

ENHANCING VISION-LANGUAGE REASONING VIA REINFORCEMENT LEARNING WITH SCALABLE MULTIMODAL QA SYNTHESIS

Anonymous authors

Paper under double-blind review

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 enhance visual-language reasoning through RL for different domains?* To address this challenge, we make three key efforts: **(1)** A novel *Scalable Multimodal QA Synthesis* pipeline that autonomously generates domain-aware, reasoning-centric question-answer (QA) pairs directly from images across different domains. **(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 simple baseline incorporating a hybrid reward mechanism that combines rule-based verification with model-based assessment to optimize RL training efficiency across different task domains. Through comprehensive exploration of RL on our dataset, we demonstrate that the **WeThink** dataset significantly improves performance across diverse MLLM benchmarks. Furthermore, we highlight that our automated data pipeline can continuously increase data diversity, further boosting 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 diverse visual question answering. Closed-source models like OpenAI’s o3 (OpenAI, 2025) and Kimi k1.5 (Team et al., 2025) 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 (Guo et al., 2025) have pioneered text-centric reasoning models by integrating reinforcement learning (RL) with verifiable rewards. However, these models (Chu et al., 2025; Muennighoff et al., 2025) are inherently constrained to unimodal (*i.e.*, text-only) scenarios, leaving a critical gap in multimodal reasoning capabilities.

Recent works (Chen et al., 2025a; Deng et al., 2025b; Zhang et al., 2025a; Yang et al., 2025; Wang et al., 2025b; Huang et al., 2025; Wang et al., 2025a; Wei et al., 2025; Liu et al., 2025a) 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 enhance visual-language reasoning through RL for different domains?* Two critical aspects stand out.

- **Diverse Domain-aware and 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 further enhance visual-language reasoning across domains, it’s important to generate domains-aware and reason-focused data from diverse domains and contexts.

054
055
056 • **RL with Hybrid Rewards.** While rule-based
057 rewards (e.g., answer verification for mathe-
058 matical problems) are effective in specific do-
059 mains, they struggle to capture the complexity of
060 real-world multimodal scenarios, where answers
061 can be subjective or context-dependent. This
062 underscores the need for a hybrid reward sys-
063 tem that combines both rule-based and model-
064 based strategies, offering more nuanced, context-
065 sensitive feedback to enable RL-trained MLLMs
066 to handle diverse task domains.

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

085 Building upon our dataset, we establish a simple
086 baseline that introduces a hybrid reward mecha-
087 nism, integrating rule-based verification with
088 model-based evaluation to enhance RL training
089 efficiency across diverse task domains. Through
090 a comprehensive exploration of RL on **WeThink**,
091 we present four key findings: (1) SFT with CoT
092 supervision on our dataset enhances the perfor-
093 mance of less optimized model (e.g., Qwen2-VL-
094 7B), yielding an average improvement of 3.5%
095 across six mathematical reasoning benchmarks.
096 (2) Using our dataset, direct RL fine-tuning on
097 Qwen2.5-VL-7B (Bai et al., 2025) is sufficient
098 and even outperforms cold-start supervised fine-
099 tuning followed by RL. (3) Ablation studies on
100 our dataset show that increasing the diversity of
101 question domains through RL fine-tuning leads
102 to significant improvements across tasks, from
103 mathematical reasoning to other domains. (4)
104 The scalability of our data pipeline enables
105 continuous collection of diverse images from
106 the Internet, further enhancing model perfor-
107 mance, as shown in Fig. 1.

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

090 ◊ **Automated Data Generation Pipeline:** We propose *Scalable Multimodal QA Synthesis* that
091 autonomously generates domain-aware, reason-centric questions paired with verifiable answers
092 directly from the given images.

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

096 ◊ **Enhanced Visual-Language Reasoning Models:** We establish a simple baseline using a hybrid
097 reward mechanism and conduct a comprehensive exploration of RL on our dataset, resulting in a
098 series of models that demonstrate improved performance across diverse tasks. Furthermore, we show
099 that the scalability of our pipeline, driven by increased data diversity, leads to further performance
100 improvements.

102 2 RELATED WORK

105 2.1 MULTIMODAL LARGE LANGUAGE MODELS (MLLMs)

106 Recent years have witnessed significant advancements in Multimodal Large Language Models
107 (MLLMs), which augment traditional Large Language Models (LLMs) by enabling them to process

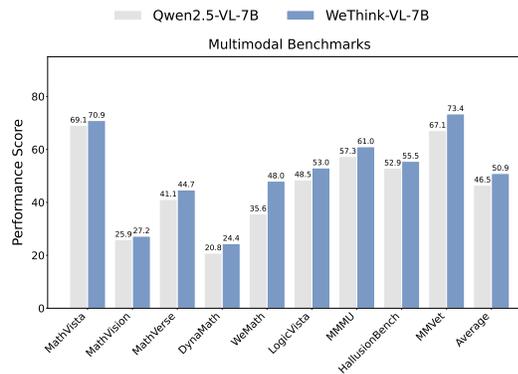


Figure 1: WeThink-VL-7B, fine-tuned on Qwen2.5-VL-7B (Bai et al., 2025) through reinforcement learning, shows significant improvements for different task domains.

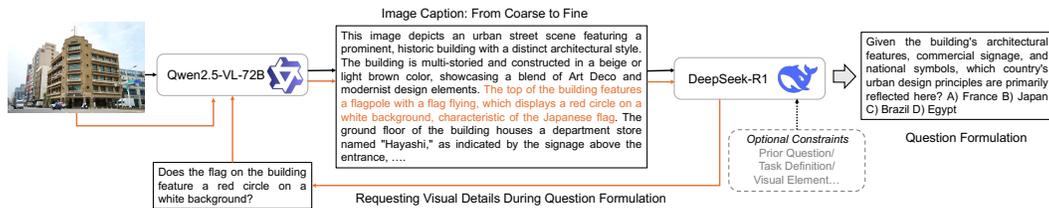


Figure 2: The automatic process of question formulation for a given image. As illustrated by the orange line, based on the coarse description provided by Qwen2.5-VL-72B, DeepSeek-R1 needs to request additional visual details (orange text) through multi-turn conversations with Qwen2.5-VL-72B, thus facilitating 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.

and comprehend information from diverse modalities, including text, images, audio, and video (Liu et al., 2023; Barrault et al., 2023; Chen et al., 2022; Li et al., 2023; Zhang et al., 2023a). The rapid evolution in this field is evidenced by the development of numerous open-source models, such as MiniGPT-4 (Zhu et al., 2023), MiniCPM-V (Yao et al., 2024), CogVLM (Wang et al., 2024c), ShareGPT4V (Chen et al., 2024a), Qwen-VL (Bai et al., 2023; Wang et al., 2024b; Bai et al., 2025), LLaVA (Liu et al., 2023; 2024b; Li et al., 2024), and InternVL (Chen et al., 2024d;c;b), alongside prominent closed-source models like Gemini (Team et al., 2023; 2024), GPT-4o (Hurst et al., 2024), Claude (Anthropic), and Grok (xAI). 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 LLMs by guiding them to articulate intermediate inferential steps prior to delivering a final answer (Wei et al., 2022; Kojima et al., 2022), has been effectively extended to the multimodal domain. In 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 (Zhang et al., 2023b; Lu et al., 2023b; Luo et al., 2024). 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 (Yang et al., 2023; Zhang et al., 2024a; Mitra et al., 2024; Zheng et al., 2023; Ni et al., 2024; Chen et al., 2024e) and Supervised Fine-Tuning (SFT) using datasets enriched with multimodal CoT examples (Xu et al., 2024; Dong et al., 2024; Thawakar et al., 2025). 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) (Zhou et al., 2025a). The integration of RL, particularly Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) or rule-based reward mechanisms (R1-style) (Guo et al., 2025), 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 (Chen et al., 2025a; Deng et al., 2025b; Zhang et al., 2025a; Yang et al., 2025; Wang et al., 2025b; Huang et al., 2025; Wang et al., 2025a; Wei et al., 2025; Liu et al., 2025a) and various specific downstream tasks. For instance, researchers have applied these RL techniques to improve scene graph understanding (Li et al., 2025; Chen et al., 2025b), visual-spatial reasoning (Zhao et al., 2025; Liao et al., 2025; Zhang et al., 2025b; Zhou et al., 2025b), referring expression comprehension (Yu et al., 2025; Shen et al., 2025; Liu et al., 2025d; Deng et al., 2025a; Liu et al., 2025b), and visual counting (Wang et al., 2025c; Tan et al., 2025; Liu et al., 2025c). While these methods have

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

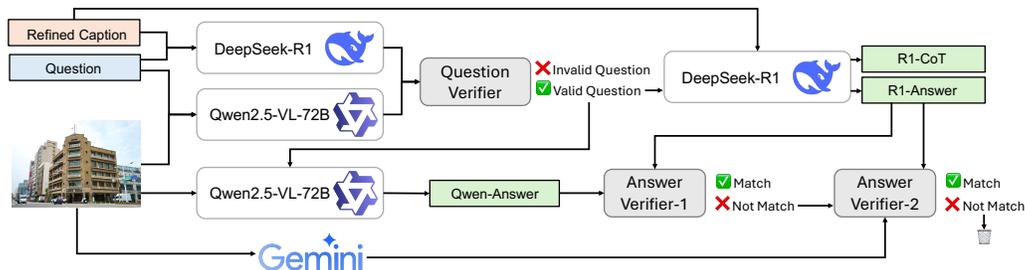


Figure 3: The automatic process of answer construction and quality control. First, DeepSeek-R1 filters out forced and open-ended questions to ensure they are verifiable. Then, using the refined caption and the valid question, DeepSeek-R1 generates a chain of thought and an answer. At the same time, Qwen2.5-VL-72B 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.

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 domain-aware, reasoning-centric question-answer (QA) pairs from the given images. Based on it, we 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**.

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 (Xu et al., 2024). These datasets cover diverse image categories, ensuring variety and complexity in the generated QA pairs, including general images (COCO (Lin et al., 2014), SAM-1B (Kirillov et al., 2023), Visual Genome (Krishna et al., 2017), GQA (Hudson & Manning, 2019), PISC (Li et al., 2017), LLaVA (Liu et al., 2023)), text-intensive images (TextVQA (Singh et al., 2019), ShareTextVQA (Chen et al., 2024a), DocVQA (Mathew et al., 2021), OCR-VQA (Mishra et al., 2019), ChartQA (Masry et al., 2022)), scientific and technical images (GeoQA+ (Cao & Xiao, 2022), ScienceQA (Lu et al., 2022), AI2D (Kembhavi et al., 2016), CLEVR-Math (Lindström & Abraham, 2022)), and images related to art and culture contexts (WikiArt (Saleh & Elgammal, 2015; Chen et al., 2024a), Web-Landmark (Kreutzer et al., 2022; Chen et al., 2024a) Web-Celebrity (Kreutzer et al., 2022; Chen et al., 2024a)).

3.2 QUESTION FORMULATION

Based on the collected images, we aim to generate domain-aware, reasoning-centric questions. The straightforward pipeline involves collaboration between two powerful models, Qwen2.5-VL-72B and DeepSeek-R1, 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: **(1) 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. **(2) 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. **(3) 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 contextually aware question.

Ability Synergy Constraints. Inspired by MM-Vet benchmark (Yu et al., 2023) 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 apply and combine the skills learned during training. Specifically, the formulation enforces mandatory reasoning capability combined with at least one complementary ability from other five-dimensional taxonomy: *a) Recognition:* General visual recognition (e.g., objects, attributes, scenes, counting, or high-level computer vision tasks); *b) Knowledge:* Use of social/visual commonsense, encyclopedic knowledge, or contextual information; *c) OCR:* Reading and reasoning over visible text (e.g., scene text, handwritten text, or embedded text in objects); *d) Spatial Awareness:* Understanding spatial relationships (e.g., object positions, directional/distance logic, layout analysis); *e) Math:* Performing arithmetic operations, solving equations, or interpreting math-specific notation. Overall, the above mechanism ensures questions inherently require: **(1) Cross-modal Reasoning Chains:** Minimum two explicit reasoning chains with comprehensive image analysis **(2) 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 precise 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 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 (Team et al., 2024) 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 **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, with the charts presented in the appendix.

Question Distribution. We use Qwen2.5-VL-72B to categorize each question into five groups: general, math, chart/table/doc, knowledge, and OCR. These categories are fairly balanced. 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.

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. The core ability is reasoning, and other abilities are also triggered depending on the semantic contents of the images. Additionally, each sample engages at least two abilities simultaneously. We also showcase top-15 ability combinations in the appendix. Notably, these combinations follow a long-tail distribution, with some ability combinations being rarer than others.

4 A SIMPLE BASELINE WITH HYBRID REWARD

4.1 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.

4.2 A SIMPLE FRAMEWORK FOR REINFORCEMENT LEARNING 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. Building on the success of DeepSeek-R1 in text-based reasoning tasks, we establish a simple yet effective framework for visual-language models to comprehensively explore

RL on our dataset. This framework incorporates a group-relative policy optimization strategy (Shao et al., 2024) and a hybrid reward system tailored to our dataset.

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} :

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

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 (Liu et al., 2024a) 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.

5 EXPERIMENTS

Due to space limitations, we provide details of benchmarks, evaluations, implementations, supplementary experiments, and case studies in the appendix B and C.

Model	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	MMMU	HallusionBench	MMVet	Average
Open-source Visual-Language Models										
LLaVA-CoT-11B	54.8	-	-	-	-	-	-	47.8	60.3	-
LLaVA-OneVision-7B	58.6	18.3	19.3	9.0	20.9	33.3	46.8	47.5	50.6	33.8
InternVL2-8B	58.3	20.0	20.4	9.2	20.2	33.6	51.2	45.0	54.3	34.7
InternVL2.5-8B	64.5	17.0	22.8	9.4	23.5	36.0	56.2	49.0	62.8	37.9
Qwen2-VL-7B	61.6	19.2	25.4	11.0	22.3	33.3	53.7	50.4	61.8	37.6
Qwen2.5-VL-7B*	69.1	25.9	41.1	20.8	35.6	48.5	57.3	52.9	67.1	46.5
Open-source Visual-Language Reasoning Models										
R1-VL-7B	63.5	24.7	40.0	-	-	-	-	-	-	-
X-Reasoner	69.0	<u>29.6</u>	-	-	-	-	56.4	-	-	-
URSA-8B-PS-GRPO	67.8	31.8	41.5	22.4	38.3	44.7	41.1	41.1	29.9	39.8
Visual-RFT*	61.4	18.9	24.5	10.5	24.6	34.0	50.9	42.3	61.2	36.5
R1-Onevision-7B*	63.1	22.2	38.2	18.4	33.2	44.6	52.1	49.7	62.5	42.7
VLAA-Thinker-7B	68.0	26.4	48.2	22.4	41.5	48.5	56.6	51.9	65.8	47.7
WeThink-VL-7B	71.6	26.7	<u>45.1</u>	<u>24.0</u>	<u>45.5</u>	<u>51.9</u>	<u>59.3</u>	55.8	<u>71.7</u>	<u>50.2</u>
WeThink-VL-7B [†]	<u>70.9</u>	27.2	44.7	24.4	48.0	53.0	61.0	<u>55.5</u>	73.4	50.9

Table 1: Comparison of various VLMs across diverse multimodal benchmarks. We evaluate the official model checkpoint provided, using VLMEvalKit, marked with a *. The best results are highlighted in **bold**, while the second-best results are underlined. A superscript [†] indicates the use of external images from the Internet to further enhance diversity.

Reward Type	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	Average
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
<i>Model</i>	63.0	24.9	43.3	25.7	31.9	45.6	39.1
<i>Rule+Model</i>	66.8	26.2	45.7	24.2	37.9	47.4	41.4

Table 2: The impact of RL with different reward types on all *math-type* questions in **WeThink**.

5.1 IMPLEMENTATION DETAILS

The implementation of our simple framework is built on EasyR1 (Zheng et al., 2025), which is based on veRL (Sheng et al., 2024). 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 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.

5.2 MAIN RESULTS

As shown in Tab. 1, WeThink-7B outperforms open-source vision-language models across various task domains, with a 4.4-point average improvement over the base model Qwen2.5-VL-7B. Compared to concurrent models, WeThink-7B shows broader improvements across different domains, not only in mathematics but also on other benchmarks. In contrast, models like R1-Onevision-7B and VLAA-Thinker-7B, which also use Qwen2.5-VL-7B, focus on domain-specific performance but suffer declines in other areas. This highlights the effectiveness of the WeThink dataset, which enhances multimodal reasoning by covering diverse domains and tasks. It also demonstrates that our pipeline brings advantages in terms of new domain data and tasks.

5.3 ABLATION STUDY

The impact of cold start during RL. Following DeepSeek-R1, we initialize RL training with the CoT SFT model as the cold-start checkpoint, trained on 120K WeThink samples using CoT prompting. For subsequent RL training, we select all math-type questions from WeThink and incorporate hybrid rewards. Fig. 4 demonstrates that using refined CoT data for cold start significantly enhances model

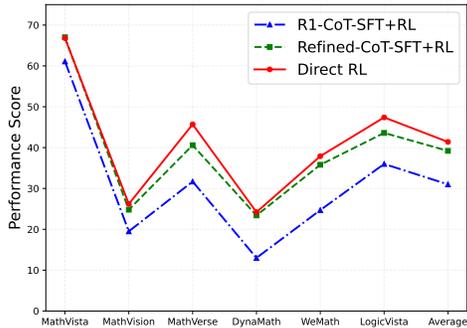


Figure 5: The impact of cold start SFT for RL on all *math-type* questions in **WeThink**.

Question Type	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	MMMU	HallusionBench	MMVet	Average
Qwen2.5-VL-7B	69.1	25.9	41.1	20.8	35.6	48.5	57.3	52.9	67.1	46.5
<i>Math</i>	66.8	26.2	45.7	24.2	37.9	47.4	57.2	56.0	72.2	48.2
<i>All</i>	71.6	26.7	45.1	24.0	45.5	51.9	59.3	55.8	71.7	50.2

Table 3: The impact of RL with different question types (*i.e.*, *math-type* and *all-type*) in **WeThink**, comparing performance across diverse multimodal benchmarks.

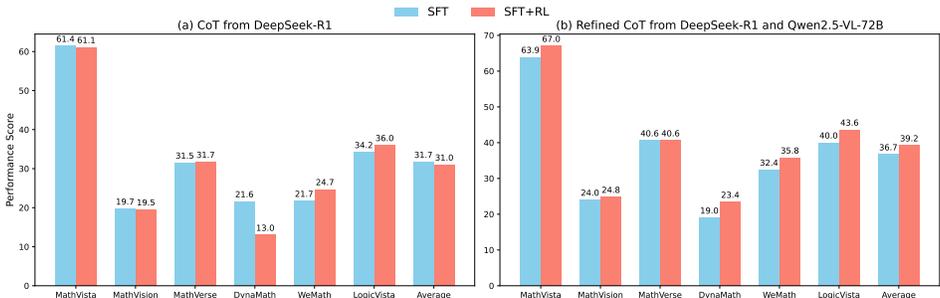


Figure 4: The impact of cold start data quality.

performance following RL. Conversely, applying RL directly to a model initialized with the original R1-CoT checkpoint yields negligible improvements. These results clearly underscore the critical importance of high-quality cold-start data for effective RL enhancement. Further reinforcing this point, Fig. 5 shows that direct RL fine-tuning on the base Qwen2.5-VL-7B model leads to better gains than the cold-start method. This suggests that when leveraging our dataset, direct RL optimization on Qwen2.5-VL-7B is a highly effective strategy.

The impact of RL with different reward types. We divide questions in WeThink 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. 2 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 enhance visual-language reasoning across domains beyond mathematics, we extend RL training to include all question types in the WeThink dataset. As shown in Tab. 3, we compare models trained exclusively on math-type questions with those trained on the full range of question types, evaluated across diverse benchmarks. The results show that using all question types not only boosts performance on math benchmarks but also improves results on 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. 1 show that performance continues to improve across diverse 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 data to further enhance performance.

6 CONCLUSION

This work advances multimodal understanding by enhancing visual-language reasoning through reinforcement learning (RL). We introduce a novel *Scalable Multimodal QA Synthesis* pipeline to generate domain-aware, reasoning-focused question-answer pairs from images across different domains. We also release the **WeThink** dataset, with over 120K multimodal QA pairs and annotated reasoning chains, to improve RL training across domains. We establish a simple baseline and conduct comprehensive exploration of RL on our dataset, incorporating a hybrid reward mechanism. Results across various MLLM benchmarks show the dataset’s effectiveness in improving task performance, while our automated pipeline ensures continuous data diversity and scalable RL training.

REFERENCES

- 486
487
488 Anthropic. Claude. <https://www.anthropic.com/>.
- 489
490 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou,
491 and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization,
492 text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- 493
494 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang,
495 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,
2025.
- 496
497 Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise
498 Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamless4t:
499 Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*,
2023.
- 500
501 Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through
502 dual parallel text encoding. In *Proceedings of the 29th international conference on computational*
503 *linguistics*, pp. 1511–1520, 2022.
- 504
505 Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang
506 Xie. Sft or rl? an early investigation into training rl-like reasoning large vision-language models.
arXiv preprint arXiv:2504.11468, 2025a.
- 507
508 Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin.
509 Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference*
on Computer Vision, pp. 370–387. Springer, 2024a.
- 510
511 Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki
512 Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training
513 for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):
1505–1518, 2022.
- 514
515 Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong
516 Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal
517 models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.
- 518
519 Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi
520 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*, 2024c.
- 521
522 Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong
523 Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning
524 for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer*
Vision and Pattern Recognition, pp. 24185–24198, 2024d.
- 525
526 Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and
527 Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In
528 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1254–1262, 2024e.
- 529
530 Zuyao Chen, Jinlin Wu, Zhen Lei, Marc Pollefeys, and Chang Wen Chen. Compile scene graphs
with reinforcement learning. *arXiv preprint arXiv:2504.13617*, 2025b.
- 531
532 Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V
533 Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation
534 model post-training. *arXiv preprint arXiv:2501.17161*, 2025.
- 535
536 Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization
537 and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint*
arXiv:2503.07065, 2025a.
- 538
539 Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker:
An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv*
preprint arXiv:2503.17352, 2025b.

- 540 Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu.
541 Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv*
542 *preprint arXiv:2411.14432*, 2024.
- 543
544 Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang
545 Zang, Pan Zhang, Jiaqi Wang, et al. VImevalkit: An open-source toolkit for evaluating large
546 multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*,
547 pp. 11198–11201, 2024.
- 548 Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang
549 Chen, Furong Huang, Yaser Yacoub, et al. Hallusionbench: an advanced diagnostic suite for entan-
550 gled language hallucination and visual illusion in large vision-language models. In *Proceedings of*
551 *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.
- 552
553 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
554 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
555 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- 556
557 Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and
558 Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models.
arXiv preprint arXiv:2503.06749, 2025.
- 559
560 Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning
561 and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer*
562 *vision and pattern recognition*, pp. 6700–6709, 2019.
- 563
564 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-
565 trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint*
arXiv:2410.21276, 2024.
- 566
567 Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi.
568 A diagram is worth a dozen images. In *Computer Vision—ECCV 2016: 14th European Conference,*
569 *Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251.
Springer, 2016.
- 570
571 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete
572 Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings*
573 *of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.
- 574
575 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
576 language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:
22199–22213, 2022.
- 577
578 Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh,
579 Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance:
580 An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational*
Linguistics, 10:50–72, 2022.
- 581
582 Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie
583 Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language
584 and vision using crowdsourced dense image annotations. *International journal of computer vision*,
585 123:32–73, 2017.
- 586
587 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan
588 Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint*
arXiv:2408.03326, 2024.
- 589
590 Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Dual-glance model for deciphering
591 social relationships. In *Proceedings of the IEEE international conference on computer vision*, pp.
592 2650–2659, 2017.
- 593
KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and
Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

- 594 Lin Li, Wei Chen, Jiahui Li, and Long Chen. Relation-r1: Cognitive chain-of-thought guided
595 reinforcement learning for unified relational comprehension. *arXiv preprint arXiv:2504.14642*,
596 2025.
- 597 Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng.
598 Improved visual-spatial reasoning via r1-zero-like training. *arXiv preprint arXiv:2504.00883*,
599 2025.
- 600
601 Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr
602 Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–*
603 *ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings,*
604 *part v 13*, pp. 740–755. Springer, 2014.
- 605
606 Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional
607 language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.
- 608
609 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
610 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*
611 *arXiv:2412.19437*, 2024a.
- 612
613 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in*
614 *neural information processing systems*, 36:34892–34916, 2023.
- 615
616 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction
617 tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
618 pp. 26296–26306, 2024b.
- 619
620 Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and
621 Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv*
622 *preprint arXiv:2504.13055*, 2025a.
- 623
624 Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-
625 zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint*
626 *arXiv:2503.06520*, 2025b.
- 627
628 Zhiyuan Liu, Yuting Zhang, Feng Liu, Changwang Zhang, Ying Sun, and Jun Wang. Othink-mr1:
629 Stimulating multimodal generalized reasoning capabilities via dynamic reinforcement learning.
630 *arXiv preprint arXiv:2503.16081*, 2025c.
- 631
632 Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi
633 Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025d.
- 634
635 Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord,
636 Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for
637 science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521,
638 2022.
- 639
640 Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng,
641 Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning
642 of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023a.
- 643
644 Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu,
645 and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models.
646 *Advances in Neural Information Processing Systems*, 36:43447–43478, 2023b.
- 647
648 Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout
649 instruction tuning with large language models for document understanding. In *Proceedings of the*
650 *IEEE/CVF conference on computer vision and pattern recognition*, pp. 15630–15640, 2024.
- 651
652 Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A bench-
653 mark for question answering about charts with visual and logical reasoning. *arXiv preprint*
654 *arXiv:2203.10244*, 2022.

- 648 Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document
649 images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*,
650 pp. 2200–2209, 2021.
- 651
- 652 Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual
653 question answering by reading text in images. In *2019 international conference on document
654 analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.
- 655
- 656 Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought
657 prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer
658 Vision and Pattern Recognition*, pp. 14420–14431, 2024.
- 659 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke
660 Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time
661 scaling. *arXiv preprint arXiv:2501.19393*, 2025.
- 662 Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. Visual-o1: Understanding am-
663 biguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. *arXiv preprint
664 arXiv:2410.03321*, 2024.
- 665
- 666 OpenAI. Openai o3 and o4-mini system card, 2025. URL [https://openai.com/index/
667 o3-o4-mini-system-card/](https://openai.com/index/o3-o4-mini-system-card/). Accessed: 2025-05-14.
- 668
- 669 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
670 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow
671 instructions with human feedback. *Advances in neural information processing systems*, 35:27730–
672 27744, 2022.
- 673 Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue,
674 Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model
675 achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.
- 676
- 677 Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the
678 right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.
- 679 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
680 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical
681 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- 682
- 683 Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun
684 Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large
685 vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.
- 686
- 687 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,
688 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint
689 arXiv:2409.19256*, 2024.
- 690
- 691 Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and
692 Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference
693 on computer vision and pattern recognition*, pp. 8317–8326, 2019.
- 694
- 695 Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and
696 Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint
697 arXiv:2503.20752*, 2025.
- 698
- 699 Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut,
700 Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly
701 capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 702
- 703 Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett
704 Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal
705 understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

- 702 Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun
703 Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with
704 llms. *arXiv preprint arXiv:2501.12599*, 2025.
- 705
706 Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan,
707 Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking
708 step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.
- 709
710 Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. V1-rethinker:
711 Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint
arXiv:2504.08837*, 2025a.
- 712
713 Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and
714 Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances
715 in Neural Information Processing Systems*, 37:95095–95169, 2024a.
- 716
717 Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu,
718 Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the
719 world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.
- 720
721 Weihai Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang,
722 Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances
in Neural Information Processing Systems*, 37:121475–121499, 2024c.
- 723
724 Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu,
725 Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for
726 multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025b.
- 727
728 Zhiqiang Wang, Pengbin Feng, Yanbin Lin, Shuzhang Cai, Zongao Bian, Jinghua Yan, and Xingquan
729 Zhu. Crowdvlm-r1: Expanding r1 ability to vision language model for crowd counting using fuzzy
group relative policy reward. *arXiv preprint arXiv:2504.03724*, 2025c.
- 730
731 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
732 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in
neural information processing systems*, 35:24824–24837, 2022.
- 733
734 Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao
735 Zhang, Yunzhuo Hao, Xuchen Song, et al. Skywork r1v2: Multimodal hybrid reinforcement
736 learning for reasoning. *arXiv preprint arXiv:2504.16656*, 2025.
- 737
738 xAI. Grok. <https://x.ai/>.
- 739
740 Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning
741 benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.
- 742
743 Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language
744 models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.
- 745
746 Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng
747 Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning
748 through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.
- 749
750 Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu,
751 Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning
752 and action. *arXiv preprint arXiv:2303.11381*, 2023.
- 753
754 Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li,
755 Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint
arXiv:2408.01800*, 2024.
- 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.

- 756 Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang,
757 and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv*
758 *preprint arXiv:2308.02490*, 2023.
- 759 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu
760 Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal under-
761 standing and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on*
762 *Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.
- 763 Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. Cocot:
764 Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs.
765 *arXiv preprint arXiv:2401.02582*, 2024a.
- 766 Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language
767 model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.
- 768 Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao.
769 R1-vl: Learning to reason with multimodal large language models via step-wise group relative
770 policy optimization. *arXiv preprint arXiv:2503.12937*, 2025a.
- 771 Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan
772 Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams
773 in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer,
774 2024b.
- 775 Wenqi Zhang, Mengna Wang, Gangao Liu, Xu Huixin, Yiwei Jiang, Yongliang Shen, Guiyang Hou,
776 Zhe Zheng, Hang Zhang, Xin Li, et al. Embodied-reasoner: Synergizing visual search, reasoning,
777 and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696*, 2025b.
- 778 Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal
779 chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023b.
- 780 Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang,
781 Xinlei Chen, Yong Li, and Wenwu Zhu. Embodied-r: Collaborative framework for activating
782 embodied spatial reasoning in foundation models via reinforcement learning. *arXiv preprint*
783 *arXiv:2504.12680*, 2025.
- 784 Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibe Yang. Ddcot: Duty-distinct chain-of-
785 thought prompting for multimodal reasoning in language models. *Advances in Neural Information*
786 *Processing Systems*, 36:5168–5191, 2023.
- 787 Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong.
788 Easyrl: An efficient, scalable, multi-modality rl training framework. [https://github.com/](https://github.com/hiyouga/EasyR1)
789 [hiyouga/EasyR1](https://github.com/hiyouga/EasyR1), 2025.
- 790 Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu.
791 Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv*
792 *preprint arXiv:2504.21277*, 2025a.
- 793 Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-
794 zero’s” aha moment” in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*,
795 2025b.
- 796 Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: En-
797 hancing vision-language understanding with advanced large language models. *arXiv preprint*
798 *arXiv:2304.10592*, 2023.
- 799 Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A
800 dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language
801 models. *arXiv preprint arXiv:2411.00836*, 2024.
- 802
803
804
805
806
807
808
809

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

Table 4: The distribution analysis of image types from **WeThink**.

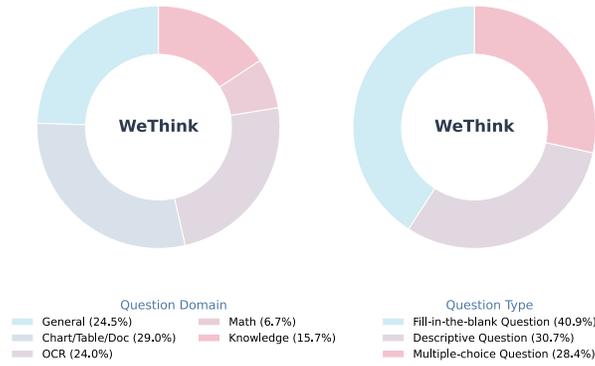
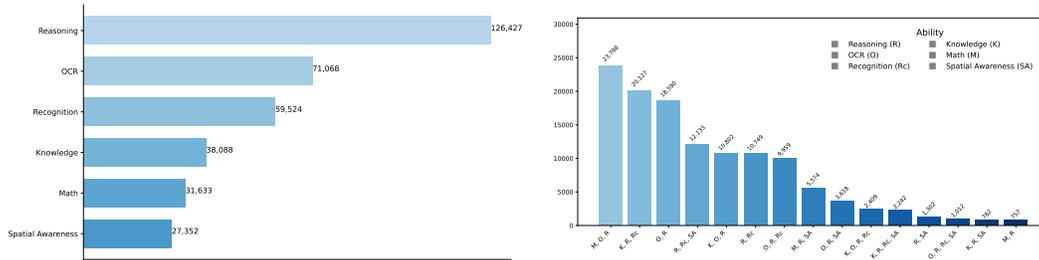


Figure 6: The distribution analysis of question domains and types from **WeThink**.



(a) The distribution of each ability

(b) The distribution of top-15 ability combinations

Figure 7: We show the difficulty distribution of question sample in the **WeThink** dataset.

A THE DETAILS OF SCALABLE MULTIMODAL QA SYNTHESIS

We formulate the question formulation designs into a prompt protocol for DeepSeek-R1, 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>`).

B EXPERIMENTAL IMPLEMENTATION DETAILS

The anonymous code and dataset are available at <https://anonymous.4open.science/r/WeThink-7C9A> and <https://huggingface.co/datasets/WeThink/WeThink-Multimodal-Reasoning-120K>.

B.1 BENCHMARKS & EVALUATION

To comprehensively evaluate multi-modal understanding and reasoning capabilities of our models, we conduct experiments across diverse benchmarks, including MathVista (Lu et al., 2023a), MathVision (Wang et al., 2024a), MathVerse (Zhang et al., 2024b), DynaMath (Zou et al., 2024), WeMath (Qiao et al., 2024), and LogicVista (Xiao et al., 2024), MMMU (Yue et al., 2024), MMVet (Yu et al., 2023), HallusionBench (Guan et al., 2024). The above benchmarks are available on the OpenCompass MLLM Leaderboard. To maintain fairness and reproducibility, we evaluate our models using VLMEvalKit (Duan et al., 2024), which is an open-source toolkit for MLLM evaluation.

B.2 SUPERVISED FINE-TUNING USING WETHINK

Implementation Details. We conduct chain-of-thought prompting using over 120K CoT-annotated diverse QA pairs from our **WeThink** dataset for direct SFT. 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 using 8 NVIDIA H20 GPUs on two instruction-tuned models: Qwen2-VL-7B and Qwen2.5-VL-7B.

Main Results. As shown in Tab. 5, 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). 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.

B.3 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."
```

B.4 BENCHMARKS & EVALUATION.

As illustrated in Sec. B.1, we conduct comprehensive evaluation across 14 MLLM benchmarks, covering six mathematical reasoning benchmarks and eight general multimodal understanding benchmarks. Below are the details:

Model	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	Average
Qwen2-VL-7B*	61.8	19.0	25.6	11.0	21.4	34.7	28.9
R1-CoT SFT	56.6	17.0	25.0	16.0	18.5	35.8	28.2
Refined-CoT SFT	59.5	20.3	32.9	15.2	25.0	41.2	32.4
Qwen2.5-VL-7B*	69.1	25.9	41.1	20.8	35.6	48.5	40.2
R1-CoT SFT	61.4	19.7	31.5	21.6	21.7	34.2	31.7
Refined-CoT SFT	63.9	24.0	40.6	19.0	32.4	40.0	36.7

Table 5: The impact of fully supervised fine-tuning on **WeThink**. * denotes the model results reproduced by us. We highlight the best average results in **bold**.

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

\diamond MathVision, which uses the full test set of MathVision

\diamond MathVerse_MINI_Vision_Only, the Test Mini split of MathVerse, using the "Vision Only" mode

\diamond DynaMath, which uses the full test set of DynaMath

\diamond WeMath, the Test Mini split of WeMath, where we report "Score (Strict)" as the main metric

\diamond LogicVista, which is the full test set of LogicVista

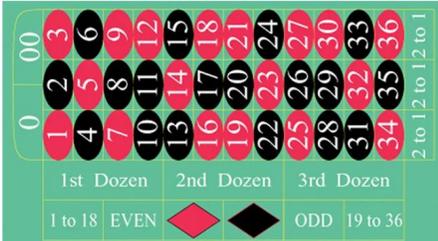
General multimodal understanding employs MMBench v1.1 (Test_CN / Test_EN), MMStar, MMMU (Val), HallusionBench, AI2D (Test), OCRBench, MMVet and RealWorldQA.

Evaluation Details. To ensure fairness and reproducibility, we conduct evaluation of our models using VLMEvalKit Duan et al. (2024), 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. B.3. We empirically found that using the default system prompt of the Qwen2.5-VL series yields better performance for OCRBench and AI2D benchmarks.

C 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

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

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:

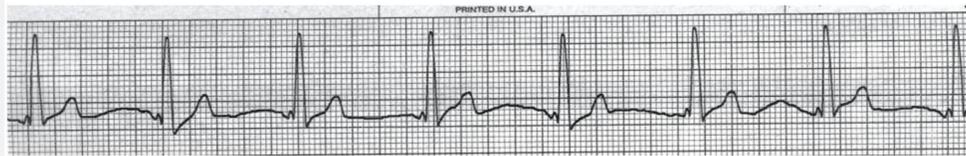
<think>To compute the probability of winning by betting on three numbers in a line, we need to understand that there are 38 possible numbers on the roulette wheel. When you bet on three specific numbers, you are essentially looking for the probability that the ball lands on one of those three numbers. 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.

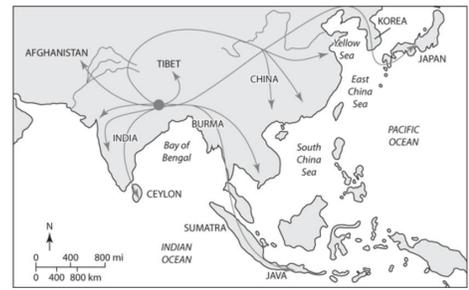
WeThink-VL-7B:

<thinkimage 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

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

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

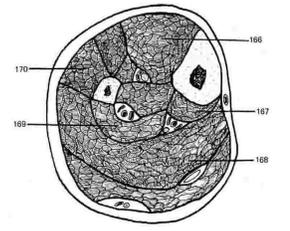
Qwen2.5-VL-7B

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.

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

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



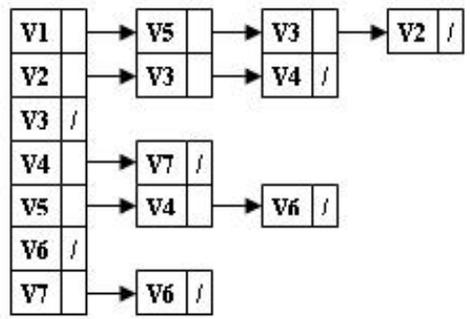
Question: The painting shown was created by:
Options:
A. Wifredo Lam
B. Frida Kahlo
C. Diego Rivera
D. Jacob Lawrence
GT Answer: D

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)

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

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 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>