

# GThinker: Towards General Multimodal Reasoning via Cue-Guided Rethinking

Yufei Zhan<sup>1,3,†,\*</sup>, Ziheng Wu<sup>2,†</sup>, Yousong Zhu<sup>1,✉</sup>,

Rongkun Xue<sup>6</sup>, Ruipu Luo<sup>2</sup>, Zhenghao Chen<sup>2</sup>, Can Zhang<sup>2</sup>, Yifan Li<sup>7</sup>, Zhentao He<sup>2</sup>,

Zheming Yang<sup>2</sup>, Ming Tang<sup>1,3</sup>, Minghui Qiu<sup>2</sup>, Jinqiao Wang<sup>1,3,4,5</sup>

<sup>1</sup> Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences

<sup>2</sup> ByteDance <sup>3</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences

<sup>4</sup> Peng Cheng Laboratory <sup>5</sup> Wuhan AI Research

<sup>6</sup> Xi'an Jiaotong University <sup>7</sup> Renmin University of China

<https://github.com/jefferyZhan/GThinker>

## Abstract

Despite notable advancements in multimodal reasoning, leading Multimodal Large Language Models (MLLMs) still underperform on vision-centric multimodal reasoning tasks in general scenarios. This shortfall stems from their predominant reliance on logic- and knowledge-based “slow thinking” strategies—while effective for domains like math and science—fail to integrate visual information effectively during reasoning. Consequently, these models often fail to adequately ground visual cues, resulting in suboptimal performance in tasks that require multiple plausible visual interpretations and inferences. To address this, we present **GThinker** (General Thinker), a novel reasoning MLLM excelling in multimodal reasoning across general scenarios, mathematics, and science. GThinker introduces Cue-Rethinking, a flexible reasoning pattern that grounds inferences in visual cues and iteratively reinterprets these cues to resolve inconsistencies. Building on this pattern, we further propose a two-stage training pipeline, including pattern-guided cold start and incentive reinforcement learning, designed to enable multimodal reasoning capabilities across domains. Furthermore, to support the training, we construct **GThinker-11K**, comprising 7K high-quality, iteratively-annotated reasoning paths and 4K curated reinforcement learning samples, filling the data gap toward general multimodal reasoning. Extensive experiments demonstrate that GThinker achieves 81.5% on the challenging comprehensive multimodal reasoning benchmark M<sup>3</sup>CoT, surpassing the latest O4-mini model. It also shows an average improvement of 2.1% on general scenario multimodal reasoning benchmarks, while maintaining on-par performance in mathematical reasoning compared to counterpart advanced reasoning models. The code, model, and data will be released soon at <https://github.com/jefferyZhan/GThinker>.

让MLLM在面对一般场景时能够更好地利用视觉信息，而不是完全依赖于外部知识

## 1 Introduction

Open-source Multimodal Large Language Models (MLLMs) [22, 43, 53, 54, 70] have made significant strides across a wide range of tasks. Leading models like Qwen2.5-VL [2] now rival closed-source counterparts such as GPT-4o [17] in performance. These advances have benefited in part from the adoption of chain-of-thought (CoT) techniques [28, 51, 59], especially in mathematics and science. With the emergence of OpenAI’s O1 model [19], several studies [45, 56, 58] have sought

\*Work done during internship at ByteDance.

†Equal Contribution.

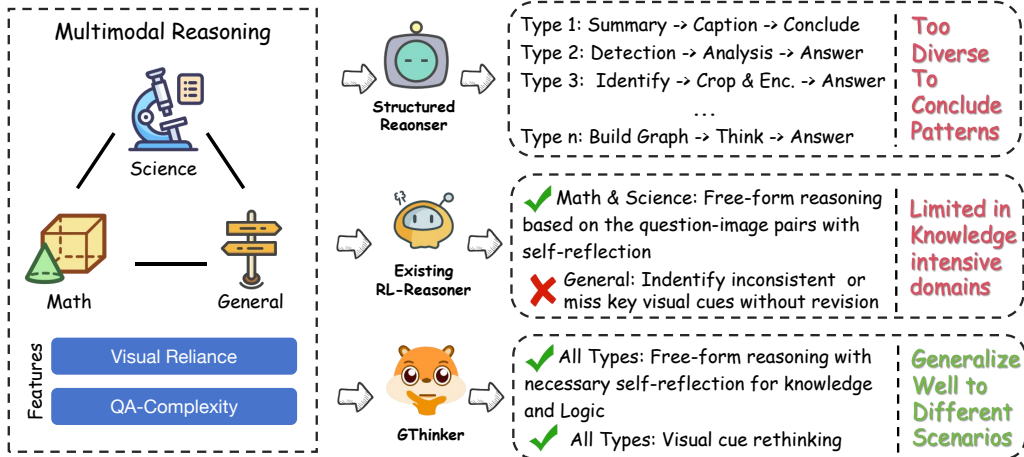


Figure 1: Multimodal reasoning methods comparison across scenarios. Multimodal Reasoning in different domains is featured with visual reliance and high question complexity, making it a challenging task. Different from previous methods, GThinker utilizes free-form thinking for different types of questions instead of a fixed structure form and enables general scenario reasoning accuracy with designed visual cue rethinking.

to transfer such slow-thinking capabilities to the multimodal reasoning domain to enhance models’ performance on complex tasks. DeepSeek-R1 [13] further introduces a new perspective, showing that outcome-reward Reinforcement Learning (RL) can awake long CoT reasoning, with promising results [4, 32, 57] in multimodal reasoning tasks involving science and mathematics.

Beyond mathematics and science, multimodal reasoning in general scenarios, which often involves visual cues and related commonsense still remains under-explored. Unlike math and science tasks, which typically follow strict logical structures and have unique answers, multimodal reasoning tasks in general scenarios are more diverse in nature. This makes it challenging to summarize a fixed CoT pattern or design an effective Process Reward Model (PRM), limiting the effectiveness of structured reasoning [56, 58] and Multimodal PRMs [25, 50]. Furthermore, general scenarios often require plausible interpretations and inferences grounded in visual content, which reduces the effectiveness of current outcome-reward-based reasoning models [4, 16, 57] that are primarily developed for math and science. As summarized in Figure 1, existing slow-thinking models frequently miss critical visual cues. When encountering plausible but inconsistent outputs, they often proceed directly to an answer without revisiting the reasoning path, unlike the reflection and verification behaviors observed in math and science domains. This suggests that in general scenarios, rethinking that integrates visual interpretation and inference cannot be effectively incentivized by RL alone, in contrast to the naturally learned reflection mechanisms in math and science reasoning tasks during pretraining [38].

To address these challenges, we propose GThinker, a novel reasoning MLLM excelling in multimodal reasoning across general scenarios, mathematics, and science. First, we introduce a new long-chain cue-driven pattern for multimodal reasoning called Cue-Rethinking. Unlike prior approaches [45, 56] that define structured CoT formats, Cue-Rethinking only requires the reasoning process to be strictly grounded in visual cues without enforcing a fixed format. After completing an initial reasoning chain, the model rethinks on the interpretations and inferences based on visual content to correct inconsistencies and arrive at the correct answer. Building on this pattern, we propose a two-stage training pipeline to enable robust multimodal reasoning. We begin by using pattern-guided cold start to train the model to learn this reasoning pattern on different tasks, and cold-start it with supervised fine-tuning. Then, we further employ an incentive RL stage to let the model explore optimal strategies for solving diverse problems across domains. To support training, we further develop a multimodal iterative annotation pipeline based on the latest advancing multimodal models like O3 [34] and construct GThinker-11k, comprised 7K cold-start data with high-quality annotated reasoning paths and 4K reinforcement learning samples, filling a key gap in multimodal reasoning fine-grained data for general scenarios.

We implement GThinker based on the advanced open-source MLLM Qwen-VL 2.5–7B and conduct extensive experiments to rigorously evaluate its effectiveness. We first benchmark GThinker against both open- and closed-source models on M<sup>3</sup>CoT [7], a challenging and comprehensive multimodal reasoning dataset spanning science, general commonsense, and mathematics. For broader validation, we include general-domain benchmarks such as MMStar [5] and RealWorld QA [55], as well as science and math-focused benchmarks including MMMU-Pro [62], MathVision [47], and MathVista [27]. GThinker demonstrates strong performance across all domains, achieving 81.5% on M<sup>3</sup>CoT—surpassing the advanced O4-mini model. On MMStar and RealWorld QA, GThinker achieves the improvement of 2.5% and 1.6%, respectively. Additionally, it performs competitively on science and math benchmarks with 40.7% on MMMU Pro and 72.7% on MathVista, matching or outperforming recent RL-enhanced approaches, further validating its effectiveness.

## 2 Related Work

### 2.1 Structured Multimodal Chain-of-Thought Reasoning

Structured Multimodal Chain-of-Thought (MCoT) reasoning builds on the Chain-of-Thought (CoT) paradigm [51], extending it to multimodal tasks using step-by-step reasoning [28, 68]. Many approaches enhance this framework with structured designs [26, 33, 69] and further improvements such as fine-grained visual grounding, context integration, or tool use [3, 12, 21, 24, 31, 39, 52]. However, these methods are often task-specific—e.g., CCoT [33] for compositional reasoning, LLaVA-Aurora [3] for spatial reasoning—and lack robustness across diverse scenarios. Recently, slow-thinking paradigms [19, 36, 44] have been proposed to improve reasoning depth. Enhanced MCoT variants like LLaVA-CoT [56], Virgo [10], and Mulberry [58] leverage long-chain generation, tree search, and self-reflection. Yet, they remain confined to structured, logic-heavy tasks and are difficult to generalize to broader settings. In contrast, GThinker adopts a free-form, cue-based thinking paradigm with further visual cue-based rethinking, moving beyond rigid structures to support open-domain multimodal reasoning. This design enables generalization across task types without sacrificing interpretability or performance.

### 2.2 Multimodal Reasoning with Reinforcement Learning

Reinforcement learning (RL) has become a powerful tool to align MLLMs and mitigate hallucinations [23, 41, 42, 61, 66, 67], and is now being explored to improve multimodal reasoning. Early approaches like LLaVA-Reasoner [65] and MPO [49] rely on rationale distillation alone and preference data to guide reasoning, while Insight-V [37] designs multi-agent systems with iterative Direct Preference Optimization. However, these methods focus on “teaching correctness” via supervised signals and human preference annotations, limiting robustness and scalability for more complex scenarios. A shift emerged with DeepSeek-R1 [13], which showed that outcome-based rewards, without fine-grained annotations, can drive reasoning through self-verification and reflection. Follow-up works [4, 6, 32, 35, 43, 57] expand this idea to the multimodal domain, leveraging verifiable reward functions or rule-based signals to improve math and science reasoning. Yet, these methods largely target well-defined tasks with unique answers. In general multimodal reasoning, models must handle ambiguity, interpret visual cues, and perform flexible inference. This limits the direct transfer of knowledge-style RL setups. Additionally, common reward models like PRMs [25, 50] struggle to capture progress in diverse tasks under general scenarios. To address this, we propose a clue-driven rethinking pattern tailored for general scenario multimodal scenarios but also accustomed to math and science settings. By further leveraging our design two-stage training, GThinker enables flexible reasoning with visual cue-based rethinking and knowledge reflection across diverse multimodal reasoning tasks.

## 3 Methodology

In this section, we provide a comprehensive description of the novel multimodal reasoning model GThinker as depicted in Figure 2. In §3.1, we first present the Cue-Rethinking Pattern, a core component built on free-form thinking to provide visual cue-driven guidance for multimodal reasoning across scenarios. Then, in §3.2, we describe Pattern-Guided Cold Start, in which we build 7k high-quality reasoning path annotated data and train the model with pattern-guided supervised fine-tuning to

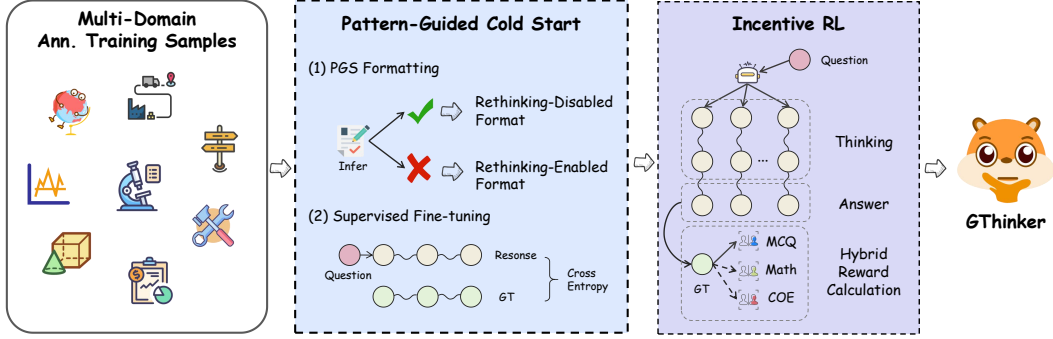


Figure 2: Overall pipeline for constructing GThinker. We collect multi-domain data covering general, math, and science tasks, and annotate it using multiple advanced MLLMs. The Pattern-Guided Cold Start phase then teaches the model the Cue-Rethinking Pattern for different question types. Finally, incentive reinforcement learning with DAPO enhances GThinker’s ability to perform adaptive and accurate multimodal reasoning across diverse scenarios.

learn how to think and rethink for different scenarios. Finally, we introduce Incentive Reinforcement Learning to generalize the multimodal reasoning capabilities of the model across diverse scenarios in §3.3.

### 3.1 Cue-Rethinking Pattern

Existing long-chain reasoning methods[45, 56] often rely on fixed, structured thinking chains tailored to specific tasks. While effective in targeted domains, their performance tends to drop sharply when applied to more general or unfamiliar scenarios. Outcome-reward models offer more flexibility, but they also fall short in general settings that require grounded, visually informed interpretations and inferences. To tackle this challenge, we introduce the Cue-Rethinking Pattern, a thinking framework that enables flexible long-chain reasoning through a combination of free-form thinking and rethinking on visual cues.

As shown in Figure 3, this process unfolds in three stages generally: initial reasoning, cue-rethinking trigger, and cue-based rethinking. During the initial stage, the model is free to reason in any form based on the question and image content itself, without structural constraints. It simply tags any referenced visual cues in the format `<vcues_*> </vcues_*>` (\* indicates the No.), which are later used for visual cues rethinking. This flexibility allows the model to apply learned reasoning strategies, such as step-by-step deduction or logical and knowledge reflection, much like how reasoning is approached in mathematical or scientific contexts, depending on the task.

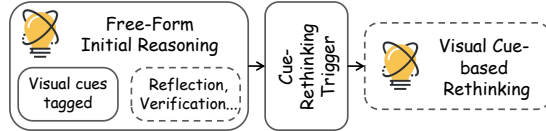


Figure 3: Toy example of the Cue-Rethinking Pattern. The dashed line indicates generation on demand.

After completing the initial reasoning, a prompt is triggered to initiate cue-based rethinking, like “*Let’s check each visual cue and corresponding reasoning before reaching the final answer*”. Importantly, we do not require immediate rethinking after visual cue identification, as doing so may disrupt the natural reasoning flow and prevent us from seeing the overall context. Then, the model revisits all previously marked visual cues, checking for inconsistencies or flaws. If problematic cues are identified, they are revised, and the model re-engages in corresponding reasoning, now grounded in the corrected cues, and concludes the final answer. This approach not only accommodates a wide range of reasoning approaches for different tasks but also addresses current limitations in handling misleading or missing visual inputs during reasoning. By combining free-form thinking with designed visual cue rethinking, this pattern delivers robust, adaptable reasoning across diverse multimodal reasoning scenarios.

反思机制：先不做约束，然后提示模型检查是否遗漏视觉线索，如果出现遗漏就修正

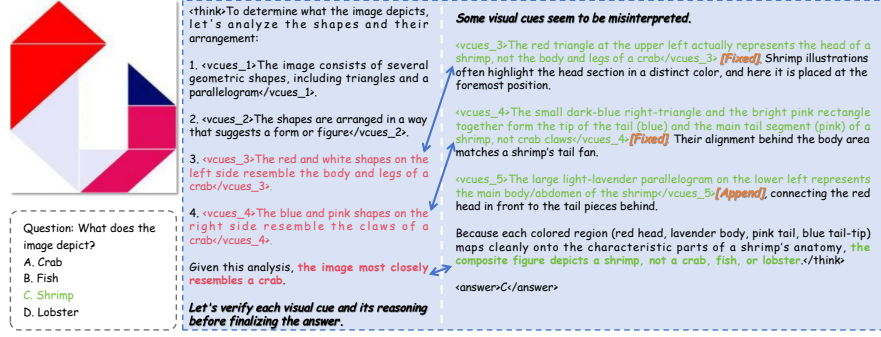


Figure 4: Constructed Data Example with Cue Rethinking. The visual cues in red are flawed ones, while the green indicates the visual cues are revised or appended.

训练文本处理：用线索-反思格式标注

### 3.2 Pattern-Guided Cold Start

Building on the Cue-Rethinking Pattern, we address how to effectively teach models to internalize and apply this reasoning pattern. While outcome-reward RL can guide models toward desired thinking, relying solely on it is still challenging and computationally intensive [13]. To overcome this, we introduce a Pattern-Guided Cold Start stage as shown in Figure 2, where the model is trained to adopt the Cue-Rethinking paradigm through supervised fine-tuning. To support this, we construct a 7K-scale dataset of annotated reasoning paths across multiple domains, using a multimodal collaborative annotation pipeline. The resulting data enables the model to learn both general problem-solving and cue-based rethinking.

**Data Construction via Multimodal Iterative Annotation.** To support domain-diverse multimodal reasoning, we collect data spanning mathematics, science, and general visual scenarios, validating each example for visual dependency and reasoning complexity. Instead of prompting models with image captions and question or structure requirements, we feed the image, question, and answer into advanced multimodal reasoning models, prompting them to reason step-by-step and identify relevant visual cues. For math and science questions, the models are allowed with self-reflection and validation; for cue-rich general questions, they are instructed to provide explicit visual references to support later rethinking. This strategy aligns well with the flexible design of Cue-Rethinking. To maximize precision, we iteratively annotate the data using several models, including GPT-4o, o1, and o3, leveraging each model’s strengths. We further extend this process to generate cue-based rethinking data. This automated pipeline results in a final dataset of 7,358 high-quality annotated samples, detailed further in the Appendix A. We provide a data example with key texts formatted in cue-rethinking in Figure 4.

**Pattern-Guided Supervised Fine-tuning.** With the annotated data, we train the model to learn the Cue-Rethinking pattern via supervised fine-tuning. Since reflection in science and math scenarios or cue-rethinking in general scenarios is one of the reasoning approaches, enforcing a single learning format could constrain the model’s robustness. To address this, we introduce **pattern-guided selective formatting** to customize the training data based on problem type. Specifically, we first run the base model on the training questions and compare its reasoning paths to the annotations. Samples with flawed visual cues are selected to form full Cue-Rethinking sequences, covering all three stages. Remaining examples are formatted as free-form reasoning paths. The model is then fine-tuned using this pattern-compiled data, enabling it to adaptively perform reasoning or rethinking as required by the question. 使用基础模型在视觉线索上有缺陷的样本进行微调

### 3.3 Incentive Reinforcement Learning 用DAPO增强训练效果

Following the Pattern-Guided Cold Start phase, the model acquires the designed reasoning pattern and learns to perform both flexible step-by-step reasoning and cue-based rethinking. Building on this foundation, we further enhance the model using outcome-reward reinforcement learning to encourage exploration and help it generalize across diverse tasks and scenarios. Given recent advances in outcome-reward reinforcement learning, we adopt the Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) algorithm [60] due to its strengths in supporting long-chain reasoning and its

efficiency in stable training. To accommodate varying task types and align with our pattern-based methodology, we design a hybrid reward computation strategy tailored to different problem categories. This training is carried out on a curated set of 4K diverse reasoning samples, enabling the model to generalize beyond the supervised data and adapt effectively to new challenges.

**Preliminaries about DAPO.** DAPO improves from the Group Relative Policy Optimization (GRPO) [40] with several enhancements to improve training efficiency, stability, and long-chain benefits, while retaining the key features such as outcome-based reward and policy optimization. As shown in the Equation 1, DAPO first employs a clip-higher strategy to address exploration limitations caused by identical responses, by adjusting the clipping threshold. It then adopts a dynamic sampling mechanism to prevent low training efficiency when all responses in a group are either entirely correct or entirely incorrect. Furthermore, it integrates Token-Level Policy Gradient Loss to encourage the model to learn high-quality reasoning patterns within long-chain responses while suppressing redundant reasoning. Lastly, the Overlong Reward Shaping strategy helps reduce the noise caused by excessively long sample sequences during training.

$$\begin{aligned} \mathcal{J}_{\text{DAPO}}(\theta) = & \mathbb{E}_{(q,a) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\text{old}}(\cdot | q)} \\ & \left[ \frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min \left( r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}) \hat{A}_{i,t} \right) \right] \quad (1) \\ \text{s.t. } & 0 < |\{o_i \mid \text{is\_equivalent}(a, o_i)\}| < G, \end{aligned}$$

where

$$r_{i,t}(\theta) = \frac{\pi_{\theta}(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \quad (2)$$

By incorporating DAPO, especially its clip-higher mechanism and token-level loss, the model is better equipped to sample diverse reasoning paths. This enables it to learn reasoning strategies such as reflective knowledge inference for math tasks or cue-based rethinking in general multimodal scenarios. As a result, the model improves the ability to dynamically select the most suitable reasoning strategy for each situation, improving both generalization and robustness across domains.

**Hybrid Reward Design.** The default DAPO setting combines format-based and accuracy-based rewards. Prior approaches often constrain QA tasks to rigid formats, such as multiple-choice, and depend on exact string matching to assess correctness. This limits the range of question types the model can handle, especially in general scenarios, and model-based verification further reduces training efficiency. To overcome these limitations, we propose a hybrid reward strategy within the constraints of verifiable rewards. We support three main question types: multiple-choice, math, and simple open-ended formats. For multiple-choice questions, we apply exact answer matching. For math problems—whether numeric or symbolic—we use Math-Verify [18] to extract and verify answers. For open-ended questions that yield concise responses (e.g., a word or short phrase), we guide the model to summarize the answer in a standardized, concise format, enabling straightforward matching during reward computation. This design expands the diversity of supported question types while preserving reward accuracy. For the format reward, we follow prior work by enforcing and verifying adherence to the think-answer structure.

**Data Construction.** To support the reinforcement learning stage, we construct a set data of 4k samples spanning math, science, and general reasoning tasks. Rather than relying solely on the 7k examples from the cold start phase, we introduce 4k samples sourced from public datasets to enhance diversity and generalization. This combined dataset offers a well-balanced and domain-spanning resource tailored for incentive RL. We provide more details about this data in the Appendix A.

## 4 Experiments

### 4.1 Implementation Details

**Training Settings.** We implement GThinker with the advanced MLLM Qwen2.5-VL-7B [2], one of the latest and most capable models at this scale, combining strong visual understanding with broad

Table 1: Main results on comprehensive multimodal reasoning benchmark M<sup>3</sup>CoT. Abbreviations used in the table: Lang. (Language), Nat. (Natural), Soc. (Social), Phys. (Physical), Temp. (Temporal), Alg. (Algebra), Geom. (Geometry), Theo. (Theory). Excluding closed-source models, values in bold represent the highest performance, while underlined values indicate the second-best performance across all models.

Model	Science			Commonsense			Mathematics			Overall
	Lang.	Nat.	Soc.	Phys.	Soc.	Temp.	Alg.	Geom.	Theo.	
<i>Closed-Source Models</i>										
Gemini-2.5 Pro [9]	97.6	91.6	75.3	92.2	81.4	94.3	81.1	78.8	61.9	85.9
O3-20250416 [34]	96.2	89.3	68.0	91.1	80.2	93.5	95.0	87.5	90.5	83.8
O4-mini-20250416 [34]	97.2	84.7	62.9	94.4	82.6	91.1	92.9	86.3	76.2	80.9
GPT-4o-20241120 [17]	96.7	72.0	58.3	91.1	76.4	82.9	21.4	31.3	23.8	67.4
<i>Open-Source Models</i>										
InternVL-2.5-8B [8]	82.5	63.7	45.2	86.7	79.8	93.4	42.8	27.5	33.3	61.8
Ovis2-8B [30]	80.6	63.1	46.2	83.3	79.3	87.8	45.0	42.5	38.9	61.9
Valley2[54]	85.3	64.4	48.4	90.0	77.7	80.5	43.6	36.3	47.6	62.8
Qwen2.5-VL-7B [2]	82.9	61.2	46.8	82.2	81.4	81.3	57.9	40.0	61.9	62.4
<i>Reasoning Models</i>										
LLaVA-CoT-11B [56]	72.0	56.4	41.7	84.4	72.3	82.1	37.9	36.3	33.3	56.0
InternVL2.5-MPO-8B [48]	92.4	75.9	61.9	85.6	82.6	94.3	55.0	43.8	76.2	73.3
Kimi-VL-A3B-Thinking [43]	86.2	64.4	39.6	91.1	78.9	89.4	13.5	15.0	14.2	58.3
MM-Eureka-7B [32]	86.7	71.5	57.3	81.1	80.2	90.2	40.0	23.8	28.6	67.4
R1-OneVision-7B [57]	74.9	66.4	51.4	84.4	72.3	85.4	30.0	31.3	42.9	61.8
VLAA-Thinker-7B [4]	91.0	70.6	58.1	78.9	78.1	87.8	45.7	35.3	28.6	68.0
GThinker-7B	92.4	90.7	68.9	82.2	81.4	94.3	73.5	62.5	81.0	81.5

general knowledge. We train the GThinker using our design two-stage pipeline, including pattern-guided cold start and incentive reinforcement learning with the constructed data. For Pattern-Guided Cold Start, we use a global batch size of 128 and a learning rate of 5e-6, training the model with the 7K reasoning path annotated data for 3 epochs. In the Incentive RL stage, we set the rollout number to 16, use a global batch size of 64, and start with a learning rate of 1e-6, training for 170 steps using the curated 4K data. Training is conducted on 4 nodes, each with 8 NVIDIA H100 GPUs. The total training time is about 9 hours. We provide more details in Appendix B.

**Evaluation Settings.** We evaluate our model against top closed-source models, including the latest O4-mini, as well as open-source base and reasoning models with comparable parameter sizes trained using diverse methodologies. The evaluation focuses on multimodal reasoning across general, mathematical, and scientific scenarios:

- **M<sup>3</sup>CoT:** A challenging benchmark that spans science, commonsense, and math domains, with each example verified to require multi-step reasoning. We primarily use this benchmark to comprehensively evaluate models’ multimodal reasoning capabilities across diverse scenarios.
- **General scenario benchmarks:** MMStar [5] and RealWorld QA [55]. These benchmarks focus on general and realistic scenarios, including parts of understanding-based reasoning tasks, and are used to evaluate multimodal reasoning capabilities.
- **Science and math scenario benchmarks:** We use MMMU-Pro [62], which covers multiple scientific subjects, to evaluate multimodal reasoning in scientific contexts. For math-specific evaluation, we adopt the widely used MathVista [27] and MathVision [47] benchmarks.

All evaluations are conducted on a single node equipped with 8 NVIDIA H100 GPUs. For M<sup>3</sup>CoT, we follow each model’s official settings and prompts and use VLMEvalKit [11] for fair evaluation. For other benchmarks, we use the results reported in their original papers. For RL-enhanced reasoning models, which primarily focus on math and science domains, we follow their released models and evaluation guidelines to conduct testing.



Table 2: Main results on math-related and multidisciplinary benchmarks, and also fine-grained understanding of multimodal benchmarks incorporating reasoning. We use the setting detailed in the evaluation settings, and for the result of Qwen2.5-VL-7B on MMMU-Pro we report the reproduced one marked in \* due to the large difference, as widely observed.

Model	MMStar	RealWorldQA	MMMU-Pro	MathVista <sub>Mini</sub>	MathVision <sub>Full</sub>
<b>Close-Source Models</b>					
Gemini-2.5 Pro	73.6	78.0	68.8	80.9	73.3
GPT-4o-20241120	65.1	76.2	54.5	63.8	31.2
<b>Open-Source Models</b>					
InternVL2.5-8B [8]	62.8	<b>70.1</b>	34.4	64.4	19.7
Ovis2-8B [30]	64.4	-	-	71.4	25.9
Valley2 [54]	62.5	67.5	-	69.1	24.9
Qwen2.5-VL-7B [2]	63.9	<u>68.5</u>	36.9*	68.2	25.1
<b>Reasoning Models</b>					
LLaVA-CoT-11B [56]	57.6	63.6	33.8	54.8	20.6
InternVL2.5-MPO-8B [48]	-	-	-	67.0	25.7
Kimi-VL-A3B-Thinking [43]	60.8	-	-	67.6	<b>36.8</b>
MM-Eureka-7B [32]	64.2	67.3	<b>40.7</b>	<b>73.0</b>	26.9
R1-Onevision-7B [57]	42.8	62.7	31.0	64.1	<u>29.9</u>
VLAA-Thinker-7B [4]	63.7	66.9	<u>39.8</u>	68.0	26.4
GThinker-7B	<b>66.4</b>	<b>70.1</b>	<b>40.7</b>	<u>72.7</u>	26.6

## 4.2 Main Results

GThinker-7B demonstrates superior multimodal reasoning, consistently outperforming advanced open-source base models and surpassing recent reasoning models on most benchmarks. On the comprehensive M<sup>3</sup>CoT benchmark depicted in Table 1, which demands balanced knowledge and visual understanding, GThinker-7B achieves 81.5% average accuracy, performing on par with the latest reasoning model O4-mini. Among the reasoning models, GThinker-7B achieves the highest performance on 8 out of 9 sets. Besides the notable progress in science and commonsense, a key advantage of our approach is evident in multimodal mathematics problems within M<sup>3</sup>CoT, where GThinker-7B successfully aligns visual elements with textual information to derive correct solutions. This contrasts sharply with models like VLAA-Thinker-7B, which, despite visual competence, struggle with the requisite text-vision integration for M<sup>3</sup>CoT’s mathematical section, while Kimi-VL-A3B-Thinking produces repeated contents, especially on the math set, reducing its overall performance. This result further underscores our method’s effectiveness in fostering robust multimodal reasoning.

Beyond M<sup>3</sup>CoT, GThinker-7B exhibits leading performance across specialized and general multimodal benchmarks requiring reasoning as demonstrated in Table 2. On challenging math benchmarks, it achieves 72.7% on MathVista (+4.5 points over baseline) and 26.6% on MathVision (+1.5 points). Similarly, on the multidisciplinary science benchmark MMMU-Pro, GThinker-7B improves by approximately 4 points. Furthermore, it shows significant gains on general benchmarks requiring fine-grained understanding and further reasoning, with 66.4% on MMStar and 70.1% on RealWorld QA. Crucially, our proposed method enhances performance across diverse domains—general, math, and science—without the typical trade-offs observed in other reasoning models. Previous leading models, by focusing heavily on knowledge long-chain CoT reasoning, often showed limited gains or even degradation on general multimodal reasoning tasks due to less emphasis on visual cues, a limitation our versatile approach overcomes.

When compared to the advancing non-thinking model GPT-4o, GThinker-7B achieves superior or competitive performance on several benchmarks, notably M<sup>3</sup>CoT, MMStar, and MathVista, despite its significantly smaller 7B backbone. While GPT-4o leads on benchmarks like RealWorldQA, MMMU-Pro, and MathVision, which heavily leverage extensive knowledge and perceptual abilities inherent in larger models, our results are compelling. The substantial gains achieved by GThinker-7B, particularly on reasoning-centric benchmarks (e.g., M<sup>3</sup>CoT), highlight the efficacy of our proposed method in significantly boosting complex reasoning capabilities, even with a more compact model



Table 3: Ablation on Data Pipeline and Quality. Lang. denotes rationales from M<sup>3</sup>CoT. GThinker refers to data from our proposed pipeline. Iter. indicates the application of our iterative annotation process.

Lang.	GThinker	Iter.	Science	Com.	Math	Overall
✓			58.9	81.8	40.2	63.5
	✓		70.6	79.1	43.2	69.6
	✓	✓	73.1	79.3	46.9	73.6

Table 4: Ablation on GThinker Components. The PGS indicate the Pattern-Guided Selection introduced in §3.2

Method	Science	Com.	Math	Overall
GThinker	82.5	83.7	70.5	81.5
- Incentive RL	73.1	79.3	46.9	73.6
- PGS Formatting	68.0	82.0	42.7	68.4
- PG Cold Start	58.8	81.7	40.2	61.5
Qwen2.5-VL-7B-Zero	63.3	81.6	49.0	64.2

architecture. This underscores the advantage of our approach in efficiently enhancing multimodal reasoning across domains.

### 4.3 Ablation Study

**Ablation on Data Pipeline and Iterative Annotation.** High-quality data is crucial for training effective multimodal reasoning models. We enhance data quality through our novel pipeline, including an iterative annotation approach (details in App. A). To validate these contributions, we ablate each component by fine-tuning the model on the constructed 7K samples, varying only the annotation source, and evaluate on the M<sup>3</sup>CoT.

As shown in Table 3, using the rationales (Lang.), which are GPT-annotated [1], yields an overall score of 63.5%. Our data generation pipeline, even without iterative annotation, significantly improves performance to 69.6% (+6.1% absolute). This demonstrates the inherent benefit of our pipeline design in producing superior data for multimodal reasoning across diverse domains. Incorporating our iterative annotation process to curate the GThinker 7k reasoning paths further boosts the overall score to 73.6%, an additional 4.0% improvement. We attribute this gain to the complementary strengths of the leading models, including GPT-4o, O1, and O3: *during the collaborative annotation iterations, visual cues and reasoning logic are more thoroughly captured, which further boosts the quality of the CoT data.*

**Ablation on GThinker Components.** Ablation on GThinker Components. To assess the contribution of each component to GThinker’s performance, we conduct ablation studies by incrementally removing modules and evaluating on M<sup>3</sup>CoT. The final row (with all modules removed) corresponds to training with the same QA pairs but without any of our proposed methods. As shown in Table 4, using the Cue-Rethinking Pattern for Pattern-Guided Cold Start—without Pattern-Guided Selection (PGS) Formatting—yields a 6.9% average improvement. Adding PGS Formatting provides a further 5.2% average gain, with science and math questions improving by 5.1% and 4.2%, respectively. In contrast, performance on commonsense questions drops by 2.7%. This is because PGS Formatting applies cue-rethinking to samples with incorrect visual cues, prompting the model to engage with misleading information and learn to reflect and reason more flexibly. While this stage introduces variability due to the diversity and ambiguity of the cues, it builds a foundation for more adaptable reasoning in later stages. Science and math tasks, which benefit from consistent patterns and structured reasoning, show more stable gains from formatting. With Incentive Reinforcement Learning added, the model achieves substantial improvements across all domains, significantly outperforming the baseline. These results show that the free-form, cue-based reasoning developed during Cold Start is effectively reinforced and leveraged in the RL stage, enhancing the model’s generalization across tasks. For comparison, we also evaluate DAPO under the same conditions. As shown in Table 4, DAPO offers limited gains in general scenarios, though it improves performance in math and science. This highlights both the rationale behind our design and the impact of each component in advancing multimodal reasoning.

## 5 Conclusion

This paper addresses the challenge of advancing multimodal reasoning in MLLMs beyond domain-specific tasks like math and science, extending toward more general scenarios. We introduce GThinker, a novel reasoning framework that excels across diverse multimodal tasks, including general, mathematical, and scientific domains. Powered by our Cue-Rethinking Pattern, GThinker moves

beyond rigid templates, enabling flexible, question-driven reasoning and robust handling of flawed visual cues through reflective and knowledge-grounded thinking. Our two-stage pipeline—Pattern-Guided Cold Start followed by Incentive Reinforcement Learning—guides the model to learn effective reasoning strategies and reinforces its ability to adapt across domains. Extensive experiments on multi-domain multimodal reasoning benchmarks show that GThinker outperforms existing reasoning MLLMs in both accuracy and cross-domain adaptability. Ablation studies further confirm the effectiveness of each core design component. We provide more discussion on limitations and broader impact in the Appendix.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. *arXiv preprint arXiv:2412.03548*, 2024.
- [4] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training rl-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025.
- [5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [6] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>, 2025. Accessed: 2025-02-02.
- [7] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*, 2024.
- [8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.
- [9] Google DeepMind. Gemini 2.5 pro preview model card. <https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro-preview.pdf>, 2025.
- [10] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*, 2025.
- [11] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024.
- [12] Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9096–9105, 2024.
- [13] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [14] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems*, 2021.
- [15] Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision*, pages 558–575. Springer, 2022.
- [16] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

- [17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [18] Hynek. Math-Verify: Math Verification Library, 2023. If you use this software, please cite it using the metadata from this file.
- [19] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Hel-  
yar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint  
arXiv:2412.16720*, 2024.
- [20] Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. In *Proceedings of the 2022 Conference on Empirical  
Methods in Natural Language Processing*, pages 582–601, 2022.
- [21] Zixi Jia, Jiqiang Liu, Hexiao Li, Qinghua Liu, and Hongbin Gao. Dcot: Dual chain-of-thought prompting  
for large multimodal models. In *The 16th Asian Conference on Machine Learning (Conference Track)*,  
2024.
- [22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang,  
Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*,  
2024.
- [23] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang,  
and Lingpeng Kong. Silk: Preference distillation for large visual language models. *arXiv preprint  
arXiv:2312.10665*, 2023.
- [24] Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuanjing Huang, and Zhongyu Wei. Vocot: Unleashing  
visually grounded multi-step reasoning in large multi-modal models. *arXiv preprint arXiv:2405.16919*,  
2024.
- [25] Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. Diving into self-evolving  
training for multimodal reasoning. *arXiv preprint arXiv:2412.17451*, 2024.
- [26] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning  
improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024.
- [27] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei  
Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation  
models in visual contexts. In *The Twelfth International Conference on Learning Representations*.
- [28] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Taffjord, Peter  
Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question  
answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- [29] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and  
Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning.  
In *The Eleventh International Conference on Learning Representations*, 2022.
- [30] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural  
embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024.
- [31] Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. Textcot: Zoom in  
for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*, 2024.
- [32] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han,  
Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning  
with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.
- [33] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought  
prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision  
and Pattern Recognition*, pages 14420–14431, 2024.
- [34] OpenAI. O3 and o4-mini system card. [https://cdn.openai.com/pdf/  
2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf](https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf), 2025. Ac-  
cessed: 2025-05-07.
- [35] Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan,  
Tianyidan Xie, Li Ge, et al. Skywork rlv: Pioneering multimodal reasoning with chain-of-thought. *arXiv  
preprint arXiv:2504.05599*, 2025.

- [36] Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, et al. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*, 2024.
- [37] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
- [38] Darsh J Shah, Peter Rushton, Somanshu Singla, Mohit Parmar, Kurt Smith, Yash Vanjani, Ashish Vaswani, Adarsh Chaluvareja, Andrew Hojel, Andrew Ma, et al. Rethinking reflection in pre-training. *arXiv preprint arXiv:2504.04022*, 2025.
- [39] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.
- [40] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [41] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.
- [42] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.
- [43] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [44] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, 2025.
- [45] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.
- [46] Huy V Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, et al. Automatic data curation for self-supervised learning: A clustering-based approach. *arXiv preprint arXiv:2405.15613*, 2024.
- [47] Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.
- [48] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.
- [49] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.
- [50] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.
- [51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [52] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.
- [53] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

- [54] Ziheng Wu, Zhenghao Chen, Ruipu Luo, Can Zhang, Yuan Gao, Zhentao He, Xian Wang, Haoran Lin, and Minghui Qiu. Valley2: Exploring multimodal models with scalable vision-language design. *arXiv preprint arXiv:2501.05901*, 2025.
- [55] xAI. Grok-1.5 vision preview. <https://x.ai/blog/grok-1.5v>, 2024.
- [56] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.
- [57] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.
- [58] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.
- [59] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.
- [60] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.
- [61] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.
- [62] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.
- [63] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *European Conference on Computer Vision*, pages 405–422. Springer, 2024.
- [64] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. In *First Conference on Language Modeling*.
- [65] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024.
- [66] Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, et al. Spa-vl: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*, 2024.
- [67] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025.
- [68] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.
- [69] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibe Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.
- [70] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

# Appendix

## Contents

<b>A</b>	<b>GThinker-11K Construction</b>	<b>2</b>
A.1	Data Preparations . . . . .	2
A.2	Multimodal Iterative Annotation . . . . .	2
A.3	Negative Reasoning Annotation . . . . .	3
A.4	Formatting . . . . .	3
A.5	Automatic Verification . . . . .	3
A.6	Reinforcement Learning Data Construction . . . . .	3
A.7	Open Source . . . . .	3
<b>B</b>	<b>Training Details</b>	<b>6</b>
B.1	System Prompt . . . . .	6
B.2	Hyper-parameters . . . . .	6
<b>C</b>	<b>Qualitative Analysis</b>	<b>7</b>
<b>D</b>	<b>Limitations</b>	<b>7</b>
<b>E</b>	<b>Broader Impact</b>	<b>7</b>



## A GThinker-11K Construction

To support the training of GThinker, we have designed a scalable data generation pipeline to construct the GThinker-11K data as we have concluded in §3.2 and §3.3, respectively. In this section, we systematically introduce the data construction process, including the 7K cold start data, as depicted in Figure 5, and 4K RL data.

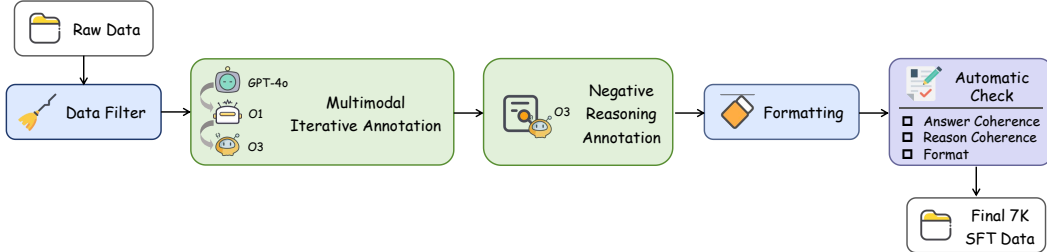


Figure 5: Data pipeline for cold start data.

### A.1 Data Preparations

Though several datasets are constructed to enhance multimodal reasoning capabilities in MLLMs [56–58] spanning diverse domains, they often present challenges such as high knowledge dependency, limited visual cues, or limited reasoning level. To extend the multimodal reasoning to general scenarios beyond knowledge-intensive math and science problems, we empirically find that the M<sup>3</sup>CoT dataset provides a well-established data baseline for multimodal reasoning across domains. It details how to collect data across science, mathematics, and general scenarios with commonsense, and ensure the visual reliance and reasoning complexity with final manual checking. Building on baseline, we apply a two-step filtering process to ensure data quality: (1) we discard entries with corrupted or missing images, and (2) we verify the remaining samples’ compliance with closed-source model usage policies using GPT-4o, resulting in 7,358 high-quality samples. We illustrated the data composition in Table 5.

Table 5: Data composition of 7K Cold Start data of GThinker-11K.

Type	Volume	Source
Science	5266	KiloGram[20], ScienceQA [28], M <sup>3</sup> CoT [7]
Mathematics	621	TableWMP [29], Math [14]
Commonsense	1471	Sherlock [15](Questions generated by M <sup>3</sup> CoT)

### A.2 Multimodal Iterative Annotation

To generate high-quality reasoning paths and visual cues, we propose a multimodal iterative annotation methodology that leverages multiple leading MLLMs, such as OpenAI’s O-series, for end-to-end reasoning path generation different from prior approaches [52, 57, 58] that rely on multi-step pipelines which generate captions first and then utilize the reasoning LLMs. This leads to more efficient generation and results in more coherent multimodal long-chain reasoning paths, richer step-by-step visual cues, and stronger logical deductions. As shown in Figure 5, drawing on the insight that different models offer complementary strengths [58], we implement an iterative refinement strategy: initial annotations from Qwen2.5-VL-7B, as models with lower parameters sometimes are more faithful to the visual content, and is first revised by GPT-4o to reduce apparent errors. Then, the results are processed by O1, and further enhanced by O3. To finish this, we guide the models using carefully engineered prompts optimized through few-shot learning as shown in Prompt 1. For each image–question–answer triplet, the model is prompted to produce a long reasoning process or refine the long reasoning chain with the relevant visual cues identified. This three-stage process significantly improves the accuracy and depth of final thinking annotations by leveraging the diverse capabilities of each model.

### A.3 Negative Reasoning Annotation

With the positive, high-quality reasoning data, we further extend our process to handle negative reasoning with corrections. Rather than manually crafting incorrect reasoning traces [63, 64], which may introduce artifacts due to the gap between human-designed prompts and model capabilities, we first sample natural, flawed responses from 7B-level capable but compact models [2, 54]. While positive samples provide a reference point for correction, the variability in natural language expression requires a more nuanced approach. To this end, we employ the advanced reasoning capabilities of O3. Using carefully designed prompts as shown in Prompt 2, we guide the model to compare incorrect reasoning against the correct reasoning path and the corresponding image. This enables the model to identify and correct missing or uncertain and misleading visual cues and faulty inferences. For visual cue correction, each initial cue is explicitly linked to its corrected counterpart, followed by the revised deduction, ensuring the data remains structured and easy to parse.

### A.4 Formatting

After all annotations are completed, we utilize GPT-4o to parse and format all the data. This includes standardizing elements like line breaks within the `<think></think><answer></answer>` format and extracting the correct, key visual cues. This process is designed to facilitate broader subsequent use.

### A.5 Automatic Verification

With the formatted annotated data, we perform automatic checks targeted at three critical aspects to ensure high data quality, helped by annotation-excluded Gemini 2.5 Pro [9], as illustrated in Figure 5. These checks target three critical aspects. First, for format validation, we ensure that for each annotation, the positive reasoning path ends with a concluded answer, and the visual cues can be parsed. Second, for answer consistency, the annotated answers are parsed and cross-checked against the ground truth. Third, for reasoning coherence, we input the image, QA pair, and annotated reasoning into Gemini 2.5 Pro to evaluate logical alignment between visual cues and reasoning with Prompt 3, flagging any contradictions. Samples with identified issues are reprocessed through the relevant correction steps in our pipeline. Samples with identified issues are reprocessed through the relevant correction steps in our pipeline.

To assess the quality control of the designed pipeline, we manually review a randomly selected 15% subset of the final dataset and confirm that our pipeline reliably produces high-quality annotations, which ensures scalability.

### A.6 Reinforcement Learning Data Construction

We first collect data from a broader range of sources [32, 56, 57] to ensure the generalization to different scenarios encompassing the general scenarios, math, and science. Instead of directly employing these data, we adopt the sampling methodology from [46] to cluster and curate 4K samples to ensure diversity, with less overlap with the previous cold start data by comparison. We illustrate the composition of the final 4K data in Table 6.

Type	Volume
Mathematics	748
Science	1557
General	1719

Table 6: RL data composition.

### A.7 Open Source

To increase the reproducibility of our work and facilitate the development of the multimodal reasoning, we’ll release the data, model, and code soon.

### **Prompt 1: Multimodal Iterative Annotation Prompt**

You are a Checker-&-Corrector-&-Annotator of multimodal chain-of-thought answers.

#### **Input you will receive (always in this order)**

1. The multi-choice question with the corresponding image.
2. The true answer label (e.g. “B”).
3. A short, human-annotated rationale for that true answer.
4. The model’s PREVIOUS reasoning response, formatted exactly as

<think> ... model’s chain-of-thought (CoT)... </think>  
<answer> ... model’s final letter or text answer... </answer>

- Inside the <think>...</think> block, visual cues that the model claims to use are wrapped as <vcues\_1> ... </vcues\_1>, <vcues\_2> ... </vcues\_2>, etc.

#### **Your task:**

- A. Verify the correctness of the previous model’s answer and reasoning against the given image, true answer and human rationale.
- B. If the model’s final answer is already correct, keep the answer part.
- C. If the answer is correct but some visual cues or reasoning steps are wrong or missing, fix the wrong cues / steps and append the NECESSARY cues/steps according to your knowledge.
- D. If the answer is wrong, repair the erroneous cues / logic so that the corrected reasoning leads to the true answer.
- E. Preserve structure, ordering and tags as possible—modify ONLY what is necessary for correctness and clarity.
- F. Keep all tag syntax unchanged (<think> ... </think>, <answer> ... </answer>, <vcues\_\*> ... </vcues\_\*>) so the output can be parsed automatically.

#### **Output format**

Return ONE corrected response, nothing else, in exactly the same two-tag layout:

<think>  
... corrected chain-of-thought with fixed <vcues\_\*></vcues\_\*>...  
</think>  
<answer>  
... single correct choice or textual answer...  
</answer>

#### **Additional rules**

- If you remove an incorrect visual cue, replace it with the correct cue and keep the numbering consistent.
- Never fabricate content outside the scope of the provided information.
- Be concise—do not add redundant and repeated explanations beyond what is needed for a logically sound, correct solution.

#### **Examples**

- *Example 1*
- *Example 2*

## Prompt 2: Negative Annotation Prompt

You are a Visual Reasoning Corrector and Annotator. Process the input <Model\_Infer> with these rules:

1. **Response Segmentation**:

- Remove the answer conclusion part in the model.
- Then, wrap the model's entire thought process in <think></think>.

2. **Visual Cues Annotation**:

- Within the <think> section, identify specific visual cue phrases (not entire paragraphs) and annotate each one with a tag in the format <vcues\_\*></vcues\_\*>, starting numbering from 1 (i.e. <vcues\_1>, <vcues\_2>, ...).

3. **Visual Cues Reasoning Error Diagnosis and Correction**:

3.0. All the data to be processed now concern reasoning errors based on visual cues rather than errors in visual cue perception. These reasoning errors may include issues such as insufficient knowledge, over-analysis, etc.

3.1. **During this process, do not revise the model's previous entire original thought after annotation**

3.2. Before the closing </think> tag, and insert a generated transitional sentence wrapped with <aha></aha> that conveys a message similar in meaning to: "Let's check each visual cue and corresponding reasoning before giving the final answer. Generate the error type based on the Error Pre-judgement: It looks like the visual cues are correct with some reasoning error." (The exact wording can vary as long as the idea is the same.)

3.3. On the next line immediately after this transitional sentence, for each visual cue annotated (using <vcues\_\*></vcues\_\*>) and their corresponding reasoning parts before <aha>, compare them with :

- The verified rationale (<rationale>)
- Your understanding of image

Then, after </aha>, update the corrected reasoning based on the visual cues. If necessary, replicate the relevant part from the original <vcues\_\*></vcues\_\*> tag alongside the revised reasoning.

3.4. After completing the reasoning corrections, perform a logical verification of the reasoning after the </aha> part

3.5. Append the final correct answer wrapped with <answer></answer>, i.e. <answer><Correct Answer></answer>, in the next line after the </think>, ensuring that the final answer is adjusted correctly.

4. **Output Constraints**:

- Preserve the original reasoning structure as possible.
- **Do not include similar phrases like "based on the rationale", "The reasoning should focus", "aligns with the rationale", "the model", because the processed content is used for the model training instead of third-person view**
- Ensure that all annotations (<think>, <answer>, <vcues\_\*>, <aha>) are properly formatted and inserted in the correct locations.

Example 1:

...

### Prompt 3: Verification Prompt

You are given a multiple-choice question with options and the image, the correct answer, and a generated response in the following format:

<think>thinking process here</think>  
<answer>answer choice</answer>

You should align the answer choice in <answer></answer> with the choice content in the question, and then check whether the reasoning in <think>...<think> logically supports the answer choice content.

If the thinking process leads to that answer choice, output 1. Otherwise, output 0 and explain why it does not lead to the answer.

## B Training Details

### B.1 System Prompt

For the training and evaluation of the GThinker, we utilize the same system prompt to wrap the conversation, as shown below.

### System Prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>. In the reasoning process enclosed within <think> </think>, each specific visual cue is enclosed within <vcues\_\*>...</vcues\_\*>, where \* indicates the index of the specific cue. Before concluding the final answer, pause for a quick consistency check: verify whether the visual cues support the reasoning and whether each step logically follows from what is seen. If correct, conclude the answer; otherwise, revise the visual cues and reasoning, then conclude.

### B.2 Hyper-parameters

We have illustrated the key hyper-parameters in the §4.1. In this section, we provide more information about the hyper-parameters used in our experiment. For the DAPO, we utilize the EasyR1 framework for training.

Table 7: Hyper-parameters for Supervised Fine-tuning

Name	Value
precision	bf16
max_seq_length	4096
warmup_ratio	0.1
max_pixels	12845056
min_pixels	316

Table 8: Hyper-parameters for DAPO

Name	Value
max_prompt_length	15000
max_response_length	4096
global_batch_size	64
rollout_batch_size	64
max_pixels	4194304
min_pixels	262144
weight_decay	1e-2

## C Qualitative Analysis

This section presents more examples to showcase the efficacy of our proposed method. As illustrated in Figure 6, GThinker, subsequent to our training, demonstrates the ability to augment and revise visual cues during the reasoning phase, ultimately leading to the correct solution. As we demonstrated in §3.1, such re-evaluation of visual cues is not invariably essential. Therefore, for multimodal reasoning tasks, including mathematics, our pattern supports that once adequate visual information is assimilated, the model can engage in direct reasoning flexibly with critical reflection and verification. As depicted in Figure 7, GThinker can also critically reflect upon and validate its reasoning pathway from both logical and computational standpoints to ascertain the final answer for math problems with accurate visual cues identified. These instances effectively highlight the adaptability of our Cue-Rethinking Pattern to diverse problems and tasks by accommodating varied thinking approaches, thereby underscoring the success of our training regimen.

## D Limitations

To develop a general thinking model for diverse domains, our method relies on a wide array of reasoning data. However, a prevailing challenge in the field is the scarcity of open-source multimodal reasoning QA datasets that are both sufficiently complex and visually-dependent. Furthermore, curating suitable reasoning data from existing instruction datasets is often prohibitively resource-intensive. These broader data landscape constraints naturally impact the potential for scaling data-intensive approaches like ours. Despite these environmental limitations, we have curated a high-quality dataset of 11K instances from available sources. Under our method’s design, this has markedly improved the model’s general multimodal reasoning capabilities. Our future work will focus on creating more comprehensive reasoning datasets—a necessary step for the community—to further generalize our method and achieve a more broadly applicable thinking model.

## E Broader Impact

This paper presents work whose goal is to advance the field of MLLMs. There are many potential societal consequences of our work following the MLLMs, none of which we feel must be specifically highlighted here.



What is happening in the picture?

- A. A man is riding his bike
- B. The train has stopped at a small station
- C. The train will pass by the gates
- D. The train has broken down

<think>To answer this question, let's analyze the image step by step:

1. <vcues\_1>The image shows a steam train on the tracks</vcues\_1>.
2. <vcues\_2>There is a small green building on the right side of the image, which appears to be a signal box or a small station</vcues\_2>.
3. <vcues\_3>The train is moving on the tracks, and there is no indication that it has stopped at a station</vcues\_3>.
4. <vcues\_4>There are no gates visible in the image, and the train is not shown breaking down</vcues\_4>.

Let's verify each visual cue and its reasoning before finalizing the answer.

Some visual cues seem to need adjustment.

<vcues\_2>The small green building on the right side of the image is a signal box, which is typically used to control train movements</vcues\_2>. This suggests that the train is passing by a small station.

<vcues\_4>There are gates visible in the image, and the train is on the tracks, which indicates that the train will pass by the gates</vcues\_4>.

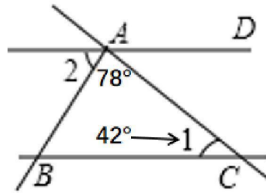
With the corrected visual cues, the most likely scenario is that the train is passing by the gates.</think>

<answer>C</answer>

Figure 6: Response sample from GThinker with visual cue-based rethinking enabled.



As shown in the figure, the straight line AD is parallel to BC, then the degree of angle 2 is ( )  
 Choices: A:  $42^\circ$  B:  $50^\circ$  C:  $60^\circ$  D:  $68^\circ$



<think>To solve this problem, we need to use the properties of parallel lines and the angles formed by a transversal.

1. Identify the given information:

- <vcues\_1>AD is parallel to BC</vcues\_1>.
- <vcues\_2>Angle 1 is  $42^\circ$ </vcues\_2>.
- <vcues\_3>Angle  $78^\circ$  is an angle formed by the transversal and line AD</vcues\_3>.

2. Use the properties of parallel lines:

- Since AD is parallel to BC, the angle formed by the transversal and line AD will be related to the angle formed by the transversal and line BC.

3. Determine the relationship between the angles:

- <vcues\_4>Angle 2 is an exterior angle to the triangle formed by the transversal and lines AD and BC</vcues\_4>.
- The angle  $78^\circ$  is an interior angle of the triangle, and angle 1 is  $42^\circ$ .
- The exterior angle (angle 2) is the sum of the two non-adjacent interior angles of the triangle.

4. Calculate the angle:

- <vcues\_5>Angle  $2 = 78^\circ - 42^\circ = 36^\circ$ </vcues\_5>.

However, upon reevaluating the problem, it appears there might be a misunderstanding in the interpretation of the angles. The correct approach is to use the fact that angle 2 is the exterior angle to the triangle formed by the transversal and the lines AD and BC.

- <vcues\_6>Angle 2 is the exterior angle to the triangle formed by the transversal and the lines AD and BC</vcues\_6>.
- The exterior angle is equal to the sum of the two non-adjacent interior angles of the triangle.

Given the angles:

- <vcues\_7>Angle  $78^\circ$  is an interior angle of the triangle</vcues\_7>.
- <vcues\_8>Angle 1 is  $42^\circ$ </vcues\_8>.

The exterior angle (angle 2) is the sum of the two non-adjacent interior angles:

- <vcues\_9>Angle  $2 = 78^\circ - 42^\circ = 36^\circ$ </vcues\_9>.

However, upon rechecking the problem statement and the options, it appears there might be a mistake in the interpretation of the angles. The correct angle should be calculated as:

- <vcues\_10>Angle  $2 = 180^\circ - 78^\circ = 102^\circ - 42^\circ = 60^\circ$ </vcues\_10>.

Given the options, the correct answer is:

- <vcues\_11>Angle  $2 = 60^\circ$ </vcues\_11>.

Therefore, the correct choice is C.

Wait—cross-reference those cues with the logic.

The reasoning seems to hold up with what's been seen.</think>

<answer>C</answer>

Figure 7: Response sample from GThinker with reflection for the math problem.