

C²-Evo: Co-Evolving Multimodal Data and Model for Self-Improving Reasoning

Xiuwei Chen¹, Wentao Hu², Hanhui Li¹, Jun Zhou¹, Zisheng Chen¹,
Meng Cao³, Yihan Zeng⁴, Kui Zhang⁴, Yu-jie Yuan⁴, Jianhua Han⁴,
Hang Xu⁴, Xiaodan Liang^{1*}

¹Sun Yat-sen University ²The Hong Kong Polytechnic University
³Peking University ⁴Huawei Noah's Ark Lab

Abstract

Recent advances in multimodal large language models (MLLMs) have shown impressive reasoning capabilities. However, further enhancing existing MLLMs necessitates high-quality vision-language datasets with carefully curated task complexities, which are both costly and challenging to scale. Although recent self-improving models that iteratively refine themselves offer a feasible solution, they still suffer from two core challenges: (i) most existing methods augment visual or textual data separately, resulting in discrepancies in data complexity (e.g., over-simplified diagrams paired with redundant textual descriptions); and (ii) the evolution of data and models is also separated, leading to scenarios where models are exposed to tasks with mismatched difficulty levels. To address these issues, we propose C²-Evo, an automatic, closed-loop self-improving framework that jointly evolves both training data and model capabilities. Specifically, given a base dataset and a base model, C²-Evo enhances them by a cross-modal data evolution loop and a data-model evolution loop. The former loop expands the base dataset by generating complex multimodal problems that combine structured textual sub-problems with iteratively specified geometric diagrams, while the latter loop adaptively selects the generated problems based on the performance of the base model, to conduct supervised fine-tuning and reinforcement learning alternately. Consequently, our method continuously refines its model and training data, and consistently obtains considerable performance gains across multiple mathematical reasoning benchmarks. Our code, models, and datasets will be released upon acceptance².

1 Introduction

Recent advancements in large language models (LLMs) have achieved remarkable progress in solving problems, including mathematics [7; 18], coding [5; 15], etc. These capabilities are enabled by advanced strategies, including chain-of-thought prompting [46], tool-augmented reasoning [13; 19; 24], *etc.* In particular, OpenAI o1 [35] and Deepseek-R1 [17] have shown that reinforcement learning plays a critical role in aligning model outputs with desired behaviors by using structured reward signals derived from correctness, consistency, or human preference. This mechanism has proven especially effective in eliciting nuanced self-verification and self-correction behavior in LLMs, thereby reinforcing the reliability and depth of their reasoning chains, particularly in mathematical and logical domains.

*Corresponding author.

²<https://github.com/chen-xw/C2-Evo>

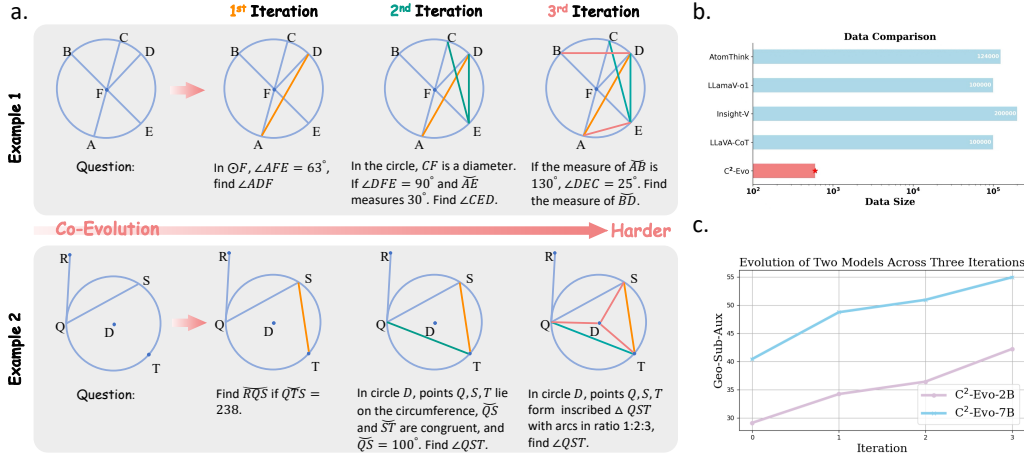


Figure 1: **a**: Co-evolution of visual-textual pairs with escalating task difficulty over three iterations. **b**: Data usage comparison among different methods. Note that the x-axis (gate count) is on a log-scale. **c**: Performance of 2B and 7B models across three iterations.

Despite these advances, achieving such strong reasoning performance remains heavily reliant on large-scale, high-quality, and complexity-aligned datasets. As task complexity increases, collecting suitable training data becomes significantly more costly and difficult, presenting a major bottleneck to further progress. This challenge has sparked growing interest in *self-improving* paradigms, where models iteratively enhance their capabilities by generating new synthetic data and refining reasoning traces.

Recent studies have shown that reasoning abilities can be substantially improved through carefully curated and progressively challenging data. For example, OpenVLThinker [11] adapts the self-improvement paradigm to the vision-language domain by iteratively alternating between supervised fine-tuning (SFT) and reinforcement learning (RL), distilling R1-style reasoning traces from text-based models into multimodal contexts.

However, despite these promising developments, existing approaches face two key limitations: (1) **Mismatched evolution of visual complexity and textual reasoning difficulty**. Prior methods often address visual and textual components in isolation. Some focus on generating visually complex scenes without meaningful reasoning tasks, while others emphasize textual complexity with simplistic visuals. This disconnect restricts the model’s ability to learn integrated cross-modal reasoning strategies. (2) **The discrepancy between model capability and task difficulty**. As models improve over the course of training, their ability to tackle more complex tasks naturally increases. However, current approaches rely on static or manually defined difficulty schedules, which do not adapt to the model’s evolving capability. This misalignment can lead to inefficient training, either under-challenging the model or overwhelming it with excessively difficult data.

To address these challenges, we propose a fully automated, adaptive multimodal learning framework (C²-Evo) that jointly evolves both the model and its training data in a closed-loop fashion, with a particular focus on geometric reasoning tasks. Table 1 summarizes the differences between our framework and state-of-the-art methods: Unlike existing methods that rely on static data, our method dynamically adjusts task complexity based on real-time assessments of model performance, ensuring a tighter coupling between model capability and data difficulty throughout the learning process. Specifically, to tackle **the challenge (1)**, we incorporate the process into a cross-modal data evolution loop, where we integrate SKETCHPAD-generated complex geometric diagrams [19] with complex problem synthesis (*cf.* Figure 1 (a)). GPT-4o is employed to generate auxiliary construction code, then executed via tools (*e.g.*, Jupyter). Formal image descriptions are produced using the Doubao model. Sub-problems (4–10 per image) are then generated following predefined guidelines and

Table 1: Comparison of evolutionary strategies.

Model	Visual Evolve	Text Evolve	Capability-Aligned
MindGYM [49]	✗	✓	-
R-CoT [10]	✓	✗	-
MAVIS [57]	✓	✗	-
OpenVLThinker [11]	-	-	✗
C ² -Evo(Ours)	✓	✓	✓

composed into complex reasoning questions. These are filtered using GPT-4o to remove misaligned cases, and validated using DeepSeek’s three-step reasoning framework to ensure consistency. This process yields a dataset with semantically aligned visual and textual components. For **the challenge (2)**, we adopt a data-model evolution loop. This loop utilizes data generated from the cross-modal data evolution loop and applies it to iteratively fine-tune the model using SFT and RL. SFT maintains output structure and coherence, while RL improves generalization through rule-based optimization. We introduce a simple error-based filtering method that evaluates sample difficulty via prediction variance over 32 generations. By selecting samples with an error rate ≥ 0.3 , which are prone to error yet still within the model’s grasp and pose a meaningful challenge, the framework ensures that task difficulty remains aligned with model capability, achieving continuous improvement over iterations (*cf.*, Figure 1 (c)).

Finally, we investigate the impact of different data strategies and iterative training regimes on model performance. Our findings offer insights into the design of effective self-improving frameworks for improving complex multimodal reasoning and guiding the progressive evolution of vision-language models.

Our contributions are summarized as follows:

- We propose a closed-loop self-improving framework, named C²-Evo, that jointly evolves training data and model capabilities.
- The proposed framework utilizes two co-evolution loops to improve the compatibility of cross-modal complexity and that between task difficulty and model capability.
- Extensive experiments demonstrate the effectiveness of different data strategies and iterative training regimes, revealing their impact on self-improving frameworks.

2 Related Work

Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) have witnessed rapid advancements in recent years, resulting in significant breakthroughs in visual understanding and cross-modal reasoning within the field of artificial intelligence. Unlike traditional Vision-Language Models [38; 22] (VLMs), which are typically trained from scratch on image-text pairs, MLLMs are generally built upon powerful pretrained text-only large language models (LLMs), and then further aligned with multimodal data, such as images and videos. Representative works such as BLIP-2 [25], LLaVA [30; 28; 29], and MiniGPT-4 [59] have demonstrated impressive zero-shot and few-shot generalization across various tasks, including image captioning, visual question answering, and image-text reasoning. LLaVA [30], for example, pioneered the use of high-quality visual instruction data generated by GPT-4 [1] to fine-tune MLLMs, achieving significant success in visual dialogue and reasoning. This method has generated significant interest in research focused on creating multimodal datasets for tuning vision instructions [4; 28]. BLIP-3 [50] extends the single image input of BLIP-2 to interleaved multimodal data, using large-scale, high-quality, and diverse curated data with training recipes and post-training strategies to beat other contemporary competitors on various visual understanding tasks. Recently, several open-source MLLMs have significantly narrowed the performance gap with proprietary models like OpenAI’s GPT-4o [21]. Among them, InternVL2 [6], Qwen2-VL [2], SPHINX [27], and MiniCPM-V [53] are particularly notable for their expanded application coverage, achieved through richer instruction-tuning datasets. In this study, we adopt a self-improving training paradigm to enhance the reasoning capabilities of MLLMs in complex scenarios.

Reinforcement Learning. Recent research has demonstrated that reinforcement learning (RL) can significantly enhance the reasoning capabilities of large language models (LLMs) such as OpenAI-o1 [35]. Some approaches have introduced RL-based mechanisms to facilitate test-time scaling, achieving notable success in tasks such as mathematical reasoning and code generation. Building on this progress, DeepSeek-R1 [17] proposed a rule-based reward strategy and adopted the Group Relative Policy Optimization (GRPO) [41] algorithm, demonstrating strong performance with only a few update steps. Motivated by the success of LLMs, recent studies [52; 20; 58; 9; 37; 34; 55; 36] have begun to explore the reasoning capabilities of MLLMs. For example, R1-OneVision [52] integrates supervised fine-tuning with RL to bridge the gap between visual perception and deep logical reasoning. Vision-R1 [20] generates cold-start initialization data and employs GRPO with hard format reward

functions to enhance the emergent reasoning capabilities of MLLMs. VisualThinker-R1-Zero [58] applies the R1 style to a base MLLM without supervised fine-tuning, surpassing traditional fine-tuning methods while exhibiting “visual aha moment” behaviors. LMM-R1 [37] adopts a staged approach that begins with textual reasoning and advances toward complex multimodal reasoning tasks. MM-EUREKA [34] introduces a novel data filtering strategy that simultaneously removes both unsolvable and trivial cases, along with rejection samples, retaining only high-confidence instances. The effectiveness of these methods in multimodal understanding tasks can be attributed to the presence of high-quality CoT datasets and the use of R1-style RL. Nonetheless, a significant drawback persists: the disconnect between the complexity of the training data and the difficulty of the tasks, particularly the inconsistency between the complexity of visual inputs and the challenges of textual reasoning. In this paper, we introduce a self-evolving training approach that alternates between RL and SFT. Our method dynamically modifies the multimodal training data to align with the model’s reasoning abilities, ensuring that the complexities of both visual and textual elements are consistent throughout the training process. In contrast to MM-EUREKA, our method prioritizes samples that are both meaningfully challenging and conditionally accessible, striking a balance between task difficulty and model learnability.

Self-Improvement. Self-improvement [14; 3; 39] is a paradigm in which models generate and train on synthetic data generated from the same or other models. While self-improvement has been widely studied in the NLP domain, several works [54; 16; 42; 33; 26; 8] have explored the approach of first generating high-quality data and subsequently fine-tuning models on this data to achieve continuous performance improvement. For example, STaR [54] introduces a bootstrapping mechanism that enhances LLM reasoning capabilities by iteratively generating and filtering “chain-of-thought” rationales, then fine-tuning the model on correct rationales to progressively improve performance. ReST [16] integrates self-generated data with offline RL, alternating between a “Grow” phase that expands the dataset by generating multiple outputs per input, and an “Improve” phase that ranks and filters these outputs using a reward model based on human preferences. Other works [33] follow a similar approach of generating synthetic data, filtering out low-quality samples, and fine-tuning models on the filtered high-quality data. Recent efforts [10; 57; 44] have introduced synthetic data generation pipelines that leverage a visual data engine to produce visual images along with corresponding reasoning tasks. However, these methods primarily focus on expanding the training distribution. Our approach shifts the emphasis from passive data enrichment to self-improvement with a co-evolution mechanism. A closely related work is OpenVLThinker [11], which introduces an iterative training paradigm in which models are exposed to progressively more complex tasks. However, their approach relies on manually defined task difficulty, which does not adapt to the evolving capabilities of the model.

3 Method

In this section, we present the details of the proposed C²-Evo framework for self-improving reasoning. The key idea behind C²-Evo is the joint evolution of multimodal data and models, thereby mitigating discrepancies not only between visual and textual modalities but also between task difficulty and model capabilities. We begin by formulating our task in Sec. 3.1, and then introduce the two core components of C²-Evo, namely the multimodal data co-evolution loop (Sec. 3.2), and the data and model co-evolution loop (Sec. 3.3), as shown in Figure 2.

3.1 Task Formulation

In this paper, we are given a pre-trained multimodal large language model (MLLM) denoted as π_θ and a dataset $\mathcal{D}$ consisting of image-question-answer triplets, namely, $\mathcal{D} = \{D_n = (I^n, Q^n, G^n) | n = 1, \dots, |\mathcal{D}|\}$, where I^n , Q^n , and G^n denote an image, questions related to I^n , and the corresponding answers of Q^n , respectively. Our goal is to improve the complex reasoning capability of π_θ by optimizing its parameters θ on $\mathcal{D}$ within T iterations. Following recent studies [10; 57; 19], we assume that the necessary external tools and models are available, such as an oracle model that can generate and execute Python code to edit images [19].

Specifically, let $t = 1, \dots, T$ denote the iteration step. At the beginning of the t -th iteration, we first conduct the multimodal data co-evolution loop to augment $\mathcal{D}_t$. This is achieved by prompting an MLLM [21] as the oracle model for solving each question in $\mathcal{D}_t$, generating step-by-step reasoning

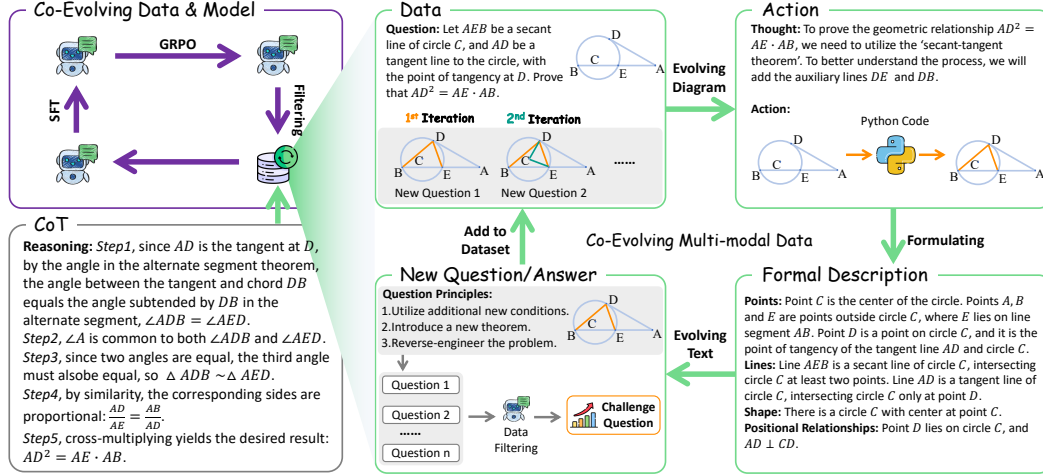


Figure 2: **The proposed C²-Evo framework.** Starting from a base dataset and a base model, a co-evolving multimodal data loop iteratively synthesizes complex paired visual and textual samples. These samples are subsequently filtered and utilized in the co-evolving data & model loop to iteratively improve the reasoning performance of the model.

Algorithm 1 C²-Evo Algorithm

Input: Seed Dataset $\mathcal{D}$, Base Model π_θ , Principles P , Number of Iterations T

Output: Improved Model $\pi_{\theta,RL}^{T+1}$, Evolved Datasets $\mathcal{D}_{T+1}$

- 1: Initialize $\pi_\theta^1 = \pi_\theta$
- 2: **for** $t = 1$ to T **do**
- 3: $\triangleright$ **Co-Evolving Multimodal Data** $\triangleright$ see §3.2
- 4: **for** $n = 1$ to $|\mathcal{D}_t|$ **do**
- 5: $Ac_t^n, C_t^n, R_t^n \leftarrow$ generate action and reasoning conditioned on D_t
- 6: $I_{t+1} \leftarrow$ apply auxiliary augmentations on I_t by executing (Ac_t^n, C_t^n)
- 7: $Q_{t+1}, A_{t+1} \leftarrow$ generate challenging questions and answers conditioned on I_{t+1} and P
- 8: $D_{t+1} \leftarrow$ add filtered tuple $(I_{t+1}, Q_{t+1}, A_{t+1})$
- 9: **end for**
- 10: $\triangleright$ **Co-Evolving Data and Model** $\triangleright$ see §3.3
- 11: $D_t^{\text{train}} \leftarrow$ select data from D_{t+1} using error rate evaluated with $\pi_{\theta,RL}^t$
- 12: $\pi_{\theta,SFT}^{t+1} \leftarrow$ update π_θ^t with SFT on D_t^{train} $\triangleright$ using Equation 1
- 13: $\pi_{\theta,RL}^{t+1} \leftarrow$ update $\pi_{\theta,SFT}^{t+1}$ with GRPO on D_t^{train} $\triangleright$ using Equation 2
- 14: **end for**

trajectories including necessary image augmentations like drawing auxiliary lines. For an arbitrary triplet $D_t^n \in \mathcal{D}_t$, the MLLM produces an action sequence Ac_t^n , corresponding Python code segment C_t^n , and a natural language reasoning trace R_t^n . The generated code C_t^n is then executed using an external image editing tool [23] to yield a more complex image I_{t+1}^n that includes auxiliary augmentations. To ensure that the questions align with I_{t+1}^n with increased complexity, more challenging questions Q_{t+1}^n and answers G_{t+1}^n regarding I_{t+1}^n are generated and curated, consequently a new multimodal triplet $D_{t+1}^n = (I_{t+1}^n, Q_{t+1}^n, G_{t+1}^n)$ is obtained. We assess the difficulty of D_{t+1}^n and if it is sufficiently challenging, we incorporate it into $\mathcal{D}_t$. Afterwards, we utilize the updated dataset $\mathcal{D}_{t+1}$ to train the base model π_θ through a combination of SFT and RL, where SFT establishes the initial reasoning structure and RL improves its generalization capability.

3.2 Co-Evolving Multimodal Data

The co-evolving multimodal data loop is introduced to mitigate the complexity discrepancy between visual and textual data, which can be further divided into an action and reasoning generation process and a challenging question generation process.

Action and Reasoning Generation. Inspired by SKETCHPAD [19], we introduce a two-stage framework designed to enable the precise construction of auxiliary lines in geometric diagrams. In the first stage, we extract the coordinates of key points using Optical Character Recognition (OCR) and other vision-based techniques³. This coordinate-level representation ensures robustness and consistency in subsequent diagram transformations, particularly under increased geometric complexity. Leveraging the extracted coordinates and image I_t (we omit the superscript n for conciseness), we prompt⁴ GPT-4o to determine whether auxiliary image augmentations are needed to facilitate problem solving. This generates a decision, denoted as *Thought*, indicating whether auxiliary augmentations are necessary and specifying their types (e.g., parallel lines, perpendicular lines, and connecting lines). Subsequently, it produces the updated image I_{t+1} with newly added auxiliary augmentations, as illustrated in Figure 2. In addition to I_{t+1} , the pair (I_t, Q_t) is also fed to GPT-4o to simultaneously generate the full reasoning trace R_t , which is used in the subsequent generation of challenging questions.

Challenging Question Generation. This step aims to match the level of problem difficulty with the degree of image complexity. Alternative to GPT-4o, we utilize the Doubao model [40] to generate a formatted description F_t conditioned on the generated images I_{t+1} , as in practice we find it provides more precise descriptions. To promote diversity in the generated sub-problems, we define a set of guiding principles P :

- 1) *Geometric Constraints*: We extract geometric constraints (e.g., perpendicularity, equality, and sub-images) from the formal description F_t .
- 2) *New Theorems and Concepts*: We incorporate relevant mathematical theorems and conceptual principles (e.g., the properties of a right triangle imply the Pythagorean theorem) that are closely related to the formal description F_t .
- 3) *Backward Reasoning*: We also include the concept (e.g., the Tangent-Secant theorem) of sub-steps in the inference trace (i.e., R_t).

Using these principles, we prompt DeepSeek-R1 with the tuples (F_t, R_t, P) to generate a diverse set of sub-problems $(q_1, q_2, \dots, q_m)$, where m typically ranges from 4 to 10, depending on the complexity of the image). To mitigate the inherent limitations of formal language in capturing visual content, we incorporate the role-playing strategy from R1-Onevision [51]. By analyzing the differences and commonalities among sub-problems $(q_1, q_2, \dots, q_m)$, we align them with corresponding sub-image elements to compose challenging geometric reasoning questions. Specifically, by combining the sub-problems $(q_1, q_2, \dots, q_m)$ with their associated sub-image descriptions F_t , we prompt DeepSeek-R1 to compose challenging questions. For example, given two sub-problems q_1 and q_2 that address different parts of an image, we combine them into a more challenging question Q_{t+1} by leveraging their thematic and contextual connections.

To obtain corresponding answers, we feed each generated question Q_{t+1} together with its formal description F_t into DeepSeek-R1 three times. For each input, the model produces a step-by-step reasoning trace R_{t+1} and the final answer A_{t+1} . We then retain only those samples with three answers $A_{t+1,1}, A_{t+1,2}, A_{t+1,3}$ are equal. We further prompt GPT-4o to filter out any questions that are inconsistent with the original image constraints, ensuring alignment between the question and the visual content. As illustrated in Figure 3, the increasing length of the reasoning trajectories serves as a reliable indicator of problem difficulty, aligning with our data evolution design.

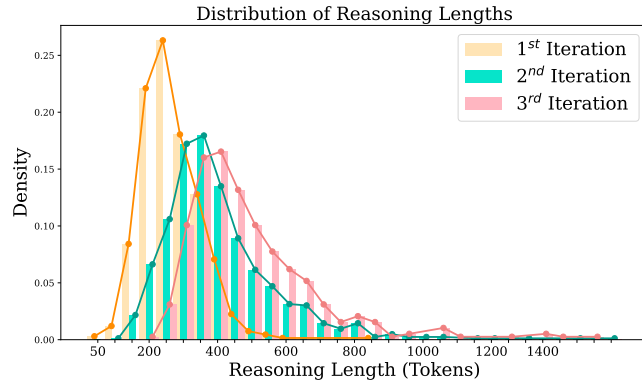


Figure 3: **Evolution of reasoning length** (tokens) over three iterations.

³https://github.com/lupantech/InterGPS/blob/main/diagram_parser

⁴All detailed prompts are provided in the Supplementary Material.

3.3 Co-Evolving Data and Model

SFT to Follow Reasoning Structure This stage aims to unify the model’s reasoning format with the structure needed during the later reinforcement learning (RL) phase, ensuring compatibility through a standardized template (*i.e.*, `<think></think><answer></answer>`). The model is trained on the previously constructed triplets (I_{t+1}, Q_t, R_t) , learning to generate the reasoning trace and final answer in a structured format. The corresponding training objective is formulated as follows:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(I,Q,R) \sim D} [\log(\pi_{\theta}(R|I, Q))]. \quad (1)$$

Group Relative Policy Optimization Based on the above obtained $\pi_{\theta, \text{SFT}}$ model, we employ two reward rules in conjunction with the GRPO algorithm to further refine the policy model. These rewards are designed as follows:

- 1) Accuracy Reward: This reward evaluates the correctness of the final answer A by processing the final answer via regular expressions and verifying it against the ground truth G .
- 2) Format Reward: To ensure consistent and well-structured reasoning trajectories, we define a reward based on the model’s adherence to the predefined format `<think> ... </think>`.

Additionally, we enforce sequential consistency within the reasoning trace by requiring that each step be explicitly arranged in the correct order. Responses that violate this ordering are penalized during training. The training objective with GRPO is defined as,

$$\mathcal{L}_{\text{RL}} = \mathbb{E} [\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon)A_t) - \beta D_{\text{KL}}[\pi_{\theta} \parallel \pi_{\text{old}}]]. \quad (2)$$

Iterative Refinements and Filtering

In the t -th iteration, we perform SFT followed by RL training using all post-filtered data, resulting in an updated model $\pi_{\theta, \text{RL}}^{t+1}$ and a new dataset $\mathcal{D}_{t+1}$ for the next round. To align the difficulty of $\mathcal{D}$ with the current capability of π_{θ} , we forward each sample through the model 32 times and compute the error rate as the proportion of times the model generated a wrong answer. As illustrated in Figure 4, we evaluate the difficulty of the data for each model obtained from RL training. We then retain only the samples with an error rate greater than 0.3 to form the training set for the next iteration, ensuring that the data complexity remains aligned with model capability.

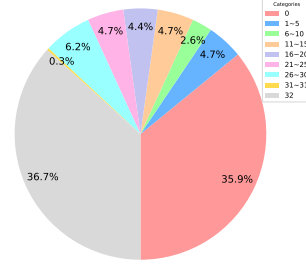


Figure 4: Assessment of third iteration training data difficulty using second iteration model $\pi_{\theta, \text{RL}}^2$.

4 Experiment

4.1 Benchmark

Dataset. We conduct experiments on the Geometry3k [32] dataset and use it as our base dataset for multimodal data evolution. Geometry3k is a mathematical benchmark dataset that comprises 3,002 geometry problems, divided into 2,101 training examples, 300 for validation, and 601 for testing. Each problem is accompanied by a corresponding geometric diagram, a natural language description, and formal language annotations.

Implementation Details. In our experiments, we adopt two state-of-the-art open-source MLLMs, *i.e.*, Qwen2-VL-2B [45] and Qwen2-VL-7B [45]. For the policy warm-up phase, we employ the LLaMA-Factory⁵ framework with a batch size of 128 and a learning rate of $1e-5$. For the RL phase, we use the VLM-R1⁶ framework. In the first iteration, we perform 32 rollouts per question, reducing to 8 in subsequent iterations. The temperature is set to the default value of 0.9, and the KL divergence coefficient β in Equation 2 is set to 0. All experiments are conducted on 32 NVIDIA V100-32GB GPUs.

We evaluate C²-Evo on several multimodal reasoning benchmarks, including Geo-Sub from Geometry3k-test [32], MathVista [31], and MathVerse [56]. To provide a more comprehensive

⁵<https://github.com/hiyouga/LLaMA-Factory>.

⁶<https://github.com/om-ai-lab/VLM-R1>.

evaluation of model performance, we construct a specialized test subset by sampling images from Geometry3K-test that require the use of auxiliary lines to solve. This results in a focused benchmark containing 274 images. When evaluated on the original images, we denote the setting as Geo-Sub. Using the corresponding images with added auxiliary lines, we refer to the setting as Geo-Sub-Aux. Further details can be found in the Supplementary Material.

4.2 Main Results

Table 2: More main experimental results. For MathVista benchmark, we have specifically compared all models on three sub-tasks that are highly related to mathematical reasoning: geometry reasoning (GEO), algebraic reasoning (ARI) and geometry problem solving (GPS). The result ([†]) is collected from R1-VL[55].

Methods	Data Amount	Geo-Sub	Geo-Sub-Aux	MathVista				MathVerse
				GEO	ARI	GPS	ALL	
<i>Closed-Source Model</i>								
GPT-4o[21]		-	-	-	-	-	63.8	39.4
<i>Reasoning Model</i>								
LLaVA-CoT-11B[48]	100K	-	-	-	-	-	54.8	-
Insight-V-8B[12]	200K	-	-	-	-	-	49.8	-
LLamaV-o1-11B[43]	>100k	-	-	-	-	-	54.4	-
AtomThink-LlamaV[47]	124K	-	-	-	-	-	59.1	33.5
Qwen2-VL-2B[45]	-	28.0	29.1	-	-	-	43.0 [†]	17.3
C ² -Evo-1 st	0.6k	32.0	34.2	38.0	40.0	38.0	49.1	16.7
C ² -Evo-2 nd	0.6k	35.6	36.4	38.0	38.0	38.0	49.3	18.5
C ² -Evo-3 rd	0.4k	38.2	42.2	40.0	40.0	38.0	50.2	19.1
Qwen2-VL-7B[45]	-	40.4	40.4	50.0	50.0	49.0	60.0	30.2
C ² -Evo-1 st	0.6k	45.5	48.7	55.0	57.0	54.0	62.3	33.6
C ² -Evo-2 nd	0.6k	46.9	50.9	56.0	58.0	56.0	62.7	34.8
C ² -Evo-3 rd	0.4k	50.9	54.9	59.0	59.0	60.0	63.4	34.9

As shown in Table 2, C²-Evo demonstrates a significant improvement over Qwen2-VL-7B and Qwen2-VL-2B across three benchmarks through three iterations, with geometric reasoning accuracy increasing notably from 40.4 to 54.9. The progressive improvement observed across three iterations highlights the effectiveness of our self-improving strategy in enhancing reasoning capabilities. Furthermore, compared to prior reasoning models, our approach achieves superior performance while leveraging less than 1% of the training data, with performance approaching that of GPT-4o.

Figure 5 illustrates the model’s behavior on a geometry problem involving a parallelogram. The response generated by Qwen2-VL-7B is relatively short but lacks a thorough reasoning process. It contains inaccuracies in both formula application and computational steps, leading to an incorrect final answer. In contrast, our C²-Evo produces a well-structured and logically coherent solution. It begins with a thorough analysis of the given conditions, followed by a step-by-step deductive process that ensures correctness at each stage, ultimately yielding the accurate final result.

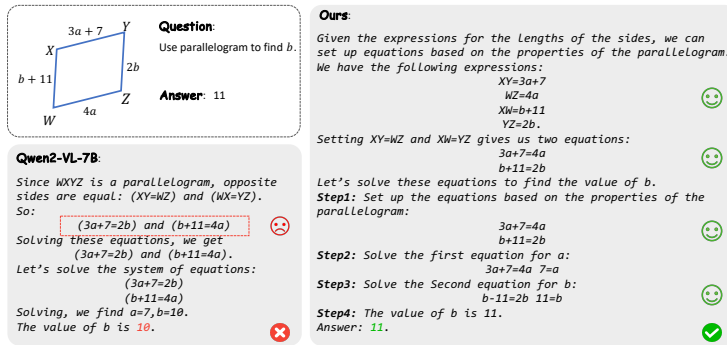


Figure 5: Qualitative comparison of geometric reasoning.

The influence of different iteration strategies. Table 3 presents the results of different iteration strategies. Fine-tuning π_{θ}^1 with additional SFT in the second iteration leads to performance degradation. In contrast, further refining the model through reinforcement learning yields improved results.

Table 3: Comparison of alternating vs. consecutive RL training schedules with progressive model updates.

Iteration	Model	SFT	RL	Geo-Sub-Aux
Second	π_{θ}^1	✓	✓	40.7
	π_{θ}	-	✓	50.6
Third	π_{θ}^2	✓	✓	42.9
	π_{θ}^2	-	✓	54.9

Table 4: Comparison of alternating vs. consecutive RL training schedules from the initial model.

Iteration	Model	SFT	RL	Geo-Sub-Aux
Second	π_{θ}^0	✓	✓	48.4
	π_{θ}	-	✓	44.0
Third	π_{θ}^0	✓	✓	46.7
	π_{θ}	-	✓	44.4

As shown in Table 4, when the second iteration model is trained via SFT and RL starting from the initial model π_{θ} , the same performance degradation is not observed. However, the resulting model underperforms compared to when the training is warm-started from π_{θ}^1 . Similar trends are observed in the third iteration.

The Influence of error-rate. As described in Section 3.3, we conduct an extensive parameter analysis over varying values of error-rate, which governs the complexity of data within each iteration. Table 5 shows the performance of models trained solely on data from the current iteration D_t , compared to those that also incorporate historical data $D_{1,\dots,t-1}$. Utilizing the complete dataset does not necessarily optimize the model’s reasoning capabilities. Specifically, as the complexity of the training data increases, relying exclusively on error-prone samples results in a decline performance. Based on this analysis, error-rate ≥ 0.3 is adopted to as the basis for our final setting.

Table 5: Comparison of error-rates based on the second iteration.

Models	Data Size ($D_t/D_{1\sim t}$)	Geo-Sub-Aux
Qwen2-VL-7B	-	38.2/40.4
Fullset	0.83K/1.48K	48.4/46.6
π_{θ}^1 - 0.3	0.48K/0.64K	49.1/ 50.9
π_{θ}^1 - 0.6	0.44K/0.54K	50.9 /47.3
π_{θ}^1 - 0.9	0.38K/0.44K	48.4/47.3
π_{θ}^1 - 1.0	0.34K/0.35K	46.2/49.5

The Influence of training data. Table 6 presents a comparison of the 7B model trained on complexified data (complex images with complex text) versus the original data (original images with complex text). The results show that jointly complexifying both images and text leads to better performance, highlighting the importance of image complexity in training.

Table 6: Training comparison between original and complex image datasets.

Model	Original Data	Complex Data
C ² -Evo-1 st	47.27	48.7
C ² -Evo-2 nd	45.82	50.9
C ² -Evo-3 rd	51.27	54.9

5 Conclusion

In this paper, we propose a closed-loop self-improving framework (C²-Evo), a multi-dimensional evolution framework that operates through two interleaved loops: **a cross-modal data evolution loop** and **a data-model co-evolution loop**. Recent studies often suffer from the decoupling of textual and visual evolution, as well as a mismatch between model capability and task difficulty. To address these limitations, we introduce a synchronized co-evolution mechanism in which auxiliary guidance is employed to progressively increase the complexity of image data, while the resulting complex visuals are used to generate increasingly challenging reasoning tasks. This ensures the joint evolution of both modalities. Furthermore, by leveraging the error-based filtering method, the model selects samples that align with the current model’s blind spots or underdeveloped reasoning capabilities, thereby maintaining consistency between data complexity and model capability through iterative training. Extensive experiments demonstrate the effectiveness of our multi-iteration evolution training strategy. The different data curation strategies and iterative mechanisms not only improve performance but also offer promising directions for future research in self-improving methods.

Limitations. While our approach uses self-iterative evolution to improve geometric reasoning and auxiliary line construction, it mainly relies on a dataset limited to mathematical reasoning problems. The scope of the data may not fully represent the diversity of real-world geometric problems. Additionally, relying on synthetic data from self-iteration may bias the model toward easier-to-evolve solutions, potentially missing less common but valid approaches to geometric problems.

Future Work. The present study is limited to the domain of geometric reasoning. Future work will explore the applicability of our framework to more diverse domains, such as mathematical function understanding and general vision-language tasks. We also intend to investigate the impact of model scale on performance, aiming to understand how architectural capacity interacts with evolutionary training dynamics.

References

- [1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
- [2] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
- [3] Bhattarai, M., Barron, R., Eren, M., Vu, M., Grantcharov, V., Boureima, I., Stanev, V., Matuszek, C., Valtchinov, V., Rasmussen, K., et al.: Heal: Hierarchical embedding alignment loss for improved retrieval and representation learning. arXiv preprint arXiv:2412.04661 (2024)
- [4] Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision. pp. 370–387. Springer (2024)
- [5] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.D.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al.: Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021)
- [6] Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
- [7] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
- [8] Costello, C., Guo, S., Goldie, A., Mirhoseini, A.: Think, prune, train, improve: Scaling reasoning without scaling models. arXiv preprint arXiv:2504.18116 (2025)
- [9] Deng, H., Zou, D., Ma, R., Luo, H., Cao, Y., Kang, Y.: Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. arXiv preprint arXiv:2503.07065 (2025)
- [10] Deng, L., Liu, Y., Li, B., Luo, D., Wu, L., Zhang, C., Lyu, P., Zhang, Z., Zhang, G., Ding, E., et al.: R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. arXiv preprint arXiv:2410.17885 (2024)
- [11] Deng, Y., Bansal, H., Yin, F., Peng, N., Wang, W., Chang, K.W.: Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. arXiv preprint arXiv:2503.17352 (2025)
- [12] Dong, Y., Liu, Z., Sun, H.L., Yang, J., Hu, W., Rao, Y., Liu, Z.: Insight-v: Exploring long-chain visual reasoning with multimodal large language models. arXiv preprint arXiv:2411.14432 (2024)
- [13] Feng, J., Huang, S., Qu, X., Zhang, G., Qin, Y., Zhong, B., Jiang, C., Chi, J., Zhong, W.: Retool: Reinforcement learning for strategic tool use in llms. arXiv preprint arXiv:2504.11536 (2025)
- [14] Fernando, C., Banarse, D., Michalewski, H., Osindero, S., Rocktäschel, T.: Promptbreeder: Self-referential self-improvement via prompt evolution. arXiv preprint arXiv:2309.16797 (2023)
- [15] Gu, A., Rozière, B., Leather, H., Solar-Lezama, A., Synnaeve, G., Wang, S.I.: Cruxeval: A benchmark for code reasoning, understanding and execution. arXiv preprint arXiv:2401.03065 (2024)
- [16] Gulcehre, C., Paine, T.L., Srinivasan, S., Konyushkova, K., Weerts, L., Sharma, A., Siddhant, A., Ahern, A., Wang, M., Gu, C., et al.: Reinforced self-training (rest) for language modeling. arXiv preprint arXiv:2308.08998 (2023)
- [17] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

- [18] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J.: Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874 (2021)
- [19] Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L.S., Smith, N.A., Krishna, R.: Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. ArXiv **abs/2406.09403** (2024)
- [20] Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-rl: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
- [21] Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
- [22] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
- [23] Jupyter, P.: Jupyter notebook, <https://jupyter.org/>
- [24] Li, C., Wu, W., Zhang, H., Xia, Y., Mao, S., Dong, L., Vulić, I., Wei, F.: Imagine while reasoning in space: Multimodal visualization-of-thought. arXiv preprint arXiv:2501.07542 (2025)
- [25] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
- [26] Liang, Y., Zhang, G., Qu, X., Zheng, T., Guo, J., Du, X., Yang, Z., Liu, J., Lin, C., Ma, L., et al.: I-sheep: Self-alignment of llm from scratch through an iterative self-enhancement paradigm. arXiv preprint arXiv:2408.08072 (2024)
- [27] Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023)
- [28] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
- [29] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024)
- [30] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
- [31] Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
- [32] Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., Zhu, S.C.: Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In: Annual Meeting of the Association for Computational Linguistics (2021)
- [33] Lupidi, A., Gemmell, C., Cancedda, N., Dwivedi-Yu, J., Weston, J., Foerster, J., Raileanu, R., Lomeli, M.: Source2synth: Synthetic data generation and curation grounded in real data sources. arXiv preprint arXiv:2409.08239 (2024)
- [34] Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Shi, B., Wang, W., He, J., Zhang, K., et al.: Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. arXiv preprint arXiv:2503.07365 (2025)
- [35] OpenAI: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
- [36] Peng, Y., Wang, X., Wei, Y., Pei, J., Qiu, W., Jian, A., Hao, Y., Pan, J., Xie, T., Ge, L., et al.: Skywork rl v: Pioneering multimodal reasoning with chain-of-thought. arXiv preprint arXiv:2504.05599 (2025)
- [37] Peng, Y., Zhang, G., Zhang, M., You, Z., Liu, J., Zhu, Q., Yang, K., Xu, X., Geng, X., Yang, X.: Lmm-rl: Empowering 3b llms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536 (2025)

- [38] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
- [39] Rosser, J., Foerster, J.N.: Agentbreeder: Mitigating the ai safety impact of multi-agent scaffolds. arXiv preprint arXiv:2502.00757 (2025)
- [40] Seed, B.: Doubao-1.5-pro system card, https://seed.bytedance.com/zh/special/doubao_1_5_pro
- [41] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
- [42] Singh, A., Co-Reyes, J.D., Agarwal, R., Anand, A., Patil, P., Garcia, X., Liu, P.J., Harrison, J., Lee, J., Xu, K., et al.: Beyond human data: Scaling self-training for problem-solving with language models. arXiv preprint arXiv:2312.06585 (2023)
- [43] Thawakar, O., Dissanayake, D., More, K., Thawkar, R., Heakl, A., Ahsan, N., Li, Y., Zumri, M., Lahoud, J., Anwer, R.M., et al.: Llamav-o1: Rethinking step-by-step visual reasoning in llms. arXiv preprint arXiv:2501.06186 (2025)
- [44] Trinh, T.H., Wu, Y., Le, Q.V., He, H., Luong, T.: Solving olympiad geometry without human demonstrations. *Nature* **625**(7995), 476–482 (2024)
- [45] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
- [46] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022)
- [47] Xiang, K., Liu, Z., Jiang, Z., Nie, Y., Cai, K., Yin, Y., Huang, R., Fan, H., Li, H., Huang, W., et al.: Can atomic step decomposition enhance the self-structured reasoning of multimodal large models? arXiv preprint arXiv:2503.06252 (2025)
- [48] Xu, G., Jin, P., Hao, L., Song, Y., Sun, L., Yuan, L.: Llava-o1: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440 (2024)
- [49] Xu, Z., Chen, D., Ling, Z., Li, Y., Shen, Y.: Mindgym: Enhancing vision-language models via synthetic self-challenging questions. arXiv preprint arXiv:2503.09499 (2025)
- [50] Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., et al.: xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872 (2024)
- [51] Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., Zhang, B., Chen, W.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *ArXiv* **abs/2503.10615** (2025)
- [52] Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., et al.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025)
- [53] Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
- [54] Zelikman, E., Wu, Y., Mu, J., Goodman, N.: Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems* **35**, 15476–15488 (2022)
- [55] Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S., Tao, D.: R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937 (2025)
- [56] Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Qiao, Y., et al.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In: European Conference on Computer Vision. pp. 169–186. Springer (2024)

- [57] Zhang, R., Wei, X., Jiang, D., Guo, Z., Li, S., Zhang, Y., Tong, C., Liu, J., Zhou, A., Wei, B., et al.: Mavis: Mathematical visual instruction tuning with an automatic data engine. arXiv preprint arXiv:2407.08739 (2024)
- [58] Zhou, H., Li, X., Wang, R., Cheng, M., Zhou, T., Hsieh, C.J.: R1-zero's "aha moment" in visual reasoning on a 2b non-sft model. arXiv preprint arXiv:2503.05132 (2025)
- [59] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)

A Experimental Supplement

A.1 Implementation Details

Table VII: Training parameters of Qwen2-VL-2B/7B model.

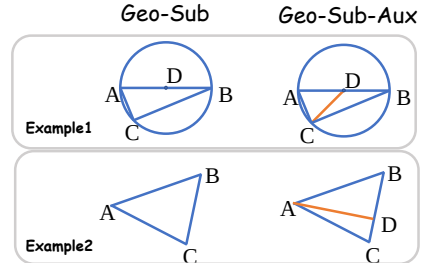
Parater	Qwen2-VL-2B-SFT	Qwen2-VL-2B-RL	Qwen2-VL-7B-SFT	Qwen2-VL-7B-RL
Learning Rate	1e-5	1e-6	1e-5	1e-6
Epochs	2	2	2	2
Batch Size	128	128	128	128
Precision	fp16	fp16	fp16	fp16
GPU	32 V100	32 V100	32 V100	32 V100
Temperature	-	0.9	-	0.9

In this section, we provide more implementation details for the C²-Evo. In Table VII, we provide the training parameters of Qwen2-VL-2B/7B during the SFT and RL stages to facilitate future work.

We evaluate C²-Evo on three reasoning benchmarks, Geo-Sub from Geometry3K-test [32], MathVista [31], and MathVerse [56]. For MathVista, we use the Test Mini split (*cf.* , around 1,000 samples). For MathVerse, we use the full dataset.

A.2 The difference between Geo-Sub and Geo-Sub-Aux.

We have introduced auxiliary lines into the Geo-Sub-Aux dataset to facilitate problem solving and improve model alignment during training. Originating from Geometry3k[32], the Geo-Sub dataset comprises 275 samples. The left image illustrates the structure of the Geo-Sub dataset (*cf.* , Figure VI left), whereas the right image showcases the Geo-Sub-Aux dataset (*cf.* , Figure VI right), highlighted by its additional auxiliary lines.



A.3 Data Difficulties Across Three Iterations.

Figure VI: Data comparison between Geo-Sub and Geo-Sub-Aux.

As shown in Figure VII, we also evaluate the difficulty of training data from previous iterations using the model training in the preceding iteration. Our observations are as follows: 1) In the evaluation of the second iteration, the first iteration model $\pi_{\theta,RL}^1$ achieved completely wrong predictions on approximately 8% of the data from the first iteration. 2) Despite the fact that the data generated in the second iteration was more challenging than that of the first iteration (*cf.* , as evidenced by the longer reasoning length in the Figure 3, the first iteration model $\pi_{\theta,RL}^1$ was still able to solve about 30% of the problems from the second iteration. 3) In the evaluation of the third iteration, when the second iteration model $\pi_{\theta,RL}^2$ was used to evaluate the data from the first iteration, there is an increase in completely wrong predictions and a decrease in fully correct ones. This suggests that the model begins to exhibit *catastrophic forgetting*. 4) Even within the second iteration data, around 30% of the questions are completely wrong , which aligns with the assumption stated in the main text that the data difficulty increases across iterations.

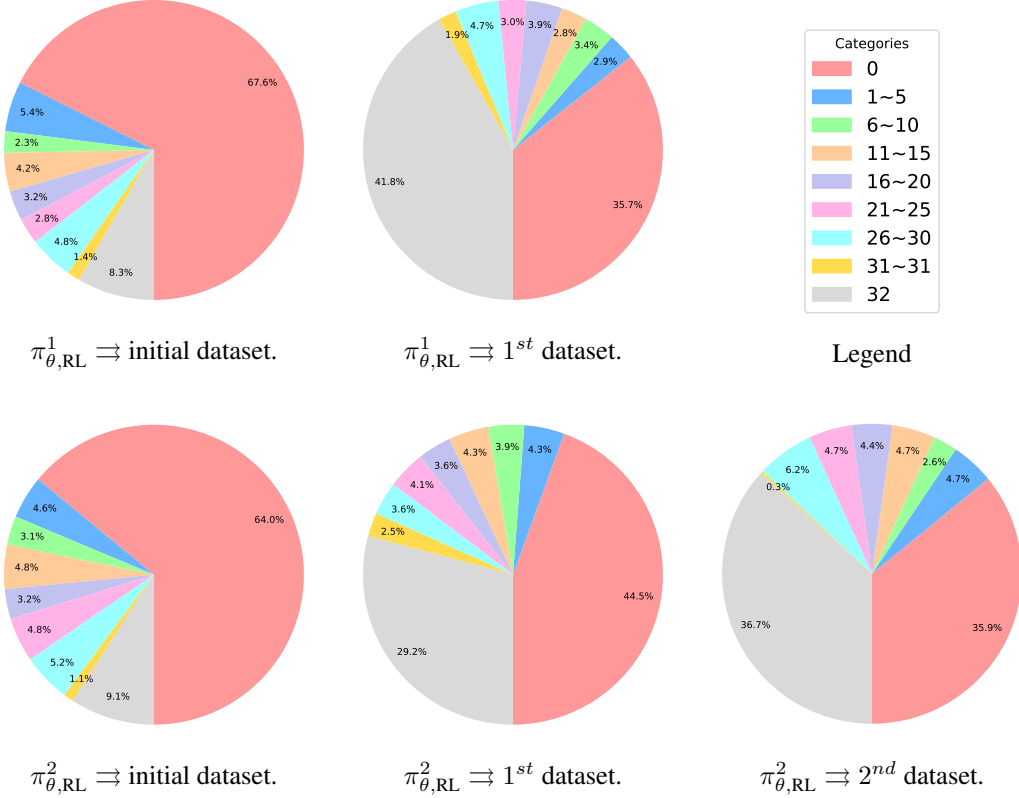


Figure VII: **Evolution of data difficulty across iterations (assessed by error-rate).** $\pi_{\theta} \Rightarrow \mathcal{D}$ denotes the difficulty distributions of dataset $\mathcal{D}$ evaluated using model π_{θ} based on error-rate.

Table VIII: Comparison of SFT vs. RL on first iteration.

Iteration	Model	SFT	RL	Geo-Sub-Aux
First	π_{θ}^0	✓		45.4
	π_{θ}^0		✓	44.7
	π_{θ}^0	✓	✓	48.7

A.4 The influence of different iteration strategies.

In this main paper, we present a comparison of different iteration strategies across the second and third iterations. To provide a more comprehensive evaluation, we further conduct experiments on different strategies in the first iteration, as shown in Table VIII. Under the same parameter settings, the results show that SFT+RL achieves the best performance in the first iteration, followed by SFT alone, while using RL-only training yields the least favorable outcomes.