Stabilizing Unsupervised Self-Evolution of MLLMs via Continuous Softened Retracing reSampling

Anonymous ACL submission

Abstract

In the unsupervised self-evolution of Multimodal Large Language Models, the quality of feedback signals during post-training is pivotal for stable and effective learning. However, existing self-evolution methods predominantly rely on majority voting to select the most frequent output as the pseudo-golden answer, which may stem from the model's intrinsic biases rather than guaranteeing the objective correctness of the reasoning paths. To counteract the degradation, we propose **C**ontinuous **S**oftened **R**etracing re**S**ampling (**CSRS**) in MLLM self-evolution. Specifically, we introduce a Retracing Re-inference Mechanism (**RRM**) that the model re-inferences from anchor points to expand the exploration of long-tail reasoning paths. Simultaneously, we propose Softened Frequency Reward (**SFR**), which replaces binary rewards with continuous signals, calibrating reward based on the answers' frequency across sampled reasoning sets. Furthermore, incorporated with Visual Semantic Perturbation (**VSP**), CSRS ensures the model prioritizes mathematical logic over visual superficiality. Experimental results demonstrate that CSRS significantly enhances the reasoning performance of Qwen2.5-VL-7B on benchmarks such as MathVision. We achieve state-of-the-art (SOTA) results in unsupervised self-evolution on geometric tasks.

1 Introduction

Self-evolution (Wu et al., 2024; Chen et al., 2025a; Jiang et al., 2025; Thawakar et al., 2025) in Multimodal Large Language Models (MLLMs) is an emerging paradigm that leverages the model's intrinsic reasoning capabilities to achieve continuous improvement through internal feedback. To alleviate the dependency on costly manual annotations, this unsupervised framework demonstrates immense potential by bypassing the labeling bottleneck and it enables autonomous evolution of the model's capablity through leveraging unlabeled data across tasks such as geometric reasoning.

Recent works have focused on generating self-improving feedback signals. For instance, MM-UPT (Wei et al., 2025) establishes an unsupervised post-training pipeline, utilizing a majority voting self-rewarding mechanism based on the GRPO (Shao et al., 2024) algorithm. Similarly, VisPlay (He et al., 2025) enhances visual reasoning capabilities and reduces hallucinations by decoupling the base MLLM into two roles: a conditional questioner and a multimodal reasoner.

However, rewards of these approaches primarily generate from the model's internal preferences, leading to a confirmation bias cycle: the model repeatedly reinforces paths it perceives as correct. This narrows the exploration horizon, eventually resulting in the loss of logical diversity and distributional shift often called "Model Collapse" (Shumailov et al., 2023; Shafayat et al., 2025; Swamy et al., 2025). Furthermore, these methods rely heavily on majority voting, which provides only coarse-grained and sparse reward signals derived entirely from the models initial biases, thereby exacerbating the model collapse.

To overcome these challenges on geometric reasoning tasks, we propose **C**ontinuous **S**oftened **R**etracing re**S**ampling (**CSRS**). As shown in Fig. 1, by establishing retracing anchor points within the maternal trajectory – initial responses and integrating softened frequency rewards with visual perturbation, CSRS effectively mitigates model collapse in unsupervised settings. Specifically, we introduce Retracing Re-inference Mechanism (**RRM**), which sets a retracing anchor point within the maternal response answers and restarts inference from anchor points to construct a re-inference answer set. Unlike static sampling in conventional methods, this mechanism compels the model to perform deep exploration at key logical decision nodes. To address the collapse triggered by major-

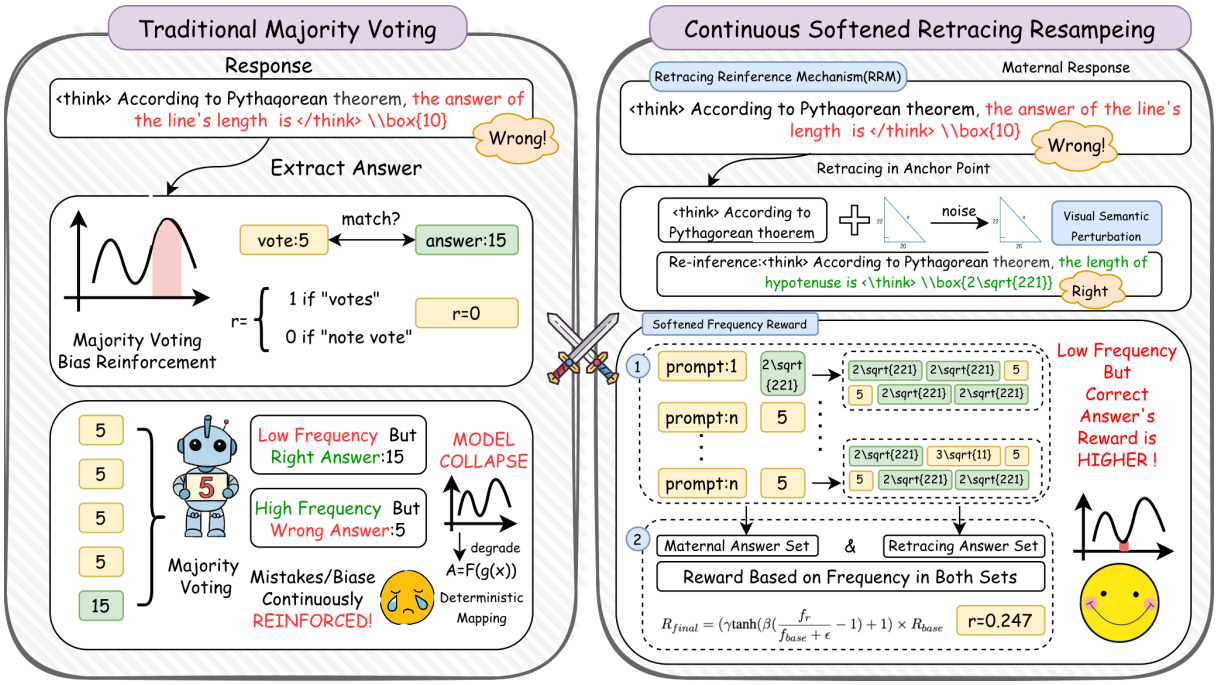


Figure 1: **Overview of our work.** During unsupervised self-evolutionary reinforcement learning, traditional majority voting methods (left) for pseudo-labeling rely solely on the model's inherent biases. This frequently leads to model collapse, where the model degenerates into a deterministic mapping and fails to explore the true solutions. To alleviate this phenomenon, our method introduces CSRS (right) to reduce the occurence of situations.

ity voting, we introduce Softened Frequency Reward (**SFR**). Rather than assigning discrete binary scores, this mechanism calculates base rewards based on the occurrence frequency of the problem's answer in the union of the re-inference and maternal sets. By utilizing the frequency variance between two sets to calibrate the reward score, we incentivize the model to explore low-probability but correct responses. Additionally, we introduce Visual Semantic Perturbation (**VSP**) during the re-inference phase, forcing the model to further prioritize invariant mathematical logic over superficially low-level image features.

We conducted a comprehensive evaluation of CSRS across four major multimodal geometric reasoning benchmarks based on Qwen2.5-VL-7B (Bai et al., 2025): MathVision (Wang et al., 2024), MathVista (Lu et al., 2023) MathVerse (Zhang et al., 2024) and We-Math (Qiao et al., 2025). Notably, compared to the baseline (MM-UPT), we gains 2.34%, 1.48%, 2.41%, 3.28% respectively on benchmarks. Experimental results show that CSRS achieves state-of-the-art (SOTA) performance in the unsupervised self-evolution. The primary contributions of this work are as follows:

(1) We propose the CSRS, which uses retracing re-inference anchor points combined with softened frequency rewards to theoretically and empirically mitigate the model collapse inherent in majority voting mechanism.

(2) Moving beyond binary rewards, CSRS dynamically amplifies the rewards for low-frequency but potentially correct long-tail reasoning paths, significantly improving the robustness of MLLM self-evolution's inference capability.

(3) CSRS achieves leading performance across four multimodal mathematical benchmarks. In particular, on MathVista, it reaches an accuracy of 68.25%, demonstrating the substantial potential of unsupervised self-evolution without the need for expensive human intervention.

2 Related Work

Reinforcement Learning (RL) (Xu and Ding, 2025; Yu et al., 2025; Lin and Xu, 2025) has emerged as a pivotal tool for unlocking the latent reasoning capabilities of Large Language Models (LLMs). These advancements span from established algorithms like PPO (Schulman et al., 2017) and DPO (Rafailov et al., 2023) to the recently introduced GRPO (Shao et al., 2024). However, these methods primarily operate within the paradigm of supervised reinforcement learning, where reliable parameter updates relies heavily on

high-quality, external ground-truth labels. This strict dependency on external signals inherently constrains the models capacity for autonomous evolution in scenarios where human annotations are expensive or unavailable.

To bridge this gap, the unsupervised self-evolution paradigm has gained significant traction as it enables models to autonomously generate pseudo-labels to guide their own iterative optimization (Liu et al., 2025; Kwan et al., 2025). Existing works attempts to substitute human labels with signals from diverse sources. AbsoluteZero (Zhao et al., 2025) uses external verifiers, such as math or code executors (e.g., MathVerify) to validate reasoning outcomes. While this reduces manual labeling effort, the feedback remains constrained to the coverage of external rules, essentially remaining a form of supervised training (Wang et al., 2025; Dong et al., 2025). Other researchs seeks complete independence from external intervention. MM-UPT (Wei et al., 2025) introduces a framework that adopts majority voting results among responses as pseudo-labels. EvolRL (Zhou et al., 2025) prioritizes the stability of majority voting while refining rewards with novelty scores to enhance exploratory diversity. MultiAgentEvolve (Chen et al., 2025b) and VisPlay (He et al., 2025) use self-play dynamics by decoupling the model into two roles: a challenging questioner and an adaptive responder.

Despite these explorations into unsupervised self-evolution, most existing approaches are restricted to providing binary (0/1) hard rewards based solely on the final answer (Kwan et al., 2025; Liu et al., 2025). Such coarse-grained feedback is insufficient for guiding the model through critical reasoning points. Worse still, these signals exacerbate a self-reinforcement trap, where the model continuously amplifies its initial distributional biases, accelerating the onset of model collapse. This degradation in signal quality ultimately leads to model collapse, causing training trajectories to deviate from valid logical reasoning.

3 Methodology

This section is structured as follows: First, Section 3.1 provides a theoretical analysis of the underlying causes of model collapse in MLLM self-evolution under conventional majority voting, followed by how to mitigate this issue through softened frequency rewards. Subsequently, Section 3.2 introduces the CSRS framework, detailing its technical components including retracing re-inference, softened frequency rewards and visual semantic perturbation strategies.

3.1 Preliminaries

3.1.1 Continuous Rewards as a Stabilizer Compared to 0-1 Rewards

This study is grounded in the Group Relative Policy Optimization (GRPO) algorithm. In an unsupervised self-evolution scenario, given a reference distribution P_{ref}, the model updates its policy P by maximizing the relative advantage of intra-group samples $A(x) = r(x) - \bar{r}$ subject to a KL divergence constraint. The objective function is:

$$\max_{P} \mathbb{E}_{x\sim P}[A(x)] - \beta \mathbb{D}_{KL}(P \| P_{ref}) \tag{1}$$

By constructing the Lagrangian functional:

$$L(P) = \sum_{x} P(x)A(x) - \beta \sum_{x} P(x) \log \frac{P(x)}{P_{ref}(x)} + \lambda \left(\sum_{x} P(x) - 1 \right) \tag{2}$$

and taking the derivative with respect to $P(x)$, we derive the ideal closed-form solution for policy iteration during self-evolution:

$$P_{n+1}(x) = \frac{P_n(x) e^{A_n(x)/\beta}}{Z} \tag{3}$$

where Z is the normalization constant, β is KL regularization weight and n denotes the number of self-evolution iteration. This derivation reveals that the essence of unsupervised self-evolution is an **exponential self-reinforcement process**, where the evolutionary trajectory of the distribution is driven by the advantage function $A_n(x)$.

However, the conventional Majority Voting (MV) mechanism defines $r_n(x)$ as a binary reward at the n-th evolution iteration, a discrete signal that triggers over-concentration of gradient updates on high frequency paths. Marginal biases in the initial distribution are rapidly polarized through iteration, causing the probability density to map extremely toward high-frequency mode regions. In the limit, the model degenerates into a deterministic mapper (Yun et al., 2025), stripped of its exploratory capacity. To quantify and mitigate this process, we define the Contrastive Factor G_n denoted as:

$$G_n = \frac{R_{n+1}}{R_n} = \exp(\eta \Delta r_n) \tag{4}$$

Algorithm 1: Continuous Softened Retracing reSampling(CSRS)

Require : Dataset $\mathcal{D}$, model θ, retracing rate ω, resample m, learning rate α, maternal num n.

Initialize optimizer $\mathcal{O}$ with θ;
for *each minibatch* $\{x\}$ *in* $\mathcal{D}$ **do**
  Sample $\mathcal{A}_n, \tau \sim \pi_\theta(| \, x)$;
  for $t = 1$ *to* n **do**
    $\tau' \leftarrow \text{Concat}(x_i^p, \text{retrace}(\tau, \omega))$;
    Visual Perturbation $x_i^i \leftarrow x + \epsilon$;
    Resample new $\mathcal{A}_m$ from τ', x_i^i;
  end
  $\mathcal{A}_{all} \leftarrow \mathcal{A}_n \cup \mathcal{A}_m$;
  for $i = 1$ *to* n **do**
    $r_{base}^i \leftarrow \mathcal{A}_n(a_i)/\mathcal{A}_{all}$;
    $r_{fre} \leftarrow \text{SFR}(\mathcal{A}_m, \mathcal{A}_{all}, r_{base}^i)$;
    $r \leftarrow r_{fre}$;
  end
  $b \leftarrow \text{ComputeBaseline}(\cdot)$;
  $g \leftarrow \nabla_\theta \log \pi_\theta(a_0 \mid x)(r - b)$;
  $\theta \leftarrow \theta + \alpha \mathcal{O}(g)$;
end
return θ

where η is $\frac{1}{\beta}$ and $R_n = \frac{P_n(x_1)}{P_n(x_2)}$. x_1 are majority samples and x_2 are long-tail samples. Under the MV mechanism, $\Delta r_{MV} = 1$ causes the contrastive factor to expand exponentially at the maximum rate of $G_{MV} = e^\eta$.

In contrast, our proposed softened frequency reward provides a damping effect by introducing continuity into the probability density. Let ρ denote the proportion of mode answers and ϵ denote the proportion of long-tail answers. Then:

$$G_{SR} = \exp(\eta(\rho - \epsilon)) < G_{MV} = \exp(\eta) \quad (5)$$

Since $0 < \rho - \epsilon < 1$, this inequality proves that the SFR mechanism effectively curtails the trend of distributional polarization by reducing the contrastive gain, thereby preserving the necessary logical diversity for the model during the self-evolution process. Full mathematical proofs for the propositions discussed in this section are deferred to Appendix A.1 in detail.

3.1.2 Retracing Re-inference Overcomes Sampling Sparsity

In long-range reasoning tasks, the candidate path space expands exponentially at a rate of $\mathcal{O}(b^L)$ where L is the reasoning length and b is the branching factor such as vocabulary size, rendering correct reasoning paths extremely rare, a phenomenon known as sampling sparsity. Under a finite computational budget, stochastic sampling struggle to cover critical logical nodes. Consequently, early-stage biases are amplified through the recursive process, exacerbating the risk of model collapse.

To address this, we introduce the concept of local stability, which posits that the robustness of a logical path should be calibrated by answer consistency at key decision points. Rather than merely narrowing the search, CSRS implements this by establishing retracing anchor points at intermediate stages to trigger local resampling. This mechanism induces a deliberate distributional divergence from the original trajectory, effectively breaking the cumulative bias of the initial policy. It expands the model's reach into the long-tail sample space, providing a vital opportunity to discover correct but low-probability paths that would otherwise be suppressed by the primary distribution. This approach effectively curtails model collapse and provides more robust gradient guidance for the MLLM self-evolution process.

3.2 Continuously Softened Retracing reSampling

3.2.1 Theoretical Assumptions

To systematically formalize our method, we establish the following fundamental assumptions about mathematical reasoning:

Semantic Uniqueness and Consistency of Reasoning. For a mathematical problem of a given complexity, the ideal reasoning path and its corresponding terminal result should remain semantically unique and consistent. We assume that if a model regenerates its reasoning paths starting from any intermediate step, the underlying mathematical logic and the final conclusion must remain invariant, regardless of linguistic variations or paraphrasing in the expression.

Existence of Critical Reasoning Pivots. We hypothesize that mathematical reasoning is not a uniform linear progression but is governed by critical reasoning pivots, retracing anchor points as human solve a problem. These pivots are key steps that shape the following reasoning and determine whether the final answer is correct.

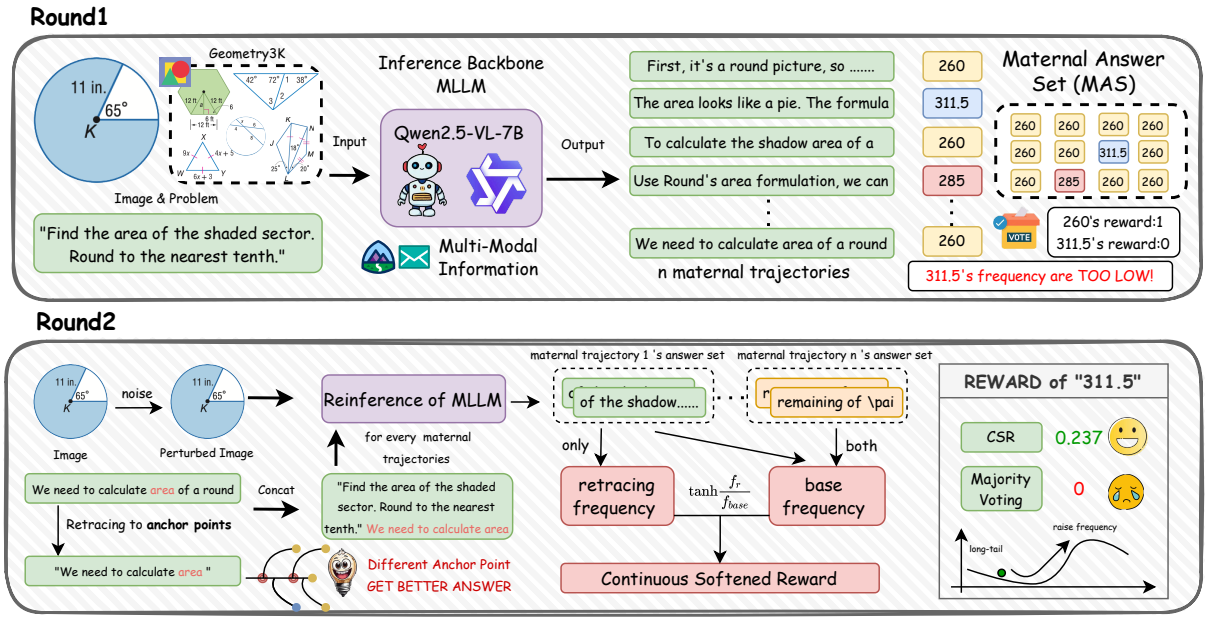


Figure 2: **Pipeline of our method**. (Round1) illustrate initially maternal trajectories and answers generated by MLLM. (Round2) shows three core components of CSRS and improvement to rewards.

3.2.2 Algorithm Details

To overcome the training instability and model collapse caused by the bias in majority voting during unsupervised self-evolution, we propose the Continuous Softened Retracing reSampling (CSRS). This framework comprises three synergistic components: Retracing Re-inference Mechanism (RRM), Softened Frequency Reward (SFR) and Visual Semantic Perturbation (VSP). The picture of our method pipeline can be seen in Fig. 2.

Retracing Re-inference Mechanism (RRM). In long-range mathematical reasoning tasks, subtle early-stage logical deviations tend to amplify exponentially as the reasoning steps progress. To capture and rectify these intermediate errors, we introduce a Retracing Re-inference Mechanism (RRM) . Let $\mathcal{D}$ denote a dataset consisting of prompt-image pairs (X_p, X_i). For a specific pair sampled from $\mathcal{D}$, the text prompt X_p is represented as a sequence of k tokens, denoted by $X_p = (x_1, x_2, \ldots, x_k)$ and the model generates n initial reasoning paths, designated as maternal trajectories denoted as $\mathcal{A}_n = \{R_{p_1}, \cdots, R_{p_n}\}$. Subsequently, a critical logical node termed the retracing anchor point, is identified within each trajectory. For every response trajectory $R_p = (y_1, y_2, \cdots, y_n)$, we truncate it at this anchor point and concatenate it with the original prompt:

$$X'_p = \text{Concat}(X_p, (y_1, y_2, \cdots, y_{\omega \text{len}(R_p)})) \quad (6)$$

where $\omega \in (0, 1)$ denotes the retracing rate, and $y_{\omega \text{len}(R_p)}$ represents the retracing anchor point located at the position $\lfloor \omega \cdot \text{len}(R_{p_n}) \rfloor$ within the initial response sequence. Then we guide the model to perform m local explorations for each prompt, resulting in a total of mn re-inference trajectories denoted as $\mathcal{A}_m = \bigcup_{i=1}^{n} \{R^1_{p_i}, R^2_{p_i}, \cdots, R^m_{p_i}\}$. This design transforms a broad global search into dense local sampling, ensuring no potential correct steps are missed at key logical forks.

Softened Frequency Reward (SFR). To mitigate the model collapse induced by the binary hard rewards of conventional majority voting, we design the Softened Frequency Reward (SFR). We first construct a comprehensive set $\mathcal{A}_{all} = \mathcal{A}_n \cup \mathcal{A}_m$ containing all $(m+1)n$ trajectory answers. For any answer a from a maternal trajectory, its base reward R_{base} is defined as its statistical frequency within the entire set:

$$R_{base}(a) = f_{base}(a) = \frac{\text{Count}(a, \mathcal{A}_{all})}{(m+1)n} \quad (7)$$

To further validate reasoning stability, we evaluate the frequency variance of the answer between the local re-inference set $\mathcal{A}_m$ and the comprehensive set $\mathcal{A}_{all}$. This variance essentially quantifies the sampling distribution divergence, enabling the model to deviate from its original biased distribution and actively explore potential long-tail samples that are often neglected during global sampling. If an answer's frequency increases after local re-inference, it validates the latent robustness

Table 1: Comparison with supervised methods. Our unsupervised method CSRS shows competitive performance against base models and current SOTA MM-UPT across various benchmarks.

Model and Methods	Unsupervised?	Training Data	MathVision	MathVerse	MathVista	WeMath	Avg
Qwen2.5-VL-3B	✗	-	22.47	34.54	62.30	57.53	44.21
Qwen2.5-VL-7B	✗	-	25.40	44.24	66.42	67.65	50.93
MM-UPT	✓	Geometry3K	26.95	44.53	66.47	68.49	51.61
MM-UPT	✓	GeoQA	26.61	44.15	65.84	68.25	51.21
MM-UPT	✓	MMR1	25.98	45.12	66.27	69.14	51.63
CSRS (Ours)	✓	Geometry3K	27.97	**46.01**	67.81	**71.77**	53.39
CSRS (Ours)	✓	GeoQA	**28.95**	45.82	**68.25**	69.32	53.09
CSRS (Ours)	✓	MMR1	27.86	45.89	67.81	70.53	53.05

Table 2: Sensitivity analysis of γ and β.

$\gamma(\beta = 5.0)$	Acc. (%)	$\beta(\gamma = 0.2)$	Acc. (%)
0.05	44.65	3.5	45.16
0.1	45.12	5.0	**46.01**
0.2	**46.01**	6.5	45.87
0.35	43.78	8.0	45.72

Table 3: Re-inference Rollout.

Size (R)	MathVista
$R = 4$	66.45
$R = 5$	**68.25**
$R = 6$	67.46
$R = 7$	66.01
$R = 8$	64.98

Table 4: Retracing Rates.

Retracing Rate	MathVista
ω=0.1	66.31
ω=0.3	67.01
ω=0.5	66.58
ω=0.7	**68.25**
ω=0.9	66.43

of that specific reasoning path. This trend signifies that the path possesses superior conditional consistency despite being a long-tail sample in the model's global distribution, prompting a hierarchical adjustment accordingly as follows:

$$R_{final} = (\gamma \tanh(\beta(\frac{f_r}{f_{base} + \epsilon} - 1)) + 1) \times R_{base} \tag{8}$$

where $f_r = \frac{\text{Count}(a, \mathcal{A}_m)}{mn}$ denotes answer's frequency in re-inference answers set and γ, β are hyperparameters. By adjusting β, the reward function becomes highly sensitive to correct but low-frequency paths in the long-tail region. Meanwhile, the saturation property of $\tanh$ caps the reward for high-frequency samples, preventing the model from over-fitting to common errors and effectively mitigating model collapse. SFR encourages exploration in these "initially low-frequency but potentially correct" regions, serveing as a proxy for process-level supervision.

Visual Semantic Perturbation. To prevent the model from cheating by exploiting superficial visual features rather than deep mathematical logic, we introduce Visual Semantic Perturbation (VSP) during the re-inference phase. By applying Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to the original image I:

$$I' = I + \epsilon \tag{9}$$

This perturbation forces the model to rely on invariant logical structures, ensuring that the self-evolution process is driven by genuine reasoning rather than visual heuristics in images.

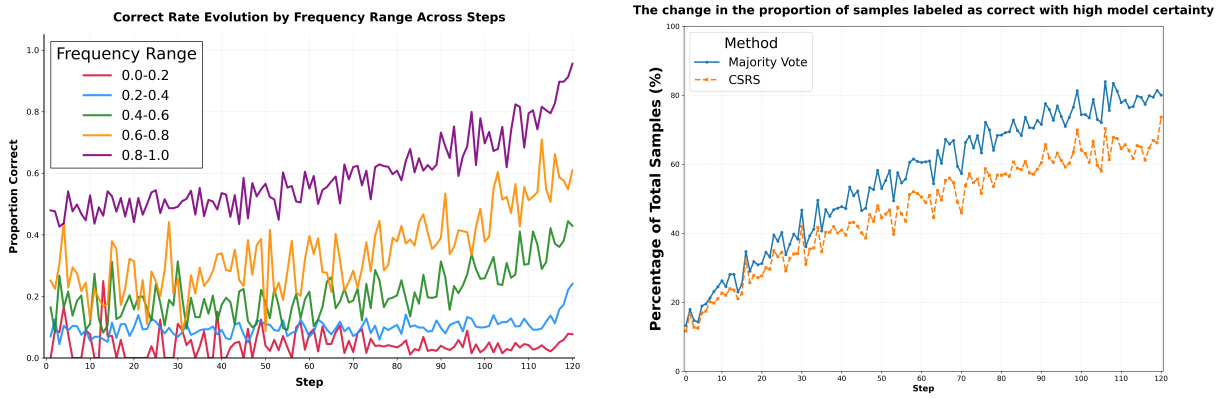
4 Experiment

In this section, we conduct extensive experiments to evaluate the performance of CSRS. We begin by detailing the experimental configuration in Section 4.1, covering the benchmarks, evaluation metrics, and specific implementation details. Subsequently, Section 4.2 presents our primary results along with a comprehensive performance analysis across various multimodal reasoning tasks. Finally, we perform ablation and visualization studies in Section 4.3 to evaluate the effectiveness of each component in our method. Specific cases study are clearly presented in Appendix A.6.

4.1 Experimental Setup

Datasets and Benchmarks. In this study, we focus on the stability of internal signals in unsupervised self-evolution. We select Geometry3K (Lu et al., 2021), GeoQA (Chen et al., 2021), MMR1 (Leng et al., 2025) as the primary training sets for the self-evolution process. To evaluate the reasoning capabilities of our model, we use four mainstream multimodal scientific reasoning benchmarks: (1) MathVision (Lu et al., 2021) , (2) MathVerse (Zhang et al., 2024) , (3) MathVista (Lu et al., 2023) , and (4) We-Math (Qiao et al., 2025) . These benchmarks provide a comprehensive evaluation ranging from formal geometric problems to complex mathematical reasoning situated in diverse real-world contexts.

Implementation Details. Our unsupervised



(a) Accuracy of answers within each frequency interval.

(b) High certainty samples distribution percentage.

Figure 3: **Visualization of the accuracy and propotion of high-confidence samples during training**. (a) Answer accuracy calculated within partitioned frequency intervals, where each sample is uniquely assigned based on its frequency. (b) The evolution of high-confidence (frequence $\in [0.8 - 1.0]$) samples proportions during training, which serves as a key indicator of model collapse.

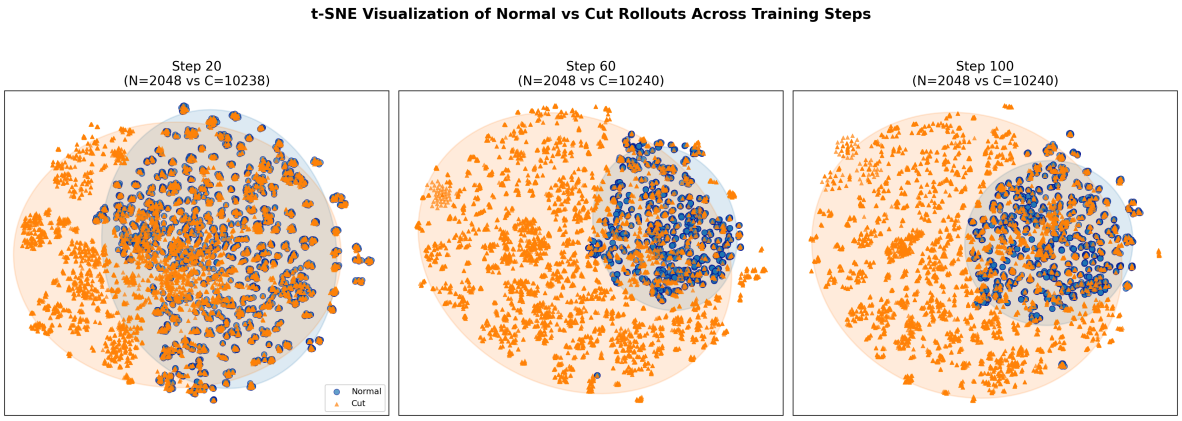


Figure 4: **The change of distribution of Maternal Answers Set and Re-inference Answers Set**. The figure shows different distribution of samples in two sets at the step of 20,60,100. **Blue dots** represent the semantic space of vanilla responses, while **orange dots** denote the semantic space of responses generated by CSRS.

post-training pipeline is implemented using the veRL (Sheng et al., 2024) framework, built upon the GRPO algorithm. Specifically, the training is conducted over 15 epoches. We use the AdamW(Loshchilov and Hutter, 2017) optimizer with an initial learning rate of 1×10^{-6} and a weight decay of 0.01. The KL divergence constraint β in GRPO is set to 0.01 to regularize policy updates. We initially generate n=8 maternal trajectories for each prompt, followed by m=5 local re-inference trajectories starting from the retracing anchor points. Other configurations are detailed in Appendix A.2.3.

4.2 Experiment Results

In this section, we evaluate **C**ontinuous **S**oftened **R**etracing re**S**ampling (**CSRS**) by comparing it with the state-of-the-art MM-UPT (Wei et al., 2025) baseline and the original Qwen2.5-VL (Bai et al., 2025) base models. To ensure a fair and rigorous comparison, both CSRS and MM-UPT are post-trained using the same Qwen2.5-VL-7B backbone under identical experimental settings. The experimental results on four benchmarks MathVision, MathVerse, MathVista and We-Math are summarized in Table 1.

As demonstrated in our results, CSRS achieves consistent performance gains over both the base models and the MM-UPT baseline. Specifically, when utilizing Qwen2.5-VL-7B as the backbone, the CSRS achieves scores of **28.95%**, **46.01%**, **68.25%**, **71.77%** respectively on four benchmarks, outperforming the MM-UPT baseline by **2.34%**, **1.48%**, **2.41%**, **3.28%** respectively and surpassing the performance of the original Qwen2.5-VL-7B . These results suggest that

CSRS is effective in complex geometric reasoning tasks, mitigating the model collapse while better exploiting high-quality reasoning paths.

4.3 Ablation Experiments and Visualization

The Synergistic Effect of RRM, SFR and VSP. As demonstrated in Table 5, removing either module results in a substantial performance degradation. This indicates that the model can overcome training instability in unsupervised learning when retracing provides diverse exploratory samples in long-tail regions. Furthermore, as illustrated in Fig. 3b, CSRS shows superior performance while maintaining the proportion of high-confidence samples at a lower level. Fig. 5 also demonstrates a slower rate of entropy reduction compared to majority voting. All results indicate that our method results in high accuracy without suffering from the degeneracy typical of self-evolution, where a model becomes overly confident in a set of responses.

Effect of Softened Frequency Reward. Our softened frequency reward plays a significant role in CSRS. Crucially, as illustrated in Fig. 3a, the proportion of correct samples across all frequency ranges exhibits a consistent upward trend as training progresses. This empirically validates our theoretical derivation in Section 3.1.

Effect of Visual Semantic Perturbation. Finally, we evaluate the impact of visual perturbation as shown in Table 5. Removing this perturbation leads to performance decay across all three benchmarks. This shows that visual perturbation prevents the model from relying on superficial visual features, compelling it to prioritize deep mathematical logic and thus enhancing the overall robustness of the reasoning process.

Ablation of hyperparameters. As shown in Table 3, the performance improves as the number of re-inference rollouts increases, reaching its peak at 5. While increasing rollouts initially facilitates the exploration of diverse information around anchored trajectories, exceeding this threshold leads to cumulative errors and introduces excessive computational overhead.

Regarding the retracing rate shown in Table 4, we observe a fluctuating yet upward trend, with performance peaking at 0.7. An excessively low retracing rate causes the search space to collapse back into the exponentially vast domain characteristic of traditional majority voting. Conversely, a high retracing rate implies the reasoning path is already functionally deterministic, offering negligible room for further refinement.

The sensitivity analysis of γ and β is presented in Table 2. We observe that the model achieves optimal accuracy when $\gamma = 0.2$ and $\beta = 5.0$. The results demonstrate that small deviations result in only minor performance fluctuations, confirming the algorithmic robustness of our SFR against hyperparameter variations.

Visualization of Maternal Answers Set and Re-inference Answer Set. We use all-MiniLM-L6-v2 (SentenceTransformers, 2024) to perform a semantic space analysis on the sets of maternal trajectories and re-inference trajectories during the training process. The corresponding t-SNE visualizations at steps 20, 60, and 100 are provided in Fig. 4. It is observed that the semantic scope of the maternal trajectories progressively contracts into a relatively confined space. However, CSRS mitigates early-stage bias and maintains a substantially larger search space around anchor points, facilitating the exploration of long-tail samples and alleviating the risk of model collapse.

Other Model Family, Stability and Anchors Strategy. To further validate the robustness of CSRS, we conducted additional experiments comparing stochastic versus fixed retracing anchors, evaluating performance across diverse model families, and assessing evolutionary stability. Detailed results for these analyses are provided in Appendix A.3, A.4, A.5. The experimental evidence consistently demonstrates that our proposed strategies are optimal, exhibiting superior robustness and stability across various architectures and self-evolution cycles on geometric reasoning tasks.

5 Conclusion

This paper presents Continuous Softened Retracing reSampling (CSRS) designed to address the challenge of model collapse on math tasks in the unsupervised self-evolution of MLLMs. CSRS introduces a retracing re-inference mechanism to effectively explore models coverage of the logical long-tail distribution samples. It implements softened frequency reward and visual perturbation dynamically calibrate the frequency variance across sampling distributions, steering the model away from over-fitting to high-frequency biases. Empirical results demonstrate that CSRS consistently outperforms contemporary unsupervised methods across multiple multimodal benchmarks.

Limitations

While CSRS demonstrates significant potential in the field of unsupervised MLLM self-evolution, we acknowledge several limitations that provide promising directions for future research:

Our current evaluation and methodology are primarily focused on mathematical reasoning tasks. It remains to be fully explored about broader, open-ended general scene understanding. Extending CSRS to diverse multimodal tasks is a key objective for our future work.

Looking ahead, we aim to further investigate the evolution of intrinsic reward mechanisms from static logical verification into dynamic, self-reflective evaluation, which would facilitate preciser guidance of reasoning quality in scenarios.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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

Jinhao Chen, Zhen Yang, Jianxin Shi, Tianyu Wo, and Jie Tang. 2025a. MathSE: Improving Multimodal Mathematical Reasoning via Self-Evolving Iterative Reflection and Reward-Guided Fine-Tuning. *arXiv preprint*. ArXiv:2511.06805 [cs].

Yixing Chen, Yiding Wang, Siqi Zhu, Haofei Yu, Tao Feng, Muhan Zhang, Mostofa Patwary, and Jiaxuan You. 2025b. Multi-Agent Evolve: LLM Self-Improve through Co-evolution. *arXiv preprint*. ArXiv:2510.23595 [cs].

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, and 1 others. 2025. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. 2025. Visplay: Self-evolving vision-language models from images. *arXiv preprint arXiv:2511.15661*.

Minqi Jiang, Andrei Lupu, and Yoram Bachrach. 2025. Bootstrapping Task Spaces for Self-Improvement. *arXiv preprint*. ArXiv:2509.04575 [cs].

Wai-Chung Kwan, Joshua Ong Jun Leang, Pavlos Vougiouklis, Jeff Z. Pan, Marco Valentino, and Pasquale Minervini. 2025. OpenSIR: Open-Ended Self-Improving Reasoner. *arXiv preprint*. ArXiv:2511.00602 [cs].

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, and 1 others. 2025. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources. *arXiv preprint arXiv:2509.21268*.

Heng Lin and Zhongwen Xu. 2025. Understanding tool-integrated reasoning. *arXiv preprint arXiv:2508.19201*.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. 2025. SPICE: Self-Play In Corpus Environments Improves Reasoning. *arXiv preprint*. ArXiv:2510.24684 [cs].

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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

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

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, and 1 others. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070.

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

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

SentenceTransformers. 2024. all-minilm-l6-v2.

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. 2025. Can Large Reasoning Models Self-Train? *arXiv preprint*. ArXiv:2505.21444 [cs].

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*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient RLHF framework. *CoRR*, abs/2409.19256.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J Andrew Bagnell. 2025. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. *arXiv preprint arXiv:2503.01067*.

Omkar Thawakar, Shravan Venkatraman, Ritesh Thawkar, Abdelrahman M. Shaker, Hisham Cholakkal, Rao Muhammad Anwer, Salman H. Khan, and Fahad Shahbaz Khan. 2025. Evolmm: Self-evolving large multimodal models with continuous rewards. *CoRR*, abs/2511.16672.

Jiahao Wang, Weiye Xu, Aijun Yang, Wengang Zhou, Lewei Lu, Houqiang Li, Xiaohua Wang, and Jinguo Zhu. 2025. Enhancing the Outcome Reward-based RL Training of MLLMs with Self-Consistency Sampling. *arXiv preprint*. ArXiv:2511.10648 [cs].

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. 2025. First SFT, Second RL, Third UPT: Continual Improving Multi-Modal LLM Reasoning via Unsupervised Post-Training. *arXiv preprint*. ArXiv:2505.22453 [cs].

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge. *arXiv preprint*. ArXiv:2407.19594 [cs].

Zhongwen Xu and Zihan Ding. 2025. Single-stream policy optimization. *arXiv preprint arXiv:2509.13232*.

Jiahao Yu, Zelei Cheng, Xian Wu, and Xinyu Xing. 2025. Gpo: Learning from critical steps to improve llm reasoning. *arXiv preprint arXiv:2509.16456*.

Jihun Yun, Juno Kim, Jongho Park, Junhyuck Kim, Jongha Jon Ryu, Jaewoong Cho, and Kwang-Sung Jun. 2025. Alignment as Distribution Learning: Your Preference Model is Explicitly a Language Model. *arXiv preprint*. ArXiv:2506.01523 [cs].

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

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025. Absolute Zero: Reinforced Self-play Reasoning with Zero Data. *arXiv preprint*. ArXiv:2505.03335 [cs].

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025. Evolving language models without labels: Majority drives selection, novelty promotes variation. *arXiv preprint arXiv:2509.15194*.

A Appendix

A.1 Proof of Section 3.1

To theoretically justify how **Continuous Softened Retracing reSampling (CSRS)** alleviates model collapse, we analyze the evolutionary dynamics of the policy under different reward mechanisms.

Proposition 1 (Self-evolution Closed-form Solution)**.** *In the GRPO optimization objective, adopting majority voting leads to distribution collapse, and its ideal closed-form solution $P(x)$ and the reference distribution $P_{ref}(x)$ satisfy the following relation:*

$$P(x) = \frac{P_{ref}(x) e^{A_n(x)/\beta}}{Z}$$

where Z is the normalization constant.

Proof. According to the Lagrange multiplier method, we construct the following functional: Under the GRPO optimization objective, the ideal closed-form solution is given by:

$$\max_P \mathbb{E}_{x\sim P}[A_n(x)] - \beta\, \mathbb{D}_{\mathrm{KL}}\big(P \parallel P_{\mathrm{ref}}\big)$$

where P denotes the optimal distribution to be found. Applying the Lagrange multiplier method yields:

$$\mathcal{L}(P) = \sum_x P(x) A_n(x) - \beta \sum_x P(x) \log \frac{P(x)}{P_{\mathrm{ref}}(x)} + \lambda\Big(\sum_x P(x) - 1\Big)$$

with λ being the Lagrange multiplier. Differentiating with respect to $P(x)$ gives:

$$\frac{\partial \mathcal{L}}{\partial P(x)} = A_n(x) - \beta\Big(\log \frac{P(x)}{P_{\text{ref}}(x)} + 1\Big) + \lambda = 0.$$

Solving this yields

$$P(x) = P_{\text{ref}}(x) e^{\frac{A_n(x)}{\beta} + C} = \frac{P_{\text{ref}}(x) e^{\frac{A_n(x)}{\beta}}}{Z}.$$

Consequently, for a MLLM that uses majority voting for selfevolution, its gradient dynamics follow the equation above, where

$$A_n(x) = \mathbf{1}(x \in M_n) - \rho_n,$$

where M_n the set of majority votes and ρ_n the average proportion of the majority. When the majority voting mechanism is used, the value of $A_n(x)$ at majority points is large, so that $e^{A_n(x)/\beta}$ is dominated by $A_n(x)$ (here β is a constant). Although the KL divergence term can be partly constrained by β, the self evolution trend still collapses towards the majority region or areas of high probability density in the models own distribution. In the limit, this distribution collapse degenerates into an extreme, fixed mapping. □

Proposition 2 (CSRS Can Relieve Model Collapse)**.** *When a reinforcement learning algorithm uses GRPO, softening the frequency reward compared with the 01 reward of majority voting alleviates the distributioncollapse phenomenon in LLM selfevolution.*

Proof. We have $A_n(x) = r(x) - \bar{r}$. For the softenedfrequency reward,

$$A_{\text{sf}}(x) = P_n(x) - \bar{P}_n.$$

Since

$$\begin{aligned}\mathbb{E}[\bar{r}] &= \mathbb{E}_{x \sim P_n}[r(x)] \\ &= \sum_{y \in P_n} P_n(y) r(y) \\ &= \sum_{y \in P_n} P_n(y) P_n(y) \\ &= \sum_{y \in P_n} P_n^2(y)\end{aligned}$$

it follows that

$$A_{\text{sf}}(x) = P_n(x) - \sum_{y \in P_n} P_n^2(y).$$

According to the gradient dynamic expression in **Proposition1**, the update with the softenedfrequency reward reads

$$P_{n+1}(x) = P_n(x) \exp\Big(\frac{\eta\big(P_n(x) - \sum_{y \in P_n} P_n^2(y)\big)}{Z}\Big).$$

Let x_{tail} denote a long tail event in the selfevolution. For majority voting, $A_{\text{MV}} = 0 - \rho = -\rho$, which gives

$$P_{n+1}(x) = P_n(x) e^{-\eta \rho_n}.$$

For the softened frequency reward, writing $P_n(x_{\text{tail}}) = \epsilon$, we obtain

$$P_{n+1}(x_{\text{tail}}) = P_n(x_{\text{tail}}) \frac{e^{\eta \epsilon}}{Z\, e^{\eta \bar{P}_n}}.$$

Define a contrast operator: let x_1 be majority samples and x_2 longtail samples. The contrast at the n-th iteration is

$$R_n = \frac{P_n(x_1)}{P_n(x_2)}.$$

The contrast at the next generation is

$$R_{n+1} = R_n \exp\big(\eta\,[r_n(x_1) - r_n(x_2)]\big).$$

Define the contrast factor as

$$G_n = \frac{R_{n+1}}{R_n} = \exp\big(\eta\, \Delta r_n\big),$$

where

$$\Delta r_n = \begin{cases} 1, & \text{if } x \in \text{Distri(MV)}, \\ \rho - \epsilon, & \text{if } x \in \text{Distri(SF)}. \end{cases}$$

Because $0 < \rho - \epsilon < 1$, we have $G_{\text{sf}} < G_{\text{mv}}$. Hence the softenedfrequency reward slows down the rate of distribution collapse. □

A.2 Implementation Details

A.2.1 Compute Resources

We conduct our experiments using NVIDIA A800-80G GPUs. The experimental time using 8 A800 for training Qwen2.5-VL-7B (Bai et al., 2025) on the Geometry3K (Wang et al., 2024),GeoQA (Chen et al., 2021), MMR1 (Leng et al., 2025) dataset using GRPO is around 36 hours.

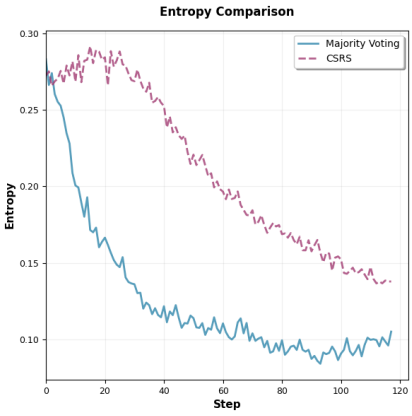


Figure 5: Entropy Change during training

Table 5: Ablation studies of different modules. The results demonstrate the incremental contribution of each component to the overall performance.

Module	M-Vis.	M-Ver.	M-Vist.	WeMath
MajorityVote	25.40	44.24	66.42	67.65
+SFR	26.52	45.10	67.15	69.42
+RRM	26.85	45.35	67.38	69.85
+VSP	25.55	44.35	66.55	67.82
+SFR+RRM	27.68	45.85	67.62	71.30
+SFR+VSP	26.70	45.22	67.28	69.65
+RRM+VSP	27.02	45.48	67.50	70.12
CSRS	**27.97**	**46.01**	**67.81**	**71.77**

A.2.2 Entropy Change

We further compare the evolution of policy entropy between our proposed CSRS and the Majority Voting baseline (MM-UPT) throughout the training process, as illustrated in Fig. 5. It is evident that the entropy in CSRS exhibits a slower and more gradual decay compared to MM-UPT. Unlike the binary reward structure of majority voting, CSRS maintains higher training stability and effectively mitigates model collapse. This gentler transition allows the model to have more access to long-tail samples, thereby facilitating the acquisition of more diverse and informative learning signals.

A.2.3 Hyperparameters

Our complete hyperparameters and system configurations are shown in Table 6.

A.2.4 Modules Ablation Studies Table

Due to space constraints, the detailed results for the module ablation study are presented in Table 5.

Table 6: Complete Hyperparameter and System Configurations.

Category	Hyperparameter (Value)
Data Configuration	
Train Batch Size	256
Max Prompt Length	1024
Max Response Length	1536
Filter Overlong Prompts	True
Truncation Strategy	"error"
Image Key	"images"
Model & Optimization	
Base Model	Qwen2.5-VL-7B-Instruct
Optimizer Learning Rate	1e-6
KL Loss Coefficient (β)	0.01
KL Loss Type	Low-Var KL
Entropy Coefficient	0
Gradient Checkpointing	Enabled
Remove Padding / Fused Kernels	Enabled
PPO / GRPO Settings	
Algorithm	GRPO
PPO Mini Batch Size	64
PPO Micro Batch Size (per GPU)	8
Rollout Group Size (G)	8
Ref Log-Prob Micro Batch (GPU)	4
Param Offload (Ref Model)	True
Sampling & Retracing (CSRS)	
Rollout Engine	vLLM
Sampling Group Size (n)	8
Re-inference Number	5
Retracing Rate	0.7
GPU Memory Utilization	0.6
Trainer Settings	
Total Epochs	15
GPUs per Node	8
Number of Nodes	1
Test Frequency	5 epochs
Save Frequency	20 steps

A.3 Model Families Experiments

In multimodal self-evolution tasks, the ability of a framework to generalize across different model families is a key indicator of its algorithmic robustness. We expand our experiments to include two additional prominent MLLM families: InternVL3-8B and ThinkLite-VL-7B. As shown in the Table 7 below, the results prove that CSRS consistently delivers significant performance gains across different model architectures and parameter scales.

A.4 Convergence and Stablity Analysis

Since the ground-truth labels are inaccessible in self-evolution tasks, the model faces the risk of converging to a local optimum. Consequently, we conducted additional long-term evolution experiments to verify whether performance degrades as training duration increases. We evaluated the model every 5 training steps on Geometry3K. As

Table 7: Comprehensive performance comparison and ablation studies after integrating CSRS.

Model / Datasets	MathVision	MathVerse	MathVista	WeMath	Avg.
Part 1: Performance of Base Models and Trained Models in ***Geometry3k***					
Qwen2.5-VL-7B(Vanilla)	25.40	44.24	66.42	67.65	50.93
+CSRS	27.97	46.01	67.81	71.77	**53.39**
InternVL3-8B(Vanilla)	29.30	45.12	71.60	73.25	54.82
+CSRS	30.58	47.23	72.98	75.15	**56.49**
ThinkLite-VL-7B(Vanilla)	25.92	46.01	68.12	67.12	51.79
+CSRS	26.98	48.25	69.76	69.32	**53.58**
Part 2: Performance of Base Models and Trained Models in ***GeoQA***					
Qwen2.5-VL-7B+CSRS	28.95	45.82	68.25	69.32	53.09
InternVL3-8B+CSRS	31.82	47.50	73.40	75.80	**57.13**
ThinkLite-VL-7B+CSRS	28.15	48.35	70.22	69.45	**54.04**
Part 3: Performance of Base Models and Trained Models in ***MMR1***					
Qwen2.5-VL-7B+CSRS	27.86	45.89	67.81	70.53	53.05
InternVL3-8B+CSRS	30.95	46.42	72.15	75.20	**56.18**
ThinkLite-VL-7B+CSRS	27.92	47.95	69.80	70.15	**53.96**

Table 8: Performance across different training steps.

Steps	M-Vis.	M-Ver.	M-Vist.	WeMath	Avg.
120	27.97	46.01	67.81	71.77	53.39
125	27.52	47.24	68.06	71.64	53.62
130	28.03	46.18	67.25	70.93	53.10
135	28.01	46.75	67.99	71.62	53.59
140	28.25	47.34	67.63	70.89	53.53

Table 9: Ablation of sampling strategies.

Method	M-Vis.	M-Ver.	M-Vist.	WeMath
CSRS (Ran.)	27.63	46.35	67.52	72.01
CSRS (Fix.)	27.97	46.01	67.81	71.77

Table 10: Efficiency and Accuracy Comparison.

Method	Relative Training Time	Acc.
CSRS Fixed	1.0×	27.97
CSRS Random	1.13×	27.63

illustrated in Table 8, CSRS maintains steady performance without the significant degradation or label contamination. This proves our framework's robustness in maintaining label purity and training stability during long-term self-evolution.

A.5 Random Re-inference VS Fixed Re-inference

As our RRM uses a fixed-anchor re-inference approach, we conduct comparative experiments on the Geometry3K dataset to verify the effectiveness of this mechanism, comparing fixed anchors with random anchor re-inference. For the stochastic retracing variant, the retracing rate ω is sampled from a uniform distribution, such that $\omega \sim \mathcal{U}(0, 1)$ at each step.

The results in Table 9 and Table 10 indicate that while accuracy remains similar, the stochastic retracing strategy results in a 1.13× computational overhead compared to our fixed-rate approach. This efficiency gap stems from two main factors. First, when the retracing rate ω is randomly chosen from the range $[0, 1]$ and happens to be small, the model is forced to re-generate the entire problem description, which wastes time and computing resources. Second, since GRPO requires all samples in a group to finish before the next step, the large variation in response lengths caused by a random ω creates a bottleneck: the entire process must wait for the longest redundant response to complete. Furthermore, in a self-evolution setting without explicit ground-truth labels, a stable training strategy is crucial to prevent the model from falling into local optima or training collapse, allowing it to reach proficiency more rapidly. In contrast, CSRS uses a fixed anchor and significantly improves training speed, which keeps the generation length consistent and stable.

A.6 Case Study

In this section, we demonstrate how CSRS operates on mathematical problems during the training process. As illustrated in following boxes, within the initial set of maternal trajectories generated by

the model, the majority of reasoning paths exhibit typical errors such as relying on superficial mathematical cues, convoluted logical reasoning, and incorrect initial directions. In this instance, only a single correct reasoning path exists. Under a conventional majority voting scheme, the model would inevitably evolve toward its own distributional bias (e.g., the incorrect answer "47"). However, with the application of CSRS, a higher frequency of correct reasoning paths emerges within the re-inference set. Furthermore, correct paths with low frequency are no longer penalized with a zero reward, thereby preserving valuable learning signals.

1. Maternal Rollouts for Problem

Maternal Response 1: To find angle BDA, we note that the inscribed angle BDC intercepts arc CB (94°), so $\angle BDC = 47^\circ$. Since $AB \parallel DC$, $\angle BDA = \angle BDC$. Therefore, $\angle BDA = 47^\circ$.
Maternal Response 2: Using cyclic quadrilateral properties, $\angle BCA$ intercepts arc CB, so $\angle BCA = 47^\circ$. Opposite angles in a cyclic quadrilateral sum to 180°, so $\angle BDA = 180^\circ - 47^\circ = 133^\circ$.
Maternal Response 3: $\angle BDC$ intercepts arc CB, giving $\angle BDC = 47^\circ$. Since $AB \parallel DC$, $\angle BDA = \angle BDC$. Therefore, $\angle BDA = 47^\circ$.
Maternal Response 4: $\angle BDC = \frac{1}{2} \cdot \widehat{CB} = 47^\circ$. Due to the parallel lines, $\angle BDA = \angle BDC$. Thus, $\angle BDA = 47^\circ$.
Maternal Response 5: $\angle BDA$ is an inscribed angle intercepting the same arc as central angle BZA. So $\angle BDA = \frac{1}{2} \cdot \angle BZA = \frac{1}{2} \times 104^\circ = 52^\circ$.
Maternal Response 6: Since $AB \parallel DC$, $\angle BAC = \angle BDC = \frac{1}{2} \cdot \widehat{CB} = 47^\circ$. And $\angle BDA = \angle BAC$, so $\angle BDA = 47^\circ$.
Maternal Response 7: In the cyclic quadrilateral, $\angle BCA$ intercepts arc CB, so $\angle BCA = 47^\circ$. The opposite angle sum gives $\angle BDA = 180^\circ - 47^\circ = 133^\circ$.
Maternal Response 8: We find arc BA from central angle BZA: $\widehat{BA} = 2 \times 104^\circ = 208^\circ$. $\angle C$ intercepts arc CB: $\angle C = 47^\circ$. Using opposite angle sum in the cyclic quadrilateral, $\angle A = 133^\circ$. Due to parallel lines, $\angle BDA = \angle C = 47^\circ$.

2. Cut rollouts for Problem

Re-inference Response 9: The central angle BZA is given as 104°. The inscribed angle BDA subtends the same arc BA. By the inscribed angle theorem, $\angle BDA = \frac{1}{2} \cdot \angle BZA = 52^\circ$. The parallel lines condition is not needed for this calculation.
Re-inference Response 15: The reasoning is as follows: The central angle BZA is 104°. Since the inscribed angle BDA subtends the same arc BA, its measure is half of the central angle, which is 52°. The information about parallel lines is not used.
Re-inference Response 18: $\angle BDA$ subtends arc BA, and central angle BZA subtends the same arc. By the inscribed angle theorem, $\angle BDA = \frac{1}{2} \times 104^\circ = 52^\circ$.
Re-inference Response 35: We use the inscribed angle theorem: The measure of an inscribed angle is half the measure of the central angle that subtends the same arc. Therefore, $\angle BDA = \frac{1}{2} \times 104^\circ = 52^\circ$.
Re-inference Response 39: The measure of $\angle BDA$ is determined by the central angle BZA (104°) since they subtend the same arc BA. Applying the inscribed angle theorem gives $\angle BDA = 52^\circ$. The parallel condition is not necessary for this solution.

The box3 and box4 illustrate the specific formatting of prompts and responses are used during the training phase. Notably, the entire reasoning trajectory is encapsulated within <think></think> tags to facilitate structured cognitive processing, while the final conclusion is presented within a \boxed{} environment for automated extraction.

3. Maternal Rollout Example

prompt:"system You are a helpful assistant. user For the pair of similar figures, use the given areas to find x. You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \\boxed.assistant"
output:"<think> To find the value of x , we need to use the properties of similar figures. The ratio of the areas of two similar figures is equal to the square of the ratio of their corresponding side lengths.1. Let the ratio of the corresponding side lengths be k. 2. The area of the smaller figure is 50 in^2. 3. The area of the larger figure is 72 in^2.4. The ratio of the areas is \frac{50}{72} .5. Since the areas are proportional to the square of the side lengths, we have k^2 = \frac{50}{72} . 6. Solving for k, we get k = \sqrt{\frac{50}{72}} = \sqrt{\frac{25}{36}} = \frac{5}{6} . 7. The side length x corresponds to the smaller figure, and the side length of the larger figure is 6 inches. 8. Therefore, x = 6 times \frac{5}{6} = 5 . So, the value of x is 5 inches.</think> \\boxed{5}"

4. Retracing Rollout Example

prompt:"system You are a helpful assistant. user Find x. Assume that segments that appear to be tangent are tangent. You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \\boxed{}. assistant <think> To find the value of x, we need to use the property of tangent segments to a circle.assistant"
output:"<think>The tangent segments from an external point to a circle are equal in length.In the given diagram, the segment that appears to be tangent from the external point to the circle is 14, and the segment that appears to be tangent to the circle from the same external point is x .Since the segments are tangent to the circle from the same external point, they are equal in length. Therefore, we can set up the equation: x = 14 So, the value of x is 14.</think> \\boxed{14}"

As a representative error analysis, box5 illustrates how logical bias propagates even after re-inference: In our observations, when re-inference is initiated from a retracing anchor point, the resulting trajectories exhibit two distinct behaviors: logical persistence and strategic reflection. If the initial reasoning direction is flawed, persistent trajectories often lead to consistent but incorrect conclusions. Conversely, the reflective paths demonstrate a self-correction capability, where the model deviates from the original erroneous logic to discover the valid solution.

5. Bad Cases

Concatenated Prompt: You are a helpful assistant. user. Find JL, if $KP = 4$ and $PM = 7$. You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}. In the diagram, we have a kite shape with the diagonals as the main focus. The diagonals of a kite are perpendicular to each other and bisect each other at right angles. Here, diagonal KS is perpendicular to diagonal JL at point P.

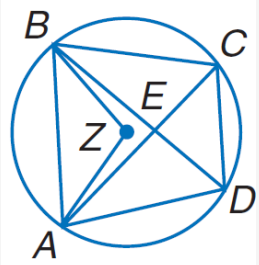
Correct Response (Maternal):
The diagonals of the kite/rhombus intersect at point P and bisect each other. Since $KP = 4$ and $PM = 7$, the total length of diagonal JL is the sum of these segments: $JL = KP + PM = 4 + 7 = 11$. $\boxed{11}$

Incorrect Response 1:
Treating the shape as a kite where the diagonals are perpendicular and bisect each other, one might incorrectly assume PM is half of JL, leading to $JL = 2 \times PM = 14$. $\boxed{14}$

Incorrect Response 2:
Assuming the figure is a rhombus, the side length JL might be calculated as the hypotenuse of the right triangle formed by the half-diagonals: $JL = \sqrt{KP^2 + PM^2} = \sqrt{65}$. $\boxed{\sqrt{65}}$

Incorrect Response 3:
Interpreting the figure as a parallelogram where diagonals bisect each other, one might incorrectly sum the segments and double them: $JL = 2 \times (KP + PM) = 22$. $\boxed{22}$

Incorrect Response 4:
Assuming symmetry incorrectly leads to $JL = KP = 4$, then the full diagonal length is doubled: $JL = 2 \times 4 = 8$. $\boxed{8}$



Question:Quadrilateral $A B C D$ is inscribed in $\odot Z$ such that $m \angle B Z A=104, m \widehat{C B}=94,$ and $\overline{A B} \| \overline{D C} .$ Find $m \angle B D A$

Answer:52

Maternal Response 1: To find angle BDA, we note that the inscribed angle BDC intercepts arc CB (94°), so angle BDC = 47°. Since AB is parallel to DC, angle BDA equals angle BDC. Therefore, angle BDA = 47°.
Maternal Response 2: Using cyclic quadrilateral properties, angle BCA intercepts arc CB, so angle BCA = 47°. Opposite angles in a cyclic quadrilateral sum to 180°, so angle BDA = 180° - 47° = 133°.
Maternal Response 3: Angle BDC intercepts arc CB, giving angle BDC = 47°. Since AB is parallel to DC, angle BDA equals angle BDC. Therefore, angle BDA = 47°.
Maternal Response 4: Angle BDC = (1/2) * arc CB = 47°. Due to the parallel lines, angle BDA = angle BDC. Thus, angle BDA = 47°.
Maternal Response 5: Angle BDA is an inscribed angle intercepting the same arc as central angle BZA. So angle BDA = (1/2) * angle BZA = (1/2)*104° = 52°.
Maternal Response 6: Since AB is parallel to DC, angle BAC = angle BDC = (1/2)*arc CB = 47°. And angle BDA = angle BAC, so angle BDA = 47°. 133Maternal Response 8: We find arc BA from central angle BZA: arc BA = 2*104° = 208°. Angle C intercepts arc CB: angle C = 47°. Using opposite angle sum in the cyclic quadrilateral, angle A = 133°. Due to parallel lines, angle BDA = angle C = 47°. 47
Maternal Response 7: In the cyclic quadrilateral, angle BCA intercepts arc CB, so angle BCA = 47°. The opposite angle sum gives angle BDA = 180° - 47° = 133°
Maternal Response 8: We find arc BA from central angle BZA: arc BA = 2*104° = 208°. Angle C intercepts arc CB: angle C = 47°. Using opposite angle sum in the cyclic quadrilateral, angle A = 133°. Due to parallel lines, angle BDA = angle C = 47°.

Low-Level Image Information Mistake: Response3 .etc
Math Relationship Mapping Mistake: Response8 .etc
Proof Direction Mistake from Start: Response6 .etc

Reinference Response 9: The central angle BZA is given as 104°. The inscribed angle BDA subtends the same arc BA. By the inscribed angle theorem, angle BDA = (1/2) * angle BZA = 52°. The parallel lines condition is not needed for this calculation.
Reinference Response 15: The reasoning is as follows: The central angle BZA is 104°. Since the inscribed angle BDA subtends the same arc BA, its measure is half of the central angle, which is 52°. The information about parallel lines is not used.
Reinference Response 18: Angle BDA subtends arc BA, and central angle BZA subtends the same arc. By the inscribed angle theorem, angle BDA = (1/2)*104° = 52°.
Reinference Response 35: We use the inscribed angle theorem: The measure of an inscribed angle is half the measure of the central angle that subtends the same arc. Therefore, angle BDA = (1/2) * 104° = 52°.
Reinference Response 39: The measure of angle BDA is determined by the central angle BZA (104°) since they subtend the same arc BA. Applying the inscribed angle theorem gives angle BDA = 52°. The parallel condition is not necessary for this solution

Frequency in Maternal Answers:0.125 !
Frequency in Reinference Answers:0.2 !

Figure 6: Case studies in geometric reasoning.