Stable-MM-R1: Anchoring Multimodal Reasoning Dynamics via Entropy-Guided Stratification

Yimeng Ye[1*] **Shuang Chen**[1*] **Wenxuan Huang**[2] **Manyuan Zhang**[2]

Kaituo Feng[2] **Zhangquan Chen**[3] **Jiayu Chen**[2] **Yucheng Zhou**[2]

Yicheng Xiao[3] **ZhiYuan Feng**[3] **Tianyu Shi**[4]

[1]Columbia University [2]CUHK [3]THU [4]McGill University

Abstract

While Reinforcement Learning (RL) effectively incentivizes reasoning in Large Language Models, current pipelines are hindered by training instability and rapid entropy collapse. These limitations often stem from "Rollout Silencing" and low-quality gradient signals in standard sampling procedures. In this work, we propose a robust, data-centric framework to stabilize RL training. We first introduce Potential-Aware Query Mining (PAQM), which filters data dynamically to focus on the "Distillation Zone"—samples with high potential for capability elicitation. Furthermore, we present Hybrid Stratified Replay (HSR), a novel mechanism that restructures batches by stratifying rollouts based on Path Entropy, a rollout-level confidence proxy, and outcome reward. Within each optimization step, HSR reuses current-policy "Stability Anchors" and "Hard Negatives" to construct high-contrast optimization groups, then clears its buffers before the next step. This approach mitigates entropy collapse while improving the utilization of learning signals under limited compute. Our method outperforms strong baselines on complex reasoning tasks, offering a principled solution for stable and efficient RL fine-tuning.

1 Introduction

Multimodal large language models (MLLMs) have recently demonstrated striking progress on complex reasoning tasks, especially when equipped with explicit reasoning traces such as chain-of-thought (CoT) (Wei et al., 2022; Chen et al., 2026c). However, reliably *eliciting* and *stabilizing* reasoning behaviors during post-training remains challenging, particularly when we move beyond supervised imitation toward reinforcement learning (RL) with verifiable rewards (Ouyang et al., 2022; Chen et al., 2026b; DeepSeek-AI et al., 2025; Zhang et al., 2025; Feng et al., 2026b). In reasoning-centric settings (e.g., math and diagram reasoning), RL is attractive because correctness can be automatically verified, enabling scalable learning signals without dense human labels (Cobbe et al., 2021; Lightman et al., 2023). Yet practitioners consistently observe that RL training can be brittle: reward curves oscillate, KL/entropy diagnostics exhibit abrupt shifts, and the policy may prematurely collapse into narrow modes that look confident but generalize poorly (Yuan et al., 2025; Yu et al., 2025).

Following recent reasoning LLM practice, we define stable training as a process where (i) reward and benchmark performance improve steadily across steps, and (ii) the model's internal state evolves smoothly without abrupt regime changes, as reflected by diagnostics such as training–inference KL divergence and sequence-level/token-level entropy (Yuan et al., 2025). Unfortunately, stability is often violated in modern RL pipelines for reasoning. A prominent failure mode is premature entropy collapse: once the policy discovers a narrow high-reward strategy, the action distribution sharpens too quickly, cutting off exploration and trapping learning in local optima. This issue becomes more severe in long-CoT reasoning where gradient signal is sparse and delayed, and even PPO-style actor–critic RL can collapse without careful value calibration (Schulman et al., 2017; Yuan et al., 2025). Moreover, group-based methods that avoid a learned critic—e.g., GRPO—reduce system complexity and have enabled strong reasoning results (Shao et al., 2024; Chen et al., 2026a; DeepSeek-AI et al., 2025; Feng et al., 2026a; Wu et al., 2025a), but still suffer from batch degeneracy: stochastic group composition can yield near-zero advantage variance (vanishing updates) or noisy baselines (destabilizing updates), especially under limited rollout budgets.

Common stabilization mechanisms are *objective-*

[*]Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 1840–1853
October 24-29, 2026 ©2026 Association for Computational Linguistics

centric, including clipped policy updates and KL or entropy control (Schulman et al., 2017; Yu et al., 2025). While effective in some regimes, these "soft constraints" can be sensitive to scaling, reward sparsity, and long-horizon CoT dynamics. Inspired by recent data-centric RL observations (e.g., advantage collapsing / rollout silencing in multimodal RL fine-tuning) (Zhu et al., 2025b), we ask a different question: *instead of modifying how RL optimizes, can we stabilize training by controlling what data the optimizer sees at each step?* We argue that a major source of instability is not the RL objective itself, but the stochastic quality of rollouts and optimization groups: easy queries waste compute with near-zero gradients, impossible queries inject high-variance noise, and mixed-quality rollouts within a group can produce unreliable advantage estimates.

To address these challenges, we propose **Stable-MM-R1**, a data-centric stabilization framework for reasoning RL that structures learning around two complementary axes: outcome (verification reward) and confidence (Path Entropy, defined as a sequence-level surprisal proxy). Unlike standard methods that rely on stochastic sampling, our approach actively manages the training distribution. First, we introduce *Potential-Aware Query Mining (PAQM)* to dynamically curate a curriculum, ensuring the model continuously operates within a "Distillation Zone" of learnable queries rather than wasting compute on the trivial or the impossible. Second, and most crucially, we employ *Hybrid Stratified Replay (HSR)* to engineer the composition of optimization groups. By pairing high-surprisal exploration targets with "Stability Anchors" (mastered samples) and "Hard Negatives" (confident errors), we construct stable, high-contrast optimization groups whenever the corresponding samples are available. Together, these components are designed to reduce the incidence of zero-signal groups and mitigate premature entropy collapse without modifying the underlying RL objective.

Our main contributions are: (1) We identify stochastic group degeneracy as one source of instability in reasoning RL, showing from a data-centric perspective how uncontrolled rollout quality can lead to vanishing gradients and premature entropy collapse. (2) We introduce Potential-Aware Query Mining (PAQM), a dynamic curriculum framework that filters data based on *Pass@K* and *Path Entropy* to concentrate updates within the "Distillation Zone." (3) We propose Hybrid Stratified Replay (HSR), a novel sampling strategy that reduces the incidence of degenerate groups through the structured inclusion of Stability Anchors and Hard Negatives when the corresponding samples are available, thereby anchoring the policy update dynamics. (4) We demonstrate that Stable-MM-R1 outperforms the evaluated open-source 7B-scale baselines on average across diverse benchmarks, while maintaining healthy Path-Entropy trajectories that avoid the collapse patterns observed in GRPO and DAPO.

2 Related Works

2.1 Multimodal Large Reasoning Models (MLRMs)

The paradigm of post-training has shifted from supervised imitation to reinforcement learning (RL), driven by the premise that reasoning capabilities emerge from exploration rather than memorization (Feng et al., 2025; Chen et al., 2025b). Pioneering works like DeepSeek-R1 (DeepSeek-AI et al., 2025) and Kimi k1.5 (Team et al., 2025) have demonstrated that RL with verifiable rewards can incentivize long-chain reasoning (CoT). This trend is rapidly extending to the multimodal domain with models like Qwen2.5-VL (Bai et al., 2025b) and LLaVA-OneVision (An et al., 2025).

Recent specialized frameworks have further expanded this frontier. SATORI-R1 (Shen et al., 2025) integrates spatial grounding with verifiable rewards to incentivize fine-grained visual reasoning by decomposing VQA into verifiable stages. Similarly, Retrv-R1 (Zhu et al., 2025a) introduces a reasoning-driven framework tailored for universal multimodal retrieval. Our work complements these by focusing on the stability of the underlying optimization process itself.

2.2 Entropy Collapse and Stabilization in RL

A critical failure mode in reasoning-oriented RL is Entropy Collapse, where the policy prematurely converges to a deterministic distribution (Yuan et al., 2025). PPO-style training constrains policy updates through clipping and can additionally use KL or entropy control, but group-based methods such as GRPO still struggle to maintain diversity (Schulman et al., 2017).

Recent works have proposed various *objective-centric* modifications to stabilize GRPO. Quantile Advantage Estimation (QAE) (Wu et al., 2025b) replaces the mean baseline with quantile estimates to prevent reward outliers from crushing entropy.

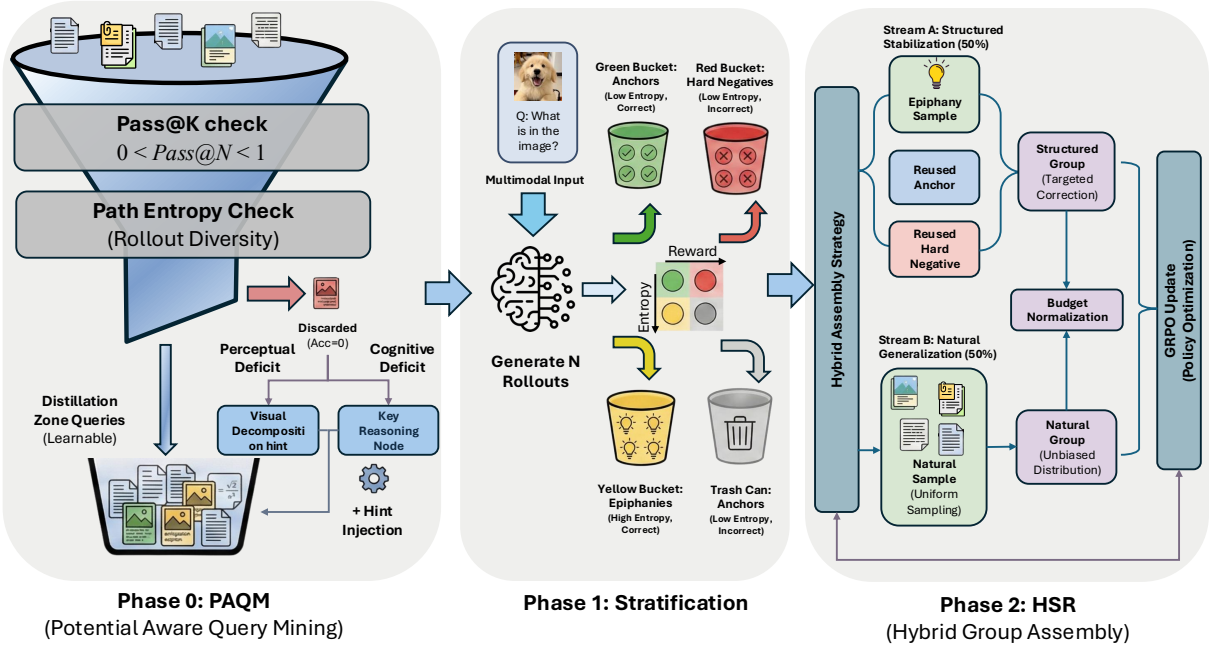


Figure 1: Overview of the Stable-MM-R1 framework. **Phase 0 (PAQM):** A dynamic curriculum is constructed via a multi-dimensional filter (Pass@K and Path Entropy). Queries in the "Impossible" zone undergo a *Modality-Decoupled Diagnosis*, receiving targeted *Visual Decomposition* or *Cognitive Key Nodes* to reclaim them into the learnable frontier. **Phase 1 (Stratification):** Generated rollouts are diagnosed using Path Entropy and reward, then classified into semantic buffers: *Anchors* (stability), *Hard Negatives* (correction), and *Epiphanies* (exploration), while stochastic *Noise* is discarded. **Phase 2 (HSR Group Assembly):** Optimization groups are assembled via a hybrid strategy. *Stream A* (Structured) explicitly reuses Anchors and Hard Negatives to stabilize the gradient variance for Epiphanies, while *Stream B* (Natural) preserves natural on-policy samples. Together, Phases 1 and 2 constitute the complete HSR pipeline.

Sharpness-Guided GRPO (GRPO-SG) (Le et al., 2025) uses probability-based token weights to downweight tokens likely to induce overly large gradients. Similarly, M-GRPO (Bai et al., 2025a) utilizes a momentum-updated reference model to anchor the policy target, reducing oscillation. Taking a control-theoretic perspective, EntroPIC (Yang et al., 2025) applies Proportional-Integral (PI) control to dynamically adjust loss coefficients, explicitly targeting a desired entropy trajectory.

2.3 Data-Centric Policy Optimization

An emerging alternative focuses on the *input data distribution* rather than the objective function. Shuffle-R1 (Zhu et al., 2025b) identifies "rollout silencing" and proposes contrastive batch organization. Concurrently, Tang et al. (2025) emphasize sample polarity, while MMR1 (Leng et al., 2025) explores variance-aware sampling.

Beyond selection, recent works explore restructuring the data itself. RoRecomp (Li et al., 2025) enhances reasoning efficiency via rollout response recomposition. Addressing gradient variance directly, Zeng et al. (2025) introduce Shrinkage Baselines for verifiable rewards to improve estimator accuracy. Unlike these approaches, our Stable-MM-R1 introduces Path Entropy, a rollout-level confidence proxy, as a dimension for data curation. By explicitly stratifying samples into stability anchors and exploration targets, we enforce training stability through structured data composition.

3 Methodology

In this section, we present Stable-MM-R1, a unified framework designed to stabilize the post-training of multimodal reasoning models. Our approach operates in two stages: first, we curate a dynamic curriculum via Potential-Aware Query Mining (Phase 0, §3.1); second, we stabilize the optimization process using Hybrid Stratified Replay, which combines diagnostic stratification (Phase 1, §3.2) with structured group assembly (Phase 2, §3.3). By addressing both data ingestion and optimization-group composition, our framework targets zero-signal groups and premature entropy collapse.

3.1 Phase 0: Potential-Aware Query Mining (PAQM)

Constructing a high-quality prompt set is the first line of defense against training instability. We introduce Potential-Aware Query Mining (PAQM), a dynamic curation pipeline that filters queries based on their "Learnability."

Mathematical Definition of Learnability. For a given prompt x, we sample $N = 16$ rollouts $\{y_1, ..., y_{16}\}$ from the current policy π_θ. We estimate the empirical success rate $\hat{p}(x) = \frac{1}{N}\sum r(y_i)$. We categorize queries into three zones:

- **Distillation Zone (Learnable):** Queries where $0 < \hat{p}(x) < 1$. These exhibit variance in outcomes, indicating the model is capable but unstable—the ideal target for RL.
- **Trivial Zone:** Queries where $\hat{p}(x) = 1$. These are mastered concepts and are downsampled to prevent overfitting.
- **Impossible Zone:** Queries where $\hat{p}(x) = 0$. These yield zero gradients in standard RL and are candidates for diagnosis.

To obtain a tractable rollout-level confidence score within the Distillation Zone, we define Path Entropy as the length-normalized sequence surprisal of a sampled rollout. Motivated by predictive-uncertainty estimation (Malinin and Gales, 2018), we compute

$$H_{\text{path}}(y|x) = -\frac{1}{|y|}\sum_{t=1}^{|y|}\log \pi_\theta(y_t|x, y_{<t}). \quad (1)$$

This sample-level quantity is a proxy for model confidence: it is not the Shannon entropy of the full predictive distribution and should not be interpreted as a direct estimate of epistemic uncertainty.

Adaptive Frontier Expansion via Modality-Decoupled Diagnosis. For queries in the Impossible Zone, we apply Modality-Decoupled Diagnosis. We probe the model using ground-truth captions to isolate the failure mode.

- **Perceptual Deficit:** If the model solves the text-only probe, we inject a Visual Decomposition Hint (e.g., coordinates, shapes) derived from an auxiliary vision model.
- **Cognitive Deficit:** If the model fails the probe, we inject a Key Reasoning Node (e.g., the first theorem step) distilled from a teacher model's trace.

This mechanism reclaims roughly 30% of hard queries back into the Distillation Zone. Note that these hints are generated offline using the proprietary teacher model Seed-1.5-VL (Guo et al., 2025) and are injected only during Phase 0 curriculum construction.

3.2 Phase 1: Diagnostic Path-Entropy–Reward Stratification

One source of instability in Group Relative Policy Optimization (GRPO) is the stochastic quality of optimization groups. To address it, we propose Hybrid Stratified Replay (HSR), which treats "Mastered Samples" and "Confident Errors" as Stability Anchors.

We categorize rollouts based on outcome correctness $r(y)$ and the Path-Entropy score $H_{\text{path}}(y|x)$. To adapt to the shifting policy distribution, we employ Dynamic Percentile Thresholding. We calculate the empirical Path-Entropy distribution $\mathcal{H}$ across all rollouts in the current global batch (covering multiple queries) to obtain batch-adaptive cutoffs. We define dynamic thresholds $\tau_{low} = \text{Percentile}(\mathcal{H}, 10)$ and $\tau_{high} = \text{Percentile}(\mathcal{H}, 90)$. Rollouts are classified into three buffers:

- **Stability Anchors** ($\mathcal{B}_{anc}$): Correct samples with low Path Entropy ($r = 1, H_{\text{path}}(y|x) < \tau_{low}$). These are reused as high-confidence positive anchors.
- **Hard Negatives** ($\mathcal{B}_{neg}$): Incorrect samples with low Path Entropy ($r = 0, H_{\text{path}}(y|x) < \tau_{low}$). These are high-confidence errors included to provide outcome contrast.
- **Epiphanies** ($\mathcal{B}_{epi}$): Correct samples with high Path Entropy ($r = 1, H_{\text{path}}(y|x) > \tau_{high}$). These are low-confidence successes emphasized as learning targets.

Samples in the "Confusion Zone" ($r = 0, H_{\text{path}}(y|x) > \tau_{high}$) are treated as low-confidence errors and excluded from the valid sample pool.

3.3 Phase 2: Hybrid Group Assembly

Standard GRPO constructs optimization groups via random sampling, which can yield low-contrast or high-variance updates (Zhu et al., 2025b; Leng et al., 2025). To balance the need for *targeted*

correction and *natural on-policy coverage*, we allocate the total group budget G into two equal streams: the Structured Stream (50%) and the Natural Stream (50%).

Stream A: Structured Stabilization (50%). This stream focuses on "Correction" by explicitly engineering the batch topology. Within the allocated budget for this stream ($G_{struct} = 0.5G$), we enforce a strict semantic ratio: 50% Epiphanies (to drive learning), 25% Stability Anchors (to anchor the baseline), and 25% Hard Negatives (to sharpen the decision boundary). Formally:

$$G_{struct} = S_{epi}^{50\%} \cup \text{Sample}(\mathcal{B}_{anc}, 25\%) \cup \text{Sample}(\mathcal{B}_{neg}, 25\%) \quad (2)$$

When all structured quotas are available, this injection grounds the selected exploration targets with positive anchors and contrasting errors within the micro-batch.

Stream B: Natural Generalization (50%). To prevent the policy from overfitting to the artificial topology of Stream A, the second stream ($G_{nat} = 0.5G$) focuses on "Robustness." We sample groups uniformly from the valid sample pool (excluding only the filtered noise). This preserves natural on-policy coverage alongside the deliberately shifted structured stream.

Clarification on Replay Scope (Intra-Step Only). A critical distinction of our framework is the scope of the replay buffers. Unlike traditional experience replay used in off-policy RL, HSR employs an Intra-Step Replay mechanism. The buffers $\mathcal{B}_{anc}$ and $\mathcal{B}_{neg}$ have a lifespan of $\tau = 1$; they are populated solely from the N rollouts generated at the *current* optimization step and are cleared immediately after the gradient computation. This ensures that all "reused" samples are drawn from the current policy π_θ, thereby maintaining on-policy consistency without requiring importance sampling corrections for stale data.

3.3.1 Theoretical Motivation: Degenerate Groups

Randomly sampled groups may become reward-degenerate when all rollouts receive the same outcome, causing group-relative advantages to vanish. HSR reduces the incidence of such zero-signal groups by constructing a structured stream containing both successful and failed rollouts whenever the corresponding samples are available. Section A.3 provides a conditional variance analysis and describes the fallback used when a structured quota cannot be satisfied.

4 Experiments

4.1 Experimental Setup

Datasets and Benchmarks. Distinct from prior multi-stage approaches, Stable-MM-R1 uses RL-only parameter updates on multimodal data, without a supervised cold-start phase; teacher-generated hints affect selected training prompts but are not used for supervised parameter updates. To support this, we construct a high-quality RL Prompt Corpus comprising approximately 189k samples by aggregating verifiable multimodal tasks from Geometry3K (Lu et al., 2021) and a curated subset of MM-Eureka (Meng et al., 2025). Furthermore, we integrate ViRL39K (Wang et al., 2025a), a large-scale dataset of verifiable visual reasoning queries, to expand the diversity of visual scenarios. Crucially, our *Potential-Aware Query Mining (PAQM)* pipeline utilizes a teacher-guided distillation strategy to generate *scaffolding hints* for hard queries. Given the multimodal nature of our data source, we exclusively employ the proprietary Vision-Language Model Seed-1.5-VL (Guo et al., 2025) to generate both perceptual descriptions and reasoning traces offline. These cached multimodal hints are injected into selected prompts during Phase 0 curriculum construction, before RL rollouts begin.

For evaluation, we employ a comprehensive suite of benchmarks covering mathematical reasoning, visual perception, and chart understanding: MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024a), MathVista (Lu et al., 2024), DynaMath (Zou et al., 2024), WeMath (Qiao et al., 2024), LogicVista (Xiao et al., 2024), MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b), CharXiv (Wang et al., 2024b), and MMStar (Chen et al., 2024).

Baselines. We benchmark Stable-MM-R1 against three categories of systems to ensure a rigorous evaluation. First, we include leading proprietary models as upper-bound references: GPT-4.1 (OpenAI, 2025), Gemini-2.5-Pro-Thinking (Comanici et al., 2025), Claude-4-Sonnet (Anthropic, 2025), and Doubao-1.5-Thinking-Vision-Pro (Guo et al., 2025). Second, we consider representative lightweight open-source MLLMs: Qwen2.5-VL-3B-Instruct (Bai et al.,

2025b), FAST-3B (Xiao and Gan, 2025), and VLAA-Thinker-3B (Chen et al., 2025a). Third, we evaluate competitive 7B-scale open-source MLLMs, including Qwen2.5-VL-7B-Instruct, OpenVLThinker-1.2-7B (Deng et al., 2025), MM-Eureka-Qwen-7B (Meng et al., 2025), MMR1-Math-v0 (Leng et al., 2025), ThinkLite-7B-VL (Wang et al., 2025b), VLAA-Thinker-7B, VL-Rethinker-7B (Wang et al., 2025a), and Vision-G1 (Zha et al., 2025). Most open-source baselines are fine-tuned from the Qwen2.5-VL family, ensuring comparability in architecture and training setup.

Implementation Details. We utilize the EasyR1 framework (Zheng et al., 2025) as our training codebase. We employ Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) as the base policy to verify scalability. We freeze the vision encoder and update only the LLM backbone during RL training. We set the global batch size to 128 and the learning rate to 1×10^{-6}. The rollout temperature is set to 1.0. For HSR, we generate $N = 16$ rollouts per query. Across in-house methods, we match the final group size ($G = 16$) and the number of gradient updates (400 steps). This controls the optimization-update budget but does not make total data-preparation and generation cost identical: PAQM uses an offline hint cache and may resample queries that initially fall in the Impossible Zone. All experiments are conducted on 8× H100 GPUs. Throughout the paper, Pass@8 denotes majority-vote accuracy over eight independently sampled responses.

Data Decontamination. Given the aggregation of multiple datasets, data leakage is a critical concern. We applied a de-duplication pipeline. For text, we used MinHash LSH to remove samples with > 0.8 Jaccard similarity to any query in the evaluation set. For images, we employed perceptual hashing (pHash) to filter out visual duplicates. These checks reduce direct and near-duplicate overlap, although they cannot rule out all semantic contamination.

Controlled Baseline Reproduction & Fairness. All in-house comparisons use the same base model, source prompt pool, final per-query rollout group size, training schedule, and evaluation protocol. PAQM and HSR are enabled only in the configurations explicitly marked in Table 2. Standard GRPO and DAPO do not use PAQM. The '+ PAQM' configuration uses the full hint-augmented curriculum with the standard GRPO objective, whereas Full Stable-MM-R1 additionally enables HSR; their comparison therefore isolates the contribution of structured batch topology under the same PAQM data pipeline. We match the final group size ($G = 16$) and update count across methods; the offline teacher and resampling overhead described above are additional costs of PAQM.

Note on Training Duration: While Stable-MM-R1 maintained stability throughout the full training schedule (approx. 400 steps), we observed that the GRPO and DAPO baselines frequently suffered from entropy collapse and performance degradation after step 250 (as visualized in Figure 3). To ensure a fair comparison, we report the peak performance achieved by these baselines prior to collapse, rather than their final converged checkpoints. This protocol ensures that the reported gaps reflect the ceiling of the baselines capabilities rather than their failure modes.

4.2 Main Results

Comparison with Baselines. Table 1 presents a comprehensive comparison against both proprietary and open-source models. Stable-MM-R1 achieves the highest average score among the evaluated open-source 7B-scale models, with an average score of 56.2%, outperforming the strongest evaluated open-source competitor Vision-G1 (46.5%) by 9.7 points. Notably, on the fine-grained visual reasoning benchmark MathVerse, our method achieves 56.3%, surpassing Vision-G1 (50.0%) by 6.3 points and indicating strong performance on fine-grained visual mathematical reasoning. On MathVision, which evaluates challenging visual mathematical reasoning, we observe an improvement of 28.0 points over the Qwen2.5-VL base model and 20.2 points over the strongest evaluated open-source baseline. While our performance on MathVista (75.5%) is slightly below Vision-G1 (-0.6%), Stable-MM-R1 outperforms the base model and the other evaluated RL baselines on average while exhibiting more stable training dynamics in Figures 2 and 3.

Stability Analysis. Figure 2 contrasts the training dynamics of the three methods. The GRPO baseline exhibits large Path-Entropy fluctuations followed by a sharp decline, while DAPO's Path Entropy gradually falls to approximately 0.2 by its final recorded step. After its initial decline, Stable-MM-R1 remains mostly within approxi-

Model	MathVerse-V	MathVision	MathVista	DynaMath-W	WeMath	LogicVista	MMMU	MMMU-Pro	CharXiv	MMStar	Avg.
Closed-Source Models											
Gemini-2.5-Pro-Thinking	*81.2*	*55.3*	*83.8*	*57.1*	78.0	*75.2*	82.0	*76.5*	*69.3*	79.7	73.8
Doubao-1.5-thinking-vision-pro	*80.4*	68.7	85.6	*60.5*	*78.0*	*71.8*	77.9	67.6	*63.4*	78.2	73.2
Claude-4-Sonnet	*66.1*	*54.6*	*70.4*	*46.9*	*63.0*	*64.4*	74.4	*60.7*	*58.4*	66.9	62.6
GPT-4.1	*59.8*	*51.8*	72.0	*48.3*	55.5	*63.8*	75.0	*65.0*	*55.8*	71.2	61.8
3B-scale MLLMs											
Qwen2.5-VL-3B-Instruct	*33.0*	21.2	62.3	*17.0*	*17.9*	35.8	53.1	31.6	*25.3*	51.1	34.8
VLAA-Thinker-3B	36.4	24.4	61.0	18.2	33.8	38.5	*49.7*	*33.3*	*28.2*	53.4	37.7
FAST-3B	*37.2*	26.8	66.2	*15.4*	*23.3*	*35.4*	*52.0*	*34.6*	*28.3*	55.2	37.4
Stable-MM-R1-3B	**47.9**	**44.9**	**67.2**	**24.4**	**41.9**	**44.1**	**56.5**	**43.9**	**34.4**	**57.2**	**46.2**
Δ *(Ours–Best Evaluated Open 3B)*	**+10.7**	**+18.1**	**+1.0**	**+6.2**	**+8.1**	**+5.6**	**+3.4**	**+9.3**	**+6.1**	**+2.0**	**+8.5**
7B-scale Models											
Qwen2.5-VL-7B-Instruct	*42.9*	25.1	68.2	*21.2*	36.2	*45.0*	58.6	38.3	*35.5*	62.1	43.3
VLAA-Thinker-7B	48.2	26.4	68.0	22.4	*29.2*	48.5	*54.6*	*41.6*	*36.1*	64.5	44.0
ThinkLite-7B-VL	*45.3*	32.9	75.1	*22.0*	*26.5*	*40.7*	55.5	*41.3*	*39.3*	63.4	44.2
OpenVLThinker-1.2-7B	*40.7*	25.9	72.3	*21.2*	*37.9*	*41.4*	58.7	42.9	*39.3*	62.9	44.3
VL-Rethinker-7B	*49.1*	32.3	74.9	27.4	*27.8*	*44.5*	56.7	41.7	*39.8*	62.7	45.7
MMR1-Math-v0	45.1	30.2	71.0	*25.2*	*33.2*	50.8	*57.1*	*43.2*	*39.3*	63.8	45.9
MM-Eureka-Qwen-7B	*49.6*	26.9	73.0	*24.0*	34.7	*46.8*	*57.3*	*43.3*	*39.5*	64.4	46.0
Vision-G1	*50.0*	31.3	**76.1**	*27.2*	29.0	50.2	53.4	41.2	*41.0*	**66.0**	46.5
Stable-MM-R1-7B	**56.3**	**53.1**	75.5	**39.3**	**49.6**	**55.3**	**67.2**	**52.3**	**47.7**	65.9	**56.2**
Δ *(Ours–Best Evaluated Open 7B)*	**+6.3**	**+20.2**	-0.6	**+11.9**	**+11.7**	**+4.5**	**+8.5**	**+9.0**	**+6.7**	-0.1	**+9.7**

Table 1: Performance comparison of various MLLMs on diverse multimodal reasoning benchmarks. **Note on Metrics:** We report Pass@8 (majority-vote) accuracy for all evaluated open-source models. Proprietary models are included as reference points and excluded from the within-scale rankings. Within each open-source model group (3B and 7B), the best results are highlighted in bold, and the second-best are underlined. MathVerse-V, DynaMath-W and WeMath-S denote the vision-only, worst, and strict settings, respectively.

Configuration	MathVista	MathVision	MMMU	Avg.
GRPO (Baseline)	70.6	50.1	61.5	60.7
DAPO (Strong Baseline)	71.8 (+1.2)	51.2 (+1.1)	62.9 (+1.4)	62.0 (+1.3)
+ PAQM (Ours)	73.5 (+2.9)	52.9 (+2.8)	65.1 (+3.6)	63.8 (+3.1)
+ HSR (Ours)	74.2 (+3.6)	53.6 (+3.5)	66.2 (+4.7)	64.7 (+4.0)
Full Stable-MM-R1	**75.5** (+4.9)	**54.7** (+4.6)	**67.2** (+5.7)	**65.8** (+5.1)

Table 2: **Ablation study.** Standard GRPO and DAPO use neither PAQM nor HSR. '+ PAQM' applies the full hint-augmented curriculum with the standard GRPO objective, '+ HSR' enables HSR without PAQM, and Full Stable-MM-R1 combines both components. The comparison between '+ PAQM' and Full isolates the contribution of HSR. Best results are **bold**.

mately 0.35–0.60 nats, and its reward stays near the late-training plateau through step 400. These trajectories show greater late-training stability for Stable-MM-R1 in this run, without implying monotonic improvement at every step.

4.3 Ablation Study

To dissect the contribution of each component, we analyze the impact of removing PAQM and HSR. Table 2 summarizes the results.

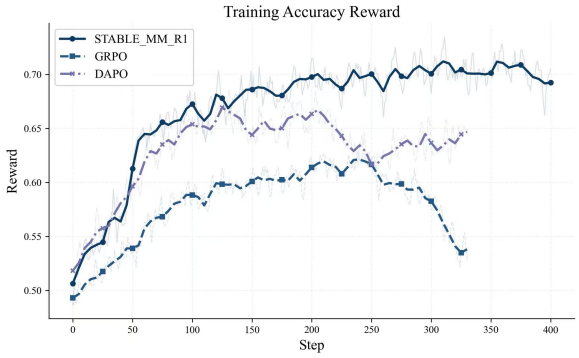
HSR vs. Hints. A key question is whether the gains stem solely from the Seed-1.5-VL hints in PAQM. The row '+ PAQM' represents GRPO trained on the same hint-augmented curriculum used by the full method. Comparing '+ PAQM' with 'Full Stable-MM-R1' therefore isolates a 2.0-point average gain from adding HSR under PAQM, showing that structured batch topology contributes beyond the hint-augmented curriculum alone.

Effectiveness of PAQM. As shown in Table 2, adding PAQM alone ('+ PAQM') improves the average score by 3.1 points over the GRPO baseline. This result is consistent with the benefit of concentrating updates on the learnable frontier.

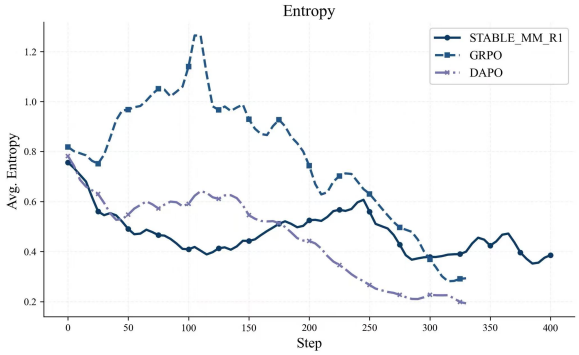
Impact of Hybrid Stratified Replay. Adding HSR without PAQM ('+ HSR') improves the average score by 4.0 points over GRPO, while combining PAQM and HSR yields the best ablation result (65.8 average). Figure 3 further shows that the full method continues to improve after the GRPO and DAPO curves peak; because the figure compares complete methods rather than individual HSR buffer components, we do not attribute this trend to any single buffer category.

5 Conclusion

In this paper, we showed that stochastic optimization-group composition can contribute to instability in RL training for reasoning. We proposed a unified data-centric framework comprising Potential-Aware Query Mining (PAQM) and Hybrid Stratified Replay (HSR). By dynamically curating learnable queries and structuring optimization groups with stability anchors and contrastive negatives, the full method mitigated entropy col-



(a) Accuracy dynamics.



(b) Path-Entropy evolution.

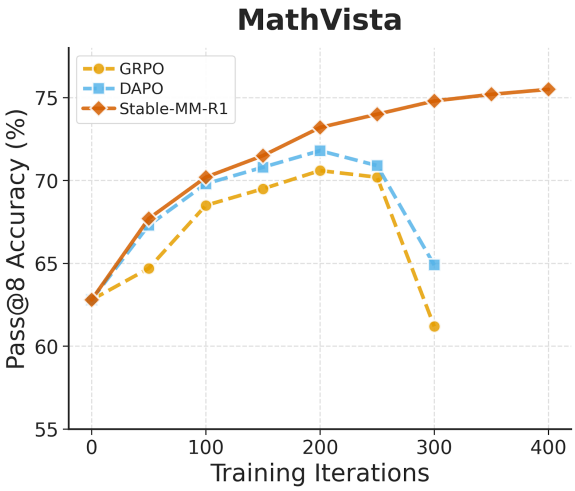
Figure 2: **Training stability analysis.** Accuracy (top) and mean Path Entropy (bottom) over training. Stable-MM-R1 maintains healthy exploration levels and avoids the collapse modes of the baselines.

lapse in our experiments.

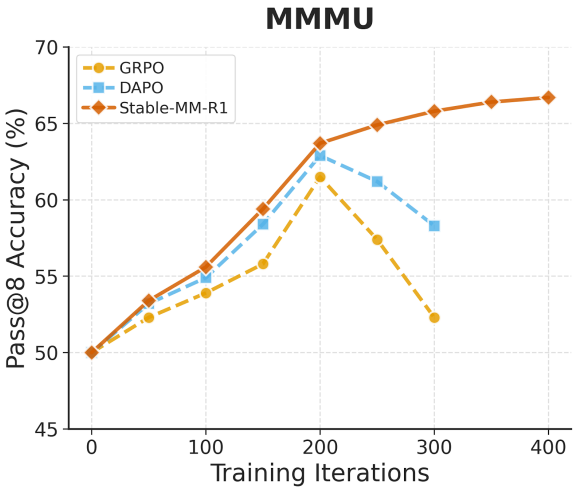
Beyond the performance gains, our results show that explicit contrast within the micro-batch—via anchors and hard negatives—can improve long-chain reasoning training without changing the underlying GRPO objective. This data-centric paradigm offers a complementary path for stabilizing future multimodal models.

Limitations

While Stable-MM-R1 improves stability, it relies on a proprietary teacher model (Seed-1.5-VL) for initial scaffolding, which may introduce distillation bias and accessibility constraints. Although our comparisons match update count and final group size, PAQM incurs additional offline teacher-generation and query-resampling costs, and HSR requires careful intra-step buffer management. Finally, our method currently focuses on verifiable reasoning; extending entropy-guided stratification to open-ended generation remains future work.



(a) MathVista Pass@8 over training.



(b) MMMU Pass@8 over training.

Figure 3: **Training dynamics comparison.** Pass@8 accuracy trajectories on (a) MathVista and (b) MMMU. While GRPO and DAPO suffer from performance collapse, Stable-MM-R1 maintains a robust upward trajectory.

Ethical Considerations

As this work introduces a framework for enhancing the reasoning capabilities of Multimodal Large Language Models (MLLMs) via Reinforcement Learning, we address several ethical implications regarding data selection, model behavior, and potential misuse.

Bias in Data Selection (PAQM). Our Potential-Aware Query Mining (PAQM) mechanism actively filters training data based on the model's current capability (Pass@K). While this improves efficiency, there is a theoretical risk of *algorithmic exclusion*. If the base model exhibits pre-existing biases

against specific linguistic dialects, cultural contexts, or visual concepts (classifying them as "Impossible"), PAQM might systematically exclude this data from the training loop, exacerbating the bias. We mitigate this by using diverse, standard academic benchmarks (STEM-focused), but we advise practitioners to audit the "Discarded" data partition when applying this method to sociologically sensitive domains.

Data Privacy and Content Safety. We inspected the training corpus to identify and mitigate potential risks related to Personally Identifiable Information (PII) and offensive content. Although our data sources (Geometry3K, MM-Eureka, and ViRL39K) primarily consist of public academic benchmarks in STEM domains where PII risks are relatively low, we employed automated keyword filtering and random-sample audits to reduce the risk of propagating sensitive personal data or toxic material into the RL fine-tuning stage.

Teacher-Student Alignment Risks. The use of teacher-generated hints (visual decompositions and reasoning nodes) facilitates rapid learning. However, this creates a dependency on the safety alignment of the teacher model. If the teacher model harbors latent jailbreak vulnerabilities or hallucinates harmful instructions, the student model may internalize these patterns via the high-reward gradient signal. Although our experiments focus on verifiable math and logic tasks where safety risks are minimal, applying Stable-MM-R1 to open-ended generation requires integrating safety-specific reward models (e.g., Rule-Based Reward Models or Constitutional AI) to prevent the reinforcement of hazardous content.

Dual Use and Automation. By stabilizing the reasoning capabilities of 7B-scale models, this work lowers the barrier for deploying powerful multimodal agents. While this democratizes access to advanced AI, it also increases the potential for misuse in automated tasks, such as solving CAPTCHAs or automating interaction in visual environments. We emphasize that the models trained in this paper are research prototypes intended for academic evaluation and should be subjected to rigorous red-teaming before real-world deployment.

Environmental Impact. RL fine-tuning is computationally intensive. Our experiments utilized 8× H100 GPUs. HSR is designed to reduce zero-signal optimization groups under a fixed step budget, although we do not report a full energy or carbon accounting. Data-centric efficiency remains an important direction for reducing the environmental cost of model post-training.

Disclosure of AI Assistance. We acknowledge the use of AI assistants for improving textual clarity and verifying code syntax. The authors reviewed the manuscript, scientific claims, and reported results and retain responsibility for the final content.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, and 4 others. 2025. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Anthropic. 2025. Introducing Claude 4. https://www.anthropic.com/research/claude-4.

Bizhe Bai, Hongming Wu, Peng Ye, and Tao Chen. 2025a. M-grpo: Stabilizing self-supervised reinforcement learning for large language models with momentum-anchored policy optimization. *arXiv preprint arXiv:2512.13070*.

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, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025a. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*. Introduces VLAA-Thinker.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Shuang Chen, Kaituo Feng, Hangting Chen, Wenxuan Huang, Dasen Dai, Quanxin Shou, Yunlong Lin, Xiangyu Yue, Shenghua Gao, and Tianyu Pang. 2026a. Opensearch-vl: An open recipe for frontier multimodal search agents. *arXiv preprint arXiv:2605.05185*.

Shuang Chen, Hangyu Guo, Yimeng Ye, Shijue Huang, Wenbo Hu, Jiayu Chen, Manyuan Zhang, Haoxi Li,

Song Guo, and Nanyun Violet Peng. 2026b. Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping. In *International Conference on Learning Representations*, volume 2026, pages 147474–147512.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. 2025b. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*.

Shuang Chen, Quanxin Shou, Hangting Chen, Yucheng Zhou, Kaituo Feng, Wenbo Hu, Yi-Fan Zhang, Yunlong Lin, Wenxuan Huang, Mingyang Song, et al. 2026c. Unify-agent: A unified multimodal agent for world-grounded image synthesis. *arXiv preprint arXiv:2603.29620*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. OpenVLThinker: Complex vision-language reasoning via iterative SFT-RL cycles. *Preprint*, arXiv:2503.17352.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. In *Advances in Neural Information Processing Systems*, volume 38, pages 99114–99137.

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. 2026a. Gen-searcher: Reinforcing agentic search for image generation. *arXiv preprint arXiv:2603.28767*.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. 2026b. One-thinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5432–5443.

Dong Guo et al. 2025. Seed1.5-VL technical report. *Preprint*, arXiv:2505.07062.

Tue Le, Linh Ngo Van, and Trung Le. 2025. Sharpness-guided group relative policy optimization via probability shaping. *arXiv preprint arXiv:2511.00066*.

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, Deli Zhao, Wei Lu, Yu Rong, Aixin Sun, and Shijian Lu. 2025. MMR1: Enhancing multimodal reasoning with variance-aware sampling and open resources. *arXiv preprint arXiv:2509.21268*.

Gang Li, Yulei Qin, Xiaoyu Tan, Dingkang Yang, Yuchen Shi, Zihan Xu, Xiang Li, Xing Sun, and Ke Li. 2025. Rorecomp: Enhancing reasoning efficiency via rollout response recomposition in reinforcement learning. *arXiv preprint arXiv:2509.25958*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

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

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. 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 (ACL)*, pages 6774–6786.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 7047–7058.

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. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

OpenAI. 2025. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`.

Long Ouyang, Jeffrey 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. 2022.

Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Chuming Shen, Wei Wei, Xiaoye Qu, and Yu Cheng. 2025. Satori-r1: Incentivizing multimodal reasoning through explicit visual anchoring. *arXiv preprint arXiv:2505.19094*.

Xinyu Tang, Yuliang Zhan, Zhixun Li, Wayne Xin Zhao, Zhenduo Zhang, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. 2025. Rethinking sample polarity in reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2512.21625*.

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, et al. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599. ArXiv preprint arXiv:2501.12599.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *Advances in Neural Information Processing Systems*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with the MATH-Vision dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 95095–95169.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025b. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024b. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. 2025a. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *Advances in Neural Information Processing Systems*, volume 38, pages 143297–143330.

Junkang Wu, Kexin Huang, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. 2025b. Quantile advantage estimation: Stabilizing RLVR for LLM reasoning. *arXiv preprint arXiv:2509.22611*.

Wenyi Xiao and Leilei Gan. 2025. Fast-Slow Thinking GRPO for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts. *Preprint*, arXiv:2407.04973.

Kai Yang, Xin Xu, Yangkun Chen, Weijie Liu, Jiafei Lyu, Zichuan Lin, Deheng Ye, and Saiyong Yang. 2025. Entropic: Towards stable long-term training of llms via entropy stabilization with proportional-integral control. *arXiv preprint arXiv:2511.15248*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. 2025. What's behind ppo's collapse in long-cot? value optimization holds the secret. *Preprint*, arXiv:2503.01491.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024a. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu,

Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024b. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Guanning Zeng, Zhaoyi Zhou, Daman Arora, and Andrea Zanette. 2025. Shrinking the variance: Shrinkage baselines for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2511.03710*.

Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. 2025. Vision-g1: Towards general vision language reasoning with multi-domain data curation. *arXiv preprint arXiv:2508.12680*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision (ECCV)*, pages 169–186. Springer.

Xiaoying Zhang, Hao Sun, Yipeng Zhang, Kaituo Feng, Chao Yang, and Helen Meng. 2025. Critique-grpo: Advancing llm reasoning with natural language and numerical feedback. *arXiv preprint arXiv:2506.03106*.

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

Lanyun Zhu, Deyi Ji, Tianrun Chen, Haiyang Wu, and Shiqi Wang. 2025a. Retrv-r1: A reasoning-driven mllm framework for universal and efficient multimodal retrieval. *arXiv preprint arXiv:2510.02745*.

Linghao Zhu, Yiran Guan, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Bin Qin, Jian Luan, Yuliang Liu, and Xiang Bai. 2025b. Shuffle-r1: Efficient rl framework for multimodal large language models via data-centric dynamic shuffle. *Preprint*, arXiv:2508.05612.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *Preprint*, arXiv:2411.00836.

A Appendix

A.1 Hyperparameter Sensitivity

We conducted a sensitivity analysis on the key hyperparameters of HSR: the stream split ratio (Structure/Natural) and the percentile thresholds (τ).

- **Stream Ratio:** We swept the Structured Stream ratio from 0% (Pure GRPO) to 100%. We observed that a ratio of 50% provides the optimal balance. Higher ratios (>75%) led to overfitting on the replay buffer, while lower ratios (<25%) failed to prevent entropy collapse.

- **Path-Entropy Thresholds:** We tested $\tau_{low} \in \{5, 10, 20\}$th percentile. Results indicate that stringent thresholds (5th percentile) overly restrict the anchor pool, leading to buffer depletion, while loose thresholds (20th percentile) allow noisy samples into the anchor set, degrading stability. The 10th/90th percentile setting proved robust across all examined benchmarks.

A.2 Algorithmic Implementation Details

In this section, we provide a step-by-step commentary on the training procedure outlined in Algorithm 1. The framework extends standard Group Relative Policy Optimization (GRPO) by introducing a dynamic curriculum (Phase 0) and a structured replay mechanism (Phase 1 & 2).

Initialization and Inputs. The algorithm requires a curated multimodal instruction dataset $\mathcal{D}$, the policy model π_θ to be optimized, and an offline hint cache $\mathcal{C}_{hint}$ generated in advance by a frozen proprietary teacher model π_{seed} (Seed-1.5-VL). For each query, we use a per-query rollout group size of $G = N = 16$. This group is subsequently partitioned between the structured and natural streams.

Phase 0: Dynamic Query Handling (Lines 3–12). Standard RL fine-tuning often wastes computational resources on queries that are currently intractable for the model.

- **Pilot Assessment:** For each sampled query x, we generate N pilot rollouts to estimate the empirical success rate $\hat{p}(x)$.

- **Intervention:** If $\hat{p}(x) = 0$ (Line 6), the query is classified as "Impossible." Rather than discarding these samples, we retrieve its precomputed visual or cognitive hint h from the offline cache $\mathcal{C}_{hint}$. The prompt is augmented ($x' \leftarrow x \oplus h$) during Phase 0 curriculum construction, and rollouts are then *regenerated* (Line 10). This mechanism converts selected "Impossible" queries into "Learnable" ones (the Distillation Zone) prior to gradient computation.

Phase 1: Path-Entropy-Based Stratification (Lines 13–24). Upon finalizing the rollouts for the current optimization step, we compute the reward r and Path-Entropy score $H_{\text{path}}(y|x)$ for every sample.

- **Dynamic Thresholding:** We pool the Path-Entropy scores of all rollouts generated for the current global batch and calculate their 10th and 90th percentiles $(\tau_{low}, \tau_{high})$. The thresholds are therefore batch-adaptive and use no samples from previous optimization steps.
- **Buffer Allocation:** Samples are stratified into three mutually exclusive buffers: $\mathcal{B}_{anc}$ (high confidence, correct), $\mathcal{B}_{neg}$ (high confidence, incorrect), and $\mathcal{B}_{epi}$ (low confidence, correct). Note that "Confusion" samples ($r = 0, H_{\text{path}} > \tau_{high}$) are implicitly filtered out, as they do not meet the criteria for any buffer.

Phase 2: Hybrid Group Assembly (Lines 25–32). To construct the final per-query optimization group G_{final}, we partition the $G = 16$ rollouts into two streams that balance stability and exploration:

- **Stream A (Structured):** We explicitly sample 50% of the total budget G from the sorted buffers. We enforce a composition of 25% Epiphanies (learning targets), 12.5% Anchors (stability), and 12.5% Hard Negatives (contrast) whenever the corresponding samples are available. This structured injection yields non-zero empirical reward variance when both outcome classes are present; otherwise, the unavailable quota falls back to natural on-policy sampling.
- **Stream B (Natural):** The remaining 50% is sampled uniformly from the valid sample pool (excluding noise). This stream acts as a regularizer, ensuring the policy does not overfit to the artificial topology of Stream A.

Optimization and Lifecycle (Lines 33–36). Finally, gradients are computed using the combined group G_{final}. A critical feature of our implementation is the Intra-Step Replay Scope (Line 36). Unlike off-policy methods with persistent replay buffers, our buffers $\mathcal{B}_{anc,neg,epi}$ are cleared immediately after every gradient step. This ensures that all optimization samples are drawn from the current policy π_θ, maintaining the on-policy consistency of the GRPO objective without requiring corrections for stale data.

Algorithm 1 Stable-MM-R1 Unified Training Framework

Require: Dataset $\mathcal{D}$; Current Policy π_θ; Offline Hint Cache $\mathcal{C}_{hint}$; Pilot Rollout Count and Final Group Size $N = G = 16$; Quantile thresholds $\delta_{low} = 0.1, \delta_{high} = 0.9$
Ensure: Optimized Policy parameters θ

1: **while** not converged **do**
2: Sample batch query $x \sim \mathcal{D}$
▷ **Phase 0: Potential-Aware Query Mining (PAQM)**
3: $\mathcal{Y}_{pilot} \leftarrow \{y_1, \ldots, y_N\} \sim \pi_\theta(\cdot|x)$ ▷ Generate pilot rollouts
4: $\hat{p}(x) \leftarrow \frac{1}{N}\sum_{i=1}^{N} r(y_i)$ ▷ Estimate empirical success rate
5: **if** $\hat{p}(x) = 0$ **then**
▷ *a. Apply Modality-Decoupled Diagnosis for Impossible Queries*
6: $h \leftarrow \textsc{LookupHint}(x, \mathcal{C}_{hint})$ ▷ Retrieve precomputed offline hint
7: $x' \leftarrow x \oplus h$ ▷ Inject hint during curriculum construction
8: $\mathcal{Y} \leftarrow \textsc{Resample}(\pi_\theta, x', N)$ ▷ Regenerate rollouts with hint
9: **else**
10: $\mathcal{Y} \leftarrow \mathcal{Y}_{pilot}$
11: **end if**
▷ **Phase 1: Diagnostic Path-Entropy–Reward Stratification**
12: $\mathcal{H}_t \leftarrow \{H_{\text{path}}(y|x) \mid y \in \mathcal{Y}_{\text{batch},t}\}$ ▷ Pool current global-batch rollouts
13: $\tau_{low} \leftarrow \textsc{Percentile}(\mathcal{H}_t, \delta_{low})$; $\tau_{high} \leftarrow \textsc{Percentile}(\mathcal{H}_t, \delta_{high})$
14: $\mathcal{B}_{anc} \leftarrow \emptyset$; $\mathcal{B}_{neg} \leftarrow \emptyset$; $\mathcal{B}_{epi} \leftarrow \emptyset$
15: **for all** $y \in \mathcal{Y}$ **do**
16: **if** $r(y) = 1 \wedge H_{\text{path}}(y|x) < \tau_{low}$ **then**
17: $\mathcal{B}_{anc} \leftarrow \mathcal{B}_{anc} \cup \{y\}$ ▷ Stability Anchor
18: **else if** $r(y) = 0 \wedge H_{\text{path}}(y|x) < \tau_{low}$ **then**
19: $\mathcal{B}_{neg} \leftarrow \mathcal{B}_{neg} \cup \{y\}$ ▷ Hard Negative
20: **else if** $r(y) = 1 \wedge H_{\text{path}}(y|x) > \tau_{high}$ **then**
21: $\mathcal{B}_{epi} \leftarrow \mathcal{B}_{epi} \cup \{y\}$ ▷ Epiphany (Target)
22: **end if**
23: **end for**
▷ **Phase 2: Hybrid Group Assembly**
▷ *b. Stream A: Structured Stabilization (50% Budget)*
24: $S_{epi} \leftarrow \textsc{Sample}(\mathcal{B}_{epi}, 0.25G)$
25: $S_{anc} \leftarrow \textsc{Sample}(\mathcal{B}_{anc}, 0.125G)$
26: $S_{neg} \leftarrow \textsc{Sample}(\mathcal{B}_{neg}, 0.125G)$
27: $G_{struct} \leftarrow S_{epi} \cup S_{anc} \cup S_{neg}$
▷ *c. Stream B: Natural Generalization (50% Budget)*
28: $G_{nat} \leftarrow \textsc{Sample}(\mathcal{Y} \setminus \text{Noise}, 0.5G)$
▷ **Optimization Update (Intra-Step Scope)**
29: $G_{final} \leftarrow G_{struct} \cup G_{nat}$
30: $\theta \leftarrow \text{GRPO_}\textsc{Update}(\theta, G_{final})$
31: $\text{Clear}(\mathcal{B}_{anc}, \mathcal{B}_{neg}, \mathcal{B}_{epi})$ ▷ Buffers are cleared immediately after step
32: **end while**

A.3 Conditional Variance Analysis of Structured Groups

This section analyzes the conditional effect of structured group composition in sparse-reward Group Relative Policy Optimization (GRPO). The analysis motivates HSR but does not assert a non-zero

update for every optimization step.

A.3.1 Degenerate Groups in Standard GRPO

Let $\mathcal{G} = \{y_1, \ldots, y_G\}$ be a group of outputs sampled from policy $\pi_\theta(\cdot|x)$. The GRPO gradient estimator is defined as:

$$\hat{g}_{GRPO} = \frac{1}{G} \sum_{i=1}^{G} \frac{r(y_i) - \bar{r}}{\hat{\sigma} + \epsilon} \nabla_\theta \log \pi_\theta(y_i|x), \quad (3)$$

where $\bar{r}$ and $\hat{\sigma}$ are the empirical mean and standard deviation of rewards within the group.

Degenerate-group probability. In a binary reward setting $r \in \{0, 1\}$, let $p_\theta(x)$ be the true success rate of the policy. The probability of sampling a *degenerate group*, in which all rewards are identical, is

$$P_{degen} = p_\theta(x)^G + (1 - p_\theta(x))^G. \quad (4)$$

In such groups, $\hat{\sigma} = 0$, so the group-relative advantages vanish or become numerically ill-conditioned depending on the implementation. When $p_\theta(x)$ is close to 0 or 1, P_{degen} increases, making zero-signal groups more likely under random sampling.

A.3.2 Conditional Variance of Structured Groups

Conditional on the relevant buffers being non-empty, HSR's structured stream includes at least k_{anc} positive samples (Stability Anchors) and k_{neg} negative samples (Hard Negatives). Let G_{struct} be the structured-group size.

Conditional lower bound. If $k_{anc} \geq 1$ and $k_{neg} \geq 1$, the empirical reward variance of the structured group is strictly positive. Let $N_+ \geq k_{anc}$ and $N_- \geq k_{neg}$ denote the numbers of positive and negative samples, with $N_+ + N_- = G_{struct}$. Then

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{G_{struct}} \left[N_+(1 - \bar{r})^2 + N_-(0 - \bar{r})^2 \right] \\ &= \frac{N_+ N_-}{G_{struct}^2} > 0. \end{aligned} \quad (5)$$

Thus, the structured group has non-zero reward variance whenever both outcome classes are available. If either quota cannot be satisfied, HSR releases the unavailable structured allocation and fills it through natural on-policy sampling. The bound is therefore conditional rather than a guarantee for every optimization step.

A.3.3 Distribution-Shift Trade-off

Structured sampling changes the empirical training distribution by emphasizing anchors, hard negatives, and high-Path-Entropy successes. We treat this as a deliberate distribution shift toward informative boundary cases rather than claiming that the resulting estimator is unbiased. Stream B retains natural on-policy samples within every optimization step, while the empirical training-dynamics results evaluate whether the resulting trade-off improves stability in practice.