V-Zero: Self-Improving Multimodal Reasoning with Zero Annotation

Anonymous ACL submission

Abstract

Recent advances in multimodal learning have significantly enhanced the reasoning capabilities of vision-language models (VLMs). However, state-of-the-art approaches rely heavily on large-scale human-annotated datasets, which are costly and time-consuming to acquire. To overcome this limitation, we introduce **V-Zero**, a general post-training framework that facilitates self-improvement using exclusively unlabeled images. V-Zero establishes a co-evolutionary loop by instantiating two distinct roles: a **Questioner** and a **Solver**. The Questioner learns to synthesize high-quality, challenging questions by leveraging a dual-track reasoning reward that contrasts intuitive guesses with reasoned results. The Solver is optimized using pseudo-labels derived from majority voting over its own sampled responses. Both roles are trained iteratively via Group Relative Policy Optimization (GRPO), driving a cycle of mutual enhancement. Remarkably, without a single human annotation, V-Zero achieves consistent performance gains on Qwen2.5-VL-7B-Instruct, improving visual mathematical reasoning by +1.7 and general vision-centric by +2.6, demonstrating the potential of self-improvement in multimodal systems.

1 Introduction

Vision-language models (VLMs) have become increasingly powerful at understanding images and solving complex reasoning problems. However, the performance of these models remains heavily dependent on extensive human annotation (Bordes et al., 2024), such as visual question-answering (VQA) datasets containing fine-grained chain-of-thought (CoT) (Wei et al., 2022) reasoning steps. Such high-quality reasoning annotations are not only costly and labor-intensive to acquire, but also typically designed for specific tasks or domains, resulting in limited generalization. In response, self-improvement methods have emerged as a compelling alternative, allowing models to iteratively enhance their capabilities using internal feedback rather than human guidance.

Self-improvement methods for large language models (LLMs) have demonstrated that models can autonomously generate training data and learning signals through self-play and role specialization interaction (Kuba et al., 2025; Dong and Ma, 2025; Liu et al., 2025a; Huang et al., 2025a). However, research on self-play for VLMs remains highly limited. Vision-Zero (Wang et al., 2025a) adapts the idea to VLMs by conducting self-play in games. However, it employs external expert models to edit and generate images during data construction. An important open question is whether VLM self-improvement can be achieved by operating directly on raw, unlabeled images without requiring annotation from humans or external models.

To address this, we propose **V-Zero**, a general post-training framework that facilitates self-improvement using exclusively unlabeled images. As shown in the Figure 1, supervised Group Relative Policy Optimization (GRPO) (Shao et al., 2024) relies on static datasets consisting of human-annotated questions and ground-truth answers for supervised learning. However, V-Zero introduces a dynamic paradigm without annotated data by establishing a co-evolutionary loop between two roles initialized from the same base model: a Questioner and a Solver. Both are trained with GRPO using internally generated reward signals. The Questioner learns to perceive an input image and generate multiple-choice questions that target the boundary of the Solver's current capabilities. Conversely, the Solver learns to answer these questions accurately. The Questioner maximizes a dual-track reasoning reward derived from the Solver's performance. By contrasting the model's initial intuition with its reasoned outputs, we encourage the Questioner to generate more challenging questions that cannot be solved by immediate guesses but require reflection and deep thinking. The Solver utilizes a binary

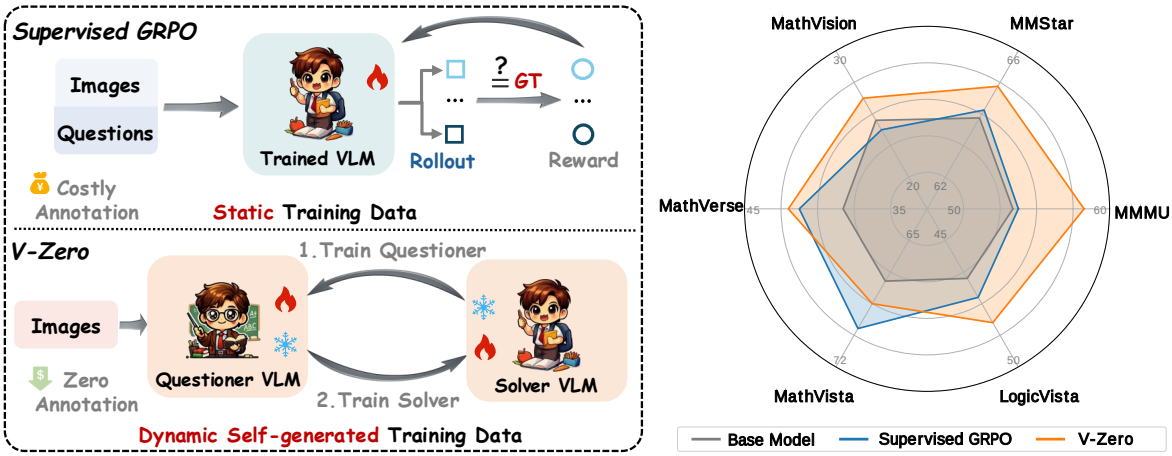

Figure 1: **Left:** Comparison of training paradigms. Supervised GRPO (top) relies on static datasets where rewards are computed by comparing outputs against external ground-truth labels. In contrast, V-Zero (bottom) achieves self-improvement from unlabeled images via a dynamic co-evolutionary loop, where reward signals are internally generated through the interaction between a Questioner and a Solver. **Right:** Performance on Qwen2.5-VL-7B-Instruct. V-Zero demonstrates broad improvements across diverse benchmarks, notably outperforming both the Base Model and the strong Supervised GRPO baseline trained on human-labeled data.

accuracy reward within a reinforcement learning with verifiable rewards (RLVR) process (Guo et al., 2025), where pseudo-labels are determined via majority voting over sampled responses. Through this zero-annotation process, the Questioner and Solver progressively refine the model's visual reasoning faculties. Our main contributions are as follows:

- We propose V-Zero, a general post-training framework that operates exclusively on image-only data without any annotation.
- We introduce a Dual-Track Reasoning Reward, which explicitly contrasts intuition and reasoning, to encourage the model to generate challenging, reasoning-intensive questions.
- Extensive experiments demonstrate that V-Zero substantially improves model performance across a range of tasks, even surpasses strong baselines trained with human-annotated supervision.

2 Related Work

2.1 Self-improvement of Large Models

Self-improvement enhances model capabilities through iterative interactions in game-based settings and problem-solving tasks.

In game-based settings, models act as participants and achieve self-improvement by training in games and receiving feedback according to explicit rules. SPC (Chen et al., 2025) frames this idea as an adversarial game between a sneaky generator injecting erroneous reasoning steps and a critic that must detect them, enabling step-level reliability learning without manual process annotation. SPIRAL (Liu et al., 2025a) further scales competitive self-improvement to multi-turn, zero-sum games by using the same model to play both roles and optimizing it through role-conditioned advantages. In the multimodal field, Vision-Zero (Wang et al., 2025a) adapts self-improvement to VLMs by building a "Who Is the Spy" game on paired original and expert-edited images, where agents exchange clues and are trained with verifiable voting outcomes.

In problem-solving scenarios, self-improvement follows a paired problem-generate-and-respond interaction. Absolute Zero (Zhao et al., 2025a) realizes this interaction in coding-style reasoning by using a code executor to validate both generated problems and candidate solutions. STP (Dong and Ma, 2025) transfers the same principle to theorem proving by pairing a conjecturer with a prover, where conjectures are iteratively generated and filtered to form a self-curated curriculum. Beyond single-domain settings, SPICE (Liu et al., 2025b) broadens self-improvement to general tasks in corpus environments, where a challenger derives document-grounded questions and a reasoner answers them, producing an evolving curriculum through interactions. R-Zero (Huang et al., 2025a) further generalizes this setup by jointly improving a challenger that pushes question difficulty and a solver that learns from its own sampled solutions.

However, current visual self-improvement remains largely constrained to the game-based sce-

nario, which requires careful human design and time-consuming data preparation. To overcome these limitations, we propose a truly zero-human-involvement training scheme for VLMs that removes multimodal data preparation and runs a self-improving Questioner-Solver loop directly on raw, unlabeled images.

2.2 Multimodal Reasoning

Recent progress in advancing multimodal reasoning lies in exploring better reasoning trajectories and improving training signal (Liu et al., 2025c).

Exploring better reasoning trajectories improves multimodal reasoning by retaining chains that best satisfy visual evidence and intermediate constraints (Yao et al., 2024; Sun et al., 2025). MM-Verify (Sun et al., 2025) builds an MM-Verifier to evaluate long CoT solutions, using verification to strengthen long-horizon multimodal reasoning, while GenPRM (Zhao et al., 2025b) proposes a generative PRM that performs explicit reasoning to judge each step, enabling scalable test-time computation. Beyond verifier-guided selection, SoTA with Less (Wang et al., 2025b) uses MCTS-guided sample selection to prioritize high-leverage supervision for data-efficient self-improvement, and Mulberry (Yao et al., 2024) extends collective MCTS to elicit reasoning and reflection trajectories with minimal reliance on humans.

Improving training signals aims to strengthen multimodal reasoning by converting verifiable intermediate constraints into explicit rewards for optimization (Deng et al., 2025; Huang et al., 2025b). Vision-R1 (Huang et al., 2025b) applies GRPO with a rule-based format and result rewards, and leverages a cold-start multimodal CoT initialization to make such simple signals effective. R1-Onevision (Yang et al., 2025) combines rule-based answer checking with strict reasoning-format constraints, thereby enhancing complex problem-solving. To address the sparsity of outcome-only supervision, R1-VL (Zhang et al., 2025) introduces dense step-wise rewards that credit matched key intermediate steps and encourage complete, logically consistent trajectories. Beyond language-centric reasoning rewards, Perception-R1 (Yu et al., 2025) designs task-specific rewards for perception outputs and performs multi-subject reward matching before scoring, providing reliable training signals for learning perception policies.

Compared with these approaches relying on large-scale annotated data, our method eliminates data annotation by training directly on unlabeled images, yet outperforms supervised training.

3 Method

3.1 Overview

We propose **V-Zero**, a general post-training framework for VLMs that operates exclusively on raw image data without any annotations, as illustrated in Figure 2. The framework establishes a co-evolutionary loop between two roles initialized from the same base model: a Questioner Q_θ and a Solver S_θ. The Questioner perceives input images to generate challenging, high-quality questions, optimized via GRPO using feedback signals from the Solver. Correspondingly, the Solver learns to answer these questions via GRPO using a difficulty-guided sampled dataset. Through iterative cycles, Q_θ and S_θ continuously co-evolve in a self-supervised manner, progressively enhancing visual perception and reasoning capabilities.

3.2 Questioner

Questioner model Q_θ learns to understand input images from an image-only dataset and generate high-quality questions that are maximally challenging for the Solver.

Image-based Question Generation Given an image input I, Q_θ fully perceives the image and extracts useful visual information, such as side lengths of geometric shapes, entries in charts, spatial relationships between objects, and other salient visual features. Q_θ then generates a concise visual description d of I, a multiple-choice question (MCQ) q, and a reference answer a_{fast} indicating the correct option for the question.

$$(q, a_{fast}, d) \sim Q_\theta(\cdot|I) \tag{1}$$

Here, a_{fast} represents the model's "direct answer" or intuition, as it is generated simultaneously with the question without extensive thinking. We specifically adopt the MCQ format for its structural advantages: it constrains the output to a fixed set of options, facilitating robust answer extraction and enabling precise, automated evaluation of the Solver's accuracy without the ambiguity of free-form generation.

Dual-Track Reasoning Reward A widely accepted principle is that, for easy problems, answers derived from intuition and careful reasoning tend to coincide; in contrast, a defining characteristic of

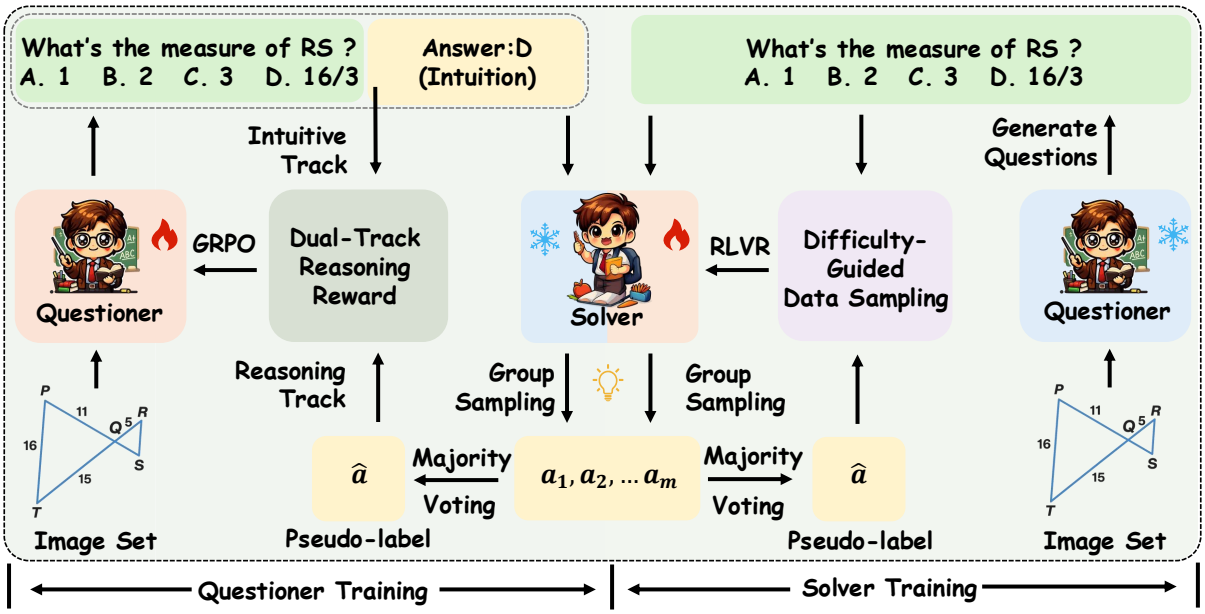

Figure 2: **Overview of V-Zero.** Our framework drives self-improvement through a co-evolutionary loop. **Left (Questioner Training):** The Questioner generates a question and an intuitive answer. A frozen Solver then samples reasoning responses to produce a pseudo-label via majority voting. The Questioner is updated via GRPO using a Dual-Track Reasoning Reward derived from the contrast between intuition and reasoning. **Right (Solver Training):** The optimized Questioner generates questions, which are paired with pseudo-labels from the Solver itself. This data undergoes difficulty-guided data sampling to filter for quality, and the Solver is then updated via RLVR.

challenging problems is that intuitive responses are often misleading, whereas correct solutions can be reached through step-by-step reasoning. Inspired by this observation, we implement a dual-track reasoning reward to encourage the Questioner Q_θ to generate such questions. We contrast two distinct answering modes:

- **Intuitive Track (a_{fast}):** The direct answer generated by the Questioner simultaneously with the question, representing the model's intuition.
- **Reasoning Track ($\hat{a}$):** The pseudo-label derived from the Solver S_θ. The Solver employs CoT reasoning to sample m responses $\{a_1, \ldots, a_m\}$. The majority vote determines $\hat{a}$, with confidence c defined as the proportion of votes matching $\hat{a}$.

The dual-track reasoning reward r_d is adaptive based on the alignment between the intuitive track and the reasoning track:

$$r_d(q) = \begin{cases} \min(c, 1-c), & \text{if } \hat{a} = a_{fast}, \\ 0.5 \cdot c, & \text{otherwise.} \end{cases} \quad (2)$$

- **Consistency Case ($\hat{a} = a_{fast}$):** When intuition aligns with reasoning, the problem is likely straightforward. In this case, we adopt an uncertainty reward. If the confidence c is high (near 1.0), it implies the question is overly simple; if c is low (near 0), it implies the question is overly ambiguous or noisy. The reward $\min(c, 1-c)$ peaks at $c = 0.5$, effectively guiding the Questioner to locate the Solver's decision boundary, targeting questions that are neither too simple nor completely ambiguous but maximally confusing for the current model.
- **Divergence Case ($\hat{a} \neq a_{fast}$):** When the reasoned answer corrects the intuitive guess, it implies the question serves as a valuable learning signal. Here, the reward is proportional to the confidence c, encouraging the generation of challenging questions that appear simple at first glance but require the Solver to engage in reflection and deep reasoning to obtain the correct answer with high confidence.
- **Prevention of Reward Hacking:** Our design naturally prevents reward hacking. Nonsensical questions lead to inconsistent reasoning (low c), yielding low reward in divergence case. Conversely, generating trivial questions to boost c causes answers to converge ($\hat{a} = a_{fast}$), triggering the uncertainty reward where high confidence yields near-zero reward. Thus, high rewards are strictly reserved for valid, challenging questions.

Final Reward with Formatting Constraints To ensure structural validity, we enforce strict formatting constraints. A generation pair (q, a_{fast})

is valid only if enclosed within distinct <*question*> and <*answer*> tags, and q is formatted as a multiple-choice question with four options. The final reward for the Questioner is:

$$r_Q = \begin{cases} r_d(q), & \text{if format is valid,} \\ -1, & \text{otherwise.} \end{cases} \quad (3)$$

3.3 Solver

Solver model S_θ learns to answer the questions proposed by Q_θ. In this phase, we focus on curating a high-quality training dataset through difficulty-guided data sampling and optimizing the model's reasoning capabilities via RLVR.

Difficulty-Guided Data Sampling Given a set of n unlabeled images $\{I_i\}_{i=1}^n$, the Questioner Q_θ first generates a corresponding multiple-choice question q_i for each image. The Solver S_θ then samples a set of m candidate answers:

$$\{a_{i,1}, a_{i,2}, \ldots, a_{i,m}\} \sim S_\theta(\cdot \mid (I_i, q_i)) \quad (4)$$

From these samples, a pseudo-label $\hat{a}_i$ is derived via majority voting, with the vote proportion recorded as a difficulty score s_i. To curate a training set that effectively drives improvement, we implement a difficulty-guided sampling strategy, keeping only pairs $(q_i, \hat{a}_i)$ where the score falls between $0.3 \leq s_i \leq 0.8$. This rule filters out samples that are either too ambiguous (low s_i, suggesting potential errors) or too simple (high s_i, offering little learning value). The final dataset D is thus constructed by selecting questions that are challenging yet solvable for the current model:

$$D = \big\{(I_i, q_i, \hat{a}_i, s_i) \;\big|\; i \in \{1, \ldots, n\}, \; 0.3 \leq s_i \leq 0.8\big\} \quad (5)$$

Binary Accuracy Reward The Solver S_θ is optimized to maximize answer accuracy on the constructed dataset D. Specifically, for each question q_i paired with its pseudo-label $\hat{a}_i$, the Solver samples a new group of candidate answers $\{a_{i,j}\}_{j=1}^G$. The reward r_j assigned to each candidate is defined as a binary value:

$$r_j = \begin{cases} 1, & \text{if } a_{i,j} = \hat{a}_i, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

The binary accuracy rewards are subsequently utilized to compute the group relative advantage $\hat{A}_{i,j}$. Specifically, the rewards within each group are standardized using z-score normalization:

$$\hat{A}_{i,j} = \frac{r_j - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\}) + \epsilon_{\text{norm}}}. \quad (7)$$

Finally, the policy is updated by minimizing the GRPO loss, which combines a clipped surrogate objective with a KL-divergence penalty:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{j=1}^{G} \min\Bigg(\frac{\pi_\theta(a_{i,j})}{\pi_{\text{old}}(a_{i,j})}\hat{A}_{i,j}, \; \text{clip}\Big(\frac{\pi_\theta(a_{i,j})}{\pi_{\text{old}}(a_{i,j})}, 1-\epsilon, 1+\epsilon\Big)\hat{A}_{i,j}\Bigg) + \beta \text{KL}(\pi_\theta || \pi_{\text{old}}). \quad (8)$$

4 Experiments

4.1 Settings

Models & Datasets We conduct experiments on two state-of-the-art VLM backbones: the widely adopted Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025b). For training data, we use images from the OpenVLThinker GRPO-medium and GRPO-hard dataset (Deng et al., 2025), comprising approximately 9K images covering multiple task types, primarily geometry with around 6K images, tables and spatial reasoning with around 3K images in total. After an initial filtering of the dataset, we train V-Zero exclusively on the image data. More details about image dataset can be found in Appendix A.1.

Baselines We compare our method against the base model and a model trained with two epochs of supervised GRPO on the same image dataset but fully human-annotated. We also include OpenVLThinker-7B as a stronger baseline, which is post-trained via supervised fine-tuning (SFT) and GRPO using human-annotated data.

Evaluation Benchmarks We evaluate V-Zero using VLMEvalKit (Duan et al., 2024) across benchmarks in two multimodal domains: general vision-centric tasks and visual mathematical reasoning. For general vision-centric evaluation, we employ MMMU (Yue et al., 2024) and MMStar (Chen et al., 2024). MMMU evaluates college-level subject knowledge and reasoning with 11.5K multimodal questions across six disciplines. MMStar tests core VLM capabilities on vision-indispensable tasks through 1,500

Methods	General Vision-Centric		Mathematical Reasoning				Avg.
	MMMU	MMStar	MathVision	MathVerse	MathVista	LogicVista	
Performance on Qwen2.5-VL-7B-Instruct							
Base Model	54.7	63.9	25.6	39.6	68.2	47.2	49.9
Supervised GRPO	55.0	64.5	25.0	42.0	70.3	47.8	50.8
OpenVLThinker-7B	53.1	59.6	**27.6**	36.8	**70.8**	45.1	48.8
V-Zero (Iter 1)	55.2	**65.7**	26.3	**43.9**	68.6	47.4	51.2
V-Zero (Iter 2)	**58.6**	65.1	27.0	42.6	69.2	**48.6**	**51.9**
Performance on Qwen2.5-VL-3B-Instruct							
Base Model	45.6	56.1	23.0	32.1	62.2	40.4	43.2
V-Zero (Iter 1)	**47.3**	54.7	**25.3**	**33.4**	**62.6**	40.2	**43.9**
V-Zero (Iter 2)	46.7	**56.4**	23.7	32.6	62.0	**41.6**	43.8

Table 1: Performance comparison on general vision-centric and mathematical reasoning benchmarks, evaluated on VLMEvalKit. All results are obtained under same settings. We report results of Solver for both Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-3B-Instruct. Best results are highlighted in **bold**.

meticulously curated samples. For mathematical reasoning, we employ MathVision (Wang et al., 2024), MathVerse (Zhang et al., 2024), MathVista (Lu et al., 2024) and LogicVista (Xiao et al., 2024). MathVision is a comprehensive benchmark comprising 3,040 high-quality visual mathematics problems sourced from real competitions. We use the Test-mini split consisting of 304 samples. MathVerse is a robust visual math benchmark containing 2,612 multi-subject problems. We report results on the Vision-Only split, as it presents the most significant challenge for the models. MathVista integrates 6,141 examples from 31 diverse datasets to evaluate fine-grained visual understanding and compositional reasoning in mathematical contexts. We use the Test-mini split, consisting of approximately 1,000 samples. LogicVista evaluates fundamental logical reasoning in visual contexts through 448 multiple-choice questions.

Training Settings Our implementation is based on the verl framework (Sheng et al., 2024). Both the Questioner and Solver are optimized via GRPO on 4 NVIDIA A800 (80GB) GPUs with a global batch size of 64, a learning rate of 1×10^{-6}, and a sampling temperature of 1.0. For the Questioner, we dedicate 2 additional GPUs to host the feedback Solver and set the group size $G = 4$. For Solver training, we set $G = 5$. Additionally, we set the sample size for majority voting to $m = 10$. More details about training settings can be found in Appendix A.2.

4.2 Main Results

Overall Performance Improvements. Table 1 presents the evaluation results of the Solver across diverse benchmarks. V-Zero consistently enhances the performance of the base models without external supervision. For Qwen2.5-VL-7B-Instruct, the iterative training yields a steady performance trajectory for the Solver, improving the average score from 49.9 to 51.9 (+2.0 points) in Iteration 2. The gains are particularly substantial on the general benchmark MMMU (+3.9 points) and the visual mathematics dataset MathVerse (+3.0 points), validating the framework's effectiveness. For the smaller Qwen2.5-VL-3B-Instruct, the model reaches its performance peak in the first iteration (+0.7 points), with notable gains on MMMU (+1.7 points) and MathVision (+2.3 points). We attribute this early saturation to the restricted model capacity and capability, which increases sensitivity to the noise in self-generated data. Furthermore, the requisite high-temperature sampling for exploration inherently introduces optimization variance, a trade-off that becomes more pronounced in smaller models with limited robustness.

Generalization Across Tasks. Our training image dataset is primarily composed of geometric figures aimed at enhancing mathematical reasoning capabilities, validated by the consistent improvements observed on mathematical reasoning benchmarks (+1.7 points on average). Importantly, these specialized gains do not compromise general capabilities. Instead, V-Zero demonstrates strong generalization, achieving even larger improvements on general vision-centric datasets (+2.6 points on average). This suggests that the rigorous visual perception required to solve complex geometric problems effectively transfers to broader domains, strengthening the Solver's fundamental visual understanding without overfitting to a specific domain.

Comparison with Supervised Baselines. V-Zero compares favorably against methods relying on ground-truth supervision. On the 7B scale, our unsupervised Solver (51.9 avg.) outperforms the Supervised GRPO baseline (50.8 avg.), which was trained using human-annotated questions and answers. This indicates that the diversity of questions generated by the Questioner may offer better exploration of the solution space than a fixed supervised dataset. Furthermore, compared to OpenVLThinker-7B, V-Zero remains highly competitive. While OpenVLThinker holds a marginal advantage in specific mathematical reasoning benchmarks like MathVista and MathVision, V-Zero achieves superior performance on LogicVista and maintains a more balanced profile across general vision-centric tasks.

4.3 Ablation Studies

To verify the effectiveness of the core components in V-Zero, we conduct comprehensive ablation studies on the Qwen2.5-VL-7B-Instruct model. We compare the full V-Zero (Iteration 2) against three variants: freezing the Questioner, removing the dual-track reasoning reward, and disabling difficulty-guided data sampling for Solver training. The results are summarized in Table 2.

Impact of Improving Questioner We first investigate the validity of the framework by freezing the Questioner at its initial state (base model) while only training the Solver. As shown in Table 2, this setting leads to a catastrophic performance drop. The Math Avg. falls to 45.4, showing negligible improvement over the Base Model (45.2). More critically, the General Avg. degrades to 58.7, which is even lower than the Base Model (59.3). This suggests that without the improvement of the Questioner, the generated questions remain static and lack sufficient challenge. Consequently, the Solver likely overfits to simple patterns or suffers from distribution shift, ultimately harming its general visual capabilities. This confirms that the dynamic co-evolution between the two roles is essential for continuous self-improvement.

Questioner Reward Design To assess the contribution of our proposed Dual-Track Reasoning Reward, we replace it with a single uncertainty reward (i.e., removing the reward of divergence case). The ablation results show a clear performance drop, with the Math Avg. dropping from 46.9 to 46.1 (-0.8 points) and the General Avg. decreasing from 61.9 to 61.3 (-0.6 points). This decline indicates that simply targeting high-uncertainty samples is insufficient. By explicitly rewarding questions where the model's intuition contrasts with its reasoning, the dual-track reasoning reward successfully encourages the Questioner to generate more challenging questions, thereby forcing the Solver to engage in deeper reasoning.

Methods	General Avg.	Math Avg.
Base Model	59.3	45.2
V-Zero (Iter 2)	**61.9**	**46.9**
w/o. Improving Questioner	58.7	45.4
w/o. Dual-Track Reward	61.3	46.1
w/o. Data Filtering	59.8	45.8

Table 2: Ablation studies on key components of V-Zero. We evaluate the impact of the improving Questioner, dual-track reasoning reward, and difficulty-guided data sampling (denoted as Data Filtering) on Qwen2.5-VL-7B-Instruct. Best results are highlighted in **bold**.

Impact of Data Filtering Finally, we examine the role of the difficulty-guided data sampling strategy (selecting samples with $0.3 \leq s \leq 0.8$). Removing this filter results in a significant performance decline, with the General Avg. dropping by 2.1 points and Math Avg. dropping by 1.1 points. This underscores the importance of data quality in the self-improvement process. Without data filtering, Solver training may be contaminated by two types of data: overly simple samples that contribute nothing to performance improvement, and noisy samples arising from ambiguous or unsolvable problems. The filtering mechanism ensures that the Solver learns exclusively from helpful data, thereby preserving training stability and efficiency.

4.4 Qualitative Analysis

We analyze the quality of the questions generated by the Questioner in the training process, focusing on two key metrics: format validity and question difficulty. The statistics are summarized in Table 3.

Format Validity As described in Section 3.2, we impose strict formatting constraints during Questioner training: a generated question sample is considered valid only if it presents a four-option multiple-choice question with correctly enclosed tags for the question and the intuitive answer. As shown in Table 3, the Base Model struggles with these structural requirements, achieving a validity rate of only 64.9%, which severely limits the uti-

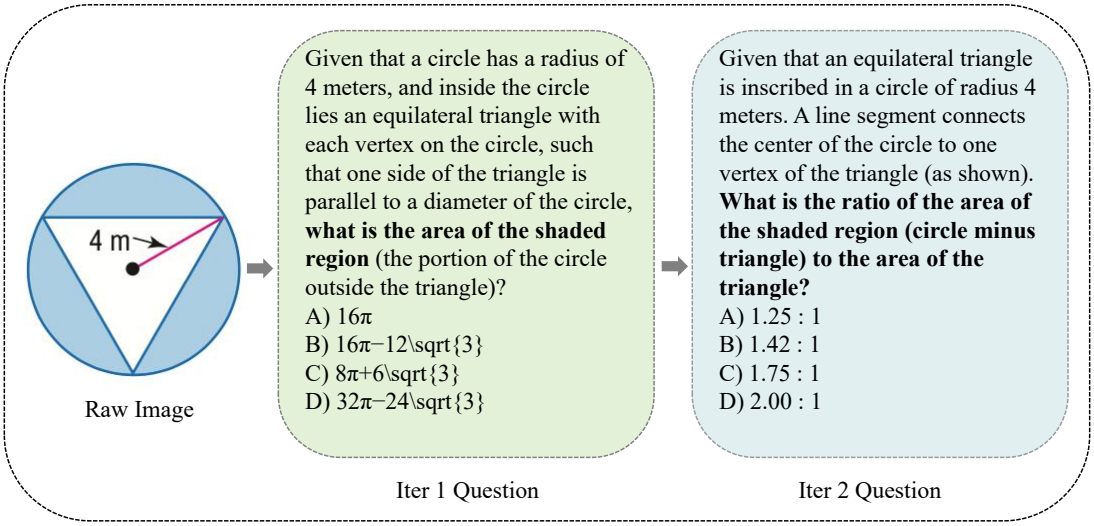

Figure 3: **Improvement of generated questions.** We visualize the questions generated by the Questioner for the same raw image (Left) across two iterations. In Iteration 1 (Middle), the model poses a straightforward area calculation problem. In Iteration 2 (Right), the question advances into a more complex ratio problem involving deeper geometric reasoning, demonstrating the progressive difficulty of the self-generated curriculum.

Questioner	Valid Rate	Difficulty
Base Model	64.9%	0.52
V-Zero (Iter 1)	99.1%	0.58
V-Zero (Iter 2)	**99.8%**	**0.60**

Table 3: Evolution of generated question quality. We report the percentage of valid questions and the average difficulty evaluated by Qwen3-VL-32B-Instruct. Best results are highlighted in **bold**.

lization of generated data. However, after just one iteration of V-Zero training, the validity rate surges to 99.1%. This near-perfect compliance demonstrates that the constraints effectively regularizes the model's output format, ensuring the production of high-quality, parsable training data.

Question Difficulty To evaluate the depth of the generated questions, we employ the advanced Qwen3-VL-32B-Instruct (Bai et al., 2025a) as a judge. We establish a scoring rule ranging from 0 to 1, categorizing questions into three levels:

- Simple (0.0-0.3): Basic object counting or simple chart reading.
- Medium (0.4-0.7): Application of geometric relations or chart trend analysis.
- Hard (0.8-1.0): Complex geometric proofs or multi-step algebraic reasoning.

We calculate the average difficulty score for all valid questions. The quantitative results indicate a clear upward trend, rising from 0.52 (Base) to 0.60 (Iter 2). This progression is shown in Figure 3: while the model in Iteration 1 poses a straightforward area calculation problem, it evolves in Iteration 2 to construct a significantly more complex ratio problem requiring multi-step derivation. This confirms that the Questioner learns to generate harder questions over time, thereby providing a curriculum of growing difficulty that continuously challenges the Solver.

5 Conclusion

In this work, we introduced V-Zero, a general framework enabling the self-improvement of VLMs utilizing exclusively unlabeled images. By establishing a dynamic co-evolutionary loop between a Questioner and a Solver, V-Zero effectively bypasses the bottleneck of costly human annotations. In this process, the Questioner learns to generate increasingly challenging questions by targeting the gap between intuition and reasoning, while the Solver iteratively improves by resolving these generated questions. Both roles are optimized via GRPO. Extensive experiments demonstrate that this paradigm not only yields significant gains in visual mathematical reasoning but also generalizes well to broader vision-centric tasks. These results suggest that constructing general, internal feedback loops is a viable path toward more general multimodal capabilities without reliance on costly human supervision.

6 Limitations

First, due to compute constraints, our experiments are limited to the Qwen2.5-VL series. Our future work will extend V-Zero to more model architectures. Second, experimental results indicate that the range of performance improvement is constrained by the model's intrinsic capacity, and that the optimization process exhibits fluctuations rather than steady growth across iterations. These fluctuations are partly due to the requirement to encourage high-degree exploration for unlocking the model's latent potential.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, and 1 others. 2024. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Jiaqi Chen, Bang Zhang, Ruotian Ma, Peisong Wang, Xiaodan Liang, Zhaopeng Tu, Xiaolong Li, and Kwan-Yee K Wong. 2025. SPC: evolving self-play critic via adversarial games for LLM reasoning. *arXiv:2504.19162*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*.

Kefan Dong and Tengyu Ma. 2025. STP: Self-play LLM theorem provers with iterative conjecturing and proving. In *Forty-second International Conference on Machine Learning*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

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

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. 2025a. R-zero: Self-evolving reasoning LLM from zero data. *arXiv:2508.05004*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025b. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv:2503.06749*.

Jakub Grudzien Kuba, Mengting Gu, Qi Ma, Yuandong Tian, and Vijai Mohan. 2025. Language self-play for data-free training. *arXiv:2509.07414*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, and 1 others. 2025a. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. *arXiv:2506.24119*.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. 2025b. SPICE: self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*.

Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. 2025c. Diving into self-evolving training for multimodal reasoning. In *Forty-second International Conference on Machine Learning*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

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. *arXiv preprint arXiv: 2409.19256*.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2025. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14100–14115. Association for Computational Linguistics.

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.

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. 2025a. Vision-zero: Scalable vlm self-improvement via strategic gamified self-play. *arXiv:2509.25541*.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025b. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv:2504.07934*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv:2503.10615*.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and 1 others. 2024. Mulberry: Empowering MLLM with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv:2412.18319*.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, and 1 others. 2025. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv:2504.07954*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-VL: learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv:2503.12937*.

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, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025a. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv:2505.03335*.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and 1 others. 2025b. Genprm: Scaling test-time compute of process reward models via generative reasoning. *arXiv:2504.00891*.

A Appendix

A.1 Dataset

The images we use are sourced from two datasets: OpenVLThinker-GRPO-hard[1], which contains approximately 6K geometric images, and OpenVLThinker-GRPO-medium[2], which comprises around 3K images of maps, charts, and object spatial relationships. From these, we select 4,000 images and train only on the raw, unlabeled image data.

A.2 Training Settings

V-Zero framework is based on **verl** codebase. We employ vLLM (Kwon et al., 2023) as the rollout engine during training, the inference temperature is set to 1.0 across all processes. The The maximum number of tokens for the response is set to 2048 for Questioner training and 4096 for Solver training. We use a KL divergence coefficient of 0.01 to constrain policy drift. To manage memory efficiency, we employ Fully Sharded Data Parallel (FSDP) with parameter offloading. A single V-Zero iteration requires about 9 hours to complete.

A.3 Evaluation Details

We conduct all evaluations using the **VLMEvalKit** codebase, a standardized open-source framework for VLM assessment. We use the default generation temperature 0.0 (greedy decoding) to eliminate randomness and maximize the determinism of the results. The maximum generation length of token is set to 2,048.

For the answer extraction and matching process (i.e., the "Judge" role), we employ Qwen3-VL-8B-Instruct. We selected this model as the evaluator because the task involves primarily straightforward answer parsing and option matching, for which a lightweight yet capable 8B model offers an optimal trade-off between accuracy and evaluation efficiency.

[1] https://huggingface.co/datasets/ydeng9/OpenVLThinker-grpo-hard

[2] https://huggingface.co/datasets/ydeng9/OpenVLThinker-grpo-medium

A.4 Prompts

Prompt for Question Generation

System Prompt:
You are a professional question designer.
User Prompt:
Create a multiple-choice question based on the image. Let's think step by step. First, you must fully perceive the image, extracting any valuable visual information from it and generate a detailed visual description of the image.
Then, write a multiple-choice question that includes necessary conditions.
- Make sure the question provides sufficient information to be answered. Use phrases like "If..." or "Given that..." to state condition shown in visual description if it's a geometry question.
- The question must include four options, one of which is the correct answer.
- Provide the correct answer to the generated question without thinking. It must be one of A/B/C/D, and MUST BE enclosed within <answer> </answer> tags.
- Any question type other than multiple-choice is STRICTLY FORBIDDEN!

Your MUST response in this format:

<description>
{Visual description you extract from the image}
</description>

<question>
{Write a complete multiple-choice question that states all necessary conditions clearly, followed by exactly 4 answer options A B C D}
</question>

<answer>
{Correct answer option A/B/C/D}
</answer>

DO NOT output anything else—no explanations, no extra markup.

Prompt for Answer Generation

System Prompt:
You are a professional question solver.
User Prompt:
Solve the multiple-choice question based on the provided image. Let's think step by step.
First, you must fully perceive the image, extracting any valuable visual information from it.
Then, solve the question, give the correct choice option.

The reasoning process MUST BE enclosed within <think> </think> tags.
The final answer MUST BE single option A/B/C/D put in `\boxed{}`.

Prompt for Question Difficulty Evaluation

System Prompt:
You are a professional question evaluator.
User Prompt:
Score the difficulty of the following visual reasoning question on a scale from 0.0 (easiest) to 1.0 (hardest), where:

0.1–0.4 (Simple): Basic object counting or simple chart reading;
0.4–0.7 (Medium): Application of geometric relations or chart trend analysis;
0.7–1.0 (Hard): Complex geometric proofs or multi-step algebraic reasoning.

Output only a single number between 0.0 and 1.0. Do not include any other text, explanation, or formatting.