Targeted Exploration via Unified Entropy Control for Reinforcement Learning

Anonymous ACL submission

Abstract

Recent advances in reinforcement learning (RL) have improved the reasoning capabilities of large language models (LLMs) and vision-language models (VLMs). However, the widely used Group Relative Policy Optimization (GRPO) consistently suffers from entropy collapse, causing the policy to converge prematurely and lose diversity. Existing exploration methods introduce additional bias or variance during exploration, making it difficult to maintain optimization stability. We propose Unified Entropy Control for Reinforcement Learning (UEC-RL), a framework that provides targeted mechanisms for exploration and stabilization. UEC-RL activates more exploration on difficult prompts to search for potential and valuable reasoning trajectories. In parallel, a stabilizer prevents entropy from growing uncontrollably, thereby keeping training stable as the model consolidates reliable behaviors. Together, these components expand the search space when needed while maintaining robust optimization throughout training. Experiments on both LLM and VLM reasoning tasks show consistent gains over RL baselines on both Pass@1 and Pass@k. On Geometry3K, UEC-RL achieves a 37.9% relative improvement over GRPO, indicating that it sustains effective exploration without compromising convergence and underscoring UEC-RL as a key principle for scaling RL-based reasoning in large models.

1 Introduction

Reinforcement learning (RL) has become a central paradigm in the post-training of large language models (LLMs) and vision-language models (VLMs) (GLM et al., 2024; Touvron et al., 2023). Early RLHF methods, such as Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO), align model outputs with human preferences using preference-based rewards (Schulman et al., 2017b; Rafailov et al., 2023; Zhong et al., 2024; Wang et al., 2024b). More recently, Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a scalable alternative by leveraging automatically verifiable supervision (Mroueh, 2025). Within this framework, Group Relative Policy Optimization (GRPO) was introduced as a lightweight PPO variant that removes the critic and estimates advantages via group-normalized rewards, achieving strong efficiency and competitive reasoning performance (Shao et al., 2024; Liu et al., 2024; Guo et al., 2025).

Exploration in RL is the process of guiding the policy to observe sufficiently diverse and informative samples during training, so that optimization can operate over a broader solution space instead of collapsing to suboptimal behaviors early (Sutton et al., 1998; Auer et al., 2002; Strehl and Littman, 2008; Kolter and Ng, 2009). **Entropy** quantifies the policy's uncertainty during inference and is commonly used as a proxy for exploration (Schulman et al., 2017a; Haarnoja et al., 2018; Nachum et al., 2017), yet recent studies show that GRPO suffers from entropy collapse, where policy entropy rapidly drops and responses become highly convergent, severely limiting the discovery of potential and valuable trajectories (Yue et al., 2025; Yu et al., 2025). DAPO slows entropy decay via a clip-higher strategy, but often at the cost of increased update variance and unstable training (Yu et al., 2025). Other methods introduce entropy bonuses or modified advantages (Cui et al., 2025b; Cheng et al., 2025), but the resulting exploration is frequently biased, as these approaches explicitly optimize entropy-related terms rather than task rewards. **Moreover, effective exploration should emphasize informative and high-quality diversity, rather than indiscriminately encouraging randomness**. This requires mechanisms that can suppress excessively high-entropy behaviors once

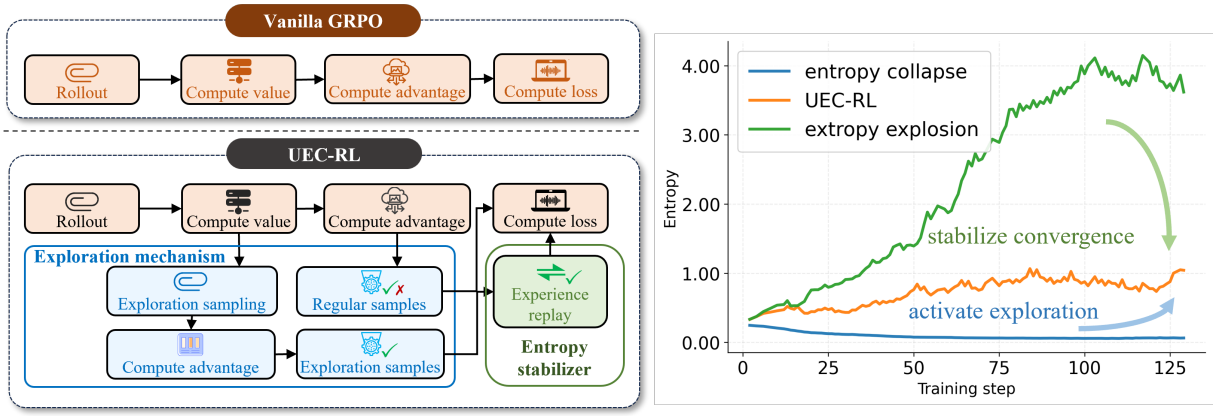


Figure 1: Illustration of UEC-RL. UEC-RL balances exploration and stabilization, keeping entropy within an optimal operating range.

exploration becomes unproductive, a capability that is largely missing in existing approaches. Overall, the field still lacks a principled mechanism for controlling entropy in both directions, exploration and stabilization.

To address this gap, we introduce UEC-RL, a framework that integrates exploration and stabilization within a single, coherent mechanism. Rather than merely slowing entropy collapse, UEC-RL actively adjusts the degree of exploration according to problem difficulty, enabling entropy to increase when deeper reasoning is required and allowing the model to access low-probability but informative trajectories that standard sampling often misses. At the same time, UEC-RL incorporates a stabilizing mechanism that restrains uncontrolled entropy growth, reinforces reliable behaviors, and guides the policy toward stable convergence as learning progresses. Through this coordinated design, UEC-RL dynamically balances exploration and exploitation throughout training.

Experiments across a wide range of LLM and VLM reasoning benchmarks show that UEC-RL delivers consistent gains over RL baselines. On the challenging Geometry3K dataset, it achieves a 37.9% relative improvement in accuracy while maintaining more appropriate training dynamics. UEC-RL also provides robust improvements on Pass@k, demonstrating that UEC-RL is an effective principle for advancing RL-based reasoning in large-scale models. Our contributions are summarized as follows:

- We introduce UEC-RL, which provides principled bidirectional entropy regulation, enabling both controllable entropy increase for deep exploration and controllable entropy stabilization for reliable training. As shown in Fig. 1, UEC-RL includes:
  - A targeted exploration mechanism that activates high-entropy reasoning specifically on difficult problems, allowing the model to uncover low-probability but informative trajectories that standard sampling rarely reaches.
  - A controllable entropy stabilizer that amplifies reliable gradients, suppresses unbounded exploration, and guides the policy toward stable convergence.
- We demonstrate consistent improvements across LLM and VLM reasoning benchmarks, including strong gains on Geometry3K and Pass@k evaluations, confirming the effectiveness and generality of the proposed entropy-control paradigm.

2 Related Work

Recent advances in aligning large language models (LLMs) and vision-language models (VLMs) have been driven by reinforcement learning techniques (OpenAI, 2023; Team et al., 2024; Zhu et al., 2023; Wei et al., 2023; Liu et al., 2023; Team et al., 2025; Yang et al., 2025), most notably RLHF and policy optimization methods such as DPO, PPO, and GRPO (Rafailov et al., 2024; Ouyang et al., 2022; Schulman et al., 2017b; Shao et al., 2024). More recently, Reinforcement Learning with Verifiable Rewards (RLVR) has shown strong performance in reasoning-intensive domains by removing learned reward models and relying on structured, verifiable supervision (Lambert et al., 2024; Guo et al., 2025). In this setting, optimization efficiency and training

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, O=\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|q)}$$
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}(r_{i,t}(\theta),1-\epsilon,1+\epsilon)\hat{A}_{i,t}\right]-\beta D_{KL}[\pi_\theta||\pi_{ref}]\right\},$$
$$\text{where } r_{i,t}(\theta)=\frac{\pi_\theta(o_{i,t}|q,o_{i<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i<t})}, \text{ and } \hat{A}_{i,t}=\frac{R_i-\text{mean}(\{R_i\}_{i=1}^{G})}{\text{std}(\{R_i\}_{i=1}^{G})}. \quad (1)$$

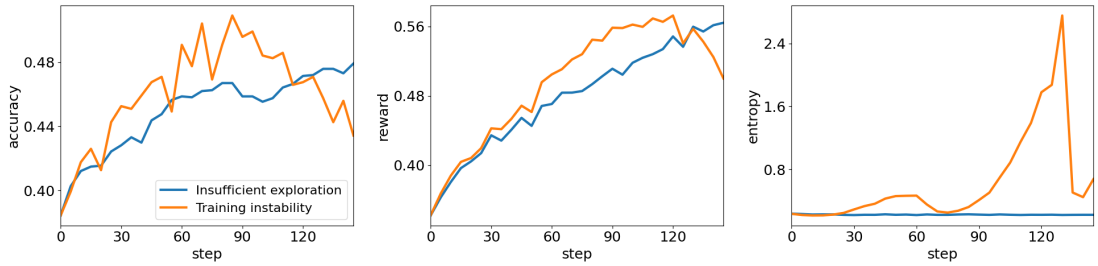


Figure 2: Two prominent optimization issues of GRPO on Geometry3K: insufficient exploration (entropy collapse) and unstable training dynamics.

dynamics become critical, particularly for GRPO-based methods that normalize rewards within rollout groups.

Exploration is a core component of reinforcement learning, enabling policies to escape local optima and discover high-reward behaviors (Sutton et al., 1998). In policy gradient methods, entropy is commonly used as a proxy for exploration to encourage behavioral diversity (Schulman et al., 2017b). However, recent studies show that GRPO-based training often suffers from entropy collapse, resulting in highly convergent samples and insufficient exploration (Yue et al., 2025; Yu et al., 2025). Existing remedies can be broadly grouped into two categories. The first modifies policy updates via clipping strategies (Yu et al., 2025; Hao et al., 2025; Su et al., 2025). By relaxing the upper clipping bound, these methods can slow entropy contraction, but often amplify update variance and destabilize training. The second category promotes exploration through entropy bonuses, such as entropy regularization or entropy-shaped advantages (Adamczyk et al., 2025; Li et al., 2025; Hou et al., 2025; Tan and Pan, 2025; Cui et al., 2025b; Cheng et al., 2025). While effective at increasing diversity, these approaches rely on coarse regularization and may introduce optimization bias, since exploration is driven by entropy-related objectives rather than task rewards.

3 Preliminaries

3.1 RL baseline: GRPO

GRPO has been widely used to improve the reasoning capabilities of large language models, particularly in mathematical problem-solving. Unlike PPO, GRPO removes the critic and estimates advantages using group-normalized rewards, resulting in significantly improved computational efficiency. Given a question q with a verifiable answer a, GRPO samples G responses $O = \{o_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$ and updates the policy by maximizing the group-relative objective shown in Eq. (1). By normalizing rewards within each sampled group, GRPO provides a lightweight and scalable alternative to PPO and has demonstrated strong performance across diverse reasoning benchmarks.

3.2 Limitations of GRPO

Despite its empirical success, GRPO presents notable shortcomings when applied to complex reasoning tasks. Fig. 2 illustrates two representative failure modes on the Geometry3K dataset, which were observed under identical settings, revealing GRPO's difficulty in maintaining adequate exploration and stable optimization dynamics.

Entropy collapse. As illustrated in Fig. 2, policy entropy often decreases rapidly during training, causing the model to converge prematurely to low-diversity behaviors. This collapse is especially common in text-only reasoning, where train-

Algorithm 1 UEC-RL

Input: Dataset $\mathcal{D}$, policy π_θ.
Hypers: $G, G', t', s', f_{\text{replay}}, A_0$.
Init queue $\mathcal{B}_{\text{replay}}$ by size s';
repeat
  Sample minibatch $\mathcal{B}_{\text{data}} \subset \mathcal{D}$;
  for each $(q, a) \in \mathcal{B}_{\text{data}}$ **do**
    Sample $O \leftarrow \{o_i\}_{1:G} \sim \pi_{\theta_{\text{old}}}$;
    Compute $R_i, \hat{A}_{i,t}$ of O;
    if $\max_i R_i > 0$ **then**
      $O_R \leftarrow \{o_i \in O : \hat{A}_{i,t} \neq 0\}$;
    else
      Sample $O' \leftarrow \{o_i\}_{1:G'} \sim \pi^{t'}_{\theta_{\text{old}}}$;
      Compute $R_i, \hat{A}_{i,t}$ of O';
      $O_H \leftarrow \{o_i \in O' : \hat{A}_{i,t} > 0\}$;
      $\mathcal{O}_S \leftarrow O_H \cup \{o_i \in O : \hat{A}_{i,t} > A_0\}$;
      Push $\mathcal{O}_S$ into $\mathcal{B}_{\text{replay}}$;
    end if
  end for
  $\mathcal{O}_{\text{eff}} \leftarrow \bigcup O_R \cup \bigcup O_H$;
  Compute $\pi_{\theta_{old}}$ of $\mathcal{O}_{\text{eff}}$;
  Update actor using $\mathcal{O}_{\text{eff}}$;
  if global_step $\bmod f_{\text{replay}} = 0$ **then**
    Sample $\mathcal{O}_S \subset \mathcal{B}_{\text{replay}}$ and update actor;
  end if
until convergence

ing is relatively stable and the absence of entropy-increasing mechanisms prevents the model from exploring low-probability but informative trajectories. Consequently, the model activates only shallow reasoning patterns while deeper knowledge remains underutilized.

Training instability. A second failure mode commonly appears in more complex settings such as multimodal reasoning, as illustrated in Fig. 2, where optimization exhibits sharp fluctuations. When sampled outputs vary greatly in correctness or reasoning difficulty, the group-normalized reward used by GRPO no longer provides sufficient variance reduction. As a result, gradient updates become brittle, causing unstable learning dynamics and occasionally leading to sudden performance degradation or even collapse.

These two issues share a structural root cause: GRPO lacks a principled mechanism for regulating entropy in either direction. It cannot actively increase entropy to enhance exploration on difficult prompts, nor can it stabilize entropy in high-variance regimes to ensure reliable convergence. This limitation motivates the development of a unified framework capable of dynamic, bidirectional entropy control.

4 Methodology

To address the entropy collapse and instability issues observed in GRPO, we introduce **UEC-RL**, which enables both controlled entropy increase for exploration and entropy reduction for stable convergence. This section first presents the targeted exploration mechanism, which adaptively activates high-entropy reasoning on difficult prompts, followed by the entropy-reducing replay mechanism that stabilizes learning and improves sample efficiency. The overall UEC-RL training procedure is summarized in Algorithm 1.

4.1 Targeted Exploration Mechanism

UEC-RL enhances GRPO's limited exploratory capacity through a unified mechanism that (1) identifies prompts requiring deeper search, (2) expands the sampling space only when necessary, and (3) selectively retains informative trajectories.

Expanding the exploration space. If none of the initial G rollouts solve a prompt, the prompt $\bar{\mathcal{D}}$ is marked as difficult, indicating insufficient exploration under the current policy.

$$\bar{\mathcal{D}} = \Big\{(q,a) \ : o_i \sim \pi_{\theta_{\text{old}}},\ i = 1, \ldots, G, \\ \#\{\, i : \ R(a, o_i) > 0 \,\} = 0\Big\}.$$

For such prompts, UEC-RL temporarily samples from a softened distribution with temperature t', increasing the chance of uncovering low-probability but informative reasoning paths while leaving easy prompts unaffected.

Exploring informative trajectories. From all collected samples, UEC-RL retains only two types of trajectories: regular samples O_R with nonzero advantage and valuable samples O_H obtained under expanded sampling:

$$O_R = \Big\{\{o_i\}_{1:G} \sim \pi_{\theta_{old}} : (q,a) \sim \mathcal{D},\ \hat{A}_{i,t} \neq 0\Big\};$$
$$O_H = \Big\{\{o_i\}_{1:G'} \sim \pi^{t_a}_{\theta_{old}} : (q,a) \sim \bar{\mathcal{D}}, \hat{A}_{i,t} \geq 0\Big\}.$$

Low-advantage exploratory samples are filtered out because they tend to introduce noisy gradients that hinder optimization and generalization. These retained samples constitute the effective optimization

Table 1: Comparison of Pass@1 accuracy on text and multimodal reasoning benchmarks. UEC-RL consistently outperforms the RL baselines. For AIME24 and AIME25, each question is repeated 32 times.

Benchmarks	**AIME24**	**AIME25**	**MATH**	**GSM8K**	**Minerva**	**Average**
Qwen2.5-math-7B	15.2	5.39	65.5	65.4	47.3	39.76
GRPO	25.8	9.27	77.6	87.1	29.0	45.75
DAPO	24.3	8.54	78.3	87.6	34.2	46.59
Entropy-Adv	26.7	9.90	78.9	86.8	35.0	47.46
URC-RL	**28.5**	**10.7**	**80.4**	**87.9**	**35.7**	**48.64**
Δ vs. GRPO	**+2.7**	**+1.4**	**+2.8**	**+0.8**	**+6.7**	**+2.89**

Benchmarks	**MathVision**	**MathVerse**	**MathVista**	**We-Math**	**Average**
Qwen2.5-VL-7B-Instruct	24.87	43.83	66.30	62.87	49.47
GRPO	**29.11**	47.51	72.60	67.53	54.19
DAPO	27.92	48.48	72.30	69.08	54.45
Entropy-Adv	27.86	48.63	71.80	68.62	54.23
UEC-RL	28.82	**49.34**	**73.40**	**69.48**	**55.26**
Δ vs. GRPO	**-0.29**	**+1.83**	**+0.80**	**+1.95**	**+1.07**

set and enable exploration to be increased in a targeted manner.

By integrating difficulty detection, adaptive search expansion, and selective retention into a coherent procedure, UEC-RL activates exploratory behavior precisely where deeper reasoning is required. The mechanism is theoretically supported by the following result:

Theorem 4.1 (Entropy Change)**.** *For a softmax policy updated by natural policy gradient with step size η,*

$$H(\pi_\theta^{k+1}) - H(\pi_\theta^k) \approx -\eta\, \mathbb{E}_{s\sim d_\mu^k} \mathrm{Cov}_{a\sim\pi_\theta^k(\cdot|s)} \left[\log \pi_\theta^k(a|s),\ A^{\pi^k}(s,a)\right].$$

This theorem was first introduced by Liu (2025), and was organized and extended by Cui et al. (2025a). Proof can be seen in Liu (2025) and Cui et al. (2025a). H indicates the policy entropy of policy model, and Cov denotes covariance, π_θ^k is the policy at step k, and $A^{\pi^k}(s,a)$ is the advantage function of action a under state s. The covariance term becomes negative when high-advantage actions receive low probability under the current policy. In such cases, updates increase the policy entropy. Using an elevated temperature $t' > 1$ amplifies this effect by further reducing the gap between high- and low-probability actions, making negative covariance more likely. As a result, UEC-RL induces controlled entropy increase specifically on difficult prompts, allowing the model to escape collapsed regimes and explore deeper reasoning trajectories that standard GRPO would fail to reach.

4.2 Controllable Entropy Stabilizer

Targeted exploration allows the policy to increase entropy on difficult prompts, but a complementary mechanism is required to prevent uncontrolled entropy growth and ensure convergence. UEC-RL introduces a controllable entropy stabilizer that repeatedly reinforces high-quality trajectories discovered during exploration.

Positive-advantage trajectories found under expanded sampling often have low initial probability and thus limited influence when used only once. Revisiting them strengthens their gradients and shifts probability mass toward correct reasoning patterns, producing a stabilizing effect on entropy.

Theorem 4.2 (Entropy Stabilization)**.** *Let (q,o) be a trajectory with $A(q,o) > 0$. If one update increases its log-likelihood,*

$$\log \pi_\theta^k(o \mid q) > \log \pi_\theta^{k-1}(o \mid q),$$

then repeating this update (e.g., via replay) yields

$$H(\pi_\theta^{k+1}) - H(\pi_\theta^k) < 0.$$

Proof. Because $A(q,o) > 0$, raising $\pi_\theta(o \mid q)$ aligns high-advantage actions with high probability, producing a positive covariance term in Theorem 4.1. A positive covariance leads to decreasing entropy. □

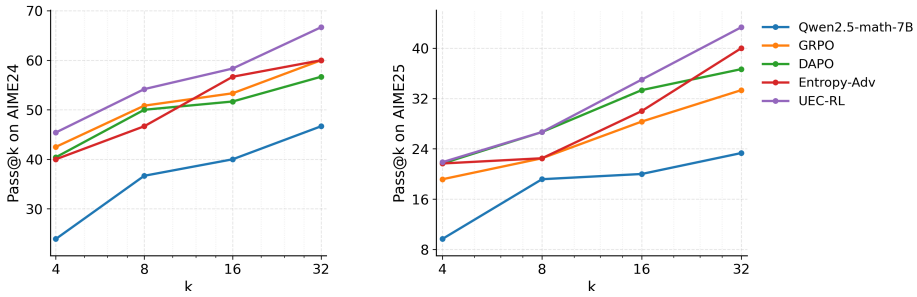


Figure 3: Pass@k performance on the AIME24 and AIME25 benchmarks. UEC-RL consistently improves the success rate across different values of k.

Geometry3K	Qwen2.5-VL-7B-Instruct	UEC-RL	GRPO	DAPO	Entropy-Adv
Accuracy	38.44	**55.41**	50.75	49.09	50.91
Δ vs. UEC-RL	-	-	-4.66	-6.32	-4.50

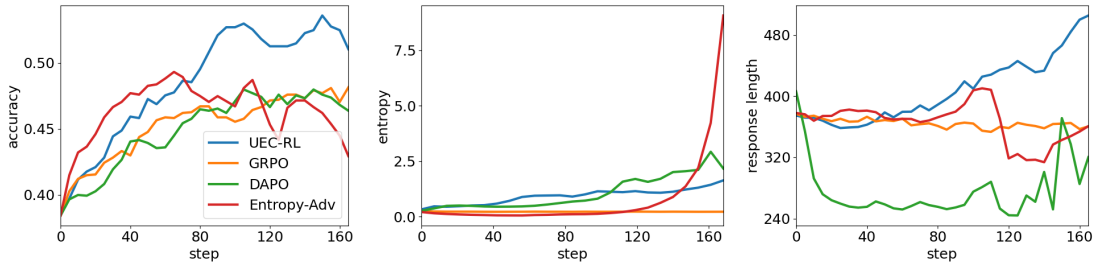


Figure 4: On the Geometry3K dataset, UEC-RL not only achieves excellent performance but also demonstrates superior exploration capability and training stability.

Thus, exploration enlarges entropy only when needed, and the stabilizer gradually decreases entropy by consolidating informative trajectories. This interplay transitions training from exploration to stable convergence, avoiding both entropy collapse and divergence. Concretely, we form a candidate set from both regular and exploratory samples ($O_R \cup O_H$), filter trajectories whose advantages exceed a threshold A_0, and keep only the most recent s' trajectories to prioritize up-to-date behaviors:

$$\mathcal{O}_S = \mathrm{Recent}_{s'}\Big(\{\, o_i \in O_R \cup O_H : \hat{A}_{i,t} > A_0 \,\}\Big),$$

where $\mathrm{Recent}_{s'}(\cdot)$ returns the s' most recent trajectories in generation order.

5 Experiments

We evaluate UEC-RL on both text-based and multimodal mathematical reasoning tasks. Our implementation is built upon EasyR1 and VeRL (Yaowei et al., 2025; Sheng et al., 2025).

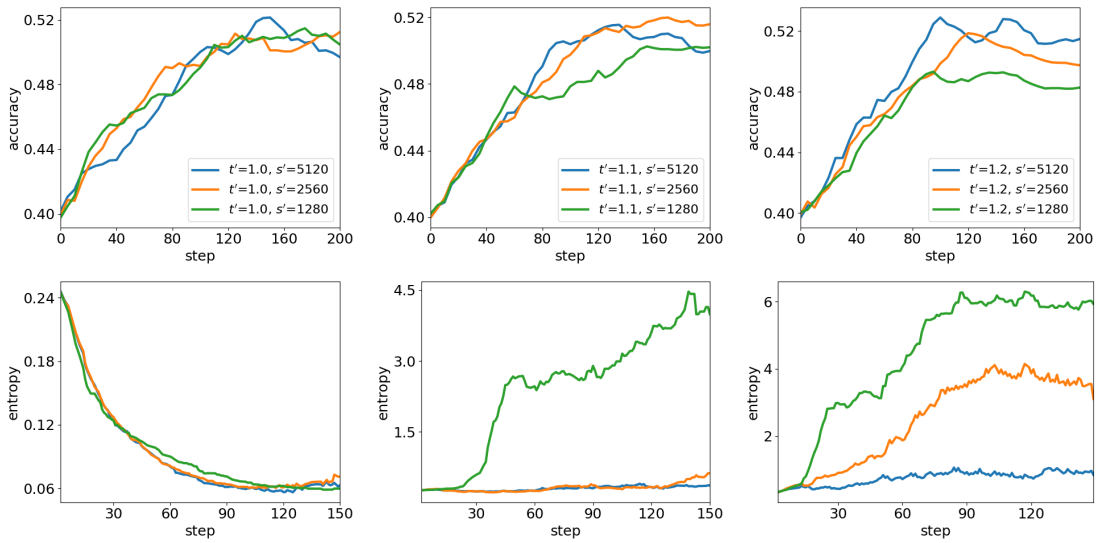
Datasets and benchmarks. We train UEC-RL on three datasets spanning both text-only and multimodal reasoning, and evaluate it on a comprehensive suite of text and multimodal benchmarks, with Geometry3K additionally used for in-domain analysis of training dynamics and ablations. Full details of the training datasets and evaluation benchmarks are provided in Appendix A.

Baseline. For comparison, we include three representative RL baselines. Additional implementation details are provided in Appendix B.

- GRPO (Shao et al., 2024), the most widely adopted RL baseline;
- DAPO (Yu et al., 2025), which enhances exploration through the clip-higher mechanism;
- Entropy-Adv (Cheng et al., 2025), which encourages exploration by augmenting the advantage with an entropy term.

6 Main Results

Table 1 summarizes Pass@1 performance on text and multimodal reasoning benchmarks. UEC-RL



Accuracy (Entropy)	$t' = 1.0$	$t' = 1.1$	$t' = 1.2$
$s' = 5120$	53.74 (0.09)	52.41 (0.31)	**55.41** (0.73)
$s' = 2560$	52.75 (0.09)	54.41 (0.31)	52.75 (2.37)
$s' = 1280$	52.41 (0.09)	51.08 (2.44)	50.25 (4.29)

Figure 5: Ablation of parameter tuning: peak accuracy and average entropy under varying t' and s'. Higher t' boosts exploration (entropy↑), larger s' aids stabilization (entropy↓), and best performance arises from the balance state of entropy.

consistently outperforms GRPO and DAPO across almost all datasets. On text reasoning benchmarks, UEC-RL achieves the best results on AIME24/25, MATH, GSM8K, and Minerva, leading to the highest overall average. The improvements further extend to Pass@k evaluation (Figure 3), where UEC-RL yields consistently stronger curves, indicating that the learned policy produces both higher-quality and more diverse reasoning trajectories.

On multimodal benchmarks, UEC-RL again achieves the highest average score, outperforming GRPO and DAPO on four multimodal benchmarks. These results show that UEC-RL remains effective even in multimodal settings, where training is typically less stable, and demonstrate that UEC-RL generalizes reliably to visually grounded reasoning tasks.

We additionally conduct an in-domain analysis on Geometry3K. As shown in Figure 4, UEC-RL achieves a substantial accuracy gain (**55.41** vs. 50.75 for GRPO), while also exhibiting more stable entropy dynamics and a smoother training trajectory. Geometry3K is a visually grounded mathematical dataset, and training on such multimodal tasks is typically far less stable than on text-only reasoning. Existing exploration strategies designed for LLMs often fail to transfer to VLMs because visual reasoning amplifies gradient variance and makes entropy harder to control. In contrast, the bidirectional entropy-control mechanism in UEC-RL enables effective exploration without destabilizing training, as reflected in both the entropy and response-length curves. These results confirm that UEC-RL successfully balances exploration and convergence even in challenging multimodal settings where conventional methods struggle.

7 Ablation Study

We conduct ablation experiments to study how the two key components of UEC-RL, namely the targeted exploration mechanism and the entropy-reducing stabilizer, contribute to training effectiveness. Our analysis consists of two parts: parameter tuning, which examines how exploration strength and stabilization capacity influence entropy dynamics, and module-level ablations, which isolate the effect of each component.

7.1 Parameter tuning.

To characterize how exploration and stabilization jointly determine model behavior, we vary the ex-

Geometry3K	UEC-RL	w/o exploration	w/o stabilizer
Accuracy	**55.41**	50.41 (-5.00)	49.58 (-5.93)

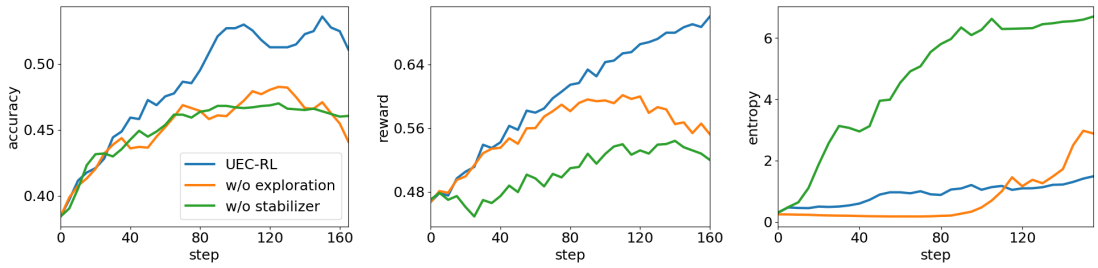


Figure 6: Module ablation of UEC-RL. Removing either exploration or stabilizer consistently degrades performance, highlighting their complementary roles in improving accuracy.

ploration temperature t' and the stabilizer budget s'. As shown in Figure 5, increasing t' enlarges the exploration space and raises policy entropy, helping the model discover deeper reasoning chains on difficult prompts. This trend is consistent with Theorem 4.1, which predicts entropy increase when exploration encourages low-probability actions.

When t' becomes large and s' is insufficient, entropy grows rapidly and accuracy degrades. In contrast, increasing s' strengthens the influence of high-quality trajectories and gradually stabilizes entropy, matching the entropy reduction effect described in Theorem 4.2. The best performance appears when exploration and stabilization operate in balance. These results confirm that exploration enables the model to escape shallow local optima, while stabilization consolidates correct behaviors and maintains entropy within a desirable range.

7.2 Module ablations.

We further disable each component to isolate its contribution. Removing the exploration module suppresses entropy growth on difficult prompts and restores entropy collapse, leading to a reduction of 5.00 accuracy points. Removing the stabilizer allows entropy to grow excessively and causes unstable reward dynamics, resulting in a reduction of 5.93 accuracy points. Figure 6 illustrates that exploration provides the necessary entropy increase to activate deeper reasoning, whereas the stabilizer prevents uncontrolled entropy drift and ensures convergence. Only the full UEC-RL framework exhibits both steady entropy regulation and consistent performance gains.

8 Conclusion

In this work, we presented UEC-RL, which addresses entropy collapse and provides a principled mechanism for bidirectional entropy regulation in RL for large language models. Unlike exploration heuristics that only partially alleviate premature convergence, UEC-RL treats entropy as an explicit optimization objective and regulates it on-policy throughout training, with the goal of preserving policy diversity while maintaining stable learning dynamics.

Empirically, UEC-RL delivers consistent improvements over strong RL baselines across both text-only and multimodal reasoning benchmarks. It improves Pass@1 performance while also strengthening Pass@k under sampling, indicating not only higher single-trajectory accuracy but also a more diverse and reliable policy distribution. These results support the central claim that effective reasoning improvements require jointly promoting exploration and stabilizing optimization, rather than relying on one-sided clip-higher or entropy bonus.

Looking forward, UEC-RL can be extended in several directions. More accurate difficulty estimation could better identify when entropy should be increased, improving exploration efficiency. Adaptive scheduling of the exploration temperature and stabilizer budget could reduce task-specific tuning while maintaining the desired entropy regime. Finally, integrating UEC-RL with multi-step verification or agent-based reasoning systems may allow entropy control at finer granularity, guiding step-level branching and consolidation to further improve reliability and scalability in complex reasoning tasks.

Ethical consideration

AI is only used for translation and language polishing in this paper.

Limitations

While UEC-RL enables stable and controllable exploration, its effectiveness depends on selecting appropriate values for the exploration temperature t' and stabilizer budget s'. Together, these hyperparameters determine the entropy range maintained during training. However, the entropy level that yields optimal performance is typically moderate, for example, around 0.5, and achieving it often requires task-specific hyperparameter configurations.

This variability arises because tasks differ substantially in difficulty: harder datasets require stronger exploration to escape local optima, whereas easier or lower-variance tasks benefit from tighter stabilization. Consequently, the (t', s') combination that preserves a desirable entropy regime is not universal but instead depends on the intrinsic difficulty and variance structure of the training set. This makes UEC-RL relatively sensitive to hyperparameter choices, and achieving consistent performance across domains may require task-specific tuning.

Developing adaptive or self-regulating strategies that automatically calibrate (t', s') based on task difficulty remains an important direction for future research.

References

Jacob Adamczyk, Volodymyr Makarenko, Stas Tiomkin, and Rahul V Kulkarni. 2025. Average-reward reinforcement learning with entropy regularization. *arXiv preprint arXiv:2501.09080*.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256.

Jie Cao and Jing Xiao. 2022. 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.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.

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. *arXiv preprint arXiv:2110.14168*.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025a. The entropy mechanism of reinforcement learning for reasoning language models. *Preprint*, arXiv:2505.22617.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025b. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr.

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. 2025. Rethinking entropy interventions in rlvr: An entropy change perspective. *arXiv preprint arXiv:2510.10150*.

Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. 2025. Advancing language model reasoning through reinforcement learning and inference scaling. *arXiv preprint arXiv:2501.11651*.

HuggingFaceH4. 2025. AIME 2024 Dataset (AIME I & II). https://huggingface.co/datasets/HuggingFaceH4/aime_2024.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

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

Daesik Kim, Seonhoon Kim, and Nojun Kwak. 2018. Textbook question answering with multimodal context graph understanding and self-supervised open-set comprehension. *arXiv preprint arXiv:1811.00232*.

J Zico Kolter and Andrew Y Ng. 2009. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pages 513–520.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Xianzhi Li, Ethan Callanan, Xiaodan Zhu, Mathieu Sibue, Antony Papadimitriou, Mahmoud Mahfouz, Zhiqiang Ma, and Xiaomo Liu. 2025. Entropy-aware branching for improved mathematical reasoning. *arXiv preprint arXiv:2503.21961*.

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. *arXiv preprint arXiv:2305.20050*.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Jiacai Liu. 2025. How does rl policy entropy converge during iteration? *https://zhuanlan.zhihu.com/p/28476703733*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.

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

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Youssef Mroueh. 2025. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. 2017. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

John Schulman, Xi Chen, and Pieter Abbeel. 2017a. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*.

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

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Alexander L Strehl and Michael L Littman. 2008. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331.

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. 2025. Ce-gppo: Coordinating entropy via gradient-preserving clipping policy optimization in reinforcement learning. *arXiv preprint arXiv:2509.20712*.

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

Hongze Tan and Jianfei Pan. 2025. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy. *arXiv preprint arXiv:2508.04349*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, and 1 others. 2024b. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*.

Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. 2023. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. 2025a. Unsupervised post-training for multi-modal llm reasoning via grpo. *arXiv preprint arXiv:2505.22453*.

Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. 2025b. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. 2024. Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Datasets and benchmarks

Training datasets. We train UEC-RL on three datasets that cover different modalities and difficulty levels:

- **DAPO-17K**: a large-scale out-of-domain mathematical reasoning dataset designed to evaluate RL-based alignment algorithms for LLMs (Yu et al., 2025).
- **Multimodal dataset (6k)**: sampled from the multimodal corpora introduced in (Wei et al., 2025b,a), spanning a wide range of diagram, geometry, chart, and table problems. The dataset aggregates established resources including Geometry3K (Lu et al., 2021a), GeoQA (Chen et al., 2021), GeoQA-Plus (Cao and Xiao, 2022), Geos (Seo et al., 2015), AI2D (Kembhavi et al., 2016), TQA (Kim et al., 2018), FigureQA (Kahou et al., 2017), TabMWP (Lu et al., 2022b), ChartQA (Masry et al., 2022), IconQA (Lu et al., 2021b), Clevr-Math (Lindström and Abraham, 2022), M3CoT (Chen et al., 2024), and ScienceQA (Lu et al., 2022a).
- **Geometry3K**: an in-domain geometric reasoning dataset used for detailed evaluation (Lu et al., 2021a).

Evaluation benchmarks. We assess UEC-RL across three categories of benchmarks:

- **Text reasoning benchmarks.** We evaluate Pass@1 on five widely used reasoning benchmarks: AIME24 (HuggingFaceH4, 2025), AIME25 (HuggingFaceH4, 2025), MATH (Lightman et al., 2023), GSM8K (Cobbe et al., 2021), and Minerva (Lewkowycz et al., 2022). These benchmarks span competition-level problems (AIME), formal mathematics (MATH), school-level word problems (GSM8K), and scientific reasoning (Minerva), providing a comprehensive assessment of textual reasoning ability.
- **Multimodal reasoning benchmarks.** We further evaluate on four challenging multimodal benchmarks: MathVision (Wang et al., 2024a), MathVerse (Zhang et al., 2024), MathVista (Lu et al., 2023), and We-Math (Qiao et al., 2024). These benchmarks cover diverse visual formats—including diagrams, charts, tables, and multi-image compositions—and require integrating visual and symbolic reasoning.
- **Geometry3K in-domain dataset.** To better understand the behavior of entropy-controlled RL, we conduct an in-depth analysis on Geometry3K (Lu et al., 2021a), including accuracy curves, entropy dynamics, response length behavior, and ablation studies.

B Implementation details

We follow the default EasyR1 configuration unless otherwise noted. Table 2 summarizes the hyperparameters for GRPO, DAPO, Entropy-Adv, and UEC-RL. For UEC-RL, difficult prompts trigger expanded exploration with $G' = 20$ and temperature $t' = 1.2$. Trajectories with absolute advantages greater than 1 are stored in a replay buffer of size 5120, and replay is performed every 5 optimization steps. For each experiment setting, we run a single training run, save checkpoints every 10 optimization steps, and report the maximum performance achieved on each benchmark over all saved checkpoints.

Table 2: Summary of implementation and evaluation details for all compared methods.

Settings of Training	**GRPO**	**DAPO**	**Entropy-Adv**	**UEC-RL**
Training settings				
Hardware	8×A800 GPUs (40GB)			
Policy model init	Qwen2.5-VL-7B-Instruct and Qwen2.5-Math-7B			
Hardware	$8 \times$ A800 (40GB)			
Max response length	8192			
Batch size	512			
Primary rollout G	5			
Learning rate	1×10^{-6}			
Temperature (training)	1.0			
ϵ_{low}	0.2	0.2	0.2	0.2
ϵ_{high}	0.2	0.3	0.2	0.2
Entropy bonus	–	–	$(\beta, \kappa) = (0.4, 2)$	–
Additional rollout G'	–	–	–	20
Exploration temperature t'	–	–	–	1.2
Replay buffer size s'	–	–	–	5120
Replay frequency	–	–	–	every 5 steps
Replay criterion	–	–	–	$\hat{A} > 1$
Settings of evaluation				
Max response length (eval)	8192			
Temperature (eval)	0.2			
Top-p (eval)	0.95			