# MMR1: Enhancing Multimodal Reasoning with Variance-Aware Sampling

**Sicong Leng**[1,2,*] **Jing Wang**[1,*] **Jiaxi Li**[3,*] **Hao Zhang**[2,*] **Zhiqiang Hu**[2]
**Boqiang Zhang**[2] **Yuming Jiang**[2] **Hang Zhang**[2] **Xin Li**[2]
**Deli Zhao**[2] **Wei Lu**[1] **Yu Rong**[2] **Aixin Sun**[1,†] **Shijian Lu**[1,†]

[1]Nanyang Technological University [2]DAMO Academy, Alibaba Group
[3]Singapore University of Technology and Design

## Abstract

*Large multimodal reasoning models have achieved rapid progress, but their advancement is constrained by two major limitations: the absence of open, large-scale, high-quality long chain-of-thought (CoT) data, and the instability of reinforcement learning (RL) algorithms in post-training. Group Relative Policy Optimization (GRPO), the standard framework for RL fine-tuning, is prone to gradient vanishing when reward variance is low, which weakens optimization signals and impairs convergence. This work makes three contributions: (1) We propose Variance-Aware Sampling (VAS), a data selection strategy guided by Variance Promotion Score (VPS) that combines outcome variance and trajectory diversity to promote reward variance and stabilize policy optimization. (2) We release large-scale, carefully curated resources containing ∼1.6M long CoT cold-start data and ∼15k RL QA pairs, designed to ensure quality, difficulty, and diversity, along with a fully reproducible end-to-end training codebase. (3) We open-source a family of multimodal reasoning models in multiple scales, establishing standardized baselines for the community. Experiments across mathematical and logical reasoning benchmarks demonstrate the effectiveness of both the curated data and the proposed VAS. Comprehensive ablation studies and analyses provide further insight into the contributions of each component. In addition, we theoretically establish that reward variance lower-bounds the expected policy gradient magnitude, with VAS serving as a practical mechanism to realize this guarantee.*

## 1. Introduction

Recent advances in large language and multimodal reasoning models have markedly improved performance on complex tasks such as mathematics and science domains. Reinforcement learning (RL) plays a central role in these developments by optimizing models with process- or outcome-based rewards [32, 33, 72]. Group Relative Policy Optimization (GRPO; Shao et al. [59]) has emerged as a widely adopted RL framework due to its efficiency and scalability, and has been successfully applied to both language models [12] and multimodal models [46, 62, 68, 77]. However, GRPO is inherently susceptible to *gradient vanishing*: when sampled rewards have low variance, relative advantages collapse toward zero, weakening optimization signals and destabilizing training [55, 56]. This issue persists across both unimodal and multimodal contexts, posing a fundamental challenge to effective RL optimization.

In parallel, the progress of multimodal reasoning research is influenced by the limited availability of open, large-scale, high-quality long chain-of-thought (CoT) data. Compared with text-only reasoning, where multiple datasets are publicly accessible [11, 49], multimodal training often relies on more restricted resources, which may constrain reproducibility and further development. Recent studies have made progress by exploring heuristic data curation [5, 19, 46], reward design modifications [60, 62], and training adjustments [9, 84]. While these approaches improve downstream performance, challenges related to stable GRPO optimization and the broader availability of curated multimodal reasoning data remain underexplored.

In this work, we introduce **V**ariance-**A**ware **S**ampling (**VAS**), a dynamic data selection strategy to mitigate gradient vanishing in GRPO-based training for multimodal reasoning models. Our approach is grounded in the theoretical insight that *reward variance provides a lower bound on the expected policy gradient magnitude*. Increasing reward variance, therefore, offers a principled means to stabilize training and strengthen policy optimization. Specifically, VAS employs the **V**ariance **P**romotion **S**core (**VPS**), which evaluates each prompt's potential to induce reward variance. VPS consists of two complementary components: the **O**utcome **V**ariance **S**core (**OVS**), which favors prompts yielding a balanced mix of correct and incorrect responses to maxi-

[†]Correspondence [*]Equal Contributions

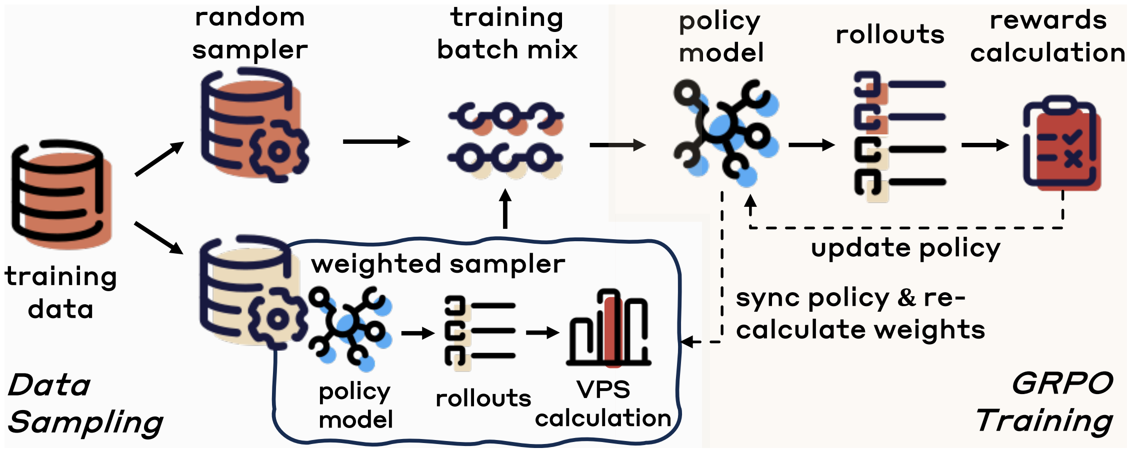

Figure 1. **Overview of the Variance-Aware Sampling (VAS) framework.**

mize expected reward variance, and the **T**rajectory **D**iversity **S**core (**TDS**), which encourages diversity among reasoning trajectories, thereby raising the lower bound of variance and sustaining informative gradient signals even under sparse or noisy correctness feedback. Depicted in Figure 1, VAS constructs each training batch from two subsets: one sampled with probabilities proportional to VPS, emphasizing prompts with higher potential to induce reward variance, and another drawn uniformly at random to maintain broad data coverage. This design aims to balance targeted promotion of reward variance with general exposure to the training distribution. By introducing outcome- and trajectory-level variability into GRPO's group-based comparisons, VAS reduces the risk of weak gradients and contributes to more stable optimization. A theoretical analysis is presented in §4.

We validate VAS on a range of multimodal mathematical and logical reasoning benchmarks. Experiments show that VAS improves convergence, stability, and downstream performance. Ablation studies further demonstrate that OVS and TDS contribute complementary benefits: OVS enhances expected reward variance by balancing outcomes, while TDS increases trajectory diversity to support more consistent gradient updates. Beyond methodology, we curate and release large-scale datasets for both supervised fine-tuning and RL. The supervised dataset ($\sim$1.6M) emphasizes long chain-of-thought reasoning paired with verified short answers, while the RL dataset ($\sim$15k) is constructed to capture diverse levels of difficulty and domain coverage. Both datasets are curated with explicit attention to quality, difficulty, and diversity, ensuring their value for multimodal reasoning research. Together with these resources, we provide a reproducible codebase and open models at multiple scales, offering standardized baselines for future research.

## 2. Related Work

Building on the success of rule-based RL [12, 25], recent work explores RL with verifiable rewards, following a pipeline that conducts optional SFT activation then applies RL [1, 57], *e.g.*, GRPO [59], with fine-grained training recipes. In the multimodal domain, various approaches refine this pipeline through specific reward design [60, 62], sample diversification [46, 68], hyperparameter tuning [19, 62, 78], and advanced RL strategies [9, 53, 84]. Nevertheless, they often overlook the gradient vanishing problem inherent in GRPO-based training, resulting in unstable optimization and slow convergence [56]. Some studies attempt to alleviate it by filtering samples with moderate pass rates [46, 68], yet they remain largely heuristic and lack comprehensive experimental validation or theoretical grounding.

Recent studies have examined gradient vanishing in a principled manner, analyzing how training objectives degrade under reward sparsity, variance reduction, and optimization dynamics [16, 36, 65, 90]. A range of remedies has been proposed, including reward rescaling [17, 31], entropy regularization [35], and improved sample selection [30, 73, 85], which primarily operate by adjusting RL algorithms or reward mechanisms. In contrast, our work takes an orthogonal perspective by mitigating gradient vanishing through data sampling during training. Supported by both theoretical grounding and extensive empirical validation, our approach complements existing GRPO variants [16, 35, 66, 81] and can be naturally combined with them to further improve training stability and effectiveness in reasoning.

## 3. Variance-Aware Sampling Framework

This section details the Variance-Aware Sampling (VAS) framework. §3.1 defines the Variance Promotion Score (VPS), comprising Outcome Variance Score (OVS) and Trajectory Diversity Score (TDS), which guides dynamic data sampling. §3.2 then outlines the sampler implementation, including VPS updates and sample selection during training.

### 3.1. Variance Promotion Score

Let $x$ be a prompt with ground-truth answer $\bar{y}$ from the train set. In GRPO framework, the model generates $N$ responses $\{y_i\}_{i=1}^N$ for each $x$. A task-specific verifier $\mathrm{V}(x, y_i, \bar{y}) \in \{0, 1\}$ evaluates each response, returning 1 if $y_i$ matches $\bar{y}$ and 0 otherwise. The **pass rate** of $x$ is defined as:

$$\mathrm{P}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{V}(x, y_i, \bar{y}).$$

Inspired by Foster and Foerster [10], we directly calculate the Outcome Variance Score (OVS) as:

$$\mathrm{OVS}(x) = \mathrm{P}(x)(1 - \mathrm{P}(x)),$$

which corresponds to the Bernoulli variance of correctness across responses. It is maximized at $\mathrm{P}(x) = 0.5$, where correct and incorrect outputs are balanced. We further define the Trajectory Diversity Score (TDS) to characterize variability in reasoning processes. Let $\mathrm{Diversity}(\{y_i\}_{i=1}^N)$ be a diversity function over sequences (*e.g.*, inverse self-BLEU or distinct-n):

$$\mathrm{TDS}(x) = \mathrm{Diversity}(\{y_i\}_{i=1}^N),$$

where a higher value reflects greater diversity among sampled trajectories. The Variance Promotion Score (VPS) is computed as a weighted combination of OVS and TDS:

$$\mathrm{VPS}(x) = \alpha \cdot \mathrm{OVS}(x) + \beta \cdot \mathrm{TDS}(x),$$

where $\alpha, \beta > 0$ balance their contributions. OVS increases the expected reward variance, while TDS provides a lower bound by encouraging trajectory diversity. Together, they are intended to strengthen the magnitude and consistency of gradient signals in GRPO training.

### 3.2. Dynamic Sampler

The dynamic sampler prioritizes prompts with higher VPS, which are expected to induce greater reward variance during training. At each sampling step, the training batch is constructed from two subsets: one from a VPS-based weighted sampler and another from a random sampler. A mix ratio hyperparameter $\lambda \in [0, 1]$ controls the proportion of samples drawn from each source, with $\lambda$ specifying the fraction of the batch selected from the weighted sampler. VPS scores are periodically updated to reflect changes in the policy. After every $T_{\text{update}}$ steps, $N$ responses are resampled for each prompt, and OVS and TDS are recomputed. The update interval $T_{\text{update}}$ trades off computational cost against adaptation speed. Algorithm 1 summarizes the VAS procedure.

**Algorithm 1** Variance-Aware Sampling (VAS) for GRPO

**Require:** Dataset $\mathcal{D}$; batch size $B$; rollouts per prompt $N$; VPS update interval $T_{\text{update}}$; mix ratio $\lambda \in [0, 1]$

**Initialize** policy parameters $\theta$
**for** each prompt $x \in \mathcal{D}$ **do** ▷ Initial VPS estimation
  Sample $N$ rollouts $\{y_i\}_{i=1}^N$ from $\pi_\theta(\cdot \mid x)$
  Compute pass rate $P(x)$, $\mathrm{OVS}(x)$, $\mathrm{TDS}(x)$, and $\mathrm{VPS}(x)$
**end for**
**for** training step $t = 1, \ldots, T$ **do**
  **if** $t \bmod T_{\text{update}} = 0$ **then** ▷ Periodic VPS refresh
    **for** each prompt $x \in \mathcal{D}$ **do**
      Sample $N$ rollouts $\{y_i\}_{i=1}^N$; update $P(x)$, $\mathrm{OVS}(x)$, $\mathrm{TDS}(x)$, $\mathrm{VPS}(x)$
    **end for**
  **end if**
  $B_{\text{w}} \leftarrow \lfloor \lambda B \rfloor$, $B_{\text{r}} \leftarrow B - B_{\text{w}}$
  Sample $B_{\text{w}}$ prompts from $\mathcal{D}$ *with replacement* proportional to $\mathrm{VPS}(\cdot)$
  Sample $B_{\text{r}}$ prompts from $\mathcal{D}$ uniformly at random
  $\mathcal{B} \leftarrow$ union of the two sets ▷ Construct training batch
  **for** each $x \in \mathcal{B}$ **do**
    Sample $N$ rollouts $\{y_i\}_{i=1}^N$ from $\pi_\theta(\cdot \mid x)$
    Compute GRPO loss $\mathcal{L}_{\text{GRPO}}(x, \{y_i\})$
  **end for**
  Update $\theta \leftarrow \theta - \eta \nabla_\theta \sum_{x \in \mathcal{B}} \mathcal{L}_{\text{GRPO}}(x, \{y_i\})$
**end for**

## 4. Theory

This section formalizes the intuition that prompts with higher *reward variance* yield more informative policy-gradient updates. We first establish a *Variance–Progress Theorem* for the vanilla REINFORCE algorithm [75], showing that expected improvement is linearly lower-bounded by the reward variance of the prompt. We then present a two-level decomposition of reward variance that directly motivates the *Outcome Variance Score* and *Trajectory Diversity Score* in our method. Finally, we extend the analysis to GRPO, with complete proofs and derivations provided in Supp. A.

### 4.1. Preliminaries

Let $x \in \mathcal{X}$ be a prompt and $y \in \mathcal{Y}$ a response from the policy $\pi_\theta(y \mid x)$. A reward model $R : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$ assigns a scalar reward. The optimization objective is defined as:

$$J(\theta) = \mathbb{E}_{x \sim \mathcal{D}}\, \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} [R(x, y)], \tag{1}$$

where $\mathcal{D}$ is the prompt distribution. Here, we omit the KL regularization, as it only rescales constants and does not affect the variance argument.

**Score-function identity.** For a fixed prompt $x$, the gradient can be expressed as:

$$\begin{aligned} \nabla_\theta J_x(\theta) &= \mathbb{E}_{y \sim \pi_\theta}\big[R(x, y)\, g(x, y)\big], \\ g(x, y) &= \nabla_\theta \log \pi_\theta(y \mid x), \end{aligned} \tag{2}$$

where $g(x, y)$ is the score function and $\mathbb{E}_y[g(x, y)] = 0$.

**Baselines and variance.** Subtracting a prompt-dependent baseline $b(x)$ keeps the estimator unbiased:

$$\begin{aligned} G(x) &= \mathbb{E}_y[g(x, y)(R(x, y) - b(x))], \\ \mathbb{E}[G(x)] &= \nabla_\theta J_x(\theta). \end{aligned} \tag{3}$$

The variance is minimized when $b^\star(x) = \mathbb{E}_y[R(x, y)]$. Under this choice, the variance factorizes as (Lemma A.1):

$$\mathrm{Var}[G(x)] = \mathrm{Var}_y[R(x, y)]\,\Gamma_\theta(x), \tag{4}$$

where $\Gamma_\theta(x) = \mathbb{E}_y\|g(x, y)\|^2$ is a Fisher-information term depending on the policy but not on the rewards. Proposition A.2 shows $\Gamma_\theta(x)$ is bounded above and below by model-dependent constants. Thus, the *reward variance* is the only prompt-dependent factor controlling the dispersion of stochastic gradients.

### 4.2. Variance Progress Theorem for REINFORCE

Consider a single gradient step $\theta^+ = \theta + \eta G(x)$ on prompt $x$, with learning rate $\eta > 0$.

**Assumption 1** (Smoothness and gradient lower bound)**.** *For each prompt $x$: (i) $J_x(\theta)$ is twice differentiable with $\|\nabla^2_\theta J_x(\theta)\| \le L$ for some $L > 0$; (ii) There exists $c_{\min} > 0$ such that $\|\nabla_\theta J_x(\theta)\|^2 \ge c_{\min}\ \mathrm{Var}_y[R(x, y)]$.*

The second condition links reward variance to the squared gradient norm. It follows from the positive definiteness of the Fisher information matrix under mild regularity conditions.

**Theorem 1** (Variance–Progress)**.** *Let Assumption 1 hold and use the optimal baseline $b^\star(x)$. Then for any step size $0 < \eta \le \frac{c_{\min}}{4L\Gamma_\theta(x)}$, the expected one-step gain satisfies*

$$\mathbb{E}\big[J_x(\theta^+) - J_x(\theta)\big] \;\ge\; \frac{\eta\, c_{\min}}{4}\ \mathrm{Var}_y[R(x, y)]. \tag{5}$$

**Intuition**

Prompts with higher reward variance yield stronger gradient signals. When all rollouts have similar rewards, gradients vanish; diverse outcomes or reasoning paths sustain learning progress.

**Sketch of proof.** A second-order Taylor expansion yields $J_x(\theta^+) = J_x(\theta) + \langle \nabla J_x, G(x)\rangle + \frac{1}{2}G^\top HG$. Taking expectations and using unbiasedness gives a linear term $\eta\|\nabla J_x\|^2$. The quadratic remainder is bounded by $\frac{1}{2}\eta^2 L\,\mathbb{E}\|G\|^2$. Substituting the variance factorization and the gradient lower bound, and restricting $\eta$ as stated, ensures the remainder is at most half the linear term, giving the result. The full details of the proof are provided in Supp. A.2.

### 4.3. Variance Decomposition

For binary rewards, the variance admits a natural two-level decomposition. Write each rollout $y = (y_{\mathrm{cot}}, y_{\mathrm{ans}})$ as a reasoning chain and its final answer, and let the verifier assign $R(x, y) = \mathbf{1}_{\text{correct}(y_{\mathrm{ans}})}$. Define $Z = \varphi(y_{\mathrm{cot}})$ as a representation of the reasoning chain and $p_Z(x) = \Pr(R = 1 \mid Z)$. The law of total variance gives

$$\mathrm{Var}_y[R(x, y)] = \underbrace{\mathbb{E}_Z[p_Z(x)(1 - p_Z(x))]}_{\text{intra-trajectory}} + \underbrace{\mathrm{Var}_Z[p_Z(x)]}_{\text{inter-trajectory}}. \tag{6}$$

The first term corresponds to variability in correctness given a fixed reasoning path, and is estimated by $\hat{p}(x)(1 - \hat{p}(x))$. This motivates the *Outcome Variance Score (OVS)*. The second term measures variation across reasoning paths, which can be lower-bounded by diversity metrics such as normalized edit distance or self-BLEU dispersion (detailed in Supp. A). This motivates the *Trajectory Diversity Score (TDS)*. Together, OVS and TDS provide complementary mechanisms to raise reward variance, directly connecting the theory to our method in §3.

### 4.4. From REINFORCE to GRPO

GRPO extends REINFORCE by normalizing rewards within each group of rollouts. For prompt $x$, the group mean reward $\bar{r}(x)$ serves as a baseline, and the sample standard deviation $s(x)$ whitens centered rewards:

$$\tilde{R}(x, y_i) = \frac{R(x, y_i) - \bar{r}(x)}{s(x) + \delta}.$$

The gradient estimator then multiplies $\tilde{R}$ by the importance ratio $r_\theta(y_i \mid x)$, optionally clipped to $1 \pm \varepsilon$ to control KL divergence. Without clipping, the estimator remains unbiased, and Theorem 1 applies with a rescaled learning rate bound reflecting the whitening factor. With clipping, the estimator acquires a bias of order $O(\varepsilon)$, which reduces but does not eliminate the lower bound. Thus, under both settings, prompts with higher reward variance continue to guarantee larger provable minimum improvements per update. Details and finite-sample corrections are provided in Supp. A.4.

## 5. Data Curation

Following prior work [46, 62, 68], our training pipeline includes two stages: a supervised cold-start stage followed by a reinforcement learning stage using GRPO, with each stage supported by meticulously curated datasets.

### 5.1. Cold-start Data

This stage utilizes long CoT data for fine-tuning. Summarized in Table 1, we collect QA pairs from diverse corpora, which primarily feature short answers or short CoTs. To address the limited coverage, particularly in science, in existing multimodal data, we supplement with publicly available practice problems and exams from biology, chemistry, geography, and physics. The final dataset spans five domain categories: *Math*, *General*, *Chart-Figure*, *Doc-Table*, and *Science*. To ensure balanced quality and difficulty, we apply filtering based on response correctness. For each question, multiple responses are generated via Qwen2.5-VL-72B [2] and verified by GPT-4o [50]. Samples are categorized by *pass rate*: easy ($\ge 0.8$), hard ($\le 0.2$), or medium (otherwise). Only medium and hard cases are retained. Long CoT annotations are then produced using Gemini 2.5 Pro/Flash [7], which

| Domain | # Samples | Datasets |
|---|---|---|
| Math | 1.6M | MM-MathInstruct [71], MathV360K [61], MultiMath [52], MATH [13], MathVision [69], CLEVR-Math [34], IconQA [38], MAVIS-Instruct [87], RCoT [8], GeoQA+ [4], Geometry3K [37], GeomVerse [23], Self-collected data |
| Science | 13K | ScienceQA [39], AI2D [24], CLEVR [21], Self-collected data |
| Chart-Figure | 8K | ChartQA [43], SPIQA [54], DVQA [22], PlotQA [48] |
| Doc-Table | 8K | TableMWP [40], InfoVQA [45], DocVQA [44], WikiTableQuestions [51], VisualMRC [63] |
| General | 8K | Sherlock [14], A-OKVQA [58], PISC [29], GQA [20] |

Table 1. Cold-start Data Statistics.

generates multi-step rationales followed by final answers. Annotations are preserved only when final answer matches ground truth as validated by GPT-4o.

## 5.2. RL Data

This stage employs prompts annotated with *concise*, *verifiable* short answers suitable for reward modeling. Specifically, we adopt GPT-4o to extract and rephrase open-form short answers from the original CoT annotations. The RL dataset integrates **two** complementary components. First, we select 8k math problems from the cold-start stage, retaining only hard-level items with low pass rates to ensure challenging supervision. These problems are further categorized into fine-grained types with GPT-4o (detailed in Supp. B) and uniformly sampled to ensure balanced coverage across categories. Second, we add 7k logical reasoning problems from Raven [82], MM-IQ [3], and EasyArc [64], curated to incentivize general reasoning ability beyond math. Together, these components form a 15k RL dataset emphasizing difficulty, diversity, and balanced coverage across mathematical and logical reasoning tasks.

# 6. Experiments

## 6.1. Experimental Setups

**Implementations.** In the cold-start stage, we fine-tune Qwen2.5-VL-Instruct [2] with curated long CoT data (§5.1) using the LLaMA-Factory framework [88]. Training runs for 5 epochs with AdamW, a cosine schedule, an initial learning rate of $1 \times 10^{-5}$, and a 0.1 warm-up ratio. The checkpoint with the best validation score is retained. This checkpoint initializes the policy for RL training, implemented with the EasyR1 codebase [89]. VAS (Algorithm 1) is set to $N = 32$, $\lambda = 0.5$, $\alpha = 0.8$, $\beta = 0.2$, and $T_{\text{update}} = 35$[1]. Additional hyperparameters for both cold-start and RL training are detailed in Supp. C.

**Benchmarks.** To evaluate MMR1, we adopt five widely used and challenging benchmarks focusing on mathematical and logical reasoning: MathVerse [86], MathVista [41], MathVision [70], LogicVista [76], and ChartQA [42]. These benchmarks collectively assess diverse aspects of problem-solving, including complex multi-step mathematics, visual reasoning, logical deduction, and chart-based understanding.

**Baselines.** In this work, we compare MMR1 against a broad set of MLLMs, covering both general-purpose and reasoning-oriented designs of comparable model size. *General-purpose MLLMs*: Qwen2.5-VL-Instruct-7B [2], InternVL2.5-8B [6], InternVL3-8B [91], and LLaVA-OneVision-7B (LLaVA-OV; Li et al. [28]), representing the recent state-of-the-art general-purpose MLLMs. *Reasoning-oriented MLLMs*: VL-Cogito-7B [80], MM-Eureka-8B [47], R1-VL-7B [83], R1-OneVision-7B [79], OpenVLThinker-7B [9], Vision-R1-7B [18], VL-Rethinker [67], and ThinkLite-VL-7B [74].

**Evaluation.** We adopt a unified prompt across all evaluations, requiring models to enclose final answers in "\box{}" (full prompt in Supp. D). Inference is performed using vLLM [26] for efficient generation. For benchmarks with official protocols (*e.g.*, MathVision, MMMU), we follow the original procedures. For others, math questions are assessed with Math-Verify [27] and MathRuler [15], while non-math ones use exact matching. To ensure robustness, we further 1) select the most semantically similar option when multiple-choice answers do not exactly match any candidate, and 2) employ GPT-4o [50] as an auxiliary judge for open-ended questions where exact matching or extraction fails.

## 6.2. Main Results

Reported in Table 2, our 7B model achieves state-of-the-art performance among reasoning-oriented MLLMs, reaching an average score of **58.4**, the highest across all evaluated models. It ranks first on most benchmarks, including **MathVerse** (55.4), **MathVision** (31.8), **LogicVista** (48.9), and **ChartQA** (83.7), while delivering competitive results on **MathVista** (72.0). Compared to general models with similar scales, our approach consistently yields superior results.

In addition, the 3B variant of our model demonstrates strong competitiveness with an average of 52.7. Despite its smaller scale, it matches or surpasses several 7B models (e.g., R1-VL at 47.7 and R1-OneVision at 47.8), underscoring the efficiency of our framework in resource-constrained settings.

These results highlight the complementary contributions of our pipeline: carefully curated long CoT supervision for cold-start initialization, RL to incentivize deeper reasoning, and VAS to stabilize optimization. Together, they enable consistent improvements in reasoning performance across both small- and large-scale models.

[1]Due to the resource constraint, experiments in ablation studies and analysis adopt $N = 8$, $\lambda = 0.5$, $\alpha = 0.5$, $\beta = 0.5$, and $T_{\text{update}} = 28$ unless otherwise specified.

| Model | Size | MathVerse | MathVista | MathVision | LogicVista | ChartQA | Average |
|---|---|---|---|---|---|---|---|
| | | | General-Purpose Models | | | | |
| Qwen2.5-VL | 7B | 50.4 (49.2) | 69.3 (68.2) | 28.7 (25.1) | 44.0 | 82.4 | 55.0 |
| InternVL2.5 | 8B | 40.0 (39.5) | 61.4 (64.4) | 19.9 (19.7) | 37.7 (36.0) | 73.4 | 46.5 |
| InternVL3 | 8B | 49.4 (39.8) | 68.5 (71.6) | 30.0 (29.3) | 41.3 (44.1) | 81.3 | 54.1 |
| LLaVA-OV | 7B | 33.6 (26.2) | 56.4 (63.2) | 15.9 | 30.6 | 65.0 | 40.3 |
| | | | Reasoning-Oriented Models | | | | |
| MM-Eureka | 8B | 52.3 (50.3) | 73.4 (73.0) | 29.4 (26.9) | <u>47.1</u> | 82.7 | 57.0 |
| R1-VL | 7B | 41.3 (40.0) | 61.5 (63.5) | 23.0 (24.7) | 36.3 | 76.3 | 47.7 |
| R1-OneVision | 7B | 44.0 (46.4) | 60.3 (64.1) | 22.0 (29.9) | 40.0 | 72.5 | 47.8 |
| OpenVLThinker | 7B | 48.1 (47.9) | 70.6 (70.2) | 22.0 (25.3) | 41.0 | 81.0 | 52.5 |
| VL-Rethinker | 7B | <u>54.6</u> (54.2) | <u>73.7</u> (74.9) | 30.1 (32.3) | 45.7 | <u>83.5</u> | 57.5 |
| Vision-R1 | 7B | 51.9 (52.4) | 72.1 (73.5) | – | 44.7 | 82.7 | – |
| ThinkLite-VL | 7B | 51.3 (50.7) | 72.5 (75.1) | 27.5 | 44.3 | 83.1 | 55.7 |
| VL-Cogito | 7B | 54.3 | **74.8** | <u>30.7</u> | **48.9** | 83.4 | <u>58.2</u> |
| MMR1 | 3B | 47.9 | 67.1 | 25.4 | 42.0 | 81.2 | 52.7 |
| MMR1 | 7B | **55.4** | 72.0 | **31.8** | **48.9** | **83.7** | **58.4** |

Table 2. Comparison of MMR1 with other MLLMs on math-related benchmarks. All models are reevaluated under identical conditions for fairness; values in parentheses are from the original papers. The **bold** and <u>underline</u> highlight the best and second-best scores, respectively.

| Model | MVerse | MVista | MVision | LVista | CQA | Avg |
|---|---|---|---|---|---|---|
| Qwen2.5-VL-3B | 40.4 | 63.5 | 24.3 | 38.4 | 76.8 | 48.7 |
| +Cold-start | 42.1 | 58.0 | 25.2 | 39.5 | 78.5 | 48.7 |
| +GRPO | 46.2 | 65.6 | 24.7 | 42.4 | 79.9 | 51.8 |
| +VAS (MMR1) | 47.9 | 67.1 | 25.4 | 43.1 | 81.2 | 52.9 |

Table 3. The effect of Cold-start SFT and Variance-Aware Sampling. MVerse, MVista, MVision, LVista, and CQA represent MathVerse, MathVista, MathVision, LogicVista, and ChartQA, respectively.

## 6.3. Effect of Cold-Start and VAS

Table 3 reports the effect of cold-start supervision and the proposed VAS strategy on Qwen2.5-VL-3B across several mathematics-related benchmarks. Beginning with the base model, cold-start fine-tuning on curated long-form CoT data yields consistent improvements, particularly on MathVerse and ChartQA. Building upon this, GRPO further enhances performance, surpassing the cold-start baseline and demonstrating that reinforcement learning effectively incentivizes exploration and strengthens reasoning. The introduction of VAS (MMR1) achieves the highest overall scores, delivering notable gains on MathVerse, MathVista, and LogicVista. Collectively, these results highlight the complementary roles of different components: (1) cold-start supervision provides a strong initialization by imitating high-quality reasoning trajectories; (2) reinforcement learning emphasizes exploratory behavior to further incentivize reasoning; and (3) VAS ensures stable, variance-aware training, thereby leading to more robust and effective learning outcomes.

## 6.4. Effect of VAS Hyper-parameters

We further investigate the sensitivity of VAS to its key hyperparameters, *i.e.*, mixture ratio $\lambda$, VPS update frequency $T_{\text{update}}$, number of rollouts $N$, and ratio between OVS and TDS in VPS. Results are summarized in Table 4.

**Mixture ratio.** The mixture ratio $\lambda$ controls the balance between VPS-weighted sampling and uniform random sampling in Algorithm 1 (Lines 12–14). Performance is generally robust across settings, with $\lambda = 0.5$ yielding competitive and stable results. Extremely large ratios (*e.g.*, $\lambda = 1.0$) reduce coverage of the overall dataset and lead to degraded performance, confirming the necessity of maintaining a balance between variance promotion and coverage.

**Update frequency.** The update interval $T_{\text{update}}$ specifies the frequency at which VPS scores are refreshed. Short intervals (*e.g.*, 4-7 steps) enhance adaptability to model dynamics but incur higher computational overhead. Moderate intervals (*i.e.*, 14–35 steps) strike a favorable balance, yielding robust performance. In contrast, excessively long intervals (*e.g.*, 56 steps) result in outdated VPS estimates, weakening gradient signals and degrading overall performance.

**Number of rollouts.** The rollout number $N$ affects the accuracy of variance estimation for OVS and TDS. Increasing $N$ from 8 to 16 provides marginal improvements by reducing sampling noise. However, further increases (*e.g.*, 32) offer limited gains while introducing higher computational costs.

**VPS weighting.** The VPS ratio $(\alpha, \beta)$ balances outcome variance and trajectory diversity. A balanced combination (*e.g.*, $\alpha = 0.8, \beta = 0.2$) delivers strong results, whereas relying solely on one component leads to instability. Consistent

| Param. | Value | MVerse | MVista | MVision | LVista | CQA | Avg |
|---|---|---|---|---|---|---|---|
| Mixture ratio | 0.2 | 46.3 | 67.9 | 23.2 | 40.2 | 79.9 | 51.5 |
| | 0.5 | 46.1 | 66.4 | 24.8 | 41.7 | 79.4 | 51.7 |
| | 0.8 | 46.9 | 64.8 | 24.3 | 43.8 | 79.9 | 52.0 |
| | 1.0 | 44.6 | 65.3 | 24.9 | 43.8 | 79.9 | 51.7 |
| Update freq. | 4 | 47.4 | 65.8 | 23.8 | 39.3 | 79.6 | 51.2 |
| | 7 | 46.6 | 65.7 | 24.6 | 41.5 | 78.9 | 51.5 |
| | 14 | 46.7 | 66.9 | 24.2 | 44.6 | 79.2 | 52.3 |
| | 28 | 46.1 | 66.4 | 24.8 | 41.7 | 79.4 | 51.7 |
| | 35 | 47.6 | 66.1 | 24.5 | 42.4 | 80.1 | 52.2 |
| | 56 | 44.5 | 65.5 | 23.9 | 40.2 | 80.2 | 50.9 |
| # rollout | 8 | 46.1 | 66.4 | 24.8 | 41.7 | 79.4 | 51.7 |
| | 16 | 45.9 | 65.0 | 24.5 | 41.1 | 79.8 | 51.3 |
| | 32 | 46.7 | 65.1 | 25.0 | 42.2 | 78.9 | 51.6 |
| VPS ratio (OVS, TDS) | (0.0, 1.0) | 46.8 | 64.4 | 24.3 | 40.9 | 78.8 | 51.0 |
| | (0.2, 0.8) | 46.4 | 65.6 | 25.0 | 40.6 | 79.7 | 51.5 |
| | (0.5, 0.5) | 46.1 | 66.4 | 24.8 | 41.7 | 79.4 | 51.7 |
| | (0.8, 0.2) | 46.9 | 66.8 | 23.7 | 45.1 | 79.0 | 52.3 |
| | (1.0, 0.0) | 46.5 | 65.6 | 24.5 | 39.7 | 79.6 | 51.2 |

Table 4. The effect of Variance-Aware Sampling hyper-parameters. MVerse, MVista, MVision, LVista, and CQA represent MathVerse, MathVista, MathVision, LogicVista, and ChartQA, respectively.

with the analysis in §4, it shows that outcome variance and trajectory diversity play complementary roles: OVS captures correctness variability, while TDS safeguards a lower bound on variance when correctness signals are sparse.

Overall, the results demonstrate that VAS maintains stability across a wide range of hyperparameter settings, with moderate mixture ratios, appropriate update frequencies, and balanced VPS weighting yielding the most consistent performance improvements.

## 6.5. Efficiency of VAS

Figure 2 presents a comparative analysis of the training efficiency of VAS under different mixture ratios (orange line: $\lambda = 1.0$, blue line: $\lambda = 0.5$) against the vanilla random-shuffle baseline (purple line). The evaluation considers three key indicators: *actor gradient norm*, *policy-gradient clipping fraction*, and *validation accuracy*.

**Gradient norm.** The gradient norm reflects the overall magnitude of parameter updates. Models trained with VAS consistently exhibit higher gradient norms compared to the shuffle baseline, indicating more substantial and informative updates. This empirical finding is consistent with our theoretical analysis in §4, which demonstrates that prompts with higher reward variance produce gradients characterized by greater magnitude and more reliable signal strength.

**Policy-gradient clip fraction.** The clip fraction quantifies the *frequency* with which the policy update magnitude reaches the clipping threshold in GRPO. Within stable ranges, a higher clip fraction indicates more effective learning: the model performs substantial yet constrained updates, exploits the trust region more fully, and extracts stronger signals from each batch. Depicted in Figure 2, VAS configs achieve higher and more stable clip fractions compared to the baseline, highlighting their improved sample efficiency and more effective exploration of the policy space.

**Validation accuracy.** On the held-out validation set, VAS demonstrates consistent improvements in both convergence speed and final accuracy compared to the baseline. Employing full VAS sampling ($\lambda = 1.0$) yields more rapid and stable performance gain, while the mixed configuration ($\lambda = 0.5$) also surpasses uniform sampling in convergence efficiency. Although differences between $\lambda = 1.0$ and $\lambda = 0.5$ are not pronounced on this math-focused set, it is worth noting that incorporating partial random sampling can encourage broader data coverage and reduce the risk of oversampling a limited subset of prompts. This trade-off is expected to be more beneficial in domains characterized by greater content heterogeneity or less structured reward signals.

## 6.6. Dynamics of Variance Promotion Score

Figure 3 summarizes how VPS evolve across update intervals ($t \to t{+}14$), where the *histograms* show the marginal VPS distribution and *transition matrices* of VPS assignments between step $t$ (rows) and step $t{+}14$ (columns).[2]

**Convergence of rankings.** As training progresses, the mass in the transition matrices progressively concentrates along the diagonal, while the off–diagonal dispersion diminishes. This behavior reflects the progressive stabilization of per-prompt VPS rankings, converging toward a consistent ordering. Such convergence aligns with the intended steady state of VAS, in which high-variance “frontier” prompts persist only as a comparatively small yet stable subset.

**Bidirectional movement (adaptation).** Non-negligible amount of off–diagonal mass persists in both directions (low→high and high→low), reflecting the adaptive nature of VPS under the evolving policy. Certain prompts gain informativeness as their outcome become more balanced or their trajectory patterns diversity increase, whereas others gradually lose informativeness once they are either consistently solved with ease or repeatedly lead to failure.

**High→high persistence fades.** In the early stages, a visible block appears in the bottom-right region (high→high), but this gradually weakens as training progresses. This pattern indicates that many prompts are initially challenging, remaining near the maximum OVS/TDS levels across updates. As the policy improves, however, these prompts either become polarized in correctness—lowering OVS—or converge toward more uniform trajectories—lowering TDS. Consequently, they gradually exit the high-VPS subset.

**Distributional shift toward mid-VPS.** The marginal distribution of VPS progressively evolves from a relatively

[2] $T_{\text{update}}$ is set to 14.

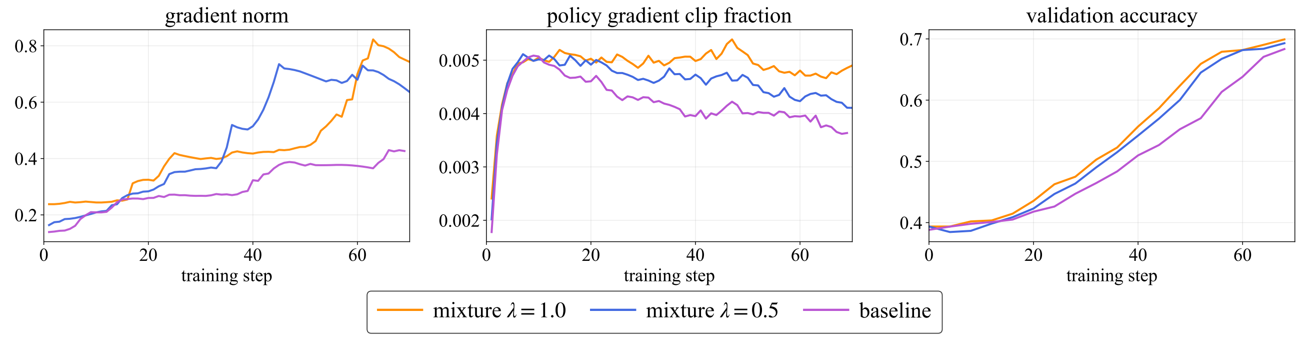


Figure 2. Training efficiency of Variance-Aware Sampling (VAS). The plots compare three settings: full VAS sampling ($\lambda = 1.0$, orange), mixed sampling with half VAS and half random ($\lambda = 0.5$, blue), and the vanilla baseline (purple). **Left:** Actor gradient norm, reflecting the magnitude of gradient signals during training. **Middle:** Policy gradient clip fraction, indicating the proportion of updates reaching the clipping boundary. **Right:** Validation accuracy, showing convergence speed and final performance.

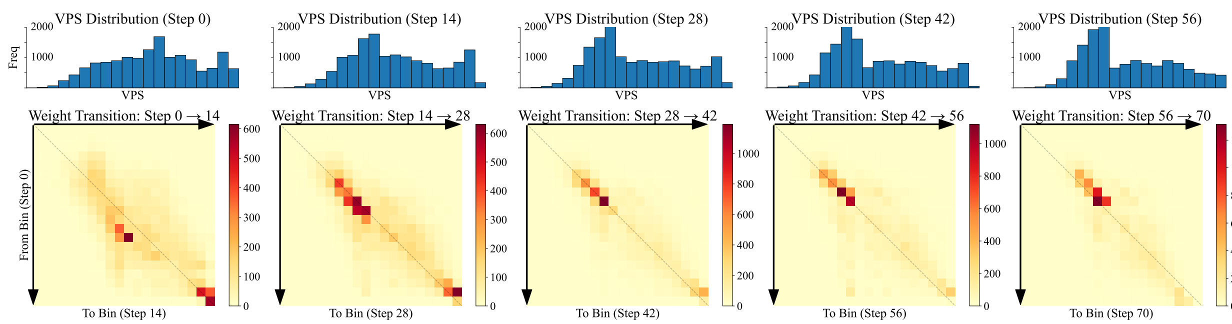


Figure 3. Dynamics of Variance Promotion Score during training. The top row illustrates the distribution of VPS values across data points at different training steps. The bottom row shows transition matrices of VPS assignments between consecutive update intervals, where each cell indicates the number of data points moving from a source bin (vertical axis) at the earlier step to a target bin (horizontal axis) at the later step. The arrows indicate the direction from lower to higher VPS bins, facilitating interpretation of upward or downward transitions.

flat shape with a pronounced upper tail to a more compact form centered around medium scores. Since $\mathrm{VPS} = \alpha\,\mathrm{OVS} + \beta\,\mathrm{TDS}$ with $\alpha(0.8) > \beta(0.2)$, this shift reflects two key dynamics: 1) as training proceeds, fewer prompts achieve near-maximum OVS (with pass rates $\approx 0.5$), and 2) TDS establishes a residual floor, anchoring many prompts at moderate VPS even when correctness becomes less stable. This emergent clustering around mid-VPS characterizes a signature of convergence: VAS increasingly narrows its focus to a frontier of prompts where reward variance remains informative and useful for further optimization.

**Asymmetry of flows.** In later intervals, transitions from higher to mid VPS bins occur more frequently than movements in the opposite direction, implying a gradual reduction in reward variance as competence improves. It aligns with the Variance–Progress principle in §4: once learning progress has been achieved on a given prompt, both its reward variance and VPS tend to diminish, thereby encouraging the sampler to redistribute probability mass toward other frontier items that remain more uncertain or informative.

Moreover, we provide a detailed qualitative demonstration in Supp. E, illustrating MMR1's reasoning process on a MathVerse problem. It presents the input question, step-by-step reasoning, and final answer, showcasing a coherent process encompassing problem analysis, solution planning, execution, verification, and exploration of alternatives that lead to the correct conclusion.

## 7. Conclusion

In this work, we investigate the challenge of gradient vanishing in reinforcement learning for multimodal reasoning. We introduce Variance-Aware Sampling (VAS), a sampling strategy that exploits outcome variance and trajectory diversity to prioritize informative prompts while maintaining broad data coverage. Grounded in theoretical analysis and supported by extensive empirical evaluation, VAS enhances training stability and improves the effectiveness of reinforcement learning in multimodal reasoning tasks. We will also release the large-scale, carefully curated cold-start datasets, codes, and well-tuned models, which we hope will provide valuable resources for benchmarking and advancing future research.

**Acknowledgment** This study is funded by the Ministry of Education Singapore, under the Tier-2 project scheme with project number MOET2EP20123-0003.

## References

[1] 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)*, pages 12248–12267, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. 4, 5

[3] Huanqia Cai, Yijun Yang, and Winston Hu. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models. *ArXiv*, abs/2502.00698, 2025. 5

[4] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. 5

[5] Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Deli Zhao, Anh Tuan Luu, and Yu Rong. Geopqa: Bridging the visual perception gap in mllms for geometric reasoning. *ArXiv*, abs/2509.17437, 2025. 1

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *ArXiv*, abs/2412.05271, 2025. 5

[7] Google DeepMind. Gemini 2.5: Our most intelligent ai model, 2025. 4

[8] Linger Deng, Linghao Zhu, Yuliang Liu, Yu Wang, Qunyi Xie, Jingjing Wu, Gang Zhang, Yingying Zhu, and Xiang Bai. Theorem-validated reverse chain-of-thought problem generation for geometric reasoning. *ArXiv*, abs/2410.17885, 2024. 5

[9] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *ArXiv*, abs/2503.17352, 2025. 1, 2, 5

[10] Thomas Foster and Jakob Foerster. Learning to reason at the frontier of learnability. *ArXiv*, abs/2502.12272, 2025. 3

[11] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. *ArXiv*, abs/2506.04178, 2025. 1

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638, 2025. 1, 2

[13] 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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 5

[14] Jack *Hessel, Jena D *Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning. In *ECCV*, 2022. 5

[15] hiyouga. Mathruler. `https://github.com/hiyouga/MathRuler`, 2025. 5

[16] Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models. *ArXiv*, abs/2501.03262, 2025. 2

[17] Chenghua Huang, Lu Wang, Fangkai Yang, Pu Zhao, Zhixu Li, Qingwei Lin, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. Lean and mean: Decoupled value policy optimization with global value guidance. *ArXiv*, abs/2502.16944, 2025. 2

[18] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *ArXiv*, abs/2503.06749, 2025. 5

[19] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *ArXiv*, abs/2503.06749, 2025. 1, 2

[20] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[21] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary

visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[22] 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 (CVPR)*, 2018. 5

[23] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. In *AI for Math Workshop @ ICML 2024*, 2024. 5

[24] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251, Cham, 2016. Springer International Publishing. 5

[25] Kimi. Kimi k1.5: Scaling reinforcement learning with llms. *ArXiv*, abs/2501.12599, 2025. 2

[26] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, page 611–626, New York, NY, USA, 2023. Association for Computing Machinery. 5

[27] Hynek Kydlíček. Math-Verify: Math Verification Library, 2025. 5

[28] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *ArXiv*, abs/2408.03326, 2024. 5

[29] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Dual-glance model for deciphering social relationships. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2669–2678, 2017. 5

[30] Xuefeng Li, Haoyang Zou, and Pengfei Liu. Limr: Less is more for rl scaling. *ArXiv*, abs/2502.11886, 2025. 2

[31] Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. ReMax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 29128–29163. PMLR, 2024. 2

[32] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From system 1 to system 2: A survey of reasoning large language models. *ArXiv*, abs/2502.17419, 2025. 1

[33] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[34] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *ArXiv*, abs/2208.05358, 2022. 5

[35] Jiacai Liu, Chaojie Wang, Chris Yuhao Liu, Liang Zeng, Rui Yan, Yiwen Sun, Yang Liu, and Yahui Zhou. Improving multi-step reasoning abilities of large language models with direct advantage policy optimization. *ArXiv*, abs/2412.18279, 2024. 2

[36] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *ArXiv*, abs/2503.20783, 2025. 2

[37] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, Online, 2021. Association for Computational Linguistics. 5

[38] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 5

[39] 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. In *Advances in Neural Information Processing Systems*, pages 2507–2521. Curran Associates, Inc., 2022. 5

[40] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and A. Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *ArXiv*, abs/2209.14610, 2022. 5

[41] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[42] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 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, Dublin, Ireland, 2022. Association for Computational Linguistics. 5

[43] 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. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. 5

[44] Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209, 2021. 5

[45] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa.

In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1697–1706, 2022. 5

[46] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *ArXiv*, abs/2503.07365, 2025. 1, 2, 4

[47] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *ArXiv*, abs/2503.07365, 2025. 5

[48] Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 5

[49] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Workshop on Reasoning and Planning for Large Language Models*, 2025. 1

[50] OpenAI. Gpt-4o system card. *AvrXiv*, abs/2410.21276, 2024. 4, 5

[51] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China, 2015. Association for Computational Linguistics. 5

[52] Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. Multimath: Bridging visual and mathematical reasoning for large language models. *ArXiv*, abs/2409.00147, 2024. 5

[53] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *ArXiv*, abs/2503.07536, 2025. 2

[54] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. SPIQA: A dataset for multimodal question answering on scientific papers. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 5

[55] Noam Razin, Hattie Zhou, Omid Saremi, Vimal Thilak, Arwen Bradley, Preetum Nakkiran, Joshua M. Susskind, and Etai Littwin. Vanishing gradients in reinforcement finetuning of language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[56] Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, and Sanjeev Arora. What makes a reward model a good teacher? an optimization perspective. *ArXiv*, abs/2503.15477, 2025. 1, 2

[57] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. 2

[58] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 2022. 5

[59] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. 1, 2

[60] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. *ArXiv*, abs/2504.07615, 2025. 1, 2

[61] 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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680, Miami, Florida, USA, 2024. Association for Computational Linguistics. 5

[62] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *ArXiv*, abs/2503.20752, 2025. 1, 2, 4

[63] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. *ArXiv*, abs/2101.11272, 2021. 5

[64] Mert Unsal and Aylin Akkus. Easyarc: Evaluating vision language models on true visual reasoning. *ArXiv*, abs/2506.11595, 2025. 5

[65] Jean Vassoyan, Nathanaël Beau, and Roman Plaud. Ignore the KL penalty! boosting exploration on critical tokens to enhance RL fine-tuning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6108–6118, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 2

[66] Huaijie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, and Yi Wu. Offline reinforcement learning for LLM multi-step reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8881–8893, Vienna, Austria, 2025. Association for Computational Linguistics. 2

[67] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *ArXiv*, abs/2504.08837, 2025. 5

[68] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *ArXiv*, abs/2504.08837, 2025. 1, 2, 4

[69] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 5

[70] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 5

[71] Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. MathCoder-VL: Bridging vision and code for enhanced multimodal mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2505–2534, Vienna, Austria, 2025. Association for Computational Linguistics. 5

[72] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1

[73] Tianlong Wang, Junzhe Chen, Xueting Han, and Jing Bai. Cpl: Critical plan step learning boosts llm generalization in reasoning tasks. *ArXiv*, abs/2409.08642, 2024. 2

[74] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *ArXiv*, abs/2504.07934, 2025. 5

[75] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. 3

[76] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *ArXiv*, abs/2407.04973, 2024. 5

[77] Yun Xing, Xiaobin Hu, Qingdong He, Jiangning Zhang, Shuicheng Yan, Shijian Lu, and Yu-Gang Jiang. Boosting reasoning in large multimodal models via activation replay. *arXiv preprint arXiv:2511.19972*, 2025. 1

[78] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *ArXiv*, abs/2503.10615, 2025. 2

[79] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *ArXiv*, abs/2503.10615, 2025. 5

[80] Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, Hao Zhang, and Yu Rong. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *ArXiv*, abs/2507.22607, 2025. 5

[81] Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, Tiantian Fan, Zhengyin Du, Xiang Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Ru Zhang, Xin Liu, Mingxuan Wang, Yong-Xu Wu, and Lin Yan. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *ArXiv*, abs/2504.05118, 2025. 2

[82] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[83] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *ArXiv*, abs/2503.12937, 2025. 5

[84] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *ArXiv*, abs/2503.12937, 2025. 1, 2

[85] Kechi Zhang, Ge Li, Jia Li, Yihong Dong, Jia Li, and Zhi Jin. Focused-DPO: Enhancing code generation through focused preference optimization on error-prone points. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9578–9591, Vienna, Austria, 2025. Association for Computational Linguistics. 2

[86] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186, Berlin, Heidelberg, 2024. Springer-Verlag. 5

[87] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and Hongsheng Li. Mavis: Mathematical visual instruction tuning with an automatic data engine. *ArXiv*, abs/2407.08739, 2024. 5

[88] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand, 2024. Association for Computational Linguistics. 5

[89] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. `https://github.com/hiyouga/EasyR1`, 2025. 5

[90] Jin Peng Zhou, Kaiwen Wang, Jonathan Chang, Zhaolin Gao, Nathan Kallus, Kilian Q. Weinberger, Kianté Brantley, and Wen Sun. $q\sharp$: Provably optimal distributional rl for llm post-training. *ArXiv*, abs/2502.20548, 2025. 2

[91] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun

He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv*, abs/2504.10479, 2025. 5