2105.14517.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024b. URL https://arxiv.org/abs/2312.14238.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards, 2025. URL https://arxiv.org/abs/2502.01456.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022. URL https://arxiv.org/abs/2210.10760.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale, 2024. URL https://arxiv.org/abs/2412.05237.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL https://arxiv.org/abs/2402.14008.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework, 2024. URL https://arxiv.org/abs/2405.11143.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero, 2025.

Shengyi Huang, Anssi Kanervisto, Antonin Raffin, Weixun Wang, Santiago Ontañón, and Rousslan Fernand Julien Dossa. A2c is a special case of ppo, 2022. URL https://arxiv.org/abs/2205.09123.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5648–5656, 2018.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251. Springer, 2016.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *Deep Reinforcement Learning Meets Structured Prediction, ICLR 2019 Workshop*. OpenReview.net, 2019.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*, 2025.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL https://arxiv.org/abs/2310.02255.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. URL https://arxiv.org/abs/2203.10244.

Fanqing Meng and Lingxiao Du. R1-multimodal-journey. https://github.com/FanqingM/R1-Multimodal-Journey, 2025. Accessed: 2025-02-12.

OpenAI. Introducing openai o1. https://openai.com/o1/, 2024. Accessed: 2024-10-02.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

YingZhe Peng, Gongrui Zhang, Xin Geng, and Xu Yang. Lmm-r1. https://github.com/TideDra/lmm-r1, 2025. Accessed: 2025-02-13.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models, 2024. URL https://arxiv.org/abs/2406.17294.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL https://arxiv.org/abs/2501.12599.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024a. URL https://arxiv.org/abs/2402.14804.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization, 2024b. URL https://arxiv.org/abs/2411.10442.

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

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint, 2024. URL https://arxiv.org/abs/2312.11456.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient. https://hkust-nlp.notion.site/simplerl-reason, 2025. Notion Blog.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024. URL https://arxiv.org/abs/2403.14624.