SSPO: Subsentence-level Policy Optimization

Kun Yang [*1], **Zikang chen** [*1], **Yanmeng Wang** [1], **Zhigen Li** [1]
[1]Ping An Technology
{yangkun219, chenzikang003}@pingan.com.cn

Abstract

As a significant part of post-training of the Large Language Models (LLMs), Reinforcement Learning from Verifiable Reward (RLVR) has greatly improved LLMs' reasoning skills. However, some RLVR algorithms, such as GRPO (Group Relative Policy Optimization) and GSPO (Group Sequence Policy Optimization), are observed to suffer from unstable policy updates and low usage of sampling data, respectively. The importance ratio of GRPO is calculated at the token level, which focuses more on optimizing a single token. This will be easily affected by outliers, leading to model training collapse. GSPO proposed the calculation of the response level importance ratio, which solves the problem of high variance and training noise accumulation in the calculation of the GRPO importance ratio. However, since all the response tokens share a common importance ratio, extreme values can easily raise or lower the overall mean, leading to the entire response being mistakenly discarded, resulting in a decrease in the utilization of sampled data. This paper introduces SSPO, which applies sentence-level importance ratio, taking the balance between GRPO and GSPO. SSPO not only avoids training collapse and high variance, but also prevents the whole response tokens from being abandoned by the clipping mechanism. Furthermore, we apply sentence entropy to PPO-CLIP to steadily adjust the clipping bounds, encouraging high-entropy tokens to explore and narrow the clipping range of low-entropy tokens. In particular, SSPO achieves an average score of 46.57 across five datasets, surpassing GRPO (43.01) and GSPO (44.42), and wins state-of-the-art performance on three datasets. These results highlight SSPO's effectiveness in leveraging generated data by taking the essence of GSPO but rejecting its shortcomings.

1 Introduction

In recent years, Reinforcement Learning has greatly improved the reasoning capabilities of large language models (LLMs). Given verifiable rewards, language models are able to solve sophisticated problems, such as Mathematics and programming, through large-scale RL. These RL methods allow LLMs to explore and apply a long chain-of-thought to tackle difficult problems. With increasing response length, the reasoning performance of LLMs has been markedly enhanced. GRPO (Shao et al., 2024), the main algorithm of RLVR, exhibits stability issues while training large language models, often resulting in high-variance training noise, leading to model collapse. In order to address these challenges, GSPO (Zhao et al., 2025), an excellent RL algorithm modified by GRPO, introduces a response-level important ratio and becomes more efficient and stable compared to GRPO.

However, applying a response-level importance sampling ratio to RLVR training might cause other problems. Firstly, due to all the response tokens sharing a common importance sampling ratio, that calculated based on the average importance weight of each token in response, these tokens might all be clipped by the PPO-clip mechanism when a few tokens possess extremely high or low values of importance ratio. This will result in a high clipping rate and mass waste of sampled data. Secondly,

[*]Equal contribution.

when there are a few abnormal values of tokens' importance weight in a response, these anomalies might be preserved in the gradient after taking the average. These preserved tokens possibly affect the training stability and training efficiency.

In this paper, we present SSPO, a novel RL algorithm designed to address the core shortcomings caused by GSPO. First of all, we introduce a sub-sentence-level importance weight by splitting the whole response into a few sentences. Each token in a sentence possesses one importance sampling weight. In this way, it can prevent the whole response from being abandoned by PPO-clip (Schulman et al., 2017), and reduce the clipping fractions. Also, we introduce entropy-adaptive clipping to dynamically adjust the clipping bounds depending on the sentence entropy, encouraging high-entropy tokens to explore further and limiting the clipping range of low-entropy tokens. In this way, entropy collapse has been effectively avoided, and the training stability of RLVR has been significantly improved.

Finally, we use levels 3-5 of MATH (Hendrycks et al., 2021) datasets as the training datasets and chose five common mathematical reasoning benchmarks to evaluate performance. We conducted our experiments on 2 models and used GRPO, GSPO, GMPO (Zheng et al., 2025), and Dr.GRPO (Liu et al., 2025) as baselines. The experimental results show that SSPO achieved the highest average score among these five datasets on the 1.5B and 7B models.

2 Related Works

2.1 GRPO reinforcement learning algorithm

In recent years, reinforcement learning algorithms have achieved great success in the field of training large models, especially the successful application of the GRPO algorithm on DeepSeek-R1 (Guo et al., 2025), which has led to a qualitative leap in inference-related tasks. Our work is based on the improvement of the GRPO algorithm, introducing a sentence-level importance ratio and sentence-level dynamic importance ratio clipping on the basis of the original GRPO algorithm. This addresses the issues of oscillation, unstable training, and entropy collapse that are prone to occur during the training process of the GRPO algorithm

2.2 Importance ratio

Importance sampling is widely used in reinforcement learning algorithms, especially in the GRPO algorithm. The importance ratio of GRPO is calculated at the token level, which focuses more on optimizing each token, but may be affected by some outliers. This approach leads to high variance and training noise, which gradually accumulates as the response length increases, ultimately leading to model training collapse. GSPO proposed the calculation of the importance ratio of response levels, which solves the problem of high variance and noise accumulation in the calculation of GRPO's importance ratio. However, because the entire response token shares the same importance ratio, some extreme values can easily raise or lower the overall average value of the sample, resulting in the entire response token being mistakenly discarded by the clipping mechanism, lowering the utilization of sampled data, resulting in data waste. Our work proposes an importance ratio calculation method based on truncated sentences, where each sentence in a response shares the same importance ratio, achieving a balance between GRPO and GSPO. This approach retains the advantages of GSPO while addressing its shortcomings.

Concurrent work proposed the LPO (Li et al., 2025) 10 days before our submission date. LPO proposed Sentence granularity Importance Sampling, which is similar to ours, and validated its effectiveness in a one-trillion-parameter model. However, LPO applies *token-level* normalization, meaning that advantage estimation $\hat{A}_i$ is averaged over the total token length $|y_i|$, which means $\hat{A}_i$ is influenced by response length. In contrast to LPO, we hold the idea that advantage estimation $\hat{A}_i$ is independent of response length. Therefore, SSPO averages $\hat{A}_i$ on *response-level*, which is the same as GSPO. Additionally, LPO uses a standard fixed *PPO-Clip* (Schulman et al., 2017), while SSPO introduces a novel *adaptive entropy clipping* mechanism (Sec. 4.2), which is designed to dynamically adjust the magnitude of updates and mitigate entropy collapse.

2.3 Clipping mechanism

The PPO-CLIP algorithm was proposed in PPO to restrict the confidence interval and improve the stability of reinforcement learning training by limiting the importance ratio. It has also been extended to GRPO. CISPO (Chen et al., 2025) observed that the gradient of the clipped token does not participate in backpropagation and cannot obtain gradient updates, so it proposed soft clipping gradient. DCPO (Yang et al., 2025) uses dynamic clipping to dynamically set different clipping ranges based on the probability of different tokens. Our algorithm introduces the clipping base entropy, taking the entropy of the token as an important factor in calculating the clipping range, which better stabilizes the changes in entropy in reinforcement learning and alleviates the problem of entropy collapse.

3 Preliminaries

Notation In this paper, we use x to denote a query and D to denote the query set. Given a response y to a query x, $|y|$ denotes the number of tokens in y.

3.1 Group Relative Policy Optimization(GRPO)

Shao et al.(2024) proposed GRPO. The value model was removed by calculating the relative advantage of each response within a set of responses to the same query. GRPO optimizes the following objective:

$$\mathscr{J}_{GRPO}\left(\theta\right) = E_{\mathrm{x}\sim\mathrm{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(.|x)}[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min(w_{i,t}\left(\theta\right)\hat{A}_{i,t}, clip(w_{i,t}\left(\theta\right),1-\epsilon,1+\epsilon)\hat{A}_{i,t}] \tag{1}$$

Where G is the number of responses generated for each query, θ_{old} is the old policy, ϵ is the clipping range of the importance ratios. The importance ratio $w_{i,t}\left(\theta\right)$ and the advantage $\hat{A}_{i,t}$ of token $y_{i,t}$ are:

$$w_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x,y_{i,<t})}, \quad \hat{A}_{i,t} = \hat{A}_i = \frac{r(x,y_i) - \mathrm{mean}(\{r(x,y_i\}_{i=1}^{G})}{\mathrm{std}(\{r(x,y_i\}_{i=1}^{G})} \tag{2}$$

3.2 Group Sequence Policy Optimization(GSPO)

Zheng et al.(2025) proposed GSPO. GSPO defines the importance ratio based on sequence likelihood and performs response-level clipping. GSPO optimizes the following objective:

$$\mathscr{J}_{GSPO}\left(\theta\right) = E_{\mathrm{x}\sim\mathrm{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(.|x)}[\frac{1}{G}\sum_{i=1}^{G}\min(s_i\left(\theta\right)\hat{A}_i, clip(s_i\left(\theta\right),1-\epsilon,1+\epsilon)\hat{A}_i] \tag{3}$$

The importance ratio $s_i\left(\theta\right)$ and advantage $\hat{A}_i$ are:

$$\hat{A}_i = \frac{r(x,y_i) - \mathrm{mean}(\{r(x,y_i\}_{i=1}^{G})}{\mathrm{std}(\{r(x,y_i\}_{i=1}^{G})} \tag{4}$$

$$s_i(\theta) = (\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)})^{\frac{1}{|y_i|}} = exp(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}log\frac{\pi_\theta(y_{i,t}|x,y_i<t)}{\pi_{\theta_{old}}(y_{i,t}|x,y_i<t)}) \tag{5}$$

4 Methodology

4.1 Subsentence-level important ratio

The importance ratio of GRPO is token-level and is easily affected by outliers, leading to unstable training. GSPO designs the calculation of the importance ratio at the response level, and uses length normalization to reduce the policy gradient variance, reduce the impact of extreme tokens, and ensure the stability of training. However, the sequence-level importance ratio calculation method results in a large proportion of clipped tokens, which reduces sample utilization. In order to maintain the stable advantages of GSPO training and increase sample utilization, we propose subsentence-level

importance ratio. We segment the samples into multiple fragments by line break or double line break(we regard as the smallest semantic unit) and calculate the importance ratio for each fragment. Without considering clips, the optimization objectives are as follows:

$$\mathscr{J}_{SSPO}(\theta) = E_{\mathrm{x}\sim\mathrm{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(.|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{j=1}^{N_{seg}(y_i)} |y_{i,j}| s_{i,j}(\theta)\hat{A}_{i,j} \right] \tag{6}$$

We adopt group-based advantage estimation:

$$\hat{A}_{i,j} = \hat{A}_i = \frac{r(x,y_i) - \mathrm{mean}(\{r(x,y_i\}_{i=1}^{G})}{\mathrm{std}(\{r(x,y_i\}_{i=1}^{G})} \tag{7}$$

$y_{i,j}$ represents the j-th segment after the i-th sample is semantically segmented. $N_{seg}(y_i)$ represent the number of segments in y_i, Similarly to GSPO, we define the importance ratio $s_{i,j}(\theta)$ as follows:

$$s_{i,j}(\theta) = (\frac{\pi_\theta(y_{i,j}|x)}{\pi_{\theta_{old}}(y_{i,j}|x)})^{\frac{1}{|y_{i,j}|}} = exp(\frac{1}{|y_{i,j}|}\sum_{t=1}^{|y_{i,j}|} log\frac{\pi_\theta(y_{i,j,t}|x,y_{i,j,}<t)}{\pi_{\theta_{old}}(y_{i,j,t}|x,y_{i,j,}<t)}) \tag{8}$$

The gradient objective of SSPO is as follows:

$$\nabla_\theta \mathscr{J}_{SSPO}(\theta) = \nabla_\theta E_{\mathrm{x}\sim\mathrm{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(.|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{j=1}^{N_{seg}(y_i)} |y_{i,j}| s_{i,j}(\theta)\hat{A}_{i,j} \right] \tag{9}$$

$$= E_{\mathrm{x}\sim\mathrm{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(.|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{j=1}^{N_{seg}(y_i)} |y_{i,j}| s_{i,j}(\theta)\hat{A}_{i,j}\nabla_\theta log s_{i,j}(\theta) \right] \tag{10}$$

$$= E_{x\sim D,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{j=1}^{N_{seg}(y_i)} |y_{i,j}| \left(\frac{\pi_\theta(y_{i,j}|x)}{\pi_{\theta_{old}}(y_{i,j}|x)}\right)^{\frac{1}{|y_{i,j}|}} \hat{A}_{i,j}\cdot\frac{1}{|y_{i,j}|}\sum_{t=1}^{|y_{i,j}|}\nabla_\theta \log \pi_\theta(y_{i,j,t}|x,y_{i,j,<t}) \right] \tag{11}$$

$$= E_{x\sim D,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{j=1}^{N_{seg}(y_i)} |y_{i,j}| \left(\frac{\pi_\theta(y_{i,j}|x)}{\pi_{\theta_{old}}(y_{i,j}|x)}\right)^{\frac{1}{|y_{i,j}|}} \hat{A}_{i}\cdot\frac{1}{|y_{i,j}|}\sum_{t=1}^{|y_{i,j}|}\nabla_\theta \log \pi_\theta(y_{i,j,t}|x,y_{i,j,<t}) \right] \tag{12}$$

For comparison, GRPO's gradient objective is as follows:

$$\nabla_\theta \mathscr{J}_{GRPO}(\theta) = \nabla_\theta E_{\mathrm{x}\sim\mathrm{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(.|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} w_{i,t}(\theta)\hat{A}_{i,t} \right] \tag{13}$$

$$= E_{x\sim D,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i\cdot\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\left(\frac{\pi_\theta(y_{i,j}|x,y_{i<t})}{\pi_{\theta_{old}}(y_{i,j}|x)}\right)\nabla_\theta \log \pi_\theta(y_{i,t}|x,y_{i<t}) \right] \tag{14}$$

The weight of each token gradient of GRPO is different, which causes the problem of high variance of the policy gradient. The gradient target of SSPO is ultimately similar to GSPO. The gradient weights of all tokens in a fragment are equal, eliminating noise interference between tokens, reducing the variance of the policy gradient, and making training more stable.

4.2 Adaptive Entropy Clipping Mechanism

In reinforcement learning, clipping is used to control the magnitude of policy updates to ensure that changes in the old and new policies are within a certain range and enhance the stability of training. However, the clipping mechanism will inhibit the probability growth of low-probability tokens and, at the same time, tend to maintain the original high-probability tokens, resulting in rapid model fitting and entropy collapse. In order to alleviate the phenomenon of entropy collapse and encourage exploration, we propose a dynamic entropy clipping method. For each token o_t of the output o, when the vocabulary size is V pairs, the entropy of the current token under the old strategy is:

$$\mathcal{H}_t = -\sum_{v \in V} \pi_{\theta_{old}}(v|x, o_{<t}) log \pi_{\theta_{old}}(v|x, o_{<t}) \tag{15}$$

Since our important ratio is in the semantic segment dimension, we define the entropy of the semantic segment dimension as follows:

$$\mathcal{H}_{i,j} = \frac{1}{|y_{i,j}|} \sum_{t=1}^{|y_{i,j}|} \mathcal{H}_t \tag{16}$$

We define the dynamic entropy clipping method as follows:

$$\epsilon_{high} = 1 + \alpha + \mathcal{H}_{i,j} \tag{17}$$

$$\epsilon_{low} = \begin{cases} 0.3, \mathcal{H}_{i,j} > 1 \\ 1.3 - \mathcal{H}_{i,j}, 0.5 \leqslant \mathcal{H}_{i,j} \leqslant 1 \\ 0.8, \mathcal{H}_{i,j} < 0.5 \end{cases} \tag{18}$$

When the entropy of the subsentence is relatively large, we expand the clip interval range, allowing tokens with high entropy to participate more in the gradient update. This prevents them from being clipped and increases the update intensity. When the entropy of the subsentence is relatively small, it means that the model has learned these tokens well. Therefore, we reduce the clipping interval range to limit the update intensity of these tokens. Through semantic segmentation and adaptive entropy clipping, our final objective is as follows:

$$\mathscr{J}_{SSPO}(\theta) = E_{\mathrm{x}\sim \mathrm{D}, \{y_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(.|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{j=1}^{N_{seg}(y_i)} |y_{i,j}| \min(s_{i,j}(\theta) \hat{A}_i, clip(s_{i,j}(\theta), \epsilon_{low}, \epsilon_{high}) \hat{A}_{i,j} \right] \tag{19}$$

5 Experimental Setup

5.1 Model

We experimented with two mathematical models Qwen2.5-Math-1.5B and Qwen2.5-Math-7B (Yang et al., 2024).

5.2 Training

Following Dr.GRPO, we choose MATH Level 3-Level 5 as our training datasets, which contain 8523 mathematical problems. For each question, we sample 8 rollouts and set the model's response length at 3000 tokens and set the batch size to 128. All problems were processed using the Qwen-Math template for both training and evaluation. In all experiments of the RL algorithms, we use the veRL (Sheng et al., 2024) framework to support our experiments.

5.3 Baseline Settings

We use GRPO, GSPO, GMPO, and Dr.GRPO as our baseline.

Table 1: Comparison of SSPO and other RL Algorithms on 5 mathematical reasoning benchmarks

Method	AIME24	AMC23	MATH	MIN.	OLY.	Avg
Qwen2.5-Math-1.5B						
GRPO	16.67	54.2	72.6	32.35	39.67	43.01
GSPO	20.0	51.49	74.6	34.56	**41.16**	44.42
Dr.GRPO	20.0	53.0	74.2	25.7	37.6	42.1
GMPO	20.0	53.0	**77.6**	30.1	38.7	43.9
SSPO (w/o entropy clip)	**23.3**	56.63	74.2	32.72	39.52	45.72
SSPO	**23.3**	**57.83**	75.4	**35.29**	41.01	**46.57**
Qwen2.5-Math-7B						
GRPO	33.3	**67.47**	79.0	40.07	45.91	53.15
GSPO	33.3	65.06	80.8	**42.28**	47.1	53.75
Dr.GRPO	**43.3**	62.7	80.0	30.1	41.0	51.4
GMPO	**43.3**	61.4	**82.0**	33.5	43.6	52.7
SSPO (w/o entropy clip)	33.3	65.06	81.6	**42.28**	**47.7**	53.99
SSPO	36.67	66.27	81.8	**42.28**	47.25	**54.85**

5.4 Evaluation

We evaluated our models on five widely used mathematical reasoning benchmarks: MATH-500 (Hendrycks et al., 2021), a subset of 500 problems from MATH datasets including algebra, geometry, and number theory; AMC (Li et al., 2024) consists of 83 intermediate-difficulty multiple-choice problems; AIME24 (Li et al., 2024) contains 30 olympiad problems from American Invitational Mathematics Examination 2024; Minerva (Lewkowycz et al., 2022) comprises 272 graduate-level multi-step reasoning problems; Olympiad Bench (He et al., 2024) owns 675 high-difficulty Olympiad questions.

6 Result and Analysis

6.1 Main Results

Table 1 summarizes the performance of various RL optimization methods, including GRPO, Dr.GRPO, GSPO, GMPO, and our proposed SSPO, on five mathematical reasoning benchmarks: MATH500, AMC23, AIME24, Minerva, and Olympiad. We evaluate these datasets basically on greedy decoding(Avg@1) to intuitively present the effectiveness of our method. We evaluate all approaches in five datasets every 10 steps in both 1.5B and 7B models, and choose the best score as our final score. Among the Qwen2.5-Math-1.5B and Qwen2.5-Math-7B models, SSPO offers superior performance relative to other methods.

6.1.1 Overall Surpass of SSPO

In the Qwen2.5-Math-1.5B-Insturct model, SSPO achieves the highest average scores on five datasets(46.57), exceeding GRPO(43.01) and GSPO(44.42). Compared to the token-level importance ratio, GSPO applies response-level importance ratio and shows a significant advance(+1.41). SSPO, using the subsentence-level importance ratio, presents a further improvement compared to GSPO(+2.15). Also, according to the results of Dr.GRPO(42.1) and GMPO(43.9), SSPO also offers great improvement (+4.47 and +2.67, respectively). SSPO shows advantages in model scaling. For Qwen2.5-Math-7B-Instruct, SSPO also delivers a strong result of 54.85 under 5 datasets, offering notable improvements over GRPO(53.15) and GSPO(53.75). Regarding Dr.GRPO and GMPO, SSPO outperforms Dr.GRPO by(+3.45) and GMPO by(+2.15). These results highlight notable improvements across different model capacities.

6.1.2 Ablation study

We conducted an ablation study on both Qwen2.5-Math-1.5B and Qwen2.5-Math-7B to assess the contribution of subsentence-level importance ratio and the adaptive entropy clipping mechanism.

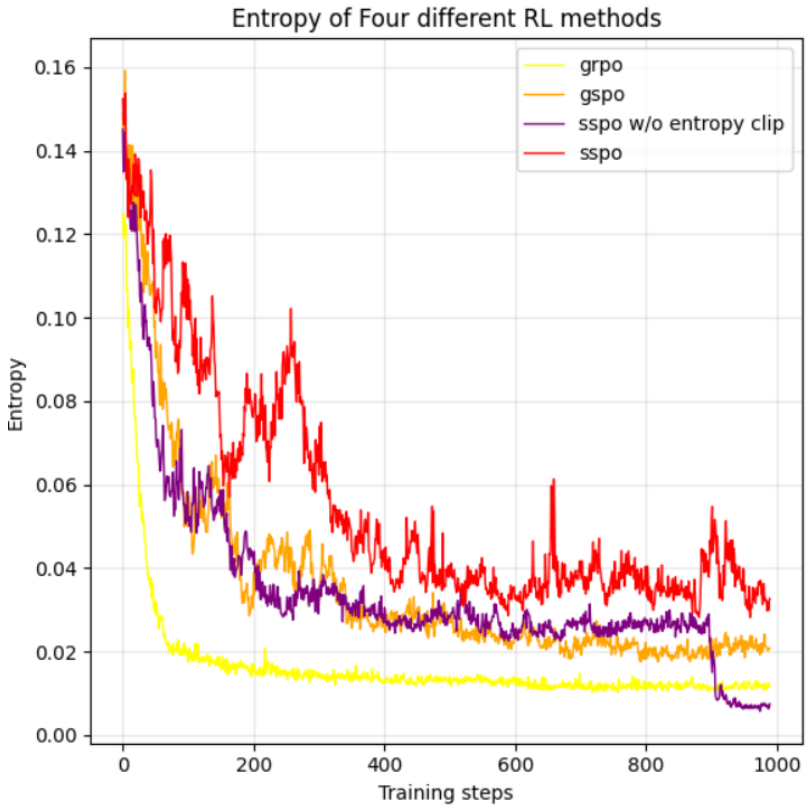

Figure 1: The training entropy of SSPO, SSPO (w/o entropy clip), GSPO and GRPO

As shown in Table 1, we can see that by applying the adaptive entropy clipping mechanism, the performance of both 1.5B and 7B models improves.

6.2 Subsentence-level importance ratio

The experiment method SSPO(w/o entropy clip) indicates that we only modify the importance ratio without applying adaptive entropy clipping to SSPO. It is obvious that the subsentence-level importance ratio (45.72) outperforms both the token-level(43.01) and response-level (44.42) in the Qwen2.5-Math-1.5B model. This gap becomes narrow in Qwen2.5-Math-7B, outperforming GRPO by(+0.84) and GSPO by(+0.24).

6.3 Adaptive entropy clipping mechanism

With our adaptive entropy clipping mechanism, it demonstrates a further improvement compared to SSPO(w/o entropy clip) by (+0.85) and (+0.86) in Qwen2.5-MATH-1.5B and Qwen2.5-MATH-7B, respectively. As shown in Figure 1, we can easily know that compared to other approaches, the tokens' entropy decreases less. It shows that adaptive entropy clipping is helpful in preventing entropy from collapsing, keeping the RL training much more stable. When the token entropy remains higher, it enables the model to continue exploring and achieve great performance.

7 Conclusion

We propose SSPO, a stable variant of GRPO. By applying subsentence-level importance ratio and adaptive entropy clipping, SSPO not only allows models to improve the performance of mathematical benchmarks, but also alleviates the entropy collapse, encouraging models to explore. Our experiments indicate that SSPO outperforms GSPO, GRPO, Dr.GRPO, and GMPO in both 1.5B and 7B models under five mathematical datasets. This work makes a contribution to developing more reliable and

scalable reinforcement learning for LLMs, and future work will focus on extending SSPO to other domains, such as programming and semantic reasoning.

References

Chen, A., Li, A., Gong, B., Jiang, B., Fei, B., Yang, B., Shan, B., Yu, C., Wang, C., Zhu, C., et al. (2025). Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., et al. (2024). Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. (2022). Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Li, A., Liu, B., Hu, B., Li, B., Zeng, B., Ye, B., Tang, C., Tian, C., Huang, C., Zhang, C., et al. (2025). Every activation boosted: Scaling general reasoner to 1 trillion open language foundation. *arXiv preprint arXiv:2510.22115*.

Li, J., Beeching, E., Tunstall, L., Lipkin, B., Soletskyi, R., Huang, S., Rasul, K., Yu, L., Jiang, A. Q., Shen, Z., et al. (2024). Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. (2025). Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. (2024). Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., et al. (2024). Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Yang, S., Dou, C., Guo, P., Lu, K., Ju, Q., Deng, F., and Xin, R. (2025). Dcpo: Dynamic clipping policy optimization. *arXiv preprint arXiv:2509.02333*.

Zhao, Y., Liu, Y., Liu, J., Chen, J., Wu, X., Hao, Y., Lv, T., Huang, S., Cui, L., Ye, Q., et al. (2025). Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*.

Zheng, C., Liu, S., Li, M., Chen, X.-H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al. (2025). Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.