

WeThink: Toward General-purpose Vision-Language Reasoning via Reinforcement Learning

Jie Yang¹ Feipeng Ma^{1,3} Zitian Wang Dacheng Yin¹ Kang Rong¹ Fengyun Rao¹ Ruimao Zhang^{2,✉}
¹WeChat Vision, Tencent ²Sun Yat-sen University ³University of Science and Technology of China

<https://github.com/yangjie-cv/WeThink>

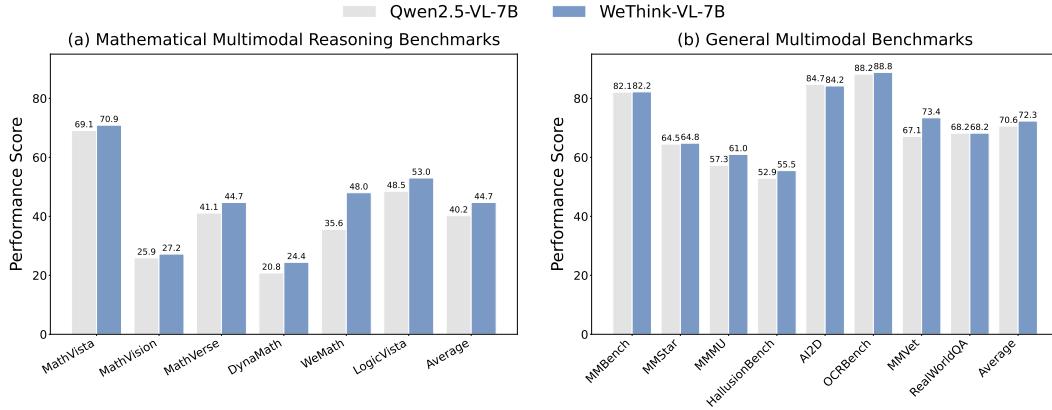


Figure 1: WeThink-VL-7B, fine-tuned on Qwen2.5-VL-7B [1] through reinforcement learning, shows significant improvements in tasks from mathematical reasoning to general challenges.

Abstract

Building on the success of text-based reasoning models like DeepSeek-R1, extending these capabilities to multimodal reasoning holds great promise. While recent works have attempted to adapt DeepSeek-R1-style reinforcement learning (RL) training paradigms to multimodal large language models (MLLM), focusing on domain-specific tasks like math and visual perception, a critical question remains: ***How can we achieve the general-purpose visual-language reasoning through RL?*** To address this challenge, we make three key efforts: (1) A novel *Scalable Multimodal QA Synthesis* pipeline that autonomously generates context-aware, reasoning-centric question-answer (QA) pairs directly from the given images. (2) The open-source WeThink dataset containing over 120K multimodal QA pairs with annotated reasoning paths, curated from 18 diverse dataset sources and covering various question domains. (3) A comprehensive exploration of RL on our dataset, incorporating a hybrid reward mechanism that combines rule-based verification with model-based assessment to optimize RL training efficiency across various task domains. Across 14 diverse MLLM benchmarks, as shown in Fig. 1, we demonstrate that our WeThink dataset significantly enhances performance, from mathematical reasoning to diverse general multimodal tasks. Moreover, we show that our automated data pipeline can continuously increase data diversity to further improve model performance.

1 Introduction

Visual-Language Reasoning has emerged as a pivotal capability for multimodal large language models (MLLMs), enabling tasks ranging from complex mathematical problem-solving to general visual question answering. Closed-source models like OpenAI’s o3 [2] and Kimi k1.5 [3] have demonstrated remarkable performance in visual-language reasoning, sparking significant interest within the open-source community. In contrast, recent open-source initiatives such as DeepSeek-R1 [4] have pioneered text-centric reasoning models by integrating reinforcement learning (RL) with verifiable rewards. However, these models [4, 5, 6] are inherently constrained to unimodal (*i.e.*, text-only) scenarios, leaving a critical gap in multimodal reasoning capabilities.

Recent works [7, 8, 9, 10, 11, 12, 13, 14, 15] have attempted to adapt DeepSeek-R1-style RL training paradigms to MLLMs, focusing primarily on domain-specific tasks like mathematical reasoning and visual perception. Yet, a key question persists: ***How can we achieve the general-purpose visual-language reasoning through RL?*** Two critical aspects stand out.

- **Diverse Reason-centric Data.** Recent DeepSeek-R1-style methods rely on pre-collected question-answer (QA) datasets for cold-start supervised fine-tuning (SFT) with Chain-of-Thought (CoT) annotations or for reformulating answers to calculate accuracy rewards in RL. However, these QA pairs often lack the multi-step reasoning needed for robust visual-language reasoning. Additionally, some methods are dependent on domain-specific question types, which limits their scalability across various domains. To enable more general visual-language reasoning, it’s important to use diverse, reason-focused data from a wide range of domains and contexts.

- **RL with Hybrid Rewards.** While rule-based rewards (*e.g.*, answer verification for mathematical problems) are effective in specific domains, they struggle to capture the complexity of general multimodal scenarios, where answers can be subjective or context-dependent. This underscores the need for a hybrid reward system that combines both rule-based and model-based strategies, offering more nuanced, context-sensitive feedback to enable RL-trained MLLMs to handle diverse task domains.

To address the data aspect, we propose an novel ***Scalable Multimodal QA Synthesis*** pipeline that can autonomously generate context-aware, reason-centric questions paired with verifiable answers directly from the given images. It can benefit from diverse data sources, including open-source datasets and various resources across the Internet, enabling the continuous enhancement of data diversity. To further contribute to the field, we open-source the WeThink dataset, which contains over 120K multimodal QA pairs with explicit reasoning paths. Curated from 18 distinct public image datasets, WeThink encompasses a broad range of question domains and types, requiring integrated abilities such as *reasoning*, *OCR*, *recognition*, *math*, *knowledge*, and *spatial awareness*, thereby enhancing general multimodal reasoning capabilities.

Building upon our dataset, we conduct a comprehensive exploration of RL on WeThink, introducing a hybrid reward mechanism that integrates rule-based verification with model-based evaluation to enhance RL training efficiency across diverse task domains. Through extensive experiments, we present four key findings: **(1)** SFT with CoT supervision on our dataset enhances the performance of less optimized model (*e.g.*, Qwen2-VL-7B), yielding an average improvement of 3.5% across six mathematical reasoning benchmarks. **(2)** In our scenarios, direct RL fine-tuning on Qwen2.5-VL-7B [1] is sufficient and even outperforms cold-start supervised fine-tuning followed by RL. **(3)** With our dataset, increasing the diversity of question domains through RL fine-tuning leads to significant improvements across tasks, from mathematical reasoning to general tasks, as shown in Fig. 1. **(4)** The scalability of our data pipeline enables continuous collection of diverse images from the Internet, further enhancing model performance.

In summary, the contributions of this work are three-fold:

- **Automated Data Generation Pipeline:** We propose an novel ***Scalable Multimodal QA Synthesis*** pipeline that autonomously generates context-aware, reason-centric questions paired with verifiable answers directly from the given images.

- **Diverse Reason-centric Dataset:** We open-source the WeThink dataset, containing over 120K multimodal QA pairs with explicit reasoning paths, curated from 18 distinct public image datasets. It spans various question domains and types, enhancing multimodal reasoning capabilities in models.

- ◊ **General-Purpose Visual-Language Reasoning Models:** We conduct a comprehensive exploration of reinforcement learning on our dataset and present a series of models that improve performance in tasks from mathematical reasoning to general challenges. Additionally, we show that our pipeline’s scalability, driven by increasing data diversity, further improves model performance.

2 Related Work

2.1 Multimodal Large Language Models (MLLMs)

Recent years have witnessed significant advancements in Multimodal Large Language Models (MLLMs), which augment traditional Large Language Models (LLMs) by enabling them to process and comprehend information from diverse modalities, including text, images, audio, and video [16, 17, 18, 19, 20]. The rapid evolution in this field is evidenced by the development of numerous open-source models, such as MiniGPT-4 [21], MiniCPM-V [22], CogVLM [23], ShareGPT4V [24], Qwen-VL [25, 26, 1], LLaVA [16, 27, 28], and InternVL [29, 30, 31], alongside prominent closed-source models like Gemini [32, 33], GPT-4o [34], Claude [35], and Grok [36]. These efforts highlight ongoing progress in architectural designs, pre-training strategies, and instruction tuning techniques. Despite these strides, fostering robust reasoning capabilities across diverse domains and tasks continues to be a primary area of research.

2.2 Chain-of-Thought Prompting for Multimodal Reasoning

Chain-of-Thought (CoT) prompting, a technique that significantly enhances the reasoning capabilities of Large Language Models (LLMs) by guiding them to articulate intermediate inferential steps prior to delivering a final answer [37, 38], has been naturally and effectively extended to the multimodal domain. In Multimodal Large Language Models (MLLMs), the application of CoT not only demonstrably improves performance on complex reasoning tasks but also offers enhanced interpretability into the model’s intricate decision-making processes [39, 40, 41]. A variety of strategies have been developed to elicit, generate, and leverage CoT reasoning in MLLMs. These include designing structured reasoning templates or programmatic approaches to systematically guide the CoT process [42, 43, 44, 45, 46, 47] and Supervised Fine-Tuning (SFT) using datasets enriched with multimodal CoT examples [48, 49, 50]. As these strategies often generate pre-defined or limited thought processes, there is a growing focus on integrating them with reinforcement learning, to encourage exploration of diverse problem-solving strategies, and ultimately develop more sophisticated and genuinely intelligent multimodal reasoning capabilities.

2.3 Reinforcement Learning for Multimodal Reasoning

Reinforcement Learning (RL) has emerged as a transformative approach for enhancing reasoning capabilities in Multimodal Large Language Models (MLLMs) [51]. The integration of RL, particularly Reinforcement Learning from Human Feedback (RLHF) [52] or rule-based reward mechanisms (R1-style) [4], aims to align MLLM outputs with desired reasoning patterns and task objectives. Recently, several works have successfully adapted and extended the R1-style RL training paradigm into MLLMs. These efforts have primarily focused on exploring how R1-style RL can enhance MLLM capabilities in math-centric multi-modal reasoning [7, 8, 9, 10, 11, 12, 13, 14, 15] and various specific downstream tasks. For instance, researchers have applied these RL techniques to improve scene graph understanding [53, 54], visual-spatial reasoning [55, 56, 57, 58], referring expression comprehension [59, 60, 61, 62, 63], and visual counting [64, 65, 66]. While these methods have demonstrated promising results within their respective scopes, few have focused on leveraging RL to broadly enhance the general multi-modal understanding and reasoning abilities of MLLMs.

3 WeThink Dataset with Scalable Multimodal QA Synthesis

This section presents an automated *Scalable Multimodal QA Synthesis* pipeline, designed to generate context-aware, reasoning-centric question-answer (QA) pairs from the given images. We also introduce WeThink, a dataset carefully curated to encompass diverse question domains, types, and integrated abilities. Below, we describe the processes of data collection, question formulation, answer construction, and quality control, and conclude by presenting the data characteristics of WeThink.

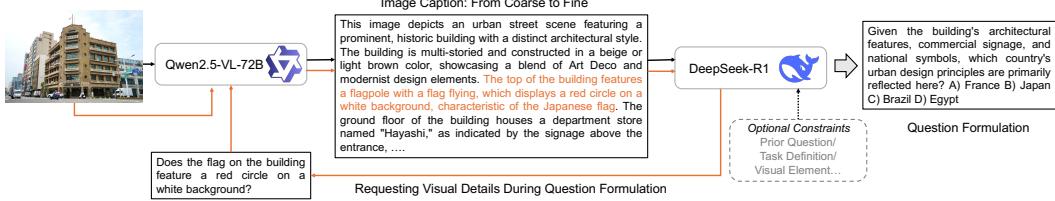


Figure 2: The automatic process of question formulation for a given image. As illustrated by the **orange** line, based on the coarse description provided by Qwen2.5-VL-72B, DeepSeek-R1 needs to request additional visual details (**orange** text) through multi-turn conversations with Qwen2.5-VL-72B, thus facilitating the generation of context-aware, reasoning-centric questions. We also highlight that the process can condition various constraints through prompts, such as prior questions (if available), task definition, and visual focus, to control the type and focus of the questions.

3.1 Data Collection

Our pipeline is designed to autonomously generate high-quality QA pairs directly from the given images. These images can come from open-source datasets or various sources across the Internet. Here, to show its effectiveness, we collect open-source images to publicly release the generated QA pairs. Specifically, we sample images from 18 distinct datasets, as used in LLaVA-CoT [48], shown in Tab. 1. These datasets cover diverse image categories, ensuring variety and complexity in the generated QA pairs, including general images (COCO [67], SAM-1B [68], Visual Genome [69], GQA [70], PISC [71], LLaVA [16]), text-intensive images (TextVQA [72], ShareTextVQA [24], DocVQA [73], OCR-VQA [74], ChartQA [75]), scientific and technical images (GeoQA+ [76], ScienceQA [77], AI2D [78], CLEVR-Math [79]), and images related to art and culture contexts (WikiArt [80, 24], Web-Landmark [81, 24] Web-Celebrity [81, 24]).

3.2 Question Formulation

Based on the collected images, we aim to generate context-aware, reasoning-centric questions. The straightforward pipeline involves collaboration between two powerful models, Qwen2.5-VL-72B [1] and DeepSeek-R1 [4], to analyze images and generate questions. In this workflow, the visual-language model Qwen2.5-VL-72B first provides a detailed description of the image, after which language-only model DeepSeek-R1 analyzes the description, reflects on its content, and synthesizes relevant questions based on the analysis. However, two critical challenges arise: **(1)** incomplete visual understanding by Qwen2.5-VL-72B, **(2)** uncontrolled complexity and reasoning focus of question generation by DeepSeek-R1. To address these issues, we carefully design the question formulation process with two core strategies: *Multi-turn Information Refinements* and *Ability Synergy Constraints*, along with *Optional Contextual Constraints*.

Multi-turn Information Refinements. As shown in Fig. 2, given that image descriptions provided by Qwen2.5-VL-72B may sometimes be insufficient or erroneous, we implement a multi-turn information refinement mechanism to address such shortcomings, including three stages:

- **Coarse Description Generation:** Qwen2.5-VL-72B extracts global features from the input image and generates an initial description that provides a broad overview of the main visual elements. This serves as the semantic anchor for subsequent multi-turn dialogues.
- **Dynamic Detail Mining:** To generate context-aware, reasoning-centric questions, DeepSeek-R1 identifies information gaps based on the initial coarse description. It then generates follow-up questions to request more detailed visual information from Qwen2.5-VL-72B. This process ensures that the questions address all relevant aspects of the image, including the reasoning needed for the final question formulation.
- **Context Integration:** As each piece of supplementary information is gathered during the multi-turn dialogue, it is integrated into the evolving description. Qwen2.5-VL-72B records this information and synthesizes it into a final, fine-grained description, which serves as the basis for generating a comprehensive and contextually aware question.

Ability Synergy Constraints. Inspired by MM-Vet benchmark [82] that evaluates the model’s integrated capabilities, we propose to incorporate multi-ability constraints into the question formulation phase. This approach aims to create more complex questions that better reflect the model’s ability to

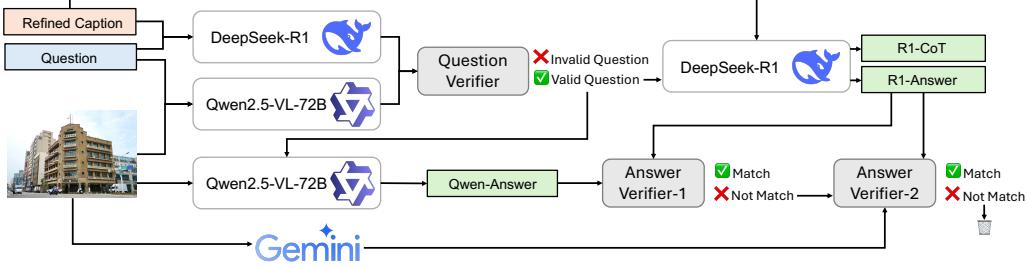


Figure 3: The automatic process of answer construction and quality control. First, DeepSeek-R1 filters out forced and open-ended questions to ensure they are verifiable. Then, using the refined caption and the valid question, DeepSeek-R1 generates a chain of thought and an answer. At the same time, Qwen2.5-VL-72B generates an answer based on the image. If their answers match, the result is kept; if not, Gemini re-evaluates DeepSeek-R1’s answer, discarding incorrect responses and keeping only the correct one with its chain of thought.

apply and combine the skills learned during training. Specifically, we present formulation protocol in Prompt. A-III, with more detail illustrations in Appendix A. The formulation enforces mandatory **reasoning** capability combined with at least one complementary ability from other five-dimensional taxonomy: 1) *Recognition*: General visual recognition (e.g., objects, attributes, scenes, counting, or high-level computer vision tasks); 2) *Knowledge*: Use of social/visual commonsense, encyclopedic knowledge, or recent/contextual information; 3) *OCR*: Reading and reasoning over visible text (e.g., scene text, handwritten text, or embedded text in objects); 4) *Spatial Awareness*: Understanding spatial relationships (e.g., object positions, directional/distance logic, layout analysis); 5) *Math*: Performing arithmetic operations, solving equations, or interpreting math-specific notation. Overall, the above mechanism ensures questions inherently require:

- **Cross-modal Reasoning Chains**: Minimum two explicit reasoning chains with comprehensive image analysis
- **Semantics-Driven Ability Selection**: Automated activation of relevant abilities based on image content information

For instance, a question requiring **object recognition + historical knowledge** might ask: “Given the architectural style of the building’s columns shown, what historical period does this structure represent?” This combines visual feature extraction with architectural history knowledge.

Optional Contextual Constraints. To achieve better precision and control over the generated questions, particularly in terms of their type and focus, we can optionally condition the generation process with contextual constraints. Fortunately, open-source collections often include QA pairs, and web images typically come with textual descriptions or captions. For instance, these may include prior questions to guide related queries, a clear task definition to direct the goal, and visual cues to highlight specific areas of the image. By leveraging these optional constraints, the pipeline can generate more targeted, relevant, and user-intended questions in a controlled manner.

3.3 Answer Construction and Quality Control

For the generated questions, the process of answers construction follows a structured approach consisting of three key stages: *Preliminary Question Filtering*, *Answer Construction and Quality Control*, as well as *CoT Refinement*. As shown in Fig. 3, each stage is carefully designed to ensure the accuracy and reliability of the answers through a multi-model verification framework.

Preliminary Question Filtering. Considering the instability of question formulation, our first step is to filter out questions that are unverifiable, ambiguous, or irrelevant to the image. In practice, we perform two rounds of verification. Firstly, DeepSeek-R1 uses a refined image caption to filter out invalid questions. Then, we also apply the visual-language model Qwen2.5-VL-72B to analyze the image further and filter out additional invalid questions. The remaining questions are categorized into three types: multiple-choice (MC), fill-in-the-blank (FIB), and descriptive (DES).

Answer Construction and Quality Control. The next stage is to generate and verify answers across different question types. For **MC** and **FIB** questions, which can be verified using rules, DeepSeek-R1 generates answers based on a refined image description, while Qwen2.5-VL-72B

Table 1: The distribution analysis of image types from WeThink.

Image Type	Source Dataset	Images
General Images	COCO	30786
	SAM-1B	12014
	Visual Genome	4414
	GQA	3483
	PISC	1148
Text-Intensive Images	LLaVA	150
	TextVQA	17571
	ShareTextVQA	429
	DocVQA	5805
	OCR-VQA	6485
Scientific & Technical	ChartQA	22865
	GeoQA+	4607
	ScienceQA	3236
	AI2D	12024
Art & Culture	CLEVR-Math	434
	WikiArt	401
	Web-Landmark	256
	Web-Celebrity	319

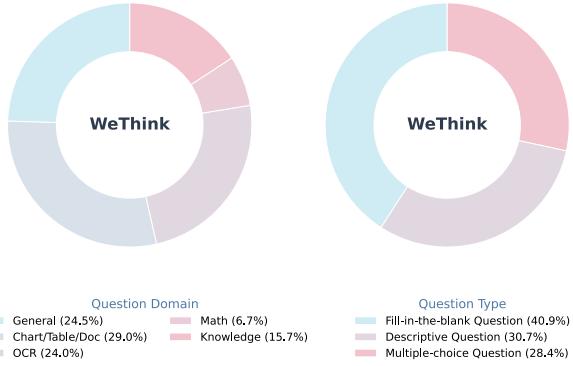


Figure 4: The distribution analysis of question domains and types from WeThink.

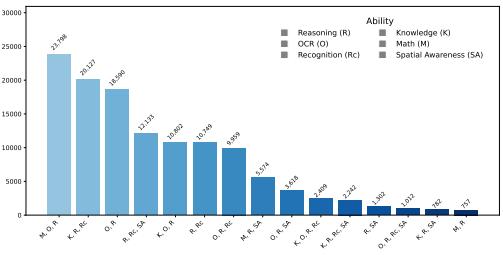
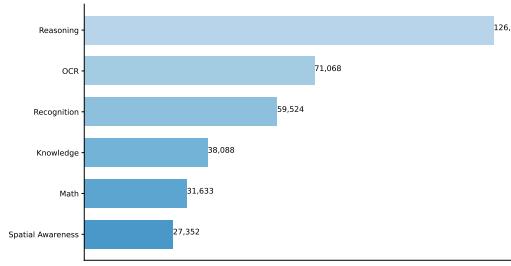


Figure 5: We show the difficulty distribution of question sample in the WeThink dataset.

generates answers from the image content. These answers are then compared for alignment, and if they match, they are considered reliable. In cases of discrepancies, a secondary evaluation by another powerful visual language model Gemini 2.5 Pro [33] is performed to re-evaluate and discard incorrect answers. For **DES** questions, which often require longer, more detailed answers to stimulate reasoning and interpretation of complex visual data, DeepSeek-R1 generates the final answer, and Qwen2.5-VL-72B directly verifies its correctness. If the answer is confirmed, it is retained; otherwise, Gemini re-assesses and filters out incorrect QA pairs.

CoT Refinement. During the answer construction process, DeepSeek-R1 naturally generates a chain-of-thought (CoT) for each question. However, we observed that these CoTs are often overly lengthy and contain redundancies. Our subsequent experiments also revealed that these CoTs are suboptimal for both direct SFT and as cold start data for RL training. To address this, we refine the CoTs by incorporating both the image and the final answer into the QwenVL2.5-72B. This refinement process yields more concise CoTs, allowing us to more effectively investigate how CoT quality influences both SFT and its role as cold-start data for RL training.

3.4 Data Characteristics

Based on the above processes, we have constructed a new dataset named WeThink from open-source images, which offers over **120K** comprehensive multimodal question-answer pairs with explicit reasoning paths. As a diverse and scalable resource, WeThink was carefully curated to encompass a broad range of question domains, types, and required integrated abilities. To better understand the dataset’s structure and focus, we will analyze the following two critical aspects:

Question Distribution. We use Qwen2.5-VL-72B to categorize each question into five groups: general, math, chart/table/doc, knowledge, and OCR. As shown in Fig. 4, these categories are fairly balanced, with math being the least common. The questions are also divided into three types: multiple-choice, fill-in-the-blank, and descriptive. The first two types are suitable for RL training with rule-based rewards, while the third is used for RL training with model-based rewards. Each question type is designed for different scenarios and includes reasoning paths.

Table 2: The impact of fully supervised fine-tuning on WeThink. * denotes the model results reproduced by us. Numbers with underlines indicate models compared to. We highlight the best average results in **bold**, improvements in **red**, and decreases in **blue**.

Model	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	Average
Qwen2-VL-7B	61.6	19.2	25.4	11.0	22.3	33.3	28.8
Qwen2-VL-7B*	61.8	19.0	25.6	11.0	21.4	34.7	28.9
Qwen2-VL-7B-SFT (R1-CoT)	56.6 (-5.2)	17.0 (-2.0)	25.0 (-0.6)	16.0 (+5.0)	18.5 (-2.9)	35.8 (+1.1)	28.2 (-0.6)
Qwen2-VL-7B-SFT (Refined-CoT)	59.5 (-2.3)	20.3 (+1.3)	32.9 (+7.3)	15.2 (+5.2)	25.0 (+3.6)	41.2 (+6.5)	32.4 (+3.5)
Qwen2.5-VL-7B	68.1	25.4	41.1	21.8	36.2	47.9	40.1
Qwen2.5-VL-7B*	69.1	25.9	41.1	20.8	35.6	48.5	40.2
Qwen2.5-VL-7B-SFT (R1-CoT)	61.4 (-7.7)	19.7 (-6.2)	31.5 (-9.6)	21.6 (-0.8)	21.7 (-13.9)	34.2 (-14.3)	31.7 (-8.5)
Qwen2.5-VL-7B-SFT (Refined-CoT)	63.9 (-5.2)	24.0 (-1.9)	40.6 (-0.5)	19.0 (-1.8)	32.4 (-3.2)	40.0 (-8.5)	36.7 (-3.5)

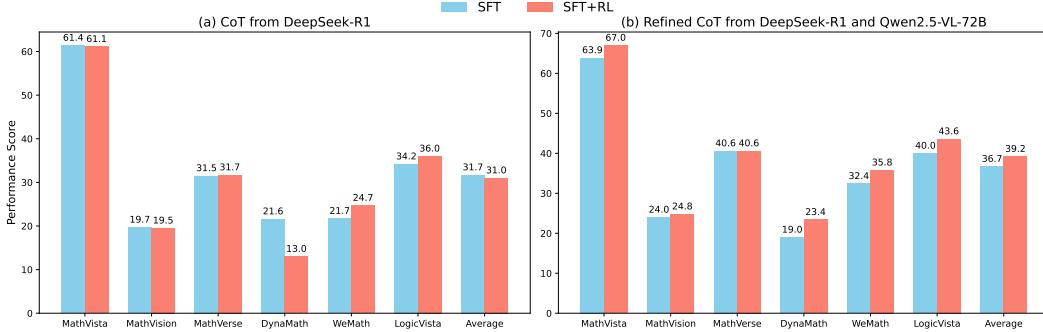


Figure 6: The impact of cold start data quality on performance.

Required Ability Distribution. The questions in the WeThink dataset are designed to integrate multiple abilities, thereby controlling the difficulty level and stimulating training across various model capabilities. As shown in the Fig. 5a, the core ability is reasoning, as every sample activates this ability. However, other abilities are also triggered depending on the semantic contents of the images. Additionally, each sample engages at least two abilities simultaneously, showcasing top-15 ability combinations, as illustrated in Fig. 5b. Notably, these combinations follow a long-tail distribution, with some ability combinations being rarer than others.

4 Experiments

4.1 Benchmarks & Evaluation

To comprehensively evaluate general multi-modal understanding and reasoning capabilities of our models, we conduct experiments across diverse 14 benchmarks, including **(1) Mathematical Reasoning:** We evaluate on six benchmarks: MathVista [83], MathVision [84], MathVerse [85], DynaMath [86], WeMath [87], and LogicVista [88], **(2) General Multimodal Understanding:** This includes MMMU [89] (college-level tasks), MMBench v1.1 [90] (bilingual multimodal benchmark), MMVet [82] (integrated tasks), MMStar [91] (multimodal understanding), HallusionBench [92] (multimodal hallucination), AI2D [78] (science diagrams), and OCRBench [93] (OCR capabilities), RealWorldQA [94] (real-world comprehension). The above benchmarks are available on the Open-Compass MLLM Leaderboard. To maintain fairness and reproducibility, we evaluate our models using VLMEvalKit [95], which is an open-source toolkit for MLLM evaluation. *We present the details of the benchmarks and implementation of the evaluation in Sec. C.2.*

4.2 Exploring Visual-Language Reasoning with Supervised Fine-Tuning on WeThink

Methods & Implementation Details. Inspired by recent works [48] that explore explicit reasoning processes through chain-of-thought (CoT) prompting during supervised fine-tuning (SFT), we conduct experiments using over 120K CoT-annotated diverse QA pairs from our WeThink dataset for direct SFT. *We present the problem definition and detailed optimization formula of SFT in Appendix B.1 and B.2.* We investigate the impact of two types of CoT (*i.e.*, the original R1-CoT and the refined CoT) on powerful open-source models. In practice, we perform full-parameter fine-tuning for 1 epoch

Table 3: The impact of RL with different reward types on all ***math-type*** questions in WeThink.

Reward Type	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	Average
Qwen2.5-VL-7B	68.1	25.4	41.1	21.8	36.2	47.9	40.1
Qwen2.5-VL-7B*	69.1	25.9	41.1	20.8	35.6	48.5	40.2
<i>Rule</i>	65.9	25.1	42.6	24.0	39.1	45.2	40.3 (+0.1)
<i>Model</i>	63.0	24.9	43.3	25.7	31.9	45.6	39.1 (-1.1)
<i>Rule+Model</i>	66.8	26.2	45.7	24.2	37.9	47.4	41.4 (+1.2)

Table 4: The impact of RL with different question types (*i.e.*, ***math-type*** and ***all-type***) in WeThink, comparing performance across both mathematical and general multimodal benchmarks.

Question Type	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	Average
Qwen2.5-VL-7B	68.1	25.4	41.1	21.8	36.2	47.9	40.1
Qwen2.5-VL-7B*	69.1	25.9	41.1	20.8	35.6	48.5	40.2
<i>Math</i>	66.8	26.2	45.7	24.2	37.9	47.4	41.4 (+1.2)
<i>All</i>	71.6	26.7	45.1	24.0	45.5	51.9	44.1 (+3.9)

Question Type	MMBench V1.1	MMStar	MMMU	HallusionBench	AI2D	OCRBench	MMVet	RealWorldQA	Average
Qwen2.5-VL-7B	82.2	64.1	58.0	51.9	84.3	88.8	69.7	68.4	70.9
Qwen2.5-VL-7B*	82.1	64.5	57.3	52.9	84.7	88.2	67.1	68.2	70.6
<i>Math</i>	79.5	63.3	57.2	56.0	83.5	88.3	72.2	67.8	71.0 (+0.4)
<i>All</i>	82.0	64.3	59.3	55.8	84.3	88.9	71.7	67.7	71.8 (+1.2)

using 8 NVIDIA H20 GPUs on two instruction-tuned models: Qwen2-VL-7B and Qwen2.5-VL-7B. The system prompt is shown in Appendix Tab. C.1.

Results. As shown in Table 2, the experimental results reveal two key findings: (1) ***The quality of the SFT CoT is critical.*** The original R1-CoT is overly long and redundant, as it only directs the model to mimic the reasoning templates in the annotated CoT structure. Even with less optimized models like Qwen2-VL-7B, fine-tuning leads to improvements on certain benchmarks (e.g., a 5% improvement on DynaMath and a 1.1% improvement on LogicVista). However, for the more advanced Qwen2.5-VL-7B model, we observe significant performance degradation across all benchmarks. (2) ***Our CoT data is particularly beneficial for less optimized models.*** Specifically, for Qwen2-VL-7B, the fine-tuning results in an average improvement of 3.5%. In contrast, applying direct SFT to the well-optimized Qwen2.5-VL-7B leads to a drop in performance.

4.3 Exploring Visual-Language Reasoning with Reinforcement Learning on WeThink

Previous experiments relying solely on CoT supervision reveal its insufficiency in exploring visual-language reasoning, particularly for powerful base models. Reinforcement Learning (RL), as a more advanced training paradigm compared to CoT, has shown its effectiveness in the text-only domain, particularly with the open-source DeepSeek-R1 model. This motivates us to investigate how to harness WeThink for effective RL training in the visual-language context, as detailed below:

Methods & Implementation Details. We adapt the DeepSeek-R1-style framework by introducing a group-relative policy optimization algorithm (GRPO) [96] and a hybrid reward system. The hybrid reward system combines accuracy and format rewards. The accuracy reward is divided into rule-based rewards for multiple-choice and fill-in-the-blank questions in WeThink, and model-based rewards for descriptive answers, assessed using the DeepSeek-V3 judge model [97]. The format reward checks for proper thinking and answer formatting in the response. The final reward combines accuracy and format rewards, with coefficients adjusting their relative importance. ***Please refer to Appendix B.3 for the method details of GRPO algorithm and reward functions.*** Our implementation is built on EasyR1 [98], which is based on veRL [99]. We perform 5 samples per query with a temperature setting of 1.0. We select Qwen2.5-VL-7B-Instruct as our base model and perform full-parameter RL fine-tuning, with rollout

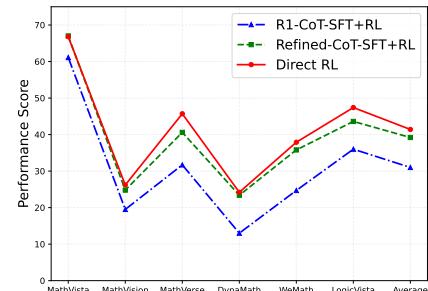


Figure 7: The impact of cold start SFT for RL on all ***math-type*** questions in WeThink.

Table 5: Comparison of different VLMs on mathematical multimodal reasoning benchmarks. We report results from other models on the OpenCompass leaderboard [95] and reproduce the results of our base model, Qwen2.5-VL-7B, denoted with *. † indicates the use of external images from the Internet to further enhance diversity.

Model	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	Average
LLaVA-OneVision-7B	58.6	18.3	19.3	9.0	20.9	33.3	26.6
InternVL2-8B	58.3	20.0	20.4	9.2	20.2	33.6	26.9
InternVL2.5-8B	64.5	17.0	22.8	9.4	23.5	36.0	28.9
Qwen2-VL-7B	61.6	19.2	25.4	11.0	22.3	33.3	28.8
Qwen2.5-VL-7B	68.1	25.4	41.1	21.8	36.2	47.9	40.1
Qwen2.5-VL-7B*	69.1	25.9	41.1	20.8	35.6	48.5	40.2
WeThink-VL-7B	71.6	26.7	45.1	24.0	45.5	51.9	44.1 (+3.9)
WeThink-VL-7B†	70.9	27.2	44.7	24.4	48.0	53.0	44.7 (+4.5)

Table 6: Comparison of different VLMs on general multimodal understanding benchmarks.

Model	MMBench V1.1	MMStar	MMMU	HallusionBench	AI2D	OCRBench	MMVet	RealWorldQA	Average
LLaVA-OneVision-7B	76.8	56.7	46.8	47.5	82.8	69.7	50.6	69.9	62.6
InternVL2-8B	79.4	61.5	51.2	45.0	83.6	79.4	54.3	64.2	64.8
InternVL2.5-8B	82.5	63.2	56.2	49.0	84.6	82.1	62.8	69.5	68.7
Qwen2-VL-7B	81.0	60.7	53.7	50.4	83.0	84.3	61.8	68.5	67.9
Qwen2.5-VL-7B	82.2	64.1	58.0	51.9	84.3	88.8	69.7	68.4	70.9
Qwen2.5-VL-7B*	82.1	64.5	57.3	52.9	84.7	88.2	67.1	68.2	70.6
WeThink-VL-7B	82.0	64.3	59.3	55.8	84.3	88.9	71.7	67.7	71.8 (+1.2)
WeThink-VL-7B†	82.2	64.8	61.0	55.5	84.2	88.8	73.4	68.2	72.3 (+1.7)

and training batch sizes set to 512 and 128, respectively. For ablation studies, we use 8 NVIDIA H20s for all experiments, while 32 NVIDIA H20s is used for full-scale training on WeThink. The DeepSeek-V3 judge model is deployed on 16 NVIDIA H20s and participates in the RL training through API calls for reward computation. The system prompt is shown in Appendix Tab. C.1.

The impact of cold start during RL. DeepSeek-R1 indicates that cold start SFT is instrumental in facilitating RL training. In our experiments, the previous CoT SFT model serves as the model after cold start training. For subsequent RL training, we select all math-type questions from WeThink and incorporate hybrid rewards. As shown in Fig. 6, we observe that using refined CoT as cold start data leads to substantial improvements after RL training. In contrast, applying RL to the model trained with the original R1-CoT yields minimal gains. This clearly illustrates that high-quality cold start data is critical for enhancing RL training. Furthermore, Fig. 7 shows that applying RL directly to Qwen2.5-VL yields better improvements than the cold start method. This could be due to two factors: (1) In our scenarios, Qwen2.5-VL-7B already has strong capabilities to follow our format and requirements, which aligns with previous findings [7], and (2) the quality of our cold start data still requires improvement. Because the cold start SFT initially results in a performance drop, RL does provide some improvements, but the gains are not highly significant.

The impact of RL with different reward types. As previously demonstrated, in our scenarios, without cold start SFT, directly applying RL with hybrid rewards on math-type questions in WeThink proves to be the most effective. Here, we divide these questions into rule-based reward questions (*i.e.*, multiple-choice and fill-in-the-blank questions) and model-based reward questions (*i.e.* descriptive questions), and train each model for the same steps. Our results in Tab. 3 show that hybrid rewards yield the best performance, with an average improvement of 1.2% across six benchmarks.

The impact of RL with different question types. To further improve the performance of general visual-language tasks beyond mathematics, it is crucial to integrate a wider variety of data sources that span multiple domains and contexts. Thus, we extend our training to include all question types in WeThink for full-scale RL training with hybrid rewards. As shown in Tab. 4, we compare models trained on math-type questions with those trained on all question types, across 14 diverse benchmarks covering both mathematical and diverse general tasks. The results show that, firstly, the all-type question trained model significantly improves performance on mathematical reasoning benchmarks, with an average increase from 41.4% to 44.1%. Secondly, it substantially enhances performance across general tasks.

The impact of RL with increasing data diversity. To demonstrate the scalability of our data generation pipeline, we collect approximately 20K in-the-wild images from the Internet. Following the same data construction process, we generate new QAs for these images and incorporate them into the training. The results presented in Tab. 5 and 6 show that performance continues to improve across

both mathematical reasoning and general benchmarks. However, the addition of new data leads to a slight decline in some benchmarks, such as MathVista and MathVerse. This suggests that RL training process is sensitive to changes in the data distribution for specific benchmarks. Nonetheless, the average improvements show our pipeline’s scalability in incorporating more diverse data to further enhance model performance.

5 Conclusion

This work advances multimodal reasoning by tackling the challenge of general-purpose visual-language reasoning through reinforcement learning (RL). By introducing a novel *Scalable Multi-modal QA Synthesis* pipeline, we can generate context-aware, reasoning-centric question-answer pairs directly from the given images. We also release the WeThink dataset, containing over 120K multimodal QA pairs with annotated reasoning chains, to enhance RL training across diverse domains. We conduct comprehensive exploration of RL on our dataset, incorporating a hybrid reward mechanism, optimizes training efficiency and performance across domains. The results across 14 diverse MLLM benchmarks demonstrate the effectiveness of our dataset in improving performance from mathematical reasoning to general challenges. Furthermore, the scalability of our automated data pipeline ensures continuous improvement in data diversity, paving the way for scalable RL training across domains and general multimodal reasoning enhancement.

Limitations. *Firstly*, our proposed data pipeline relies on multiple large models, requiring more GPU resources for deployment. *Secondly*, the pipeline generates all questions solely based on the model’s interpretation of the image content, leading to inherent randomness and instability in the generated questions. As a result, some questions need be filtered out due to quality issues. *Thirdly*, although we employ multiple powerful models for question filtering and answer verification, our dataset still contains noise and wrong cases.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [2] OpenAI. Openai o3 and o4-mini system card, 2025. Accessed: 2025-05-14.
- [3] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.
- [4] 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.
- [5] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.
- [6] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.
- [7] 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.
- [8] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Open-vlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025.
- [9] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

- [10] 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.
- [11] 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.
- [12] 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.
- [13] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhua Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.
- [14] Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, Xuchen Song, et al. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*, 2025.
- [15] Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025.
- [16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [17] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.
- [18] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.
- [19] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.
- [20] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- [21] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [22] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- [23] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogylm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.
- [24] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024.
- [25] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

- [26] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [28] 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.
- [29] 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.
- [30] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [31] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [32] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [33] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [34] 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.
- [35] Anthropic. Claude. <https://www.anthropic.com/>.
- [36] xAI. Grok. <https://x.ai/>.
- [37] 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.
- [38] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [39] 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.
- [40] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36:43447–43478, 2023.
- [41] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15630–15640, 2024.

- [42] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.
- [43] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*, 2024.
- [44] 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.
- [45] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.
- [46] Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. *arXiv preprint arXiv:2410.03321*, 2024.
- [47] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1254–1262, 2024.
- [48] 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.
- [49] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.
- [50] 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.
- [51] Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 2025.
- [52] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [53] Lin Li, Wei Chen, Jiahui Li, and Long Chen. Relation-r1: Cognitive chain-of-thought guided reinforcement learning for unified relational comprehension. *arXiv preprint arXiv:2504.14642*, 2025.
- [54] Zuyao Chen, Jinlin Wu, Zhen Lei, Marc Pollefeys, and Chang Wen Chen. Compile scene graphs with reinforcement learning. *arXiv preprint arXiv:2504.13617*, 2025.
- [55] Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. *arXiv preprint arXiv:2504.12680*, 2025.
- [56] Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. Improved visual-spatial reasoning via r1-zero-like training. *arXiv preprint arXiv:2504.00883*, 2025.
- [57] Wenqi Zhang, Mengna Wang, Gangao Liu, Xu Huixin, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, et al. Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696*, 2025.

- [58] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.
- [59] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.
- [60] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.
- [61] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.
- [62] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.
- [63] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.
- [64] Zhiqiang Wang, Pengbin Feng, Yanbin Lin, Shuzhang Cai, Zongao Bian, Jinghua Yan, and Xingquan Zhu. Crowdvlm-r1: Expanding r1 ability to vision language model for crowd counting using fuzzy group relative policy reward. *arXiv preprint arXiv:2504.03724*, 2025.
- [65] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.
- [66] Zhiyuan Liu, Yuting Zhang, Feng Liu, Changwang Zhang, Ying Sun, and Jun Wang. Othink-mr1: Stimulating multimodal generalized reasoning capabilities via dynamic reinforcement learning. *arXiv preprint arXiv:2503.16081*, 2025.
- [67] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.
- [68] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.
- [69] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- [70] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- [71] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2659, 2017.
- [72] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

- [73] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.
- [74] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.
- [75] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.
- [76] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th international conference on computational linguistics*, pages 1511–1520, 2022.
- [77] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, 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.
- [78] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.
- [79] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.
- [80] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.
- [81] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.
- [82] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- [83] 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. *arXiv preprint arXiv:2310.02255*, 2023.
- [84] Ke Wang, Junting 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.
- [85] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.
- [86] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024.
- [87] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, MiaoXuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.
- [88] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

- [89] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [90] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.
- [91] 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? *arXiv preprint arXiv:2403.20330*, 2024.
- [92] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
- [93] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.
- [94] X.AI Corp. Grok-1.5 vision preview: Connecting the digital and physical worlds with our first multimodal model. 2024.
- [95] 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.
- [96] 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.
- [97] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [98] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. <https://github.com/hiyouga/EasyR1>, 2025.
- [99] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*, 2024.

A The Details of Scalable Multimodal QA Synthesis

We formulate the question formulation designs into a prompt protocol for DeepSeek-R1, as illustrated in Prompt. A, which is structured into five parts:

Input Information: The input consists of potentially insufficient or erroneous preliminary visual information from Qwen2.5-VL-72B. DeepSeek-R1 relies on Qwen2.5-VL-72B to continuously refine this visual information and use it as updated input for generating reasoning-based questions through a multi-turn process.

Core Requirements: For each generated question, it is essential that multiple abilities are triggered, ensuring the depth and complexity of reasoning. Typically, this involves at least two reasoning steps, making the question both logical and comprehensive. Furthermore, the question design must rely on thorough image analysis to maintain clarity, depth, and completeness.

Question Designs: To ensure high-quality output, we have established specific guidelines for the selection of abilities. Every question must include reasoning as a mandatory component. In addition, depending on the image content, the question should also incorporate at least one of the following abilities: recognition (feature extraction), knowledge (external knowledge application), OCR (text recognition), spatial awareness (geometric or positional reasoning), or math (numerical reasoning). Moreover, we can optionally condition the generation process with contextual constraints, to achieve better control questions' type and focus.

Information Request Protocol: In practice, when the image description provided by Qwen2.5-VL-72B is insufficient, we activate the Information Request Protocol, which allows for up to three rounds of clarification requests to ensure that the necessary visual information is complete for generating the subsequent questions.

Output Specifications: Finally, the output of the question generation process must adhere to strict format specifications, including clarification requests for insufficient information (e.g., <clarify>...</clarify>) and properly formatted valid questions (e.g., <q>...</q>).

Question Formulation Protocol for Visual Reasoning

I. Input Information

Image Caption: Potentially insufficient/erroneous preliminary visual information

II. Core Requirements

- Multi-ability triggering (≥ 2 combinations)
- Explicit reasoning chains (≥ 2 logical leaps)
- Comprehensive image analysis requirement

III. Question Design

Ability Synergy Constraints

Mandatory: Reasoning

A. Recognition

B. Knowledge

Feature extraction & classification

Domain knowledge application

Others (Select ≥ 1): C. OCR

D. Spatial Awareness

Text recognition & processing

E. Math

Geometric & positional reasoning

Numerical derivation

Optional Contextual Constraints

Prior Question, Task Definition, Visual Cues, etc.

IV. Information Request Protocol

Information Clarification Mechanism

Insufficient Caption → Activate Information Request Protocol

- Max conversation rounds: 3
- Request for more information: <clarify>...</clarify>

V. Output Specifications

Format Requirements

- Request More Details: <clarify>...</clarify>
- Valid Question Formulation: <q>...</q>

B Method Details

B.1 Problem Definition

Given a multi-modal input consisting of a question q and an image I , our goal is to generate the correct answer a by reasoning over both the textual and visual inputs. This reasoning process is mathematically modeled as a sequential conditional probability:

$$P(a | q, I) = \prod_{t=1}^T P(a_t | q, I, a_{<t}),$$

where a_t is the t -th token of the model's output, representing a reasoning step, and $a_{<t}$ is the sequence of previously generated tokens. The model is expected to produce a logically consistent reasoning chain that integrates both the question and image, using elements such as mathematical formulas, contextual clues, and visual features. These reasoning steps should progressively lead to the final, accurate answer, bridging the textual and visual inputs in a structured manner.

B.2 Supervised Fine-Tuning with Chain-of-Thought Prompting

One of the most straightforward methods to enable the model to generate explicit reasoning steps is to use chain-of-thought (CoT) prompting during supervised fine-tuning. This approach requires CoT-annotated data:

$$D_{SFT} = \{(q_i, I_i, r_i, a_i)\},$$

which consists of question-image pairs (q_i, I_i) , along with the corresponding intermediate reasoning steps r_i and the final answer a_i . The model is then trained to generate the reasoning steps r in sequence, which ultimately leads to the correct final answer. We optimize the model directly using maximum likelihood estimation for generating reasoning steps:

$$\mathcal{L}_{SFT} = - \sum_{(q, I, r, a) \in D_{SFT}} \sum_{t=1}^T \log P(r_t | q, I, r_{<t}),$$

where r_t represents the t -th reasoning step and $r_{<t}$ is the sequence of previous reasoning steps. This process ensures that the model learns to generate a coherent and accurate chain of reasoning steps, which together lead to the correct final answer a .

B.3 Reinforcement Fine-tuning with Hybrid Reward

While supervised fine-tuning with chain-of-thought prompting provides explicit step-by-step supervision, reinforcement learning offers a complementary paradigm for optimizing reasoning generation

through reward signals. Inspired by the success of DeepSeek-R1 in text-based reasoning tasks, we adapt this framework for visual-language models by introducing a group-relative policy optimization strategy [96] and a hybrid reward system.

Group-Relative Policy Optimization eliminates value function dependency through reward normalization within response groups. For each question-image pair (q, I) , we sample G reasoning paths $\{o_1, \dots, o_G\}$ from the current policy π_θ . The advantage function is computed as:

$$\hat{A}_{i,t} = \frac{R_i - \mu(\{R_j\}_{j=1}^G)}{\sigma(\{R_j\}_{j=1}^G)}$$

where μ and σ denote the group mean and standard deviation of final rewards. The objective function combines clipped policy updates with KL regularization against the reference policy π_{ref} :

$$J_{\text{GRPO}}(\theta) = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^{|o_i|} \sum_{t=1}^{|o_i|} \min \left(r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_{i,t} \right) - \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right]$$

where $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q, I, o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|q, I, o_{i,<t})}$ is the importance sampling ratio. This approach stabilizes training while encouraging exploration of high-reward reasoning paths.

Hybrid Reward System integrates accuracy and format rewards, similar to DeepSeek-R1, with accuracy reward further divided into rule-based reward and model-based reward to handle different types of answers. For example, in our WeThink dataset, rule-based reward is employed for multiple-choice and fill-in-the-blank questions, while model-based reward is used for descriptive questions that require longer descriptive answers.

- **Rule-Based Reward** (R_{rule}): For multiple-choice and fill-in-the-blank questions, we apply exact string matching between the predicted answer and the ground truth. This is done with text normalization for case and punctuation insensitivity:

$$R_{\text{rule}} = \mathbb{I}(\text{normalize}(a_{\text{pred}}) = \text{normalize}(a_{\text{true}}))$$

where \mathbb{I} is an indicator function, returning 1 for a true condition (exact match) and 0 for false (no match).

- **Model-Based Reward** (R_{model}): For descriptive questions, we use the DeepSeek-V3 [97] judge model to assess answer correctness, assigning rewards based on the clarity and correctness of the response:

$$R_{\text{model}} = \begin{cases} 1 & \text{Definitely correct} \\ 0.5 & \text{Ambiguous/Partially correct} \\ 0 & \text{Definitely incorrect} \end{cases}$$

- **Format Reward** (R_{format}): To ensure the reasoning process is structured correctly, the format reward checks whether the response includes valid thinking and answer blocks, such as `<think></think>` and `<answer></answer>`:

$$R_{\text{format}} = \mathbb{I}(\text{Valid thinking and answer blocks})$$

where \mathbb{I} returns 1 for valid (correct) formatting and 0 for invalid (incorrect) formatting.

The final reward can be computed as:

$$R = \alpha_{\text{accuracy}} \cdot R_{\text{accuracy}} + \alpha_{\text{format}} \cdot R_{\text{format}}$$

where R_{accuracy} is either R_{rule} or R_{model} , depending on the type of question. The α coefficients control the relative importance of the accuracy and format components. We empirically set α_{accuracy} to 0.7 and α_{format} to 0.3.

C Experimental Implementation Details

C.1 System Prompt

To structure the reasoning process during training, we use the following system prompt for both supervised fine-tuning and reinforcement fine-tuning settings as follow:

System Prompt

"You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE enclosed within <answer> </answer> tags."

C.2 Benchmarks & Evaluation.

As illustrated in Sec. 4.1, we conduct comprehensive evaluation across 14 MLLM benchmarks, covering six mathematical reasoning benchmarks and eight general multimodal understanding benchmarks. Below are the details:

Mathematical reasoning employs the following benchmarks: ◊ MathVista_MINI, which is the Test Mini split of the MathVista dataset

- ◊ MathVision, which uses the full test set of MathVision
- ◊ MathVerse_MINI_Vision_Only, the Test Mini split of MathVerse, using the "Vision Only" mode
- ◊ DynaMath, which uses the full test set of DynaMath
- ◊ WeMath, the Test Mini split of WeMath, where we report "Score (Strict)" as the main metric
- ◊ LogicVista, which is the full test set of LogicVista

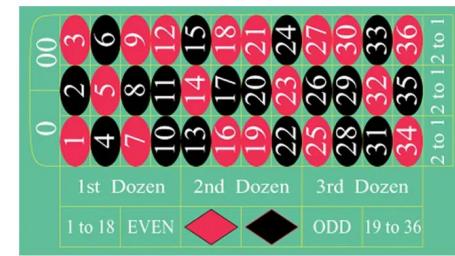
General multimodal understanding employs MMBench v1.1 (Test_CN / Test_EN), MMStar, MMMU (Val), HallusionBench, AI2D (Test), OCRBench, MMVet and RealWorldQA.

Evaluation Details. To ensure fairness and reproducibility, we conduct evaluation of our models using VLMEvalKit [95], an open-source toolkit designed for MLLM evaluation. All evaluations are conducted using 8 NVIDIA A800s. We follow the Qwen2.5-VL series settings, where the minimum number of pixels is set to 1280*28*28, the maximum number of pixels is set to 16384*28*28, and the `use_custom_prompt` option is disabled. The `max_new_tokens` parameter is set to 2048 by default, in alignment with Qwen2.5-VL series. The system prompt is provided in Tab. C.1. We empirically found that using the default system prompt of the Qwen2.5-VL series yields better performance for OCRBench and AI2D benchmarks.

D Case Study: From Math to General Challenges

Below, we conduct case study to investigate how reasoning contributes to performance improvements in diverse scenarios, including *Mathematics, Medicine, Humanities, Science, Art, Engineering* and *Daily Scenes*. These examples show that WeThink-VL-7B exhibits superior logical and analytical capabilities, leading to correct answer.

Mathematics



Question: The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range. Compute the probability of winning the following type of bets: Betting on three numbers in a line, as in 1-2-3

Options:

- A. 1/38
- B. 2/38
- C. 3/38
- D. 6/38

GT Answer: C

Qwen2.5-VL-7B

To determine the probability of winning a bet on three numbers in a line (such as 1-2-3) in roulette, we need to understand the total number of possible outcomes and the number of favorable outcomes.

1. Total number of possible outcomes: There are 38 numbers on the roulette wheel.

2. Favorable outcomes: There is only 1 way to win by betting on the exact three numbers 1-2-3.

The probability P of winning is given by the ratio of the number of favorable outcomes to the total number of possible outcomes:

$$P = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}} = \frac{1}{38}$$

Therefore, the correct answer is:

[A. 1/38](#)

WeThink-VL-7B:

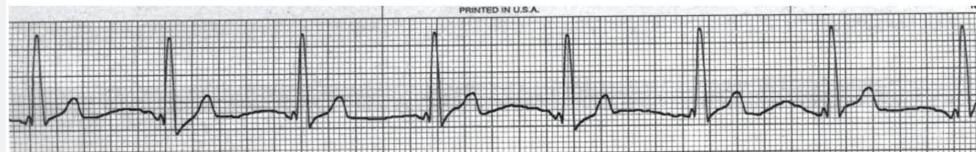
<think> To compute the probability of winning by betting on three numbers in a line, we need to understand that there are 38 possible numbers on the roulette wheel. When you bet on three specific numbers, you are essentially looking for the probability that the ball lands on one of those three numbers. The probability of the ball landing on any specific number is 1/38. Since you are betting on three specific numbers, the probability of winning is the sum of the probabilities of the ball landing on each of those three numbers. However, since these are mutually exclusive events (the ball can't land on more than one number at a time), we simply add the probabilities:

$$\text{Probability} = \frac{1}{38} + \frac{1}{38} + \frac{1}{38} = \frac{3}{38}$$

So, the probability of winning by betting on three numbers in a line is 3/38. </think>

<answer> [C. 3/38](#) </answer>

Medicine



Question: What is this rhythm?

Options:

- A. Normal Sinus Rhythm
- B. Sinus Bradycardia
- C. Accelerated Junctional Rhythm
- D. NSR with 1st Degree AV Block
- E. Idioventricular Rhythm
- F. Paced with Failure to Capture

GT Answer: C

Qwen2.5-VL-7B

The rhythm in the provided ECG appears to be irregular with a slow heart rate. The P waves are not clearly visible, and the QRS complexes are wide and irregular. This suggests a rhythm that is not sinus rhythm. Given the options provided, the most likely diagnosis is:

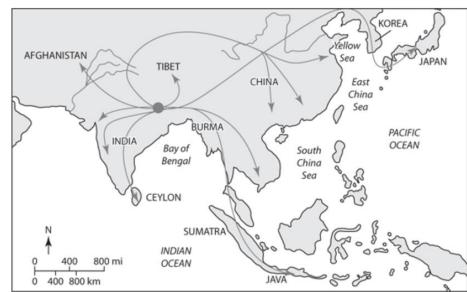
[E. Idioventricular Rhythm](#)

Idioventricular rhythm is characterized by wide, irregular QRS complexes and a slow heart rate, which matches the appearance of the rhythm in the ECG.

WeThink-VL-7B:

<think>The image provided is an ECG (Electrocardiogram) which is used to monitor the electrical activity of the heart. The rhythm in the ECG shows regular, narrow QRS complexes with no P waves preceding them, which is characteristic of junctional rhythms. The QRS complexes are regular and the heart rate appears to be relatively slow, which is not typical for normal sinus rhythm or bradycardia. The absence of P waves and the regularity of the QRS complexes suggest a junctional rhythm. The regularity and the narrow QRS complexes also rule out idioventricular rhythm, which is typically irregular and wide QRS complexes. The paced rhythm with failure to capture is not applicable as there are no artificial pacing spikes visible. The rhythm is not consistent with first-degree AV block as the QRS complexes are narrow and regular. Therefore, the most likely rhythm is an accelerated junctional rhythm. </think>

<answer> C. Accelerated Junctional Rhythm </answer>

Humanities

Question: A historian would use the above map to illustrate which of the following?

Options:

- A. Major trade routes in the classical era
- B. The spread of communism from 1900 to today
- C. One aspect of cultural diffusion
- D. The monsoonal wind patterns of Asia

GT Answer: C

Qwen2.5-VL-7B

The correct answer is:

A. Major trade routes in the classical era

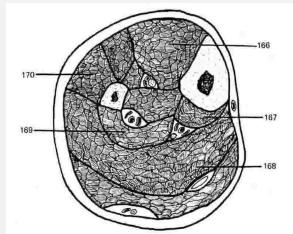
The map shows various routes connecting different regions, which is indicative of trade and cultural exchange patterns. The presence of arrows and the geographical context suggest the movement of goods and ideas, which is characteristic of major trade routes in the classical era.

WeThink-VL-7B:

<think> The map shows various routes connecting different parts of Asia, including China, India, and the islands of Southeast Asia. These routes are likely to represent trade or cultural exchanges rather than wind patterns or the spread of communism. The presence of arrows and the interconnected nature of the lines suggest movement and exchange, which is characteristic of trade routes or cultural diffusion. The map does not specifically indicate monsoonal wind patterns, as it lacks the necessary meteorological data and is more focused on geographical and cultural connections. </think>

<answer> C. One aspect of cultural diffusion </answer>

Science



Question: In the Section of left leg, identify the 170 structure.

Options:

- A. Tibialis anterior
- B. Tibialis posterior
- C. Flexor hallucis longus
- D. Peroneus longus

GT Answer: D

Qwen2.5-VL-7B

The structure labeled as 170 in the image appears to be the tibialis posterior muscle, which is located in the posterior compartment of the leg. Therefore, the correct answer is:

B. Tibialis posterior

WeThink-VL-7B:

<think> The image provided is a cross-sectional view of a leg, likely representing the anatomy of the lower leg. The structure labeled as 170 appears to be located on the outer side of the section, which corresponds to the lateral side of the leg. In the anatomy of the leg, the tibialis posterior is typically located on the medial side, the tibialis anterior on the anterior side, the flexor hallucis longus is deep within the leg, and the peroneus longus is on the lateral side, which matches the location of the structure labeled 170. </think>

<answer> The correct answer is D. Peroneus longus. </answer>

Art



Question: The painting shown was created by:

Options:

- A. Wifredo Lam
- B. Frida Kahlo
- C. Diego Rivera
- D. Jacob Lawrence

GT Answer: D

Qwen2.5-VL-7B:

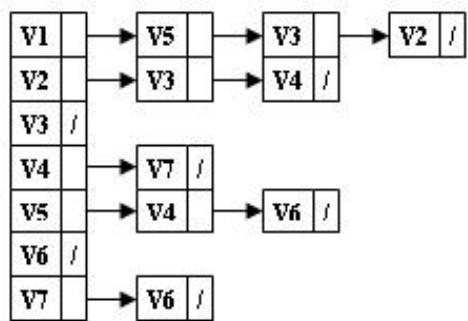
The painting shown is "The Four Stages of a Man" by Wifredo Lam. Therefore, the correct answer is:
[A. Wifredo Lam](#)

WeThink-VL-7B:

<think> The image depicts a painting with a series of figures engaged in various activities, such as reading, writing, and working. The style of the painting, with its flat, almost cut-out appearance of the figures and the use of a single, prominent line that seems to connect the different scenes, is characteristic of the work of Jacob Lawrence. Jacob Lawrence is known for his series of paintings that often depict historical events or social issues, and his style often includes a similar use of bold lines and a focus on the human figure. The other artists listed have distinct styles: Wifredo Lam is known for his surrealist works, Frida Kahlo for her self-portraits and symbolic works, and Diego Rivera for his large-scale murals and political themes. The style of this painting does not align with the known styles of the other artists. </think>

<answer> [D. Jacob Lawrence](#) </answer>

Engineering



Question: The adjacency list of a directed graph is given as follows. Starting from the vertex V1 and traversing according to the depth-first search method, the obtained vertex sequence is:

Options:

- A. V1,V5,V4,V7,V6,V2,V3
- B. V1,V2,V3,V4,V7,V6,V5
- C. V1,V5,V4,V7,V6,V3,V2
- D. V1,V5,V6,V4,V7,V2,V3

GT Answer: C

Qwen2.5-VL-7B:

To determine the correct vertex sequence obtained by traversing the graph using a depth-first search (DFS) starting from vertex V1, we need to follow the adjacency list and the DFS rules:

1. Start at V1.
2. Visit V1 and then follow the edges to its neighbors.
3. Continue this process, visiting each vertex and its neighbors, ensuring that each vertex is visited only once.

Let's follow the adjacency list step-by-step:

1. Start at V1.
2. V1 points to V5, so visit V5.
3. V5 points to V4, so visit V4.
4. V4 points to V7, so visit V7.
5. V7 points to V6, so visit V6.
6. V6 has no further neighbors, so backtrack to V5.
7. Since V5 has already been visited, backtrack to V4.
8. Since V4 has already been visited, backtrack to V1.
9. Since V1 has already been visited, backtrack to V2.
10. V2 has no further neighbors, so backtrack to V3.
11. V3 has no further neighbors, so backtrack to V1.

The sequence of vertices visited is: V1, V5, V4, V7, V6, V2, V3.

Thus, the correct answer is: [A. V1,V5,V4,V7,V6,V2,V3](#)

WeThink-VL-7B:

<think> To determine the correct vertex sequence obtained by traversing the graph from vertex V1 using a depth-first search (DFS) method, we need to follow the adjacency list and the rules of DFS. DFS typically involves visiting a vertex, then exploring as far as possible along each branch before backtracking.

Starting from V1:

1. V1 -> V5 (first child)
2. V5 -> V4 (first child)
3. V4 -> V7 (first child)
4. V7 -> V6 (first child)
5. V6 -> (no further children, backtrack)
6. V4 -> (already visited, backtrack)
7. V5 -> V3 (next child)
8. V3 -> V2 (first child)
9. V2 -> (no further children, backtrack)
10. V3 -> (already visited, backtrack)
11. V5 -> (already visited, backtrack)
12. V1 -> (already visited, backtrack)

The sequence obtained is: V1, V5, V4, V7, V6, V3, V2.

Let's match this with the given options:

- A. V1,V5,V4,V7,V6,V2,V3
- B. V1,V2,V3,V4,V7,V6,V5
- C. V1,V5,V4,V7,V6,V3,V2
- D. V1,V5,V6,V4,V7,V2,V3

The correct sequence is: V1, V5, V4, V7, V6, V3, V2, which matches option C. </think>

<answer> C. V1,V5,V4,V7,V6,V3,V2 </answer>

Daily Scenes



Question: What fruit is to the right of plums?

GT Answer: orange

Qwen2.5-VL-7B

The fruit to the right of the plums in the image is a kiwi.

WeThink-VL-7B:

<think> The image shows a variety of fruits and vegetables. The plums are located in the middle of the image, slightly to the left. To the right of the plums, there are several oranges. The oranges are clearly visible and are positioned next to the plums. </think>

<answer> The fruit to the right of the plums is an orange. </answer>



Question: According to the positive sequence of the images, does the man close the door?

GT Answer: Yes

Qwen2.5-VL-7B

No, the man does not close the door in the sequence of images provided. The images show the man standing in front of a door, but there is no indication that he closes the door. The sequence appears to show the man facing the door, possibly preparing to open it, but the door is not closed in the images.

WeThink-VL-7B:

<think> The sequence of images shows a man in a white shirt. In the first image, he is facing someone or something off-camera. In the second image, he is seen from the side, and it appears he is moving towards a door. In the third image, he is seen from behind, and it looks like he is interacting with the door, possibly closing it. The progression of the images suggests a movement towards the door and a possible action of closing it. </think>

<answer> Yes, according to the positive sequence of the images, the man appears to close the door.
</answer>