Syn-GRPO: Self-Evolving Data Synthesis for MLLM Perception Reasoning

Qihan Huang[1], Haofei Zhang[1,†], Rong Wei[2], Yi Wang[1], Rui Tang[2], Mingli Song[1], and Jie Song[1]

[1] Zhejiang University, Hangzhou, China
{qh.huang,haofeizhang,y_w,brooksong,sjie}@zju.edu.cn
[2] Manycore Tech Inc., Hangzhou, China
{guangmo,ati}@qunhemail.com

Abstract. RL (reinforcement learning) methods (*e.g.*, GRPO) for multimodal LLM perception ability has attracted wide research interest owing to its remarkable generalization ability. Nevertheless, existing reinforcement learning methods still face the problem of **low data quality**, where data samples cannot elicit diverse responses from MLLMs, thus restricting the exploration scope for MLLM reinforcement learning. Some methods attempt to mitigate this problem by imposing constraints on entropy, but none address it at its root. Therefore, to tackle this problem, this work proposes Syn-GRPO (**Syn**thesis-**GRPO**), which employs an online data generator to synthesize high-quality training data with diverse responses in GRPO training. Specifically, Syn-GRPO consists of two components: (1) data server; (2) GRPO workflow. The data server synthesizes new samples from existing ones using an image generation model, featuring a decoupled and asynchronous scheme to achieve high generation efficiency. The GRPO workflow provides the data server with the new image descriptions, and it leverages a diversity reward to supervise the MLLM to predict image descriptions for synthesizing samples with diverse responses. Experiment results across three visual perception tasks demonstrate that Syn-GRPO improves the data quality by a large margin, achieving significant superior performance to existing MLLM perception methods, and Syn-GRPO presents promising potential for scaling long-term self-evolving RL. Our code is available at *https://github.com/hqhQAQ/Syn-GRPO*.

1 Introduction

MLLM (Multimodal LLM) perception ability has drawn widespread research for its role as the prerequisite for all core functions of MLLMs. Recently, RL (reinforcement learning) methods (*e.g.*, GRPO [6,20]) fit well with MLLM perception training, owing to its strong generalization and visual perception tasks' inherent verifiable labels. Therefore, numerous reasoning methods (*e.g.*, VLM-R1 [22], Visual-RFT [13], VisionReasoner [12]) based on GRPO have significantly promoted MLLM perception ability.

[1] † Corresponding author.

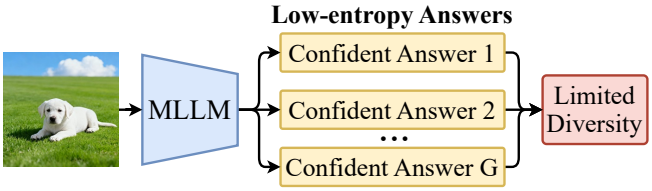


Fig. 1: In GRPO training, the MLLM generates non-diverse answers with low entropy, restricting the exploration space of RL.

Despite these developments, existing RL methods for MLLM perception still encounter the problem of **low data quality**. Low data quality refers to the issue where data samples fail to stimulate diverse responses from MLLMs, thereby limiting the exploration space for MLLM reinforcement learning, as shown in Figure 1. To investigate this problem, this work analyzes the *entropy* and *diversity* during GRPO training for perception reasoning. Specifically, the entropy reflects the uncertainty of the MLLM's output distribution, and the diversity measures the variety of multiple responses for each sample. Figure 2 shows that the entropy and diversity of the MLLM exhibit a trend of rapid decline (referred to as entropy collapse [28] and diversity collapse), indicating the low quality of visual perception data and resulting in extremely low training efficiency. Existing LLM RL methods (*e.g.*, CLIP-Higher [28], CLIP-Cov [4], KL-Cov [4], Entropy Adv. [3]) attempt to mitigate this problem from GRPO itself; however, they still cannot overcome the limitations of low-quality data at the root.

Therefore, to break the data limit, this work proposes Syn-GRPO (**Syn**thesis-**GRPO**), which leverages an online data generator to synthesize high-quality training data with diverse responses during GRPO training. Syn-GRPO aims to synthesize new data to replace the old data when processing each training batch, through a self-evolving data synthesis approach. To this end, Syn-GRPO consists of two components: **(1) data server**; **(2) GRPO workflow**.

The data server synthesizes new image data based on the old image data in the visual perception datasets, featuring **(1) foreground consistency** and **(2) decoupled design**. For the first feature, the data server first adopts a foreground segmentation model to remove irrelevant background from the image, and then uses an outpainting model to generate a new background. The foreground consistency preserves the invariance of visual perception task labels after generation, ensuring accurate supervisory signals during RL training. Meanwhile, by preserving the foreground, this method also alleviates the problem of excessive distribution discrepancy between new and old data. For the second feature, the data server adopts a data generation scheme that is decoupled and asynchronous from the GRPO workflow, connected by a unified API. This separation enables independent management of the data server and GRPO workflow while preserving efficient communication between them.

The GRPO workflow requires the MLLM to predict two outputs in addition to the original reasoning process and final answer: **(1) new image description** and **(2) predicted diversity**. Specifically, the new image description is the

text prompt for the new sample, which will be sent to the data server along with the old image to generate the new image. The predicted diversity represents the MLLM's estimation of the response diversity for the input sample. However, generating images from the predicted image descriptions does not necessarily yield highly diverse responses. To tackle this problem, the GRPO workflow computes a diversity reward by comparing the predicted diversity with the ground-truth diversity from rollout responses. In this manner, the diversity reward effectively supervises the MLLM to distinguish whether input samples have diverse responses, thereby enhancing its ability to generate image descriptions for samples with diverse responses. Additionally, this work identifies a diversity drift problem, where the declining trend of the ground-truth diversity causes a shift in the predicted diversity, and accordingly proposes a diversity smoothing strategy to mitigate it.

We perform comprehensive experiments to validate the performance of the proposed Syn-GRPO. Specifically, we apply Syn-GRPO to three types of visual perception tasks: REC (Referring Expression Comprehension), OVD (Open-Vocabulary Object Detection), and ISR (Indoor Scene Refinement). The experiment results demonstrate that Syn-GRPO improves the training efficiency of original GRPO by a large margin, achieving significantly superior performance to existing MLLM perception methods. Furthermore, we reveal two intriguing phenomena: (1) as the size of the original training dataset increases, Syn-GRPO exhibits a stable performance improvement trend; (2) as training progresses through iterations, the generated data shows an increasingly complex trend. These phenomena indicate that Syn-GRPO, a self-evolving data synthesis approach, has the potential to achieve long-term scalability in RL.

To sum up, the main contributions of this work can be summarized as follows:

• We identify and analyze the problem of low data quality in RL training for MLLM perception.

• We propose Syn-GRPO to tackle this problem, with a data server to efficiently synthesize data in an asynchronous manner, and a GRPO workflow to achieve self-evolving data synthesis through a proposed diversity reward.

• Experiment results across three visual perception tasks demonstrate that Syn-GRPO significantly outperforms existing MLLM perception methods, and Syn-GRPO exhibits promising potential for long-horizon scalability in RL.

2 Related Work

Reinforcement Learning for MLLM Perception. After the emergence of DeepSeek-R1, researchers discovered that reinforcement learning (*e.g.*, GRPO) can significantly improve the reasoning abilities of LLMs and MLLMs by triggering chain-of-thought (CoT) reasoning, with many studies applying this method to various domains of LLMs and MLLMs. Among these, RL is particularly well-suited for MLLM perception tasks, as such tasks typically come with verifiable labels (*e.g.*, bounding box annotations). Specifically, VLM-R1 [22] and Visual-RFT [13] first demonstrate that GRPO significantly outperforms SFT on visual

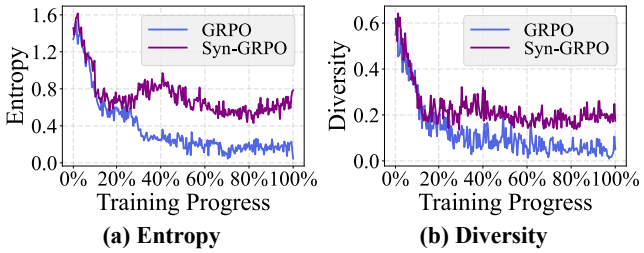


Fig. 2: Entropy collapse and diversity collapse of GRPO for Qwen2.5-VL-3B on the visual perception task (REC). Note that 20% training progress corresponds to one training epoch.

perception tasks by directly extending GRPO through carefully designed task-specific rewards. Built upon GRPO, SATORI-R1 [21] (three-stage decomposition), Visionary-R1 [25] (caption-reason-answer format), UniVG-R1 [2] (difficulty-aware weight adjustment), VisionReasoner [12] (unified framework with non-repeat reward), and Rex-Thinker [8] (HumanRef-CoT dataset) improve MLLM perception by addressing distinct issues and enhancing framework & structured reasoning. However, these methods still suffer from low data quality, resulting in suboptimal performance.

Entropy Mechanism of LLM Reinforcement Learning. Recently, the entropy mechanism has become a key research focus in LLM RL, because of its close connection with the exploration space of LLM RL. DAPO [28] first identified that during GRPO training, the entropy of the model's predicted tokens rapidly decreases. After this, CLIP-Higher [28] (higher GRPO sampling ratio clipping threshold), Clip-Cov [4] (clip high-covariance tokens), KL-Cov [4] (KL penalty on high-covariance tokens), and Entropy Adv. [3] (entropy-included GRPO advantage) attempt to address this issue by imposing constraints on entropy. Despite their advances, these methods remain fundamentally constrained by data quality and exhibit degraded performance on visual perception tasks where data samples follow more uniform formats.

Data Synthesis for LLM Reinforcement Learning. Synthetic data holds great potential to address the problem of missing data, and data synthesis methods have emerged for LLM RL. Specifically, PROMPTCOT [31] (emulate expert problem designers for math problem synthesis), Genetic-Instruct [15] (Instructor & Coder & Judge-LLM collaborative instruction-code synthesis), METASYNTH [19] (multi-agent meta-prompting), and TaskCraft [24] (automated workflow for multi-tool verifiable agent tasks) advance data synthesis via distinct collaborative or workflow-driven strategies. Absolute Zero [30] and R-Zero [7] propose an additional generation scheme to synthesize data along with the GRPO training. However, these data synthesis methods are either inapplicable to visual perception tasks or offline-generated, lacking self-evolution and failing to adaptively synthesize data.

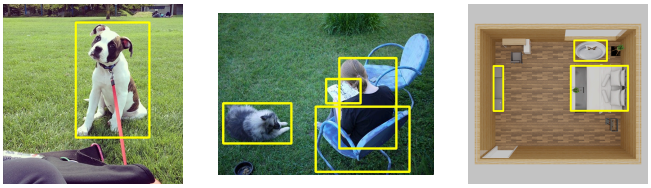


Fig. 3: Samples of three visual perception tasks (REC, OVD, and ISR) with the bounding box annotations.

3 Method

3.1 Preliminaries

Supervised Fine-tuning (SFT). SFT trains the LLM on curated query-output pairs to improve its instruction-following ability. The aim of SFT is to maximize the following objective:

$$\mathcal{J}_{\text{SFT}}(\theta) = \mathbb{E}[q, o \sim P(Q,O)] \left(\frac{1}{|o|} \sum_{t=1}^{|o|} \log \pi_\theta(o_t | q, o_{<t}) \right),$$

where q, o is the query-output pair sampled from the SFT dataset $P(Q,O)$, θ denotes the model parameters, $\pi_\theta(o_t|q, o_{<t})$ represents the logit of the model predicting the next token o_t from q and previous tokens $o_{<t}$.

Group Relative Policy Optimization (GRPO). GRPO calculates the advantage $\hat{A}_{i,t}$ for the t-th token using the average reward of multiple sampled outputs, eliminating the additional value function V_ψ in PPO. Specifically, GRPO samples a group of G outputs $\{o_1, o_2, ..., o_G\}$ from the old model $\pi_{\theta_{\text{old}}}$, then optimizes the model by maximizing the following objective (**min & clip operations** omitted here):

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)]$$
$$\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t} - \beta D_{\text{KL}}[\pi_\theta || \pi_{\text{ref}}] \right\},$$

where $D_{\text{KL}}[\pi_\theta || \pi_{\text{ref}}]$ serves as a regularization term that prevents the new model π_θ from deviating too far from the original model π_{ref} (the model before training). As an ORM method, GRPO provides the reward r_i at the end of each output o_i ($r_i = 1$ if the reasoning result is correct, otherwise $r_i = 0$), and sets the advantage $\hat{A}_{i,t}$ of all tokens in o_i as the normalized reward ($\mathbf{r} = \{r_1, r_2, ..., r_G\}$, mean$(\cdot)$ denotes the average, and std$(\cdot)$ denotes the standard deviation):

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}. \tag{1}$$

Task-Specific Rewards. This work applies Syn-GRPO to three types of visual perception tasks (REC, OVD, and ISR) (Figure 3), each with a task-specific accuracy reward.

REC (Referring Expression Comprehension) aims to localize the object in an image from a given referring expression. Following VLM-R1 [22], denote q, o as the query-output pair, b^{gt} as the ground-truth bounding box, $f_{\text{rec}}(o)$ as the predicted bounding box extracted from o, and $\text{IoU}(\cdot)$ as the intersection-over-union metric, then the accuracy reward $\mathbf{R}_{\text{acc}}^{\text{rec}}(o)$ for REC is calculated as:

$$\mathbf{R}_{\text{acc}}^{\text{rec}}(o) = \text{IoU}(b^{\text{gt}}, f_{\text{rec}}(o)). \tag{2}$$

OVD (Open-Vocabulary Object Detection) aims to detect the objects in the image and output the corresponding bounding boxes and class labels. Following VLM-R1 [22], denote $\hat{b}^{\text{gt}} = \{(b_i^{\text{gt}}, c_i^{\text{gt}})\}_{i=1}^{N^{\text{gt}}}$ as the list of N^{gt} ground-truth bounding-boxes and class labels, $f_{\text{ovd}}(o) = \{(b_i, c_i)\}_{i=1}^{N^{\text{pred}}}$ as the list of N^{pred} predicted bounding-boxes and class labels, and $\text{mAP}(\cdot)$ as the mean average precision metric, then the accuracy reward $\mathbf{R}_{\text{acc}}^{\text{ovd}}(o)$ for OVD is calculated as (s_{ovd} aims to penalizes redundant predictions):

$$\begin{cases} \mathbf{R}_{\text{acc}}^{\text{ovd}}(o) = s_{\text{ovd}} \cdot \text{mAP}(\hat{b}^{\text{gt}}, f_{\text{ovd}}(o)). \\ s_{\text{ovd}} = \min\left(1, \dfrac{N^{\text{gt}}}{N^{\text{pred}}}\right). \end{cases} \tag{3}$$

ISR (Indoor Scene Refinement) refines the indoor scene based on top-view rendering images to enhance its aesthetic quality. This work decomposes ISR into two stages: perception and refinement, with Syn-GRPO applied in the first stage. Specifically, the perception stage equals OVD, with ISR's accuracy reward $\mathbf{R}_{\text{acc}}^{\text{isr}}(o)$ identical to $\mathbf{R}_{\text{acc}}^{\text{ovd}}(o)$. More details about ISR are in §2.2 of the appendix.

3.2 Data Quality Analysis

Although current RL methods have significantly enhanced MLLM perception, they suffer from the problem of **low data quality**. Low data quality refers to data samples failing to elicit diverse responses from MLLMs, thereby constraining the exploration space in MLLM RL. To investigate this problem, this work examines the **entropy** and **diversity** of MLLM during GRPO training.

Entropy quantifies the predictability or randomness inherent in the tokens predicted by the LLM. Following Clip-Cov & KL-Cov [4], the entropy $\mathcal{H}(q)$ corresponding to sample q is calculated as the average of the entropy of each predicted token (note that π_θ denotes the model):

$$\begin{aligned} \mathcal{H}(q) &= -\mathbb{E}_{\{o_i\}_{i=1}^{G} \sim \pi_\theta(O|q)} \left[\log \pi_\theta(o_{i,t}|q, o_{i,<t})\right] \\ &= -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\mathbb{E}_{o_{i,t}\sim\pi_\theta(O|q)}\left[\log \pi_\theta(o_{i,t}|q, o_{i,<t})\right]. \end{aligned}$$

Diversity measures the extent to which an MLLM generates diverse responses for each sample, quantified by the variance of the accuracy rewards

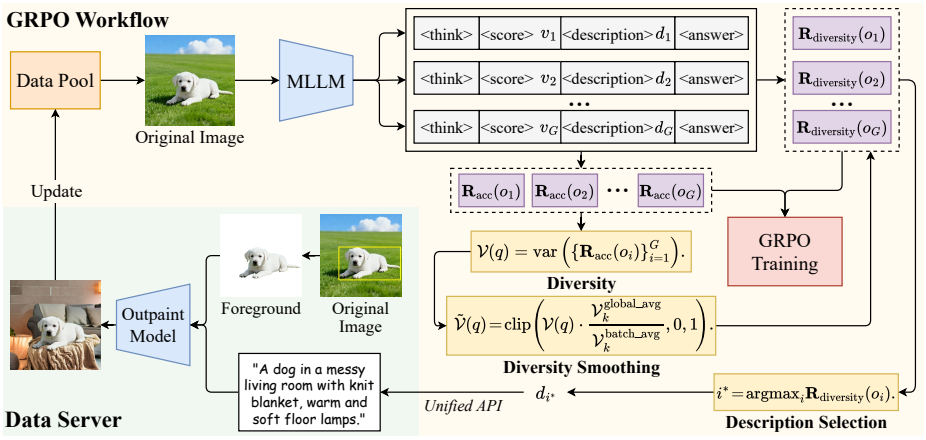


Fig. 4: Framework of Syn-GRPO. Syn-GRPO consists of a data server and a GRPO workflow. The data server generates new high-quality images with diverse responses, based on the original images and image descriptions d_{i^*}. The GRPO workflow predicts new image descriptions for the data server, and it employs a diversity reward $\mathbf{R}_{\text{diversity}}(o_i)$ to supervise the generation of these descriptions.

$\mathbf{R}_{\text{acc}}(o_i)$ across G responses. Let $\text{var}(\cdot)$ denote the variance, then the diversity $\mathcal{V}(q)$ corresponding to sample q is calculated as:

$$\mathcal{V}(q) = \text{var}\left(\{\mathbf{R}_{\text{acc}}(o_i)\}_{i=1}^{G}\right). \tag{4}$$

In practice, each metric is averaged over the samples in the training batch. As shown in Figure 2, MLLM exhibits entropy collapse and diversity collapse in the visual perception task (REC), characterized by a rapid decline in both entropy and diversity during GRPO training. Entropy collapse and diversity collapse indicate that the current visual perception datasets are low-quality, exhibiting overly uniform formatting, which constrains the RL exploration space.

3.3 Syn-GRPO

As shown in Figure 4, Syn-GRPO consists of two components: (1) data server; (2) GRPO workflow.

Data Server. Previous data synthesis methods use advanced LLMs (*e.g.*, GPT) to generate new text responses from existing images without modifying the images. The proposed data server, the first to overcome image-modality data limitations, leverages an image-outpainting model to create new images via original images and new image descriptions. Specifically, the data server replaces the old images with these new images for the training of next epoch, and the parameters of the image-outpainting model **are frozen** during training.

To accommodate visual perception tasks and enhance training efficiency, this data server incorporates two key features: **(1) foreground consistency** and **(2) decoupled design**.

Foreground consistency. In data synthesis for RL, it is essential to ensure the accuracy of labels in the synthesized data. Therefore, to preserve visual perception task labels, the data server masks out image regions beyond the ground-truth bounding boxes and then employs an outpainting model to generate new images according to the new image descriptions. Besides, the data server also employs a foreground segmentation model to more accurately remove the background. As shown in Figure 5, the labels (*i.e.*, bounding boxes) for visual perception tasks remain unchanged in the new images, ensuring their correctness. Furthermore, by modifying the original image instead of generating one from scratch, the data server can reduce discrepancies between new and old data distributions.

Decoupled design. The data server adopts a modular design decoupled from the GRPO workflow, implemented via Python's `HTTPServer` package. Through a unified communication API, the GRPO workflow requests data generation from the data server with the generation parameters and receives the results upon completion. The decoupled design enables asynchronous execution of GRPO training and data generation, enhancing training efficiency. Moreover, this design readily supports future extensions to other forms of data synthesis, promoting our method's generalization.

GRPO workflow. The GRPO workflow builds upon the original GRPO with three key modifications: **(1) New image description**, **(2) Diversity reward**, and **(3) Description selection**.

New image description. In addition to the original reasoning process and final answer, the GRPO workflow prompts the MLLM to predict a text description for generating a new high-quality image (with diverse responses) from the input image. Specifically, during GRPO training, the MLLM generates G responses $\{o_i\}_{i=1}^{G}$ for each sample q, with each o_i containing a distinct new image description d_i to form the set $\{d_i\}_{i=1}^{G}$.

Diversity reward. However, the generated samples from the predicted image descriptions $\{d_i\}_{i=1}^{G}$ do not ensure highly diverse responses. Thus, this work proposes a diversity reward to supervise the MLLM in distinguishing whether input samples elicit diverse responses, thereby **encouraging it to learn image descriptions associated with such diversity**. To this end, the diversity reward is calculated by comparing the predicted diversity with the ground-truth diversity. Specifically, the predicted diversity v_i of each response o_i is generated by the MLLM and lies in the range $[0, 1]$. Next, the ground-truth diversity $\mathcal{V}(q)$ for each sample q is calculated as the variance of accuracy rewards across G responses (Equation 4). Besides, $\mathcal{V}(q)$ is normalized to the range of $[0, 1]$ for the convenience of comparison, using its original upper bound $\frac{1}{4}$ as a reference (see §1 of the appendix). Consequently, the diversity reward $\mathbf{R}_{\text{diversity}}(o_i)$ of the i-th response o_i is calculated as:

$$\mathbf{R}_{\text{diversity}}(o_i) = 1 - |v_i - \mathcal{V}(q)|. \tag{5}$$

Moreover, this work identifies a *diversity drift* problem: the ground-truth diversity of the model evolves with training; thus, the predicted diversity of the **current model** is generated for the **prior model** as it is supervised by the ground-truth diversity of the **prior model**. Given the overall declining trend in diversity (Figure 2), this work proposes a **diversity smoothing method** that mitigates this issue by calibrating the diversity of each training batch using an exponential moving average of historical diversity. Specifically, let $\mathcal{B}_k$ denote the training batch at the k-th step, $\beta \in [0, 1]$ denote the smooth weight, $\mathcal{V}_k^{\text{batch\_avg}}$ denote the average diversity of $\mathcal{B}_k$, $\mathcal{V}_k^{\text{global\_avg}}$ denote the moving average of diversity at the k-th step, then the smoothed ground-truth diversity $\tilde{\mathcal{V}}(q)$ for sample q at the k-th step is calculated as (clip$(\cdot)$ clamps a value to a range):

$$\begin{cases} \mathcal{V}_k^{\text{batch\_avg}} = \dfrac{1}{|\mathcal{B}_k|} \displaystyle\sum_{q' \in \mathcal{B}_k} \mathcal{V}(q'). \\ \mathcal{V}_k^{\text{global\_avg}} = \beta \cdot \mathcal{V}_{k-1}^{\text{global\_avg}} + (1-\beta) \cdot \mathcal{V}_k^{\text{batch\_avg}}. \\ \tilde{\mathcal{V}}(q) = \text{clip}\left(\mathcal{V}(q) \cdot \dfrac{\mathcal{V}_k^{\text{global\_avg}}}{\mathcal{V}_k^{\text{batch\_avg}}}, 0, 1 \right). \end{cases}$$

Note that $\mathcal{V}_1^{\text{global\_avg}}$, the $\mathcal{V}_k^{\text{global\_avg}}$ at the first step, is equal to $\mathcal{V}_1^{\text{batch\_avg}}$. The diversity smoothing method aligns the current model's diversity distribution with the previous one through the ratio of $\mathcal{V}_k^{\text{global\_avg}}$ and $\mathcal{V}_k^{\text{batch\_avg}}$, enhancing reward training accuracy and stability. Finally, the updated diversity reward $\mathbf{R}_{\text{diversity}}(o_i)$ is calculated as in Equation 6, which also lies in the range $[0, 1]$.

$$\mathbf{R}_{\text{diversity}}(o_i) = 1 - |v_i - \tilde{\mathcal{V}}(q)|. \tag{6}$$

The final reward r_i for the i-th response o_i is calculated as the summation of its accuracy reward $\mathbf{R}_{\text{acc}}(o_i)$, format reward $\mathbf{R}_{\text{format}}(o_i)$, and diversity reward $\mathbf{R}_{\text{diversity}}(o_i)$. Note that $\mathbf{R}_{\text{format}}(o_i)$ supervises o_i to incorporate the reasoning process, predicted diversity, new image description, and the final answer.

$$r_i = \mathbf{R}_{\text{acc}}(o_i) + \mathbf{R}_{\text{format}}(o_i) + \mathbf{R}_{\text{diversity}}(o_i). \tag{7}$$

Description selection. For the G new image descriptions $\{d_i\}_{i=1}^G$ derived from the MLLM-generated responses $\{o_i\}_{i=1}^G$, the GRPO workflow requires selecting one for transmission to the data server. To this end, this work selects the image description with the highest $\mathbf{R}_{\text{diversity}}(o_i)$, as this response more accurately assesses the diversity of answers for the input q, leading to a more accurate image description for the new image with diverse responses. Specifically, the selected index i^* is calculated as in Equation 8, and the corresponding selected description is d_{i^*}.

$$i^* = \underset{i \in \{1,2,\ldots,G\}}{\text{argmax}} \mathbf{R}_{\text{diversity}}(o_i). \tag{8}$$

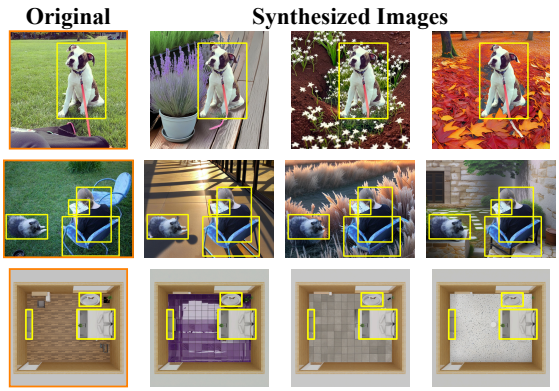


Fig. 5: Synthesized images of three visual perception tasks.

4 Experiments

Implementation details. Following existing MLLM perception methods (*e.g.*, VLM-R1 [22], Visionary-R1 [25]), we conduct the main experiments using Qwen2.5-VL-3B [1] and Qwen2.5-VL-7B [1] as base models. This work implements SynGRPO on the verl framework [23], leveraging the vLLM engine [9] for RL rollout acceleration and FSDP [32] for distributed training. In the data server, the image-outpainting model is a ControlNet [29] based on the SDXL [18] model, and the foreground segmentation model is BEN2 (Background Erase Network) [17]. For all three types of visual perception tasks, we adopt the AdamW optimizer [14] to train the model for 5 epochs, with a learning rate of 1×10^{-6}, a batch size of 20, and a rollout number G of 6. Besides, we allocate 4 GPUs for the GRPO workflow, and 1 GPU for the data server. The hyper-parameter β in the diversity reward is set to 0.7, as validated by ablation experiments.

Training datasets. For the REC task, we follow VLM-R1 to use the RefCOCO & RefCOCO+ & RefCOCOg datasets [16, 27] for training. Specifically, we randomly select 2,000 samples from their training sets and train for 5 epochs (resulting in 10,000 sample updates), which is comparable to VLM-R1's total training budget of 9,600 samples. For the OVD task, we also follow VLM-R1 to use the D^3 dataset [26] for training, and randomly select 2,000 samples likewise. For the ISR task, we utilize the 3D-FRONT dataset [5] for training, and randomly select 2,000 samples.

Test datasets. For the REC task, following VLM-R1, we evaluate on out-of-domain data (the test set of LISA-Grounding [10]) and in-domain data (validation sets of RefCOCO, RefCOCO+, and RefCOCOg). For the OVD task, we also follow VLM-R1 to evaluate on the $\text{COCO}_{\text{filter}}$ dataset [11], a filtered subset derived from the COCO validation set. For the ISR task, we randomly select 500 samples from the 3D-FRONT dataset for evaluation.

Evaluation Metric. For the REC task, we follow VLM-R1 and calculate the accuracy as the ratio of predicted bounding boxes with IoU > 0.5 relative to the ground-truth box. For the OVD task and the perception stage of ISR task, we follow VLM-R1 to utilize mAP (mean Average Precision), GP (Greedy Precision), and GR (Greedy Recall) for evaluation. Specifically, GP is the fraction of predicted boxes that match any ground-truth box, while GR is the reverse. For the refinement stage of ISR task, we employ S_{bbox} and S_{rotation} to evaluate the refinement quality, assessing object position and rotation, respectively. More details about GP, GR, S_{bbox}, and S_{rotation} are in §2.3 of the appendix.

Table 1: Experiment results of *Qwen2.5-VL-3B* on the REC task. The table has two sections: upper for existing models' performance, lower for existing methods & our method trained in the same setting. LISA abbreviates LISA-Grounding, and bold font denotes the best result.

Method	LISA	RefCOCO	RefCOCO+	RefCOCOg
Original	56.51	88.70	81.95	86.05
Visionary-R1	60.80	86.85	82.65	86.50
SATORI-R1	58.38	85.80	82.20	85.25
SFT	54.82	88.70	82.25	85.95
VLM-R1	63.14	90.55	84.30	87.10
GRPO	62.24	89.40	84.10	86.60
GRPO + Entropy Loss	64.23	90.20	82.55	85.60
GRPO + Entropy Adv.	63.69	90.90	83.85	86.40
GRPO + Offline Generation	64.23	91.10	83.40	86.30
Syn-GRPO (w/o $\mathbf{R}_{\text{diversity}}$)	64.60	91.35	83.85	85.85
Syn-GRPO	**68.28**	**92.15**	**85.30**	**87.45**

Table 2: Experiment results of *Qwen2.5-VL-7B* on the REC task. The table has two sections: upper for existing models' performance, lower for existing methods & our method trained in the same setting. LISA abbreviates LISA-Grounding, and bold font denotes the best result.

Method	LISA	RefCOCO	RefCOCO+	RefCOCOg
Original	61.34	90.00	84.20	87.20
Rex-Thinker	67.49	91.20	86.35	87.80
SFT	63.27	89.10	85.95	86.65
VLM-R1	66.71	92.60	89.40	89.00
GRPO	65.26	91.90	89.85	87.80
GRPO + Entropy Loss	66.59	91.40	87.65	87.95
GRPO + Entropy Adv.	67.25	92.05	88.30	87.30
GRPO + Offline Generation	66.95	91.80	89.85	88.55
Syn-GRPO (w/o $\mathbf{R}_{\text{diversity}}$)	67.67	92.75	89.50	88.75
Syn-GRPO	**70.14**	**93.55**	**90.65**	**89.25**

Table 3: Experiment results on the OVD task.

Method	mAP	GP (IoU=0.5)	GR (IoU=0.5)
Original	14.20	56.06	33.79
SFT	18.50	53.15	39.40
VLM-R1	21.10	67.34	43.84
GRPO	18.66	63.83	36.95
Syn-GRPO	**23.74**	**71.42**	**46.44**

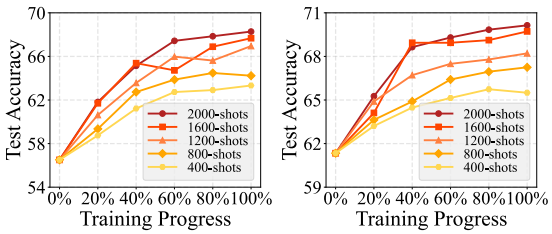


Fig. 6: Ablation experiments of data size for Syn-GRPO on the LISA-Grounding dataset.

4.1 Comparison Analysis

REC. Table 1 and Table 2 demonstrate the performance comparison of different methods on the REC task, evaluated with two base MLLMs: Qwen2.5-VL-3B and Qwen2.5-VL-7B. Several conclusions can be drawn from these two tables.

(1) In the upper portion of the table, RL methods with entropy constraints (Entropy Loss, Entropy Adv.) partially mitigate entropy collapse, preserving exploration capacity on low-quality data. However, they fail to fundamentally resolve the core data quality issue, yielding only marginal improvements over GRPO.

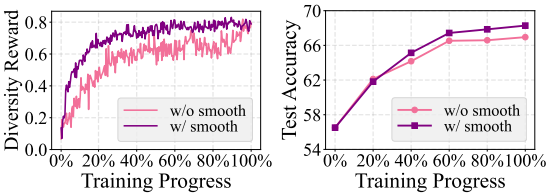
(2) In the lower portion of the table, GRPO underperforms VLM-R1, attributed to its utilization of fewer training samples (2,000 versus 9,600).

(3) In the lower portion of the table, Syn-GRPO outperforms both standard RL methods, entropy-based methods, and the offline generation method (using the descriptions from GPT-4o).

(4) Syn-GRPO exhibits more pronounced performance gains on out-of-domain data (LISA-Grounding) than on in-domain benchmarks (RefCOCO, RefCOCO+, and RefCOCOg), as the synthesized data spans a broader distribution, endowing the MLLM with more robust and comprehensive perception abilities.

Furthermore, §2.6 of the appendix indicates that Syn-GRPO also achieves better performance than the original GRPO on Qwen3-VL-8B.

OVD & ISR. Table 3 shows that Syn-GRPO also surpasses GRPO on the OVD task. Besides, §2.5 of the appendix demonstrates the superior performance of Syn-GRPO over GRPO on the first perception stage of ISR task, as well as its effectiveness in enhancing the second refinement stage of ISR.



(a) Diversity Reward **(b) Test Accuracy**

Fig. 7: Ablation experiments of the diversity smoothing method for Qwen2.5-VL-3B on the LISA-Grounding dataset.

Table 4: Ablation experiments of β for diversity smoothing on the LISA-Grounding dataset. † denotes the selected one.

Model	0.1	0.3	0.5	0.7†	0.9
Qwen2.5-VL-3B	67.43	67.31	67.85	**68.28**	67.55
Qwen2.5-VL-7B	68.88	69.60	69.84	**70.14**	69.24

4.2 Ablation Experiments

Variations in entropy & diversity. As shown in Figure 2, during the original GRPO training, the entropy and diversity of the MLLM exhibit a rapid decreasing trend. By contrast, Syn-GRPO sustains high entropy and diversity from the first epoch (20% training progress), substantially mitigating the issues of entropy collapse and diversity collapse using the newly generated samples.

Effect of data size. This experiment varies data size (400 to 2,000 samples) in the REC task, investigating the impact of data size on model performance. As shown in Figure 6, Syn-GRPO performance displays a steady upward trend with increasing data size, indicating a data scaling law and suggesting strong potential for generalization to long-term reinforcement learning.

Diversity smoothing. This experiment verifies the effect of the diversity smoothing method on the proposed diversity reward. Figure 7 (a) demonstrates that, with diversity smoothing, the diversity reward increases more steadily (with reduced fluctuations). Moreover, it also shows that diversity smoothing leads to a more rapid increase in the diversity reward. Figure 7 (b) presents that, with diversity smoothing, the test accuracy increases more rapidly and ultimately reaches a higher plateau. Furthermore, Table 4 shows that performance improves with initial increases in β but degrades when β is excessively large, as it restricts the update of smoothed diversity.

Asynchronous data synthesis. Table 5 shows that the synchronous data server design nearly doubles training time versus original GRPO, as it must generate new data for each batch before proceeding. To address this, the data server adopts a modular design decoupled from the GRPO workflow, enabling asynchronous data synthesis with negligible overhead, as shown in Table 5.

Table 5: Training time of original GRPO on the REC task, Syn-GRPO (Synchronous), and Syn-GRPO (Asynchronous).

Model	GRPO	Synchronous	Asynchronous
Qwen2.5-VL-3B	6.10 Hours	13.16 Hours	**6.45** Hours
Qwen2.5-VL-7B	13.66 Hours	28.47 Hours	**13.93** Hours

Fig. 8: Generated images show a growingly complex trend.

4.3 Visualization Analysis

Visualization of generated images. Figure 5 presents examples of generated images for three types of visual perception tasks. These visualizations show that the generated images possess high image fidelity and foreground consistency, thereby ensuring the accuracy of labels for the new data and satisfying the requirements for RL training.

Trends in generated images. Figure 8 shows the generated images from the same original image at various epochs, highlighting their increasing complexity and difficulty, thereby preserving the diversity in reinforcement learning. This phenomenon demonstrates Syn-GRPO's potential for adapting to long-term scalable RL by generating increasingly challenging images throughout RL training.

Surreal generated images. We observe an intriguing phenomenon wherein the MLLM occasionally generates surreal textual descriptions for new images, as illustrated in Figure 9. Although these images do not exist in reality, they hold potential to benefit RL training by stimulating the MLLM's reasoning capability through eliciting novel reasoning pathways from new perspectives.

Besides, we have provided more visualization analysis in §3 of the appendix.

5 Conclusion

In this work, we provide a thorough analysis of the low-quality data problem in RL for MLLM perception, where data samples fail to elicit diverse responses from MLLMs, thereby limiting the exploration scope for RL. To tackle this problem, we propose Syn-GRPO (**Syn**thesis-**GRPO**), which utilizes an online data generator to synthesize high-quality training data with diverse responses in GRPO training. Specifically, Syn-GRPO consists of a data server and a GRPO workflow. The data server leverages an image generation model to synthesize new samples with an asynchronous design. The GRPO workflow provides new image descriptions to the server and proposes a diversity reward, supervising the MLLM to

Fig. 9: Two examples of surreal generated images.

predict descriptions for high-quality images with diverse responses. Experimental results across three visual perception tasks indicate that Syn-GRPO substantially enhances data quality, significantly outperforming existing RL methods. We hope our framework can contribute to the community of MLLM reasoning.

Acknowledgements

This work was supported by the Pioneer R&D Program of Zhejiang (2026SDXT005) and National Natural Science Foundation of China (62576305).

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Bai, S., Li, M., Liu, Y., Tang, J., Zhang, H., Sun, L., Chu, X., Tang, Y.: Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. arXiv preprint arXiv:2505.14231 (2025)
3. Cheng, D., Huang, S., Zhu, X., Dai, B., Zhao, W.X., Zhang, Z., Wei, F.: Reasoning with exploration: An entropy perspective. arXiv preprint arXiv:2506.14758 (2025)
4. Cui, G., Zhang, Y., Chen, J., Yuan, L., Wang, Z., Zuo, Y., Li, H., Fan, Y., Chen, H., Chen, W., et al.: The entropy mechanism of reinforcement learning for reasoning language models. arXiv preprint arXiv:2505.22617 (2025)
5. Fu, H., Cai, B., Gao, L., Zhang, L., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., Zhang, H.: 3d-front: 3d furnished rooms with layouts and semantics. In: ICCV 2021. pp. 10913–10922. IEEE (2021)
6. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
7. Huang, C., Yu, W., Wang, X., Zhang, H., Li, Z., Li, R., Huang, J., Mi, H., Yu, D.: R-zero: Self-evolving reasoning llm from zero data. arXiv preprint arXiv:2508.05004 (2025)
8. Jiang, Q., Chen, X., Zeng, Z., Yu, J., Zhang, L.: Rex-thinker: Grounded object referring via chain-of-thought reasoning. arXiv preprint arXiv:2506.04034 (2025)
9. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (2023)

10. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: LISA: reasoning segmentation via large language model. In: CVPR 2024. pp. 9579–9589. IEEE (2024)
11. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV 2014. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014)
12. Liu, Y., Qu, T., Zhong, Z., Peng, B., Liu, S., Yu, B., Jia, J.: Visionreasoner: Unified visual perception and reasoning via reinforcement learning. arXiv preprint arXiv:2505.12081 (2025)
13. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR 2019. OpenReview.net (2019)
15. Majumdar, S., Noroozi, V., Samadi, M., Narenthiran, S., Ficek, A., Ahmad, W., Huang, J., Balam, J., Ginsburg, B.: Genetic instruct: Scaling up synthetic generation of coding instructions for large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track). pp. 208–221 (2025)
16. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR 2016. pp. 11–20. IEEE Computer Society (2016)
17. Meyer, M., Spruyt, J.: Ben: Using confidence-guided matting for dichotomous image segmentation. arXiv preprint arXiv:2501.06230 (2025)
18. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
19. Riaz, H., Bhabesh, S.S., Arannil, V., Ballesteros, M., Horwood, G.: Metasynth: Meta-prompting-driven agentic scaffolds for diverse synthetic data generation. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 18770–18803 (2025)
20. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
21. Shen, C., Wei, W., Qu, X., Cheng, Y.: Satori-r1: Incentivizing multimodal reasoning with spatial grounding and verifiable rewards. arXiv preprint arXiv:2505.19094 (2025)
22. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
23. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv: 2409.19256 (2024)
24. Shi, D., Cao, J., Chen, Q., Sun, W., Li, W., Lu, H., Dong, F., Qin, T., Zhu, K., Liu, M., et al.: Taskcraft: Automated generation of agentic tasks. arXiv preprint arXiv:2506.10055 (2025)
25. Xia, J., Zang, Y., Gao, P., Li, Y., Zhou, K.: Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. arXiv preprint arXiv:2505.14677 (2025)
26. Xie, C., Zhang, Z., Wu, Y., Zhu, F., Zhao, R., Liang, S.: Described object detection: Liberating object detection with flexible expressions. In: NeurIPS 2023 (2023)

27. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV 2016. Lecture Notes in Computer Science, vol. 9906, pp. 69–85. Springer (2016)
28. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025)
29. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV 2023. pp. 3813–3824. IEEE (2023)
30. Zhao, A., Wu, Y., Yue, Y., Wu, T., Xu, Q., Lin, M., Wang, S., Wu, Q., Zheng, Z., Huang, G.: Absolute zero: Reinforced self-play reasoning with zero data. arXiv preprint arXiv:2505.03335 (2025)
31. Zhao, X., Wu, W., Guan, J., Kong, L.: Promptcot: Synthesizing olympiad-level problems for mathematical reasoning in large language models. arXiv preprint arXiv:2503.02324 (2025)
32. Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., Desmaison, A., Balioglu, C., Damania, P., Nguyen, B., Chauhan, G., Hao, Y., Mathews, A., Li, S.: Pytorch FSDP: experiences on scaling fully sharded data parallel. Proc. VLDB Endow. **16**(12), 3848–3860 (2023)