

# The Role of Entropy in Visual Grounding: Analysis and Optimization

Shuo Li<sup>\*,1</sup>, Jiajun Sun<sup>\*,1</sup>, Zhihao Zhang<sup>1</sup>,  
Xiaoran Fan<sup>1</sup>, Senjie Jin<sup>1</sup>, Hui Li<sup>1</sup>, Yuming Yang<sup>1</sup>, Junjie Ye<sup>1</sup>,  
Lixing Shen<sup>2</sup>, Tao Ji<sup>1</sup>, Tao Gui<sup>†1,3,4</sup>, Qi Zhang<sup>1</sup>, Xuanjing Huang<sup>1</sup>  
Fudan University<sup>1</sup> Hikvision Research Institute<sup>2</sup>  
Institute of Trustworthy Embodied AI, Fudan University<sup>3</sup>  
Shanghai Key Laboratory of Multimodal Embodied AI<sup>4</sup>  
lis23@m.fudan.edu.cn, tgui@fudan.edu.cn

## Abstract

Recent advances in reinforcement learning-based fine-tuning of multimodal large language models (MLLMs) have highlighted the importance of entropy control for stable optimization. While entropy has been extensively studied in reasoning-centric settings, its behavior and optimization role in perception-oriented tasks remain insufficiently understood. In this work, we reveal that entropy exhibits fundamentally different roles in perception tasks compared to reasoning, particularly under continuous prediction and reward. Using visual grounding as a case study, we show that entropy can remain misaligned with optimization objectives due to persistent low-probability yet high-advantage samples. Building on these findings, we introduce ECVGPO (Entropy Control Visual Grounding Policy Optimization), an interpretable algorithm designed for effective entropy regulation. Through entropy control, the trade-off between exploration and exploitation is better balanced. Experiments show that ECVGPO achieves broad improvements across various benchmarks and models.

## 1 Introduction

Large language models (LLMs), such as ChatGPT (OpenAI, 2022), have shown outstanding ability in handling a wide spectrum of text-based tasks. When equipped with visual encoders like CLIP (Radford et al., 2021), these models evolve into multimodal large language models (MLLMs) that extend their strengths to vision-related domains. This integration enables them to tackle various vision-language challenges, including image captioning (Wang et al., 2020), visual question answering (Antol et al., 2015), and visual dialogue (Das et al., 2017). Driven by the success of works like OpenAI’s o1 (OpenAI, 2024b) and DeepSeek’s R1 (Guo et al., 2025), reinforcement

learning finetuning of LLMs and MLLMs become a central focus for researchers. Notably, RL-like approaches often combine Reinforcement Learning with Verifiable Rewards (RLVR) and the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) algorithm. Such methods have proven particularly effective for complex reasoning tasks, including mathematical problem solving.

In reasoning tasks, entropy is an important metric for evaluating a model’s responses. Prior studies (Wang et al., 2025; Cheng et al., 2025a) show that high-entropy tokens often cluster around specific types of words, such as transition terms (e.g., “however”, “but”), which play a critical role in the model’s reasoning process. Consequently, response entropy is frequently used as an indicator of a model’s exploration capability in reasoning tasks. However, as RLVR is applied to more non-reasoning tasks such as visual grounding (Liu et al., 2025a; Shen et al., 2025b; Fan et al., 2025), whether entropy plays a different role has not yet been investigated. Studying entropy in visual grounding can provide crucial insights into model training, leading to more reliable and robust spatial localization.

In this paper, we analyze the behavior of entropy in a visual grounding task and identify a phenomenon that is opposite to that observed in reasoning tasks. For models trained on their respective reasoning tasks, entropy tends to drop to very low values after training. In contrast, in the visual grounding task, entropy remains high even after extensive training. Notably, these high-entropy tokens are primarily associated with the coordinate values of predicted answers and the object nouns that appear in the model’s reasoning process. Our investigation reveals that, during the GRPO training, the model consistently operates under higher entropy in visual grounding scenarios. We further analyze the cause of high entropy, observing that responses with a high advantage often correspond

\* Equal contributions.

† Corresponding author.

to low generation probabilities.

Based on these findings, we propose Entropy Control Visual Grounding Policy Optimization (ECVGPO), a novel reinforcement learning method that employs self-information to quantify the confidence of the model’s generated responses and introduces a self-information penalty or encouragement in visual grounding training. By reshaping the advantages of positive samples, we leverage the generation probability of tokens (i.e., self-information) to increase or decrease their advantages, thereby encouraging the model to update more conservatively or aggressively. This enables explicit control over the model’s entropy, achieving a better balance between exploration and exploitation. We evaluate our method on two of the most popular multimodal models: Qwen2.5-VL (Bai et al., 2025a) and InternVL3 (Zhu et al., 2025). The experimental results show that ECVGPO improves the performance of the model on both the in-domain and out-of-domain evaluation benchmarks. At the same time, we conduct extensive ablation experiments to demonstrate the generalization capability of our proposed method.

In this paper, our main contributions are:

- We compare the token-level entropy distributions between reasoning and visual grounding tasks, and find that models trained with RLVR on visual grounding exhibit consistently higher entropy, mainly because responses with high advantages often correspond to low probabilities;
- We propose ECVGPO, the first method that dynamically adjusts entropy during training to achieve a better trade-off between exploration and exploitation;
- We evaluate ECVGPO on the most popular MLLMs and demonstrate broad performance gains on benchmarks.

## 2 Related Work

**Multimodal large language model** Multimodal large language models (MLLMs), exemplified by GPT-4o (OpenAI, 2024a), are emerging as a central research focus in deep learning due to their impressive cross-domain capabilities. By integrating visual encoders with powerful language models, these systems enable joint visual-linguistic reasoning. Early milestones such as CLIP (Radford et al., 2021) establish a strong link between textual

and visual representations, laying the groundwork for numerous cross-modal applications. The BLIP series (Li et al., 2022, 2023; Dai et al., 2023) extends this paradigm to tasks like visual question answering, whereas LLaVA (Liu et al., 2024b,a) adopts a lightweight projection module and two-stage training to enhance spatial grounding and overall performance. More recent efforts, including MouSi (Fan et al., 2024) and Cambrian-1 (Tong et al., 2024), leverage diverse vision encoders to improve multimodal comprehension. State-of-the-art models such as QwenVL (Bai et al., 2023; Wang et al., 2024; Bai et al., 2025a), InternLM-XComposer (Zhang et al., 2023; Dong et al., 2024), and InternVL (Chen et al., 2023, 2024b) follow design principles similar to LLaVA and consistently achieve leading results.

**Reinforcement learning with verifiable reward** Reinforcement Learning with Human Feedback (RLHF) (Bai et al., 2022) has become a cornerstone for aligning large language models (LLMs) with human intent. However, the quality and scalability of human feedback remain significant challenges. To address this, a new paradigm has emerged: Reinforcement Learning with Verifiable Rewards (RLVR) (Shao et al., 2024). Instead of relying on subjective human preferences, this method leverages a reward function that can be algorithmically verified. Pioneering works like O1 (OpenAI, 2024b) from OpenAI and R1 (Guo et al., 2025) from DeepSeek exemplify this direction, particularly in complex reasoning tasks such as mathematics (Shao et al., 2024; Yu et al., 2025; Chen et al., 2025; Jin et al., 2026). These methods often design reward functions that verify the correctness of each reasoning step or the final answer, allowing for more objective and scalable policy optimization. Recently, several studies, such as Visual-RFT (Liu et al., 2025a), VLM-R1 (Shen et al., 2025a), UniVG-R1 (Bai et al., 2025b) and ViGoRL (Sarch et al., 2025), introduce visual grounding tasks into RLVR frameworks to enhance perception-oriented capabilities.

**Entropy’s role in RLVR training** Recent research has identified entropy as a crucial metric in RLVR. For example, (Wang et al., 2025) analyzes token-level entropy distributions in Chain-of-Thought (CoT) (Wei et al., 2023) reasoning and finds that only a small fraction of tokens exhibit high entropy, acting as pivotal decision points that guide the model toward diverse reasoning trajec-

ries. Building on this, (Cui et al., 2025) reports an empirical relationship between entropy and downstream performance, indicating that policy performance is fundamentally constrained by the available policy entropy and reaches a predictable upper bound when entropy is fully exhausted. Similarly, (Cheng et al., 2025b) revisits entropy as an exploration signal in reinforcement learning, examining its role in facilitating exploratory reasoning behaviors in large language models. While these studies (Zhang et al., 2025; Tan and Pan, 2025; Su et al., 2025) provide valuable insights into the role of entropy for complex reasoning tasks, its role in visual perception tasks—such as visual grounding—remains largely unexplored.

### 3 Preliminary

In this section, we present the preliminaries, covering the training process of RLVR in the visual grounding task and highlighting the distinct roles that entropy plays in reasoning tasks versus visual grounding tasks.

#### 3.1 RLVR Training in Visual Grounding Task

In the visual grounding setting, in line with most related work (Shen et al., 2025a; Liu et al., 2025b), we utilize Group Relative Policy Optimization (GRPO), which does not require a separate critic network to estimate policy values. GRPO directly compares multiple candidate responses from the same prompt, removing the dependency on an additional value function. Details are in the Appendix A.

#### 3.2 High Entropy in Visual Grounding Task

**Token-level entropy** We begin by defining the token-level entropy, which measures the uncertainty of the model’s output distribution at each decoding step. Formally, the token-level entropy at time step  $i$  is computed over the vocabulary distribution as:

$$\mathcal{H}_{o_i} = - \sum_{v \in \mathcal{V}} \pi_{\theta}(v|q, o_{<i}) \log \pi_{\theta}(v|q, o_{<i}), \quad (1)$$

where  $\mathcal{V}$  denotes the vocabulary, and  $\pi_{\theta}(v|q, o_{<i})$  is the predicted probability of token  $v$  given the input  $q$  and the previously generated tokens  $o_{<i}$ .

**Policy entropy** Furthermore, we compute the policy entropy by definition, i.e., the average token-level entropy evaluated over the current policy

model  $\pi_{\theta}$  and the dataset  $\mathcal{D}$ :

$$\hat{\mathcal{H}}(\pi_{\theta}, \mathcal{D}) = -\mathbb{E}_{\pi_{\theta}, \mathcal{D}}[\log \pi_{\theta}(o_i|q, o_{<i})] \quad (2)$$

$$= \sum_{O \in \mathcal{D}} \sum_{o_i \in O} \mathcal{H}_{o_i} \quad (3)$$

We use LMM-R1 (Peng et al., 2025), trained on the reasoning task, and VLM-R1 (Shen et al., 2025a), trained on visual grounding, both based on Qwen2.5-VL-3B for comparison, and analyze the high-entropy tokens among the generated outputs, as shown in Figure 1. The distribution of these tokens differs substantially between the two tasks, with the disparities primarily manifesting in two aspects:

1. **Positional distribution.** In reasoning tasks, high-entropy tokens appear more frequently during the reasoning process (i.e., between `<think>` and `</think>`) and are relatively sparse in the final answer segment (`<answer>...</answer>`). After training, the number of high-entropy tokens decreases substantially. In contrast, in vision-based tasks, high-entropy tokens are prevalent in both the reasoning and answer segments, both before and after training as shown in Figure 2.
2. **Token type characteristics.** In reasoning tasks, prior to training, and consistent with the observations reported in (Wang et al., 2025), high-entropy tokens often correspond to “pivot tokens” that preserve semantic continuity within the reasoning chain. After training, all types of high-entropy tokens are greatly reduced. In visual grounding tasks, however, These tokens are mainly object nouns occurring in the thinking process and **coordinate values contained in the predicted bounding box of the answer**.

#### 3.3 Explaining the Persistent Occurrence of High Entropy

We aim to investigate one of the fundamental reasons behind the persistent occurrence of high-entropy tokens during visual grounding training.

Entropy change under policy gradient Following the analysis in (Cui et al., 2025), proof and empirical evidence on the correlation of entropy change and covariance in B, the change of entropy  $\mathcal{H}$  during training under policy gradient can be described

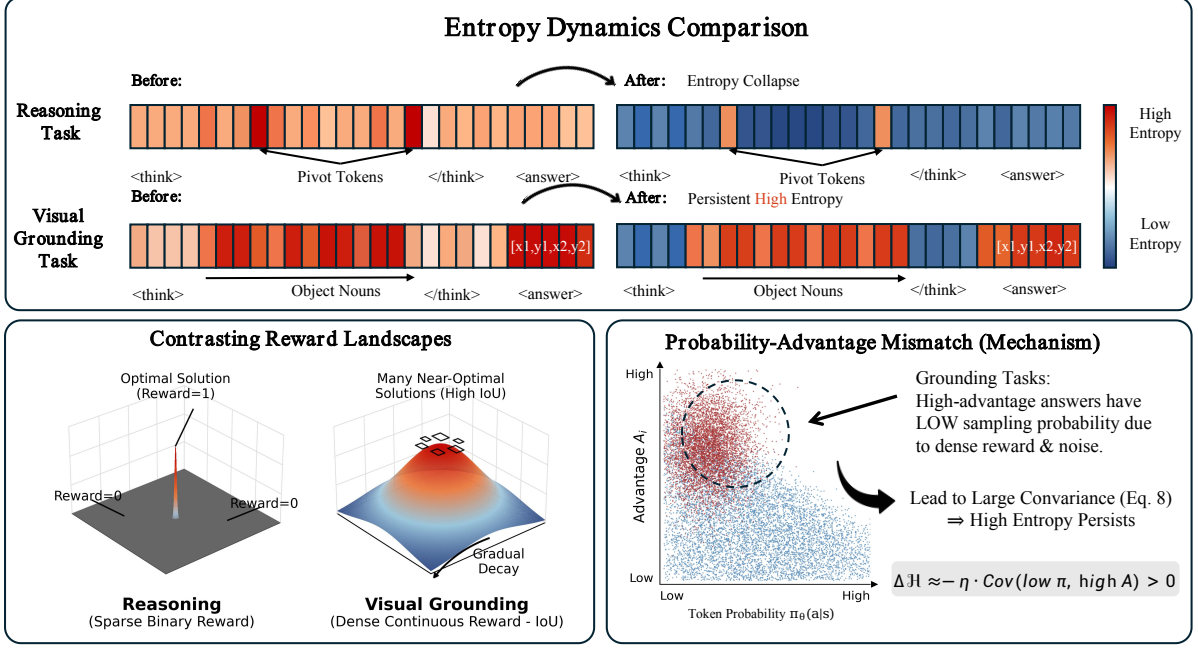


Figure 1: Overview of entropy behaviors in RLVR for Visual Grounding. (A) Contrast of token-level entropy heatmaps. Unlike reasoning tasks where entropy collapses post-training (concentrating only on "pivot tokens"), visual grounding tasks exhibit persistent high entropy, particularly on object nouns in thought and coordinates in the answer. (B) The underlying reward landscapes dictate this behavior. Reasoning tasks typically have sparse, sharp reward peaks leading to certainty, whereas visual grounding (IoU-based) features broad reward plateaus with many near-optimal solutions, encouraging uncertainty. (C) Under GRPO training, this dense reward landscape results in a critical mismatch: high-advantage responses often have low current policy probabilities (top-left quadrant).

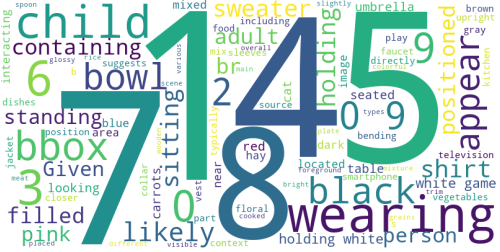


Figure 2: Word cloud of tokens with twice the average entropy.

by the following Equation 4:

$$\begin{aligned} \mathcal{H}(\pi_{\theta}^{k+1}|s) - \mathcal{H}(\pi_{\theta}^k|s) \approx \\ -\eta \cdot \text{Cov}_{a \sim \pi_{\theta}^k}(\log \pi_{\theta}^k(a|s), \pi_{\theta}^k(a|s)A(s, a)). \end{aligned} \quad (4)$$

Here,  $s$  denotes the state,  $\pi_{\theta}^k$  is the policy network at iteration  $k$ ,  $\eta$  represents the learning rate, and  $\text{Cov}$  denotes the covariance. According to this formula, we observe that **the action  $a$  with both high (or low) probability and high (or low) advantage tends to decrease entropy, whereas the opposite conditions lead to increased entropy.**

Expected low probability of high-advantage responses In the visual grounding task, the expected probability of high-dominance responses remains low due to two factors: (1) the presence of numerous responses with similarly high rewards, and (2) annotation noise, which blurs the distinction between optimal and near-optimal responses. Consider a simplified reward scenario as shown in Figure 1. In many **reasoning tasks**, the reward signal is sparse and binary. For a correct answer of 20, the reward function  $R_{\text{reasoning}}$  is defined as:

$$R_{\text{reasoning}}(a) = \begin{cases} 1, & \text{if } a = 20, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

In this case, the policy gradient will concentrate probability mass on the single correct action, yielding a sharp, low-entropy distribution.

In contrast, consider a **grounding-like task** with a numerically continuous reward. For illustration, suppose the target value is 20 and the reward is defined by a smooth decay function:

$$R_{\text{num}}(a) = \max(0, 1 - \lambda|a - 20|), \quad (6)$$

where  $\lambda > 0$  controls the decay rate. Here, the optimal action  $a_{20} = 20$  receives the maximum reward

1, but neighboring actions  $a_{19}$  and  $a_{21}$  still receive large rewards. Such a reward function encourages the policy to increase the probabilities of multiple adjacent actions with similar values.

Meanwhile, the answer coordinates generated in the visual grounding task often consist of approximately 10 tokens, and similar IoU scores can correspond to dozens or even hundreds of candidate answers. Due to the discrete nature of the tokens, these answers can be approximately viewed as different actions. That is, there is such an advantage set:

$$\{A_\delta\} = \{A(s, a) \mid |A - A_{\max}| \leq \delta\}, \quad (7)$$

where  $|\{A_\delta\}| = M, M \gg 1$ .

where  $\{A_\delta\}$  represents the advantage set of suboptimal solutions, and  $M$  denotes the total number of such suboptimal solutions.

In addition, according to the conclusion of (Chen et al., 2024a), due to the ambiguous annotation standards in the visual grounding task (e.g., an answer with IoU = 0.9 may be nearly identical to one with IoU = 0.95), it lacks the certainty present in most reasoning tasks, where the correct answer is typically a single numerical value or discrete option. Consequently, the expected advantage and probability of the optimal solution in the annotations and that of suboptimal solutions differ only marginally:

$$\mathbb{E}(A_\delta) \approx \mathbb{E}(A_{\max}), \quad (8)$$

$$\mathbb{E}(\pi(s, a_\delta)) \approx \mathbb{E}(\pi(s, a_{\max})) \leq \frac{1}{M}. \quad (9)$$

Therefore, in each training round, **sampled responses with high advantage  $A_{high}$  often have low probabilities  $\pi(a|s)$** . According to Equation 4, this phenomenon intuitively results in a larger covariance, which in turn leads to higher entropy. The experimental results in Figure 3 confirm this phenomenon. Obviously and intuitively, when the entropy of the answer is high, numerous high-entropy tokens also appear in the thinking process. In addition, we offer an empirical discussion of underlying causes in the Appendix G.

Overall, the behavior of entropy in visual grounding differs markedly from that in reasoning tasks. In visual grounding, a substantial number of high-entropy tokens persist, indicating that the model consistently operates under high uncertainty and low confidence when generating responses. This means that the entropy in the visual grounding task behaves differently; therefore, it is necessary to

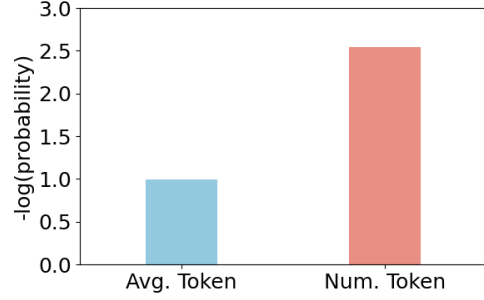


Figure 3: Comparison of the average probability of all tokens and numeric tokens in positive advantage samples during VLM-R1’s training.

design specific strategies to regulate it for a better balance between exploration and exploitation.

#### 4 Method: ECVGPO

Inspired by the findings in Section 3, we introduce the **Entropy Control Visual Grounding Policy Optimization (ECVGPO)** algorithm for visual grounding. The key idea is to adjust the entropy by adjusting the covariance of probability and advantage at training time. Unlike reasoning tasks, which only aim to increase entropy to facilitate exploration, we consider reducing entropy when it is too high, improving the model’s ability to exploit good policies. By attenuating policy updates in such cases, ECVGPO suppresses the reinforcement of unreliable patterns, thereby fostering the acquisition of more robust and generalizable representations. Appendix C provides a pseudo-code implementation of ECVGPO.

We first compute the self-information for each candidate sequence and for each token:

$$S_i = -\frac{1}{T} \sum_{t=1}^T \log \pi_\theta(o_{i,t} | q, o_{i,<t}), \quad (10)$$

$$S_{i,t} = -\log \pi_\theta(o_{i,t} | q, o_{i,<t}), \quad (11)$$

where  $o_{i,t}$  denotes the  $t$ -th token of the  $i$ -th sampled sequence  $O_i$ , and  $T$  is the sequence length.

According to the definition of self-information, higher self-information corresponds to lower confidence in the output response, and vice versa (Fu et al., 2025).

The standard deviation of the rewards for the sampled responses is computed as follow:

$$r_{std} = \sqrt{\frac{1}{|O|} \sum_{i=1}^{|O|} (r_i - \frac{1}{|O|} \sum_{i=1}^{|O|} r_i)^2}. \quad (12)$$

A high reward combined with a low standard deviation indicates that the model has limited room for further policy updates.

After that, we reshape the advantages of  $i$ -th positive samples as follows:

$$A'_i = \begin{cases} \max\left(\frac{A_{i,t}}{l_0}, A_{i,t} - \frac{S_{i,t}}{r_0}\right) & A_i > 0 \text{ \& } r_{std} < \delta \\ A_i & A_i < 0 \end{cases} \quad (13)$$

Here,  $A_i$  denotes the advantage of the  $i$ -th sample, as obtained in Equation 19. Specifically,  $A_i$  is a vector of token-level advantages, where each element corresponds to the advantage  $A_{i,t}$  of the  $t$ -th token in the sequence  $O_i$ . The coefficient  $l_0 > 1$  is the lower-bound clipping factor, which is always greater than 0 to ensure that the advantage remains positive after the reshape, thereby preventing a positive example from being transformed into a negative one. The coefficient  $|r_0| > 1$  is used to control the change in entropy. By enforcing  $r_{std} < \delta$ , we restrict the intervention to situations where the training is stable, preventing instability caused by premature adjustments.

Finally, we obtain the new training objective:

$$\mathcal{J}_{ECVGPO}(\theta) = \mathbb{E}_{\{o_i\}_{i=1}^N \sim \pi_{\theta_{old}}(q)} \left[ \frac{1}{N} \sum_{i=1}^N \left\{ \min[C_1, C_2] - \beta \mathbb{D}_{KL}[\pi_{\theta} \parallel \pi_{ref}] \right\} \right], \quad (14)$$

$$C_1 = \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} \cdot A'_i, \quad (15)$$

$$C_2 = \text{clip} \left( \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) \cdot A'_i. \quad (16)$$

In order to analyze the interpretability of ECVGPO, we consider the covariance between the log-probability and the reshaped advantage. Although the reshaping in Equation 13 is applied at the token level, we can approximate the sequence-level covariance as:

$$\text{Cov}_{a \sim \pi_{\theta}^k} \left( \log \pi_{\theta}^k(a|s), \pi_{\theta}^k(a|s) \left[ A'_i - \frac{\log \pi_{\theta}^k(a|s)}{r_0} \right] \right), \quad (17)$$

where  $A'_i = \frac{1}{T} \sum_{t=1}^T A'_{i,t}$  is the mean of the token-level reshaped advantages in sequence  $O_i$ .

This approximation allows us to interpret the effect of the token-level reshape on entropy: for positive samples, high advantage and low probability tokens contribute to the covariance, and applying  $r_0 > 0$  decreases this covariance, reducing the

model entropy. Conversely,  $r_0 < 0$  increases the covariance and thus the entropy. The token-level intervention is designed to be minimal while still aligning the policy updates with sequence-level dynamics, ensuring interpretability of ECVGPO even under token-level reshaping.

## 5 Experiment

### 5.1 Setup

**Implementation detail** We use the training splits of RefCOCO, RefCOCO+, and RefCOCOg (Kazemzadeh et al., 2014; Yu et al., 2016) as our training data, which are the most widely used datasets for the visual grounding task. We randomly selected a total of 5.6k samples from these three datasets. Our codebase is based on VLM-R1 (Shen et al., 2025a). Specifically, we set rollout = 8, temperature = 1.0, number of iterations = 1, KL divergence ratio (i.e.,  $\beta$ ) = 0.04, learning rate =  $1 \times 10^{-6}$ , batch size = 4, gradient accumulation = 1. The experiments are conducted with  $8 \times \text{H100}$  GPUs. See Appendix F for more Settings.

### 5.2 Baseline

We choose three state-of-the-art models, **Qwen2.5-VL-3B-Instruct**, **InternVL3-2B-Instruct** and **InternVL3-8B-Instruct**, as our base models. We compare our proposed method with three variants: the base model, the Supervised Fine-Tuning (SFT) model, and the GRPO-based fine-tuning model. For a fair comparison, the base model and SFT-based model are trained and evaluated in the same format as the base model’s standard documentation. GRPO-based method uses the same training data as ECVGPO, and the training settings for GRPO are kept consistent with those of ECVGPO.

**Benchmark** For evaluation, we use the validation split of RefCOCO+/g (Kazemzadeh et al., 2014; Yu et al., 2016) as the in-domain benchmark, while the test splits of LISA-Grounding (Lai et al., 2024) and RefGTA (Tanaka et al., 2019) serve as out-of-domain benchmarks. As a reasoning-centric benchmark, LISA-Grounding requires models to handle complex tasks such as detailed visual analysis, understanding nuanced referring expressions, and performing relational reasoning among objects to correctly determine the target region. Meanwhile, RefGTA is constructed from game scenes rather than the general scenarios in RefCOCO. We adopt the same subset used in VLM-R1 (Shen et al.,

2025a). These two out-of-domain test sets are employed to evaluate the generalization ability of the model under different perception and reasoning scenarios.

### 5.3 Main Result

**Comparison of ECVGPO with baseline methods** Table 1 presents the comparison between our proposed ECVGPO and other baseline methods, including SFT and GRPO, across both in-domain and out-of-domain benchmarks on Qwen2.5-VL-3B, InternVL3-2B, and InternVL3-8B. To ensure fairness, all models adopt the official question-answering templates used in the base and SFT versions. Considering the high variance of GRPO-type methods in visual grounding tasks, we report both the mean and standard deviation over four independent runs for each configuration.

As shown in the table, ECVGPO achieves superior performance compared with both GRPO and SFT across overall. On Qwen2.5-VL-3B, our method reaches an overall score of 78.37, outperforming GRPO (78.17) and clearly surpassing the SFT baseline (70.32). The improvement becomes more pronounced for larger models: ECVGPO yields a +1.44 gain on InternVL3-2B and +2.60 on InternVL3-8B, respectively. These results demonstrate that the proposed entropy-controlled mechanism enables more stable and effective policy optimization, leading to benefits over both supervised and GRPO baselines.

In terms of stability, ECVGPO shows smaller standard deviations across most benchmarks, indicating improved training consistency. This confirms that our method effectively mitigates the high-variance issue often observed in GRPO-type algorithms by adaptively controlling entropy to maintain a balanced exploration–exploitation trade-off.

Overall, ECVGPO achieves both higher accuracy and greater stability than existing baselines, proving its effectiveness and scalability for MLLMs.

**Comparison of training efficiency** As shown in Figure 4, the training time of our method is comparable to that of GRPO across all models. Specifically, for Qwen2.5-VL-3B, our method slightly reduces the training time (200 min vs. 204 min), while for InternVL3-2B and InternVL3-8B, the difference is marginal (within  $\pm 3$  minutes). This indicates that incorporating our proposed optimization introduces almost no additional computational

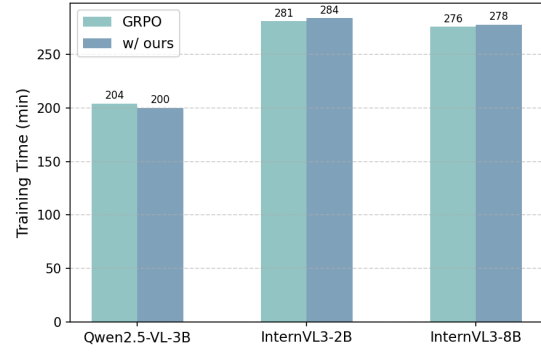


Figure 4: Comparison of the training time of GRPO and our method.

overhead, maintaining a similar level of training efficiency as the baseline GRPO. Therefore, our method maintains both effectiveness and efficiency, making it suitable for MLLMs’ reinforcement fine-tuning.

**Comparison of entropy trends** Figure 5 illustrates the trend of policy entropy under different values of  $r_0$  on the InternVL3-8B model. The results show that the change in policy entropy is fully consistent with the trend of  $1/r_0$ . Specifically, smaller (positive) values of  $r_0$  correspond to lower entropy, while larger (negative) values of  $r_0$  lead to higher entropy. This inverse relationship demonstrates that our method can effectively and smoothly control the entropy of the policy. Such controllability and consistency indicate that the proposed entropy control mechanism is both interpretable and stable, allowing the model to dynamically balance exploration and exploitation during training.

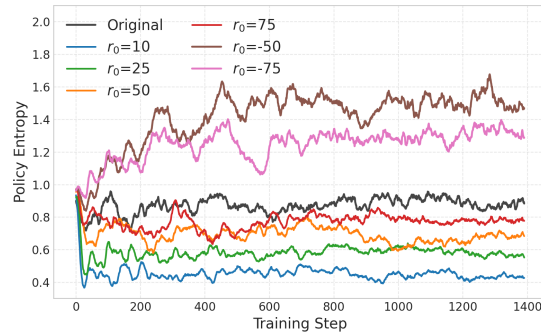


Figure 5: Trend of entropy under different parameters  $r_0$ . The experiments are conducted on the InternVL3-8B model. The lines in the figure are smoothed.

### 5.4 Ablation Study

**Comparison with other entropy-regularization methods** As shown in Table 2, we compare

| Method                 | In-Domain        |                  |                  |                  |                  |                  |                  |                  | Out-of-Domain    |                  | Overall |
|------------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|---------|
|                        | RefCOCO          | RefCOCO          | RefCOCO          | RefCOCO+         | RefCOCO+         | RefCOCO+         | RefCOCOg         | RefCOCOg         | LISA             | RefGTA           |         |
|                        | (val)            | (testA)          | (testB)          | (val)            | (testA)          | (testB)          | (val)            | (test)           | (test)           | (test)           |         |
| Qwen2.5-VL-3B-Instruct |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |         |
| Base Model             | 88.05            | 91.55            | 83.20            | 79.35            | 86.65            | 72.35            | 84.80            | 84.40            | 56.82            | 72.45            | 74.21   |
| w/ SFT                 | 88.75            | 91.85            | 84.80            | 80.35            | 87.40            | 74.15            | 85.15            | 85.25            | 42.10            | 69.75            | 70.32   |
| w/ GRPO                | 90.16 $\pm$ 0.18 | 93.00 $\pm$ 0.43 | 86.16 $\pm$ 0.41 | 83.59 $\pm$ 0.45 | 89.34 $\pm$ 0.42 | 76.53 $\pm$ 0.48 | 86.51 $\pm$ 0.35 | 86.29 $\pm$ 0.55 | 65.90 $\pm$ 1.14 | 73.90 $\pm$ 1.00 | 78.17   |
| w/ Ours                | 90.25 $\pm$ 0.39 | 93.16 $\pm$ 0.20 | 86.49 $\pm$ 0.49 | 83.75 $\pm$ 0.36 | 89.15 $\pm$ 0.20 | 76.57 $\pm$ 0.52 | 86.86 $\pm$ 0.25 | 86.49 $\pm$ 0.17 | 66.02 $\pm$ 0.24 | 74.27 $\pm$ 1.00 | 78.37   |
| InternVL3-2B-Instruct  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |         |
| Base Model             | 89.80            | 93.20            | 86.80            | 83.45            | 88.95            | 76.25            | 88.15            | 86.00            | 0.13**           | 26.50            | 49.95   |
| w/ SFT                 | 86.65            | 91.20            | 84.55            | 78.9             | 85.8             | 70.95            | 85.5             | 84.3             | 48.91            | 13.50            | 57.34   |
| w/ GRPO                | 89.21 $\pm$ 0.35 | 92.58 $\pm$ 0.28 | 86.62 $\pm$ 0.33 | 81.95 $\pm$ 0.72 | 88.82 $\pm$ 0.49 | 74.67 $\pm$ 0.55 | 87.50 $\pm$ 0.32 | 86.87 $\pm$ 0.38 | 51.12 $\pm$ 5.63 | 51.26 $\pm$ 2.48 | 68.61   |
| w/ Ours                | 89.49 $\pm$ 0.51 | 92.73 $\pm$ 0.17 | 86.50 $\pm$ 0.17 | 81.80 $\pm$ 0.40 | 88.76 $\pm$ 0.25 | 74.70 $\pm$ 0.49 | 87.56 $\pm$ 0.32 | 86.80 $\pm$ 0.25 | 59.26 $\pm$ 1.76 | 53.49 $\pm$ 1.85 | 71.21   |
| InternVL3-8B-Instruct  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |         |
| Base Model             | 93.00            | 95.05            | 88.65            | 88.40            | 92.05            | 81.35            | 89.35            | 90.05            | 0.84**           | 28.65            | 52.24   |
| w/ SFT                 | 88.3             | 91.05            | 85.05            | 81.65            | 88.60            | 75.05            | 86.20            | 87.05            | 58.38            | 33.05            | 65.54   |
| w/ GRPO                | 91.65 $\pm$ 0.65 | 94.29 $\pm$ 0.44 | 88.01 $\pm$ 0.31 | 87.33 $\pm$ 0.52 | 91.96 $\pm$ 0.60 | 80.51 $\pm$ 0.90 | 88.97 $\pm$ 0.46 | 89.86 $\pm$ 0.52 | 78.78 $\pm$ 1.45 | 48.80 $\pm$ 6.30 | 76.43   |
| w/ Ours                | 92.15 $\pm$ 0.15 | 94.42 $\pm$ 0.28 | 88.53 $\pm$ 0.13 | 88.13 $\pm$ 0.33 | 92.36 $\pm$ 0.52 | 81.18 $\pm$ 0.71 | 89.43 $\pm$ 0.41 | 90.06 $\pm$ 0.26 | 81.14 $\pm$ 0.45 | 51.29 $\pm$ 6.24 | 77.87   |

Table 1: Compare results of ECVGPO with other baseline methods on in-domain and out-of-domain benchmarks. We set  $1/r_0 = 1/50, 1/25, -1/50$  for each model. For a fair comparison, the question-answering templates used in the base and SFT models follow the official templates. Considering the high variance of GRPO-type methods on the visual grounding task, we report the mean and standard deviation over four runs. Overall represents the average score of the in-domain and out-of-domain performance. \*\* indicates that the model does not follow the format correctly. The best performances within each metric are **bolded**.

| Setting        | RefCOCO <sub>val</sub>  | RefCOCO <sub>+val</sub> | RefCOCO <sub>gval</sub> | LISA                    |
|----------------|-------------------------|-------------------------|-------------------------|-------------------------|
| GRPO           | 91.65 $\pm$ 0.65        | 87.33 $\pm$ 0.52        | 88.97 $\pm$ 0.46        | 78.78 $\pm$ 1.45        |
| w/ Clip-higher | 91.96 $\pm$ 0.34        | 87.56 $\pm$ 0.22        | 89.04 $\pm$ 0.25        | 79.43 $\pm$ 1.59        |
| w/ En. Re.     | 1.25                    | 1.15                    | 1.5                     | 0.72                    |
| w/ En. Adv.    | <b>92.40</b> $\pm$ 0.40 | 87.68 $\pm$ 0.94        | 89.21 $\pm$ 0.37        | 78.90 $\pm$ 1.05        |
| w/ Ours        | 92.15 $\pm$ 0.15        | <b>88.13</b> $\pm$ 0.33 | <b>89.43</b> $\pm$ 0.41 | <b>81.14</b> $\pm$ 0.45 |

Table 2: Result of different entropy regularization method. The best performance is **bolded**. The experiments are conducted on the InternVL-8B model. We report the mean and standard deviation over four runs.

different entropy regularization strategies under the InternVL-8B model. It can be observed that directly applying the entropy-based methods commonly used in reasoning-oriented reinforcement learning tasks (e.g., Entropy Reward (Bai et al., 2022; Yu et al., 2022) and Entropy Advantage (Cheng et al., 2025a)) is not well suited for the visual grounding task. In particular, the Entropy Reward method leads to severe training instability and performance collapse, reflected by extremely low scores across all benchmarks. This phenomenon indicates that simply increasing policy entropy fails to provide meaningful exploration in grounding scenarios, where visual information plays a dominant role. In contrast, our proposed method achieves more strong and stable results across all datasets, outperforming the GRPO baseline and the Entropy Advantage method. Notably, our approach improves performance on out-of-domain datasets such as LISA, demonstrating both its effectiveness and robustness. These results validate

that our entropy regularization design better balances exploration and stability in visual grounding.

**More result** Due to space limitations, additional ablation results, hyperparameter(e.g.  $l_0, r_0, \delta$ ) selection strategy and sensitivity analysis are provided in the Appendix H.

## 5.5 Case Study

Appendix I presents a case-by-case comparison between our proposed method and the GRPO method.

## 6 Conclusion

In this work, we present a comprehensive analysis of entropy behavior in perception-oriented reinforcement learning, focusing on the visual grounding task. Our study reveals that entropy exhibits distinct dynamics compared to reasoning tasks, where high-advantage responses often correspond to low-generation probabilities, leading to persistently high entropy. Based on these insights, we propose ECVGPO, a simple yet effective extension of GRPO that reshapes advantages according to self-information to achieve more stable and interpretable optimization. Extensive experiments on multiple multimodal large language models and benchmarks demonstrate that ECVGPO improves grounding accuracy, stability, and generalization without introducing additional training costs. We hope our findings provide a new perspective on understanding entropy in multimodal

reinforcement learning and inspire further research into perception-oriented policy optimization.

## Limitation

While ECVGPO demonstrates strong performance and stability improvements, several limitations remain. First, due to the intrinsic uncertainty caused by high entropy in visual grounding, the evaluation results can still exhibit noticeable fluctuations, which may pose challenges for reproducibility. Second, the method involves several hyperparameters that require careful tuning, making the optimization process time-consuming and somewhat inconvenient in practice. Finally, although our experiments cover multiple benchmarks, further validation on broader scenarios and diverse multimodal tasks is needed to fully assess the generality and robustness of ECVGPO.

## Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No.62521004, 62476061, 62576106, 62376061).

## References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025a. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. 2025b. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *Preprint*, arXiv:2505.14231.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Huayu Chen, Kaiwen Zheng, Qinsheng Zhang, Ganqu Cui, Yin Cui, Haotian Ye, Tsung-Yi Lin, Ming-Yu Liu, Jun Zhu, and Haoxiang Wang. 2025. Bridging supervised learning and reinforcement learning in math reasoning. *arXiv preprint arXiv:2505.18116*.
- Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S. H. Gary Chan, and Hongyang Zhang. 2024a. Revisiting referring expression comprehension evaluation in the era of large multimodal models. *Preprint*, arXiv:2406.16866.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.
- Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025a. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.
- Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025b. Reasoning with exploration: An entropy perspective on reinforcement learning for llms. *Preprint*, arXiv:2506.14758.
- 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. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *Preprint*, arXiv:2505.22617.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.
- Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, and 4 others. 2024. Internlm-xcomposer2: Mastering

- free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.
- Xiaoran Fan, Tao Ji, Changhao Jiang, Shuo Li, Senjie Jin, Sirui Song, Junke Wang, Boyang Hong, Lu Chen, Guodong Zheng, and 1 others. 2024. Mousi: Poly-visual-expert vision-language models. *arXiv preprint arXiv:2401.17221*.
- Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. 2025. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*.
- Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. [Deep think with confidence](#). *Preprint*, arXiv:2508.15260.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Senjie Jin, Peixin Wang, Boyang Liu, Xiaoran Fan, Shuo Li, Zhiheng Xi, Jiazhang Zhang, Yuhao Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2026. [Entropy is not enough: Unlocking effective reinforcement learning for visual reasoning via vision-anchored token selection](#). *Preprint*, arXiv:2606.03937.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.
- Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. [Lisa: Reasoning segmentation via large language model](#). *Preprint*, arXiv:2308.00692.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. [BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. [Llava-next: Improved reasoning, ocr, and world knowledge](#). Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Jiacai Liu. 2025. How does rl policy entropy converge during iteration? <https://zhuanlan.zhihu.com/p/28476703733>.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025a. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. [Visual-rft: Visual reinforcement fine-tuning](#). *Preprint*, arXiv:2503.01785.
- OpenAI. 2022. ChatGPT: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt/>.
- OpenAI. 2024a. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2024b. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.
- Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b llms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). *Preprint*, arXiv:2103.00020.
- Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. 2025. [Grounded reinforcement learning for visual reasoning](#). *Preprint*, arXiv:2505.23678.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025a. [Vlm-r1: A stable and generalizable r1-style large vision-language model](#). *Preprint*, arXiv:2504.07615.
- Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025b. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

- Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. 2025. Ce-gppo: Controlling entropy via gradient-preserving clipping policy optimization in reinforcement learning. *arXiv preprint arXiv:2509.20712*.
- Hongze Tan and Jianfei Pan. 2025. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy. *arXiv preprint arXiv:2508.04349*.
- Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. 2019. [Generating easy-to-understand referring expressions for target identifications](#). *Preprint*, arXiv:1811.12104.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. [Cambrian-1: A fully open, vision-centric exploration of multimodal llms](#). *Preprint*, arXiv:2406.16860.
- Haoran Wang, Yue Zhang, and Xiaosheng Yu. 2020. An overview of image caption generation methods. *Computational intelligence and neuroscience*, 2020(1):3062706.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025. [Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning](#). *Preprint*, arXiv:2506.01939.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Haonan Yu, Haichao Zhang, and Wei Xu. 2022. Do you need the entropy reward (in practice)? *arXiv preprint arXiv:2201.12434*.
- Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. [Modeling context in referring expressions](#). *Preprint*, arXiv:1608.00272.
- Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.
- Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, and 2 others. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.
- Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity. *arXiv preprint arXiv:2507.21848*.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

## Appendix

### A RLVR Training in Visual Grounding Task

#### A.1 GRPO Training

Given a query  $q$ , the policy  $\pi_\theta$  samples  $N$  candidate outputs  $\{O_1, O_2, \dots, O_N\}$ , and each candidate is evaluated by a task-specific reward function  $R(q, O_i)$ . The relative quality is measured by computing the standardized advantage for each token in the sequence. First, we calculate the sequence-level advantage for each candidate output  $O_i$  as:

$$A_i = \frac{r_i - \text{mean}\{r_1, r_2, \dots, r_N\}}{\text{std}\{r_1, r_2, \dots, r_N\}}, \quad (18)$$

where  $r_i$  is the reward for the entire sequence  $O_i$ , and  $A_i$  is the sequence-level advantage.

Then, the token-level advantage for each token  $t$  in candidate output  $O_i$  is equal to the sequence-level advantage:

$$A_{i,t} = A_i, \quad \forall t \in \{1, 2, \dots, T\}, \quad (19)$$

where  $A_{i,t}$  is the advantage for token  $t$  in the sequence  $O_i$  and  $T$  is the length of the sequence. The policy is optimized to increase the probability of responses with higher  $A_i$  while constraining deviation from the reference model via a KL penalty:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{\{O_i\}_{i=1}^N \sim \pi_{\theta_{old}}(q)} \left[ \frac{1}{N} \sum_{i=1}^N \left\{ \min[C_1, C_2] - \beta \mathbb{D}_{KL}[\pi_\theta \parallel \pi_{ref}] \right\} \right], \quad (20)$$

$$C_1 = \frac{\pi_\theta(O_i|q)}{\pi_{\theta_{old}}(O_i|q)} \cdot A_i, \quad (21)$$

$$C_2 = \text{clip} \left( \frac{\pi_\theta(O_i|q)}{\pi_{\theta_{old}}(O_i|q)}, 1 - \epsilon, 1 + \epsilon \right) \cdot A_i. \quad (22)$$

Here,  $\epsilon$  is the clipping range, and  $\beta$  controls the KL regularization strength.

#### A.2 Reward setting

For the Visual Grounding task, consistent with (Shen et al., 2025a; Liu et al., 2025b), reward  $R = R_{acc} + R_f$  consists of two reward components:

**Accuracy reward** In the visual grounding task, the objective is to predict the bounding box of the target entity described in the input query. Let  $\hat{b}$  be the predicted bounding box obtained from the model output  $y$ , and  $b_{gt}$  be the ground truth box. The accuracy reward is defined as:

$$R_{acc}(x, y) = \text{IoU}(b_{gt}, y), \quad (23)$$

where IoU denotes the Intersection-over-Union score, which measures the overlap ratio between the predicted and reference boxes, thus incentivizing accurate localization.

**Format reward** Format reward  $R_f$  checks whether the output follows the required JSON-style structure within the `<answer>` tag and contains bounding box coordinates in the format: `(<think>...</think><answer>{[x1, y1, x2, y2]}</answer>)`, assigning 1 if correct and 0 otherwise.

By jointly optimizing these rewards under the GRPO framework, the model learns to produce both accurate and well-formatted responses for visual grounding.

### B Proof of Policy Entropy Change Approximation

This section provides the detailed proof for the approximation of the policy entropy difference between two consecutive training steps, assuming a **Tabular Softmax Policy** updated via **Vanilla Policy Gradient (VPG)** with a learning rate  $\eta$ . The proof process is derived from (Liu, 2025) and (Cui et al., 2025).

**Theorem 1 (Entropy Change under Policy Gradient)** Assuming the actor policy  $\pi_\theta$  is a tabular Softmax policy updated by VPG, the difference in policy entropy at state  $s$  between two consecutive steps is approximated by:

$$\begin{aligned} \mathcal{H}(\pi_\theta^{k+1}|s) - \mathcal{H}(\pi_\theta^k|s) \approx \\ - \eta \cdot \text{Cov}_{a \sim \pi_\theta^k(\cdot|s)}(\log \pi_\theta^k(a|s), \pi_\theta^k(a|s) \cdot A(s, a)) \end{aligned} \quad (24)$$

**Proof** The proof proceeds in three main steps: 1) expressing the entropy difference in terms of logit changes, 2) deriving the logit changes under the VPG update rule, and 3) combining the results.

#### Step 1: Entropy Difference based on Logit Change

For a Softmax policy  $\pi_\theta(a|s)$ , the change in policy entropy between steps  $k$  and  $k+1$  can be approximated using a first-order Taylor expansion (or a

known result for Softmax policies) in terms of the logit difference  $z_{s,a}^{k+1} - z_{s,a}^k$ :

$$\mathcal{H}(\pi_{\theta}^{k+1}|s) - \mathcal{H}(\pi_{\theta}^k|s) \approx -\text{Cov}_{a \sim \pi_{\theta}^k(\cdot|s)} \left( \log \pi_{\theta}^k(a|s), z_{s,a}^{k+1} - z_{s,a}^k \right) \quad (25)$$

### Step 2: Logit Change under VPG Update

For a tabular Softmax policy, the logit parameter  $z_{s,a} = \theta_{s,a}$  is updated by VPG using the gradient  $\nabla_{\theta_{s,a}} J(\theta)$ :

$$z_{s,a}^{k+1} - z_{s,a}^k = \eta \cdot \nabla_{\theta_{s,a}} J(\theta)$$

The policy gradient at state  $s$  is:

$$\nabla_{\theta_{s,a}} J(\theta) = \mathbb{E}_{a' \sim \pi_{\theta}(\cdot|s)} [\nabla_{\theta_{s,a}} \log \pi_{\theta}(a'|s) \cdot A(s, a')] \quad (26)$$

Using the identity for the gradient of the log-probability of a Softmax function,  $\frac{\partial \log \pi_{\theta}(a'|s)}{\partial \theta_{s,a}} = \mathbb{I}_{a'=a} - \pi_{\theta}(a|s)$ , we have:

$$\begin{aligned} \nabla_{\theta_{s,a}} J(\theta) &= \sum_{a' \in \mathcal{A}} [\pi_{\theta}(a'|s) \cdot (\mathbb{I}_{a'=a} - \pi_{\theta}(a|s)) \cdot A(s, a')] \\ &= \pi_{\theta}(a|s) \cdot \left[ A(s, a) - \sum_{a' \in \mathcal{A}} \pi_{\theta}(a'|s) \cdot A(s, a') \right] \end{aligned} \quad (27)$$

Under the standard VPG assumption that the advantage function is centered,  $\sum_{a'} \pi_{\theta}(a'|s) \cdot A(s, a') = \mathbb{E}_{a' \sim \pi_{\theta}(\cdot|s)} [A(s, a')] = 0$ , the gradient simplifies to:

$$\nabla_{\theta_{s,a}} J(\theta) = \pi_{\theta}(a|s) \cdot A(s, a)$$

Substituting this back into the logit difference equation yields:

$$z_{s,a}^{k+1} - z_{s,a}^k = \eta \cdot \pi_{\theta}^k(a|s) \cdot A(s, a) \quad (28)$$

### Step 3: Final Substitution and Conclusion

Substituting Equation 28 into the entropy approximation in Equation 25:

$$\begin{aligned} \mathcal{H}(\pi_{\theta}^{k+1}|s) - \mathcal{H}(\pi_{\theta}^k|s) &\approx -\text{Cov}_{a \sim \pi_{\theta}^k(\cdot|s)} \left( \log \pi_{\theta}^k(a|s), \eta \cdot \pi_{\theta}^k(a|s) \cdot A(s, a) \right) \end{aligned} \quad (29)$$

Factoring the constant learning rate  $\eta$  out of the covariance, we obtain the final result:

$$\begin{aligned} \mathcal{H}(\pi_{\theta}^{k+1}|s) - \mathcal{H}(\pi_{\theta}^k|s) &\approx -\eta \cdot \text{Cov}_{a \sim \pi_{\theta}^k(\cdot|s)} (\log \pi_{\theta}^k(a|s), \pi_{\theta}^k(a|s) \cdot A(s, a)) \end{aligned} \quad (30)$$

## Experimental Evidence

Our experiments reveal a clear linear correlation between covariance and entropy difference during training. As shown in the Figure 6, the Spearman correlation coefficient increases from 0.52 globally to 0.60 after 200 steps, demonstrating a growing linear relationship between these two metrics as training progresses. The curves for covariance (orange) and entropy difference (blue) exhibit a consistent, parallel trend, further supporting the presence of this linear correlation. This finding suggests that entropy behavior directly influences the optimization process in a predictable manner, highlighting the importance of entropy regulation for model stability and performance.

### C Pseudo-code Implementation of ECVGPO

We provide Pseudo-code implementation of ECVGPO as shown in Algorithm 1.

### D Quantitative Analysis of the Misalignment between Advantage and Probability

As illustrated in Figure 7, our quantitative investigation reveals a profound global misalignment between the advantage and probability. Specifically, the observed negligible global correlation ( $r = 0.0777$ ) and minimal covariance (0.0151) indicate a statistical decoupling where rewards fail to effectively guide policy convergence. This phenomenon is further evidenced by the prevalence of high-advantage, low-probability sequences—samples residing in the lower-right quadrant of the distribution.

### E Example of a comparison between reasoning and the visual Grounding task

We provide an Example of a comparison between reasoning and the visual Grounding task as shown in Figure 8.

### F More Training Setting

Table 3 presents a detailed training setup.

### G More Discussion for Low Probability Tokens

During the thinking process, we also observe a large number of low-probability tokens. We at-

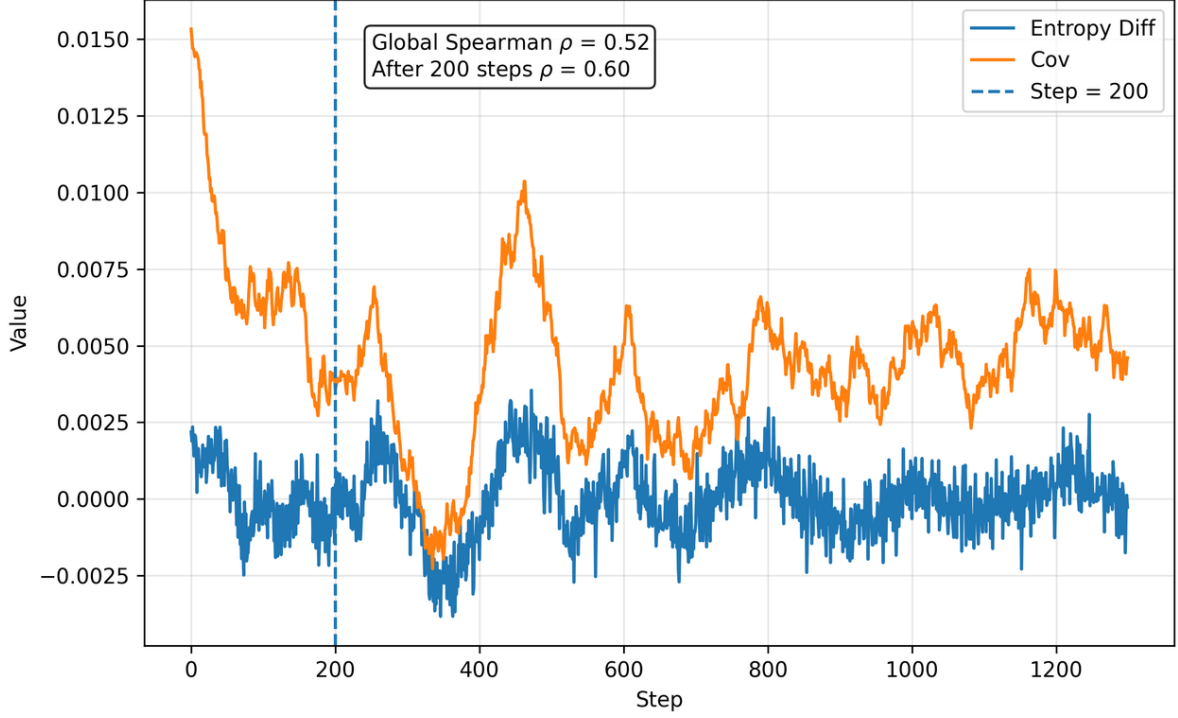


Figure 6: Comparison of covariance and entropy difference during training. The lines in the figure are smoothed.

tribute this phenomenon to two main qualitative reasons:

### G.1 Mismatch between thinking steps and rewards

In reasoning tasks, the model typically performs step-by-step thinking, where each intermediate step explicitly contributes to the final answer generation. In contrast, for visual grounding tasks, only the final coordinate prediction receives a reward, making it difficult to determine which parts of the reasoning process actually influence the final output.

### G.2 Limited in-domain performance improvement

After GRPO training, the model (especially the InternVL3 series) shows little improvement over the base model on in-domain data. This suggests that the model struggles to learn a stable and effective thinking strategy under such training settings, which may further lead to the frequent occurrence of low-probability tokens.

## H More Ablation Study

### H.1 Detailed training analysis

Figure 9 compares the reward trends of our proposed EC-GRPO method and the GRPO baseline

during training on Qwen2.5-VL-3B, InternVL3-2B, and InternVL3-8B. Across all models, ECVGPO consistently achieves higher average rewards and exhibits smoother convergence compared to GRPO. This indicates that our method enables more effective policy updates and better balances exploration and exploitation during training.

### H.2 Comparison with different training data

we experiment with different types of training data and find that RefCOCO, a general visual grounding dataset covering a wide range of scenarios, yields the best performance, as shown in Table 4. In contrast, RefGTA, which contains limited and less diverse scenarios, often leads to training instability or even collapse.

### H.3 Comparison of different coefficients

Table 5 presents the performance of ECVGPO under different  $r_0$  values on the InternVL3-8B model. Overall, the results demonstrate that the choice of  $r_0$  has a clear influence on model performance. When  $r_0 = -50$ , the model achieves the best results across all benchmarks, surpassing the GRPO baseline by a large margin (+0.5 on RefCOCO, +0.8 on RefCOCO+, +0.5 on RefCOCog, and +2.4 on LISA). This suggests that a moderate negative  $r_0$  provides a proper level of entropy increase, en-

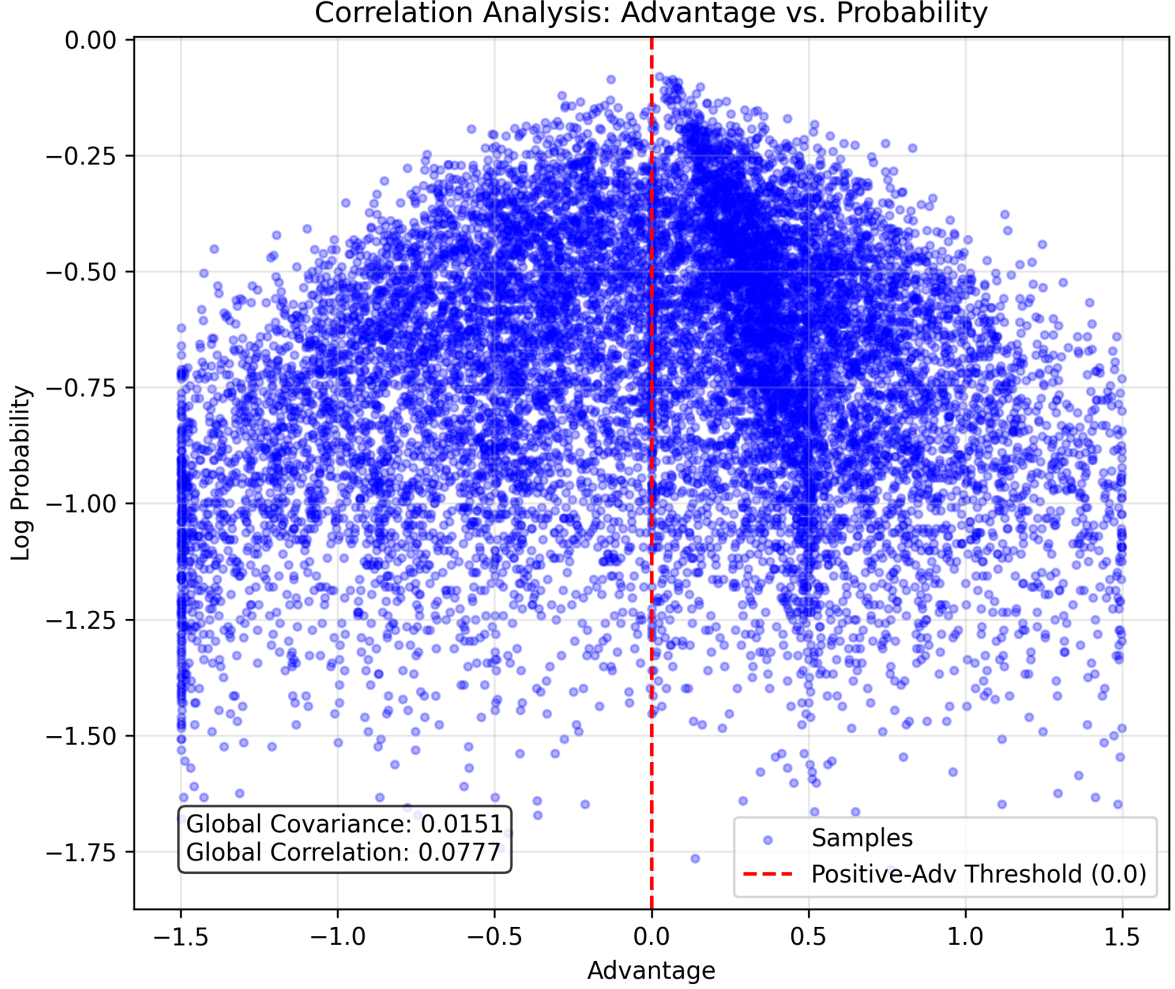


Figure 7: Scatter plot of Advantage vs. Log-Probability in Visual Grounding.

hancing exploration while maintaining stable optimization. When  $r_0$  becomes positive (e.g.,  $r_0 = 50$  or  $r_0 = 75$ ), the entropy is suppressed, leading to slightly worse performance due to insufficient exploration. These results validate that controlling entropy through  $r_0$  effectively regulates the trade-off between exploration and exploitation, and that setting  $r_0$  to a small negative value yields the best balance.

#### H.4 Comparison with different temperature

As shown in Table 9, we evaluate the impact of different sampling temperatures on model performance. Our method consistently outperforms the GRPO baseline across all temperature settings, demonstrating strong robustness and generalization. Specifically, at the standard temperature of 1.0, our method achieves better results, while maintaining competitive performance when the temperature is varied to 0.8 or 1.2.

#### H.5 Comparison with different rollout

As shown in Table 10, we investigate the effect of different rollout numbers during training. To prevent model collapse, a smaller intervention ( $r_0=75$ ) was used for rollout=16. Our method consistently surpasses the GRPO baseline across all rollout settings, indicating its robustness and stability. Notably, the improvement is most evident when the rollout is set to 8, where our approach achieves the best overall performance on all benchmarks.

#### H.6 Hyperparameter Selection and Ablation

**Analysis of  $r_0$ .** We observe that the sign of  $r_0$  should be determined by the intrinsic entropy level of the model during training. As shown in Figure 10, smaller models (e.g., Qwen2.5-VL-3B and InternVL3-2B) tend to maintain relatively higher entropy, indicating excessive randomness in policy exploration. In this case, using  $r_0 > 0$  effectively suppresses entropy and stabilizes optimization. In

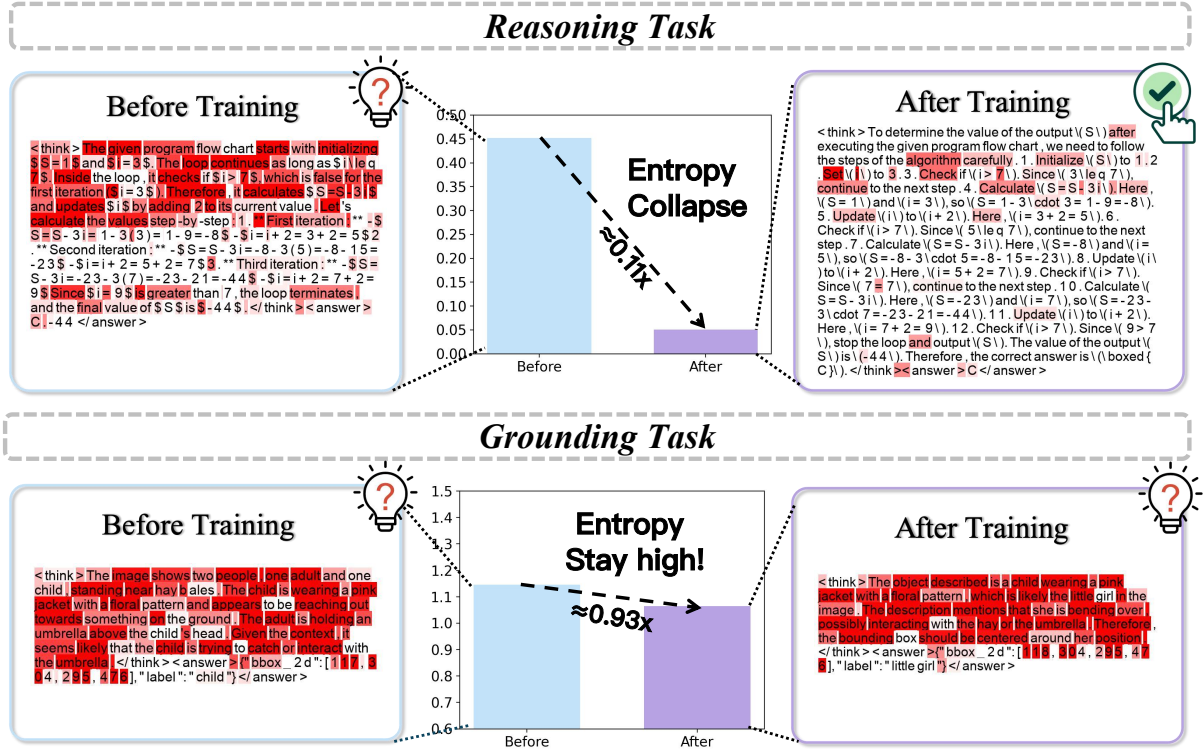


Figure 8: Comparison of entropy between reasoning and grounding tasks. Darker red indicates higher token entropy. In reasoning tasks, entropy rapidly decreases after training (entropy collapse), whereas in grounding tasks, it remains consistently high.

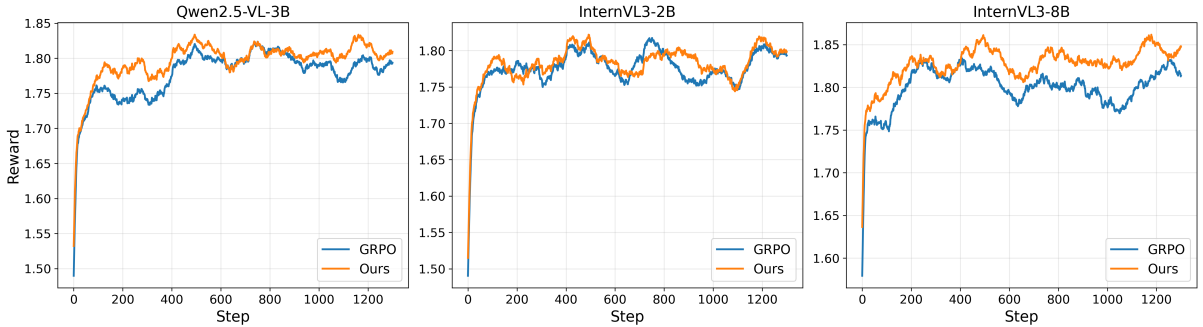


Figure 9: Comparison of reward's trend between our proposed method and the GRPO method during training. The lines in the figure are smoothed.

contrast, larger models (e.g., Qwen2.5-VL-7B and InternVL3-8B) exhibit lower entropy, suggesting insufficient exploration. Therefore, a negative  $r_0$  is adopted to encourage exploration by increasing entropy. This entropy-aware strategy provides a simple yet consistent guideline for selecting  $r_0$ , as summarized in Table 6.

**Analysis of  $\delta$ .** We determine the threshold  $\delta$  according to the intrinsic  $r_{\text{std}}$  level during stable training. As shown in Figure 11, Qwen models typically converge to  $r_{\text{std}} < 0.1$ , suggesting that  $\delta = 0.1$  is a reasonable threshold for detecting unstable optimization. In contrast, InternVL maintains higher

$r_{\text{std}}$  even after training stabilizes, indicating larger intrinsic variability. Therefore, a larger threshold is more appropriate for InternVL. As shown in Table 7, increasing  $\delta$  improves performance compared with  $\delta = 0.1$ , which is consistent with the higher stable  $r_{\text{std}}$  observed in Table 11.

**Analysis of  $l_0$ .** We further study the necessity of the lower bound  $l_0$  through an ablation experiment. As shown in Table 8, removing  $l_0$  results in inconsistent performance across benchmarks. Although the variant without  $l_0$  achieves competitive results on some datasets, it fails to maintain stable improvements overall. In contrast, introducing

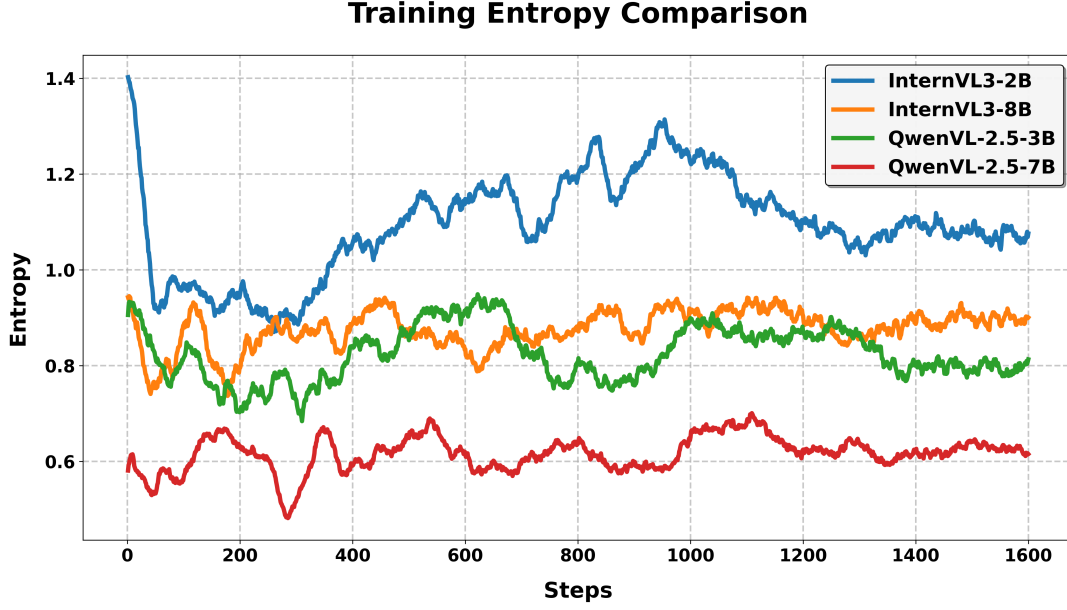


Figure 10: The entropy of different models.

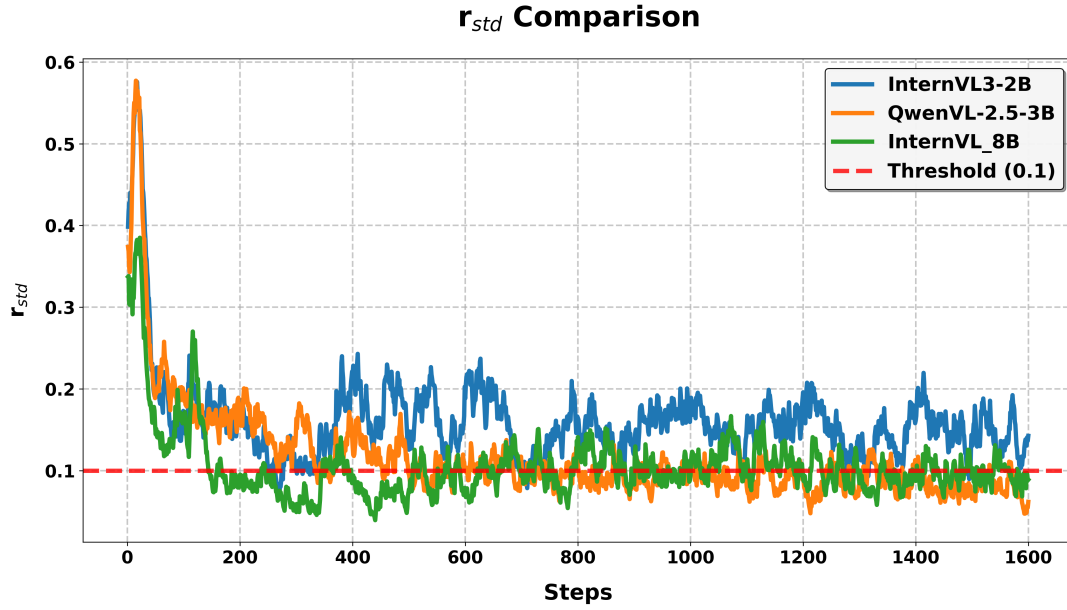


Figure 11: The  $r_{std}$  of different models.

$l_0$  consistently improves the overall performance and achieves the best results on RefCOCO, RefCOCOg, and RefGTA. This observation suggests that the lower bound plays a crucial role in stabilizing the optimization and preventing undesirable updates.

## I Case Study

Figure 12 presents a case-by-case comparison between our proposed method and the GRPO method. The results demonstrate that our method signifi-

cantly improves the model’s visual grounding capabilities.



Figure 12: Comparison between our proposed method and the GRPO method in some cases.

---

**Algorithm 1:** Entropy Control Visual Grounding Policy Optimization (ECVGPO)

---

**Input** : Dataset  $\mathcal{D}$  containing visual grounding prompts  $q$  and images; Initialized policy network  $\pi_\theta$  (and old policy  $\pi_{\theta_{old}}$ ); Reference network  $\pi_{ref}$ ; Verifiable reward function  $\mathcal{R}$  (e.g., IoU-based reward).

**Hyperparameters** : Group size  $N$ ; Stability threshold  $\delta$ ; Lower-bound clipping factor  $l_0 > 1$ ; Entropy control factor  $r_0$  ( $|r_0| > 1$ ); KL coefficient  $\beta$ ; PPO clip threshold  $\epsilon$ .

```
1 Initialize policy parameters  $\theta$ ;  
2 for each training iteration do  
3   Sample a batch of prompts  $\{q\}$  from Dataset  $\mathcal{D}$ ;  
4   for each prompt  $q$  in the batch do  
5     // Step 1: Rollout Group Responses  
6     Generate  $N$  candidate sequences  $\{O_1, O_2, \dots, O_N\}$  from current policy  $\pi_{\theta_{old}}(\cdot | q)$ ;  
7     // Step 2: Compute Rewards and Standard Deviation  
8     for each sequence  $O_i$  (with token length  $T_i$ ) do  
9       | Compute reward  $r_i = \mathcal{R}(O_i)$ ;  
10      |  $r_{avg} \leftarrow \frac{1}{N} \sum_{i=1}^N r_i$ ;  
11      |  $r_{std} \leftarrow \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i - r_{avg})^2}$ ;  
12      | // Step 3: Self-Information & Advantage Reshaping  
13      | for each sequence  $O_i$  do  
14        | Compute sequence-level self-information:  $S_i \leftarrow -\frac{1}{T_i} \sum_{t=1}^{T_i} \log \pi_{\theta_{old}}(o_{i,t} | q, o_{i,<t})$ ;  
15        | for each token  $t$  from 1 to  $T_i$  do  
16          | Compute standard GRPO advantage:  $A_{i,t} \leftarrow \frac{r_i - r_{avg}}{r_{std} + \epsilon_{eps}}$ ;  
17          | Compute token-level self-information:  $S_{i,t} \leftarrow -\log \pi_{\theta_{old}}(o_{i,t} | q, o_{i,<t})$ ;  
18          | // Apply Entropy Control for Positive Samples Under Stable Stage  
19          | if  $A_{i,t} \geq 0$  and  $r_{std} < \delta$  then  
20            | |  $A'_{i,t} \leftarrow A_{i,t} - \max\left(A_{i,t} - l_0, \frac{S_{i,t}}{r_0}\right)$ ;  
21          | else  
22            | |  $A'_{i,t} \leftarrow A_{i,t}$ ;  
23          |  
24      |  
25      | // Step 4: Policy Optimization Update  
26      |  $D_{KL}(\pi_\theta \| \pi_{ref}) \leftarrow \frac{1}{T_i} \sum_{t=1}^{T_i} \left[ \log \frac{\pi_\theta(o_{i,t} | \cdot)}{\pi_{ref}(o_{i,t} | \cdot)} - \frac{\pi_\theta(o_{i,t} | \cdot)}{\pi_{ref}(o_{i,t} | \cdot)} + 1 \right]$ ;  
27      |  $r_{i,t}(\theta) \leftarrow \frac{\pi_\theta(o_{i,t} | q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t} | q, o_{i,<t})}$ ;  
28      |  $\mathcal{L}_{CLIP}(\theta) \leftarrow \frac{1}{N} \sum_{i=1}^N \frac{1}{T_i} \sum_{t=1}^{T_i} \min\left(r_{i,t}(\theta) A'_{i,t}, \text{clip}(r_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon) A'_{i,t}\right)$ ;  
29      |  $\mathcal{L}(\theta) \leftarrow -\mathcal{L}_{CLIP}(\theta) + \beta D_{KL}(\pi_\theta \| \pi_{ref})$ ;  
30      | Update  $\theta$  by minimizing  $\mathcal{L}(\theta)$  via gradient descent;
```

---

| Hyperparameter   | QwenVL | InternVL        |
|------------------|--------|-----------------|
| Precision        |        | bf16            |
| Optimizer        |        | AdamW           |
| Image resolution | anyres | 1344*896(max)   |
| Attention type   |        | flash attention |
| Deepspeed stage  | 3      | 2               |
| $\delta$         | 0.1    | /               |
| $l_0$            |        | 25              |

Table 3: Hyperparameters of our Qwen2.5-VL’s and InternVL3’s training.

| Setting    | RefCOCO <sub>val</sub> | RefCOCO+ <sub>val</sub> | RefCOCOg <sub>val</sub>          | LISA                             |
|------------|------------------------|-------------------------|----------------------------------|----------------------------------|
| Base Model | <b>93.00</b>           | <b>88.40</b>            | 89.35                            | 0.84                             |
| w/ RefGTA  | 27.90                  | 26.65                   | 30.95                            | 4.46                             |
| w/ RefCOCO | 92.15 $\pm$ 0.15       | 88.13 $\pm$ 0.33        | <b>89.43<math>\pm</math>0.41</b> | <b>81.14<math>\pm</math>0.45</b> |

Table 4: Result of different training data. The best performance is **bolded**. The experiments are conducted on the InternVL-8B model.

| Setting   | RefCOCO <sub>val</sub>  | RefCOCO <sub>val</sub>  | RefCOCO <sub>val</sub>  | LISA                    |
|-----------|-------------------------|-------------------------|-------------------------|-------------------------|
| GRPO      | 91.65 $\pm$ 0.65        | 87.33 $\pm$ 0.52        | 88.97 $\pm$ 0.46        | 78.78 $\pm$ 1.45        |
| $r_0=-50$ | 92.15 $\pm$ 0.15        | <b>88.13</b> $\pm$ 0.33 | <b>89.43</b> $\pm$ 0.41 | 81.14 $\pm$ 0.45        |
| $r_0=-60$ | 91.78 $\pm$ 0.30        | 88.13 $\pm$ 0.28        | 89.41 $\pm$ 0.31        | 80.36 $\pm$ 0.82        |
| $r_0=-70$ | <b>92.38</b> $\pm$ 0.32 | 87.73 $\pm$ 0.30        | 89.15 $\pm$ 0.52        | <b>81.30</b> $\pm$ 1.09 |
| $r_0=50$  | 91.56 $\pm$ 0.43        | 87.91 $\pm$ 0.47        | 88.88 $\pm$ 0.16        | 79.55 $\pm$ 0.19        |
| $r_0=75$  | 91.58 $\pm$ 0.22        | 87.43 $\pm$ 0.25        | 89.40 $\pm$ 0.60        | 80.71 $\pm$ 0.76        |

Table 5: Result of different  $r_0$ . The best performance is **bolded**. The experiments are conducted on the InternVL-8B model. We report the mean and standard deviation over four runs.

| Model         | $\Delta(\text{ECVGPO-GRPO})$ | Strategy  | Entropy Diff |
|---------------|------------------------------|-----------|--------------|
| Qwen2.5-VL-3B | +0.2                         | $r_0 > 0$ | $\downarrow$ |
| Qwen2.5-VL-7B | +0.2                         | $r_0 < 0$ | $\uparrow$   |
| InternVL3-2B  | +2.6                         | $r_0 > 0$ | $\downarrow$ |
| InternVL3-8B  | +1.4                         | $r_0 < 0$ | $\uparrow$   |

Table 6: The selection strategy for the hyperparameter  $r_0$ . We report the mean value over four runs.

| Setting               | RefCOCO <sub>val</sub>  | RefCOCO+ <sub>val</sub> | RefCOCOg <sub>val</sub> | LISA                    | RefGTA                  |
|-----------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| $\delta = 0.1$        | 91.60 $\pm$ 0.13        | 87.58 $\pm$ 0.30        | 89.24 $\pm$ 0.26        | 78.60 $\pm$ 0.64        | 36.17 $\pm$ 4.57        |
| $\delta = 0.2$        | 92.01 $\pm$ 0.37        | 87.70 $\pm$ 0.39        | 89.23 $\pm$ 0.30        | 80.37 $\pm$ 1.82        | 48.06 $\pm$ 8.93        |
| $\delta = \text{N/A}$ | <b>92.15</b> $\pm$ 0.15 | <b>88.13</b> $\pm$ 0.33 | <b>89.43</b> $\pm$ 0.41 | <b>81.14</b> $\pm$ 0.45 | <b>51.29</b> $\pm$ 6.24 |

Table 7: Result of different  $\delta$ . The best performance is **bolded**. The experiments are conducted on the InternVL-8B model. We report the mean and standard deviation over four runs.

| Setting           | RefCOCO <sub>val</sub>  | RefCOCO+ <sub>val</sub> | RefCOCOg <sub>val</sub> | LISA                    | RefGTA                  |
|-------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| GRPO              | 89.21 $\pm$ 0.35        | <b>81.95</b> $\pm$ 0.72 | 87.50 $\pm$ 0.32        | 51.12 $\pm$ 5.63        | 51.26 $\pm$ 2.48        |
| Ours (w/o $l_0$ ) | 89.07 $\pm$ 0.32        | 81.74 $\pm$ 0.36        | 87.41 $\pm$ 0.21        | <b>61.46</b> $\pm$ 4.07 | 49.47 $\pm$ 7.02        |
| Ours (w/ $l_0$ )  | <b>89.49</b> $\pm$ 0.51 | 81.80 $\pm$ 0.40        | <b>87.56</b> $\pm$ 0.32 | 59.26 $\pm$ 1.76        | <b>53.49</b> $\pm$ 1.85 |

Table 8: Result of different  $l_0$ . The best performance is **bolded**. Consider that the parameter  $l_0$  is not in effect in Internvl-8B, the experiments are conducted on the InternVL-2B model. We report the mean and standard deviation over four runs.

| Setting                  | RefCOCO <sub>val</sub>  | RefCOCO+ <sub>val</sub> | RefCOCOg <sub>val</sub> | LISA                    | RefGTA                  |
|--------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| <b>temperature = 1.0</b> |                         |                         |                         |                         |                         |
| GRPO                     | 91.65 $\pm$ 0.65        | 87.33 $\pm$ 0.52        | 88.97 $\pm$ 0.46        | 78.78 $\pm$ 1.45        | 48.80 $\pm$ 6.30        |
| w/ Ours                  | <b>92.15</b> $\pm$ 0.15 | <b>88.13</b> $\pm$ 0.33 | <b>89.43</b> $\pm$ 0.41 | <b>81.14</b> $\pm$ 0.45 | <b>51.29</b> $\pm$ 6.24 |
| <b>temperature = 0.8</b> |                         |                         |                         |                         |                         |
| GRPO                     | 91.10 $\pm$ 0.43        | 87.32 $\pm$ 0.26        | 89.05 $\pm$ 0.25        | 79.02 $\pm$ 0.77        | <b>47.67</b> $\pm$ 6.21 |
| w/ Ours                  | <b>92.28</b> $\pm$ 0.29 | <b>87.86</b> $\pm$ 0.30 | <b>89.18</b> $\pm$ 0.32 | <b>80.02</b> $\pm$ 0.42 | 45.28 $\pm$ 3.59        |
| <b>temperature = 1.2</b> |                         |                         |                         |                         |                         |
| GRPO                     | 92.13 $\pm$ 0.23        | 87.36 $\pm$ 0.45        | 88.95 $\pm$ 0.51        | 77.85 $\pm$ 1.20        | 29.79 $\pm$ 2.37        |
| w/ Ours                  | <b>92.13</b> $\pm$ 0.27 | <b>87.53</b> $\pm$ 0.25 | <b>89.22</b> $\pm$ 0.42 | <b>78.37</b> $\pm$ 1.93 | <b>35.68</b> $\pm$ 4.06 |

Table 9: Result of different temperature of our method. The best performance is **bolded**. The experiments are conducted on the InternVL-8B model. We report the mean and standard deviation over four runs.

| Setting             | RefCOCO <sub>val</sub>  | RefCOCO+ <sub>val</sub> | RefCOCOg <sub>val</sub> | LISA                    | RefGTA                   |
|---------------------|-------------------------|-------------------------|-------------------------|-------------------------|--------------------------|
| <b>rollout = 8</b>  |                         |                         |                         |                         |                          |
| GRPO                | 91.65 $\pm$ 0.65        | 87.33 $\pm$ 0.52        | 88.97 $\pm$ 0.46        | 78.78 $\pm$ 1.45        | 48.80 $\pm$ 6.30         |
| w/ Ours             | <b>92.15</b> $\pm$ 0.15 | <b>88.13</b> $\pm$ 0.33 | <b>89.43</b> $\pm$ 0.41 | <b>81.14</b> $\pm$ 0.45 | <b>51.29</b> $\pm$ 6.24  |
| <b>rollout = 4</b>  |                         |                         |                         |                         |                          |
| GRPO                | 91.68 $\pm$ 0.14        | <b>87.83</b> $\pm$ 0.16 | 89.06 $\pm$ 0.27        | 78.35 $\pm$ 1.27        | 37.87 $\pm$ 7.09         |
| w/ Ours             | <b>91.74</b> $\pm$ 0.13 | 87.61 $\pm$ 0.33        | <b>89.16</b> $\pm$ 0.27 | <b>79.37</b> $\pm$ 2.34 | <b>47.29</b> $\pm$ 13.69 |
| <b>rollout = 16</b> |                         |                         |                         |                         |                          |
| GRPO                | 92.35 $\pm$ 0.39        | <b>88.38</b> $\pm$ 0.36 | <b>89.56</b> $\pm$ 0.31 | 78.64 $\pm$ 0.43        | 41.67 $\pm$ 5.91         |
| w/ Ours             | <b>92.54</b> $\pm$ 0.35 | 88.21 $\pm$ 0.25        | 89.48 $\pm$ 0.36        | <b>79.71</b> $\pm$ 0.65 | <b>52.58</b> $\pm$ 4.71  |

Table 10: Result of different rollouts of our method. The best performance is **bolded**. The experiments are conducted on the InternVL-8B model. We report the mean and standard deviation over four runs.