

# M2-Reasoning: Empowering MLLMs with Unified General and Spatial Reasoning

Inclusion AI, Ant Group\*

\* See Contributions section (Sec. 6) for full author list.

Recent advancements in Multimodal Large Language Models (MLLMs), particularly through Reinforcement Learning with Verifiable Rewards (RLVR), have significantly enhanced their reasoning abilities. However, a critical gap persists: these models struggle with dynamic spatial interactions, a capability essential for real-world applications. To bridge this gap, we introduce **M2-Reasoning-7B**, a model designed to excel in both general and spatial reasoning. Our approach integrates two key innovations: (1) a novel data pipeline that generates 294.2K high-quality data samples (168K for cold-start fine-tuning and 126.2K for RLVR), which feature logically coherent reasoning trajectories and have undergone comprehensive assessment; and (2) a dynamic multi-task training strategy with step-wise optimization to mitigate conflicts between data, and task-specific rewards for delivering tailored incentive signals. This combination of curated data and advanced training allows M2-Reasoning-7B to set a new state-of-the-art (SOTA) across 8 benchmarks, showcasing superior performance in both general and spatial reasoning domains.

Date: July 10, 2025

Model: <https://huggingface.co/inclusionAI/M2-Reasoning>

Code: <https://github.com/inclusionAI/M2-Reasoning>

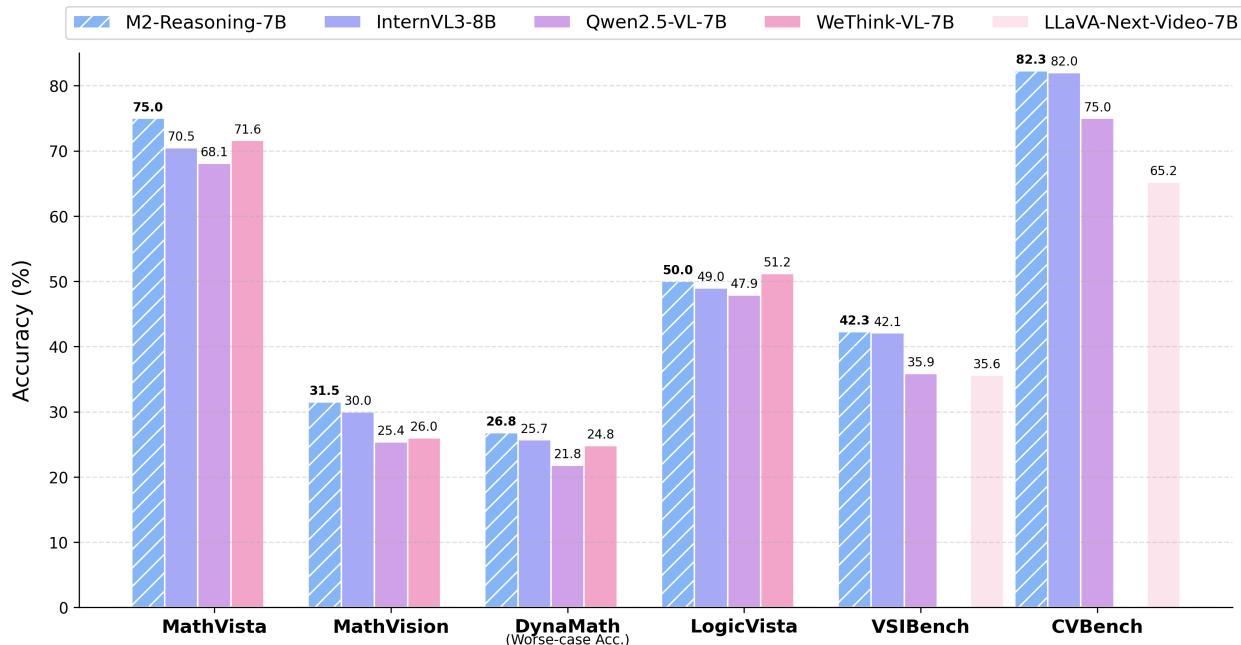


Figure 1 Benchmark performance of M2-Reasoning-7B.

# 1 Introduction

The recent advancements in Large Language Models (LLMs) have redefined the landscape of artificial intelligence, with models demonstrating remarkable capabilities in language understanding, generation, and reasoning (Grattafiori et al., 2024; Wang et al., 2024b; Yang et al., 2025a; Seed et al., 2025; Liu et al., 2024; Guo et al., 2025a; Team, 2025; Cai et al., 2024; Team et al., 2023). This progress has directly catalyzed the emergence of Multimodal Large Language Models (MLLMs), which integrate these sophisticated language faculties with visual perception (Li et al., 2024a; OpenAI, 2024; Bai et al., 2025; Zhu et al., 2025; Li et al., 2024b). Consequently, MLLMs can now adeptly handle a diverse spectrum of complex vision-language tasks, such as visual question answering, detailed image captioning, object grounding, and nuanced visual reasoning (Jaech et al., 2024; AI, 2025; Guo et al., 2025b; Team et al., 2025b).

Recent breakthroughs in leading LLMs, such as OpenAI’s o1/o3 (Jaech et al., 2024) and Deepseek-R1 (Guo et al., 2025a) have shown that a two-stage training process involving Supervised Fine-Tuning (SFT) and Reinforcement Learning with Verifiable Rewards (RLVR) is highly effective for unlocking advanced reasoning abilities. Inspired by this success, the same paradigm is increasingly being adapted for MLLMs to bolster their cognitive capabilities (Chen et al., 2025; Yang et al., 2025b; Guo et al., 2025d; Peng et al., 2025; Deng et al., 2025). However, a critical gap remains: these models largely fail to reason about the dynamic interplay of space and motion, a capability essential for meaningful real-world applications.

To address this critical gap, we introduce **M2-Reasoning-7B**, a Multimodal Large Language Model (MLLM) engineered to excel in two fundamental domains: **general reasoning**, for solving complex abstract problems, and **spatial reasoning**, for comprehending motion, orientation, and physical interactions in dynamic environments (Ray et al., 2025).

The development of M2-Reasoning is underpinned by a synergistic combination of a novel data curation strategy and an advanced training framework.

First, to overcome the scarcity of high-quality data, we established a multi-stage pipeline for data synthesis and curation. This process includes automated question-answer generation and the synthesis of detailed reasoning trajectories. Critically, all generated data undergoes a holistic evaluation that assesses not only correctness but also its quality, difficulty, and diversity. This comprehensive assessment framework is the key to enabling effective curriculum learning during the RLVR stage, as it allows us to strategically sequence the training data from simple to complex. This pipeline yielded 168K instances for cold-start, and a further 126.2K for RLVR, comprising 100K for general reasoning and 26.2K (18.7K image-based and 7.5K video-based) for spatial reasoning.

Second, to effectively train our model on this diverse dataset, we introduce a sophisticated training strategy that combines multi-task optimization with domain specialization. Our approach employs step-wise dynamic optimization to mitigate conflicts arising from data heterogeneity, and avoid suboptimal, one-size-fits-all configurations. Furthermore, we incorporate a task-specific reward formulation, which designs distinct reward functions tailored to the unique requirements of each task category. This is particularly crucial for spatial reasoning since such tailored incentive signals are essential for effectively bootstrapping the MLLM’s nascent perceptual abilities.

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

- ◆ **A High-quality Data Construction Pipeline:** We design and implement a multi-stage data synthesis and curation pipeline that generates vast amounts of reasoning data. (See Section 2)

- ◆ **A Dynamic Multi-Task Training Strategy:** We propose a sophisticated training strategy that effectively handles data heterogeneity. It features step-wise dynamic optimization to mitigate conflicts between different data sources and a task-specific reward formulation to provide tailored incentive signals. (See Section 3)
- ◆ **Unified General and Spatial Reasoning Model:** We propose M2-Reasoning-7B, an MLLM uniquely engineered for both abstract and spatial reasoning. Extensive evaluations on 8 distinct benchmarks demonstrate that, by leveraging our custom data and training pipelines, M2-Reasoning-7B establishes new state-of-the-art (SOTA) results across both general and spatial reasoning domains. (See Section 4)

## 2 Data Construction

To comprehensively activate and enhance the model’s reasoning capabilities across the tasks, we propose a **high-quality data construction pipeline** that incorporates general reasoning and spatial reasoning.

Specifically, for the general reasoning tasks (Section 2.1), we construct both cold-start and RLVR data: the former aims to help the model understand structured reasoning patterns, while the latter focuses on optimizing reasoning trajectories. For the spatial reasoning task (Section 2.2), we utilize image-text and video-based for RLVR to better capture dynamic spatial transformations and temporal dependencies.

In the following sections, we provide detailed descriptions of the data construction methodologies tailored to each task type, highlighting how they contribute to the development of multi-faceted reasoning abilities in the MLLM. Finally, an overview of the overall data configuration used in our training is provided in Section 2.3.

### 2.1 General Reasoning Data

To enhance the model’s general reasoning capabilities, we optimize the data strategy during the cold-start stage and RLVR curriculum. First, we implement a high-quality, multimodal chain-of-thought (CoT) synthesis pipeline that generates and filters reasoning chains based on answer accuracy and fine-grained reasoning quality metrics, ensuring structured and effective training. Second, we introduce a prompt difficulty scoring method for RLVR, enabling the model to learn progressively from tasks of increasing difficulty.

#### 2.1.1 High-Quality Multimodal Chain-of-Thought Synthesis Pipeline

In our preliminary experiments, we collect the open-source image-text cold-start dataset R1-OneVision (Yang et al., 2025d) and text-only data from Ling-Lite (Team, 2025) to activate the model’s general reasoning capabilities. These datasets provide a basic foundation for developing the model’s thinking ability. However, the quality and formatting of the reasoning chains contained in these datasets are not consistently controlled, which may lead to suboptimal performance when training models for complex reasoning tasks.

Previous studies (Ye et al., 2025; Xiaomi, 2025) have demonstrated that high-quality thought processes significantly contribute to effective model learning. Motivated by this, we design a comprehensive pipeline for thought chain generation and quality screening. This pipeline enables us

to construct well-structured, high-quality reasoning data, thereby enhancing the model’s ability to perform multi-step, complex reasoning.

We begin by curating a collection of publicly available datasets (Meng et al., 2025; Lu et al., 2021; Liu et al., 2025; Chen et al., 2025; Wang et al., 2025), which span a diverse range of reasoning tasks — from geometric problem-solving to logical inference in visual environments. To generate high-quality multimodal CoT data, we utilize an open-source vision-language reasoning model, WeThink-VL-7B (Yang et al., 2025b), to synthesize reasoning chains for each input.

To ensure diversity in the generated reasoning chains, we sample multiple responses per instance using a high-temperature generation strategy. This approach encourages varied thought processes and enhances both the breadth and richness of the resulting CoT dataset. Next, we assess the quality of the synthesized CoT data along two key dimensions:

- ◆ **Answer Accuracy Assessment.** We first evaluate the final answer of each reasoning chain to determine its correctness. Any reasoning chain that leads to an incorrect final answer is considered flawed in its overall thought process and is therefore discarded. To assess answer accuracy, we utilize Qwen2.5-7B-Instruct (Team, 2024), which serves as an automated evaluator for response correctness serves as an automated evaluator for response correctness.
- ◆ **Reasoning Quality Assessment.** Inspired by (Ye et al., 2025), beyond basic answer accuracy, we further establish a quality scoring framework to identify and retain high-quality reasoning chains that are more effective at eliciting deep thinking from the model. This assessments focuses on three key dimensions:
  - ◊ *Structural Correctness of Reasoning*: Assess whether the reasoning process is well-structured and clearly organized. Does it elaborate on complex steps in detail while keeping simpler ones concise, thereby demonstrating an adaptive adjustment of reasoning granularity?
  - ◊ *Cognitive Load Management*: Assess whether the reasoning effectively guides the reader through complex logic by progressively introducing concepts, clearly articulating key insights, and bridging cognitive gaps.
  - ◊ *Verification Richness*: Assess whether the reasoning process frequently validates intermediate results, examines the validity of underlying assumptions, and ensures logical consistency throughout the inference chain.

We employ Qwen2.5-7B-Instruct (Team, 2024) model, in conjunction with carefully designed prompts based on the aforementioned assessment principles, to assess the quality of the generated reasoning chains.

After this filtering process, we select a final set of 168K high-quality reasoning chains as our cold-start training data. Experimental results demonstrate that the data generated through this pipeline significantly enhances the model’s overall reasoning capabilities. The prompts used for scoring are provided in Appendix A.

### 2.1.2 Difficulty Scoring for Data in Reinforcement Learning

For the RLVR data, we quantify the difficulty of each prompt by the accuracy of model responses. Specifically, we generate ten reasoning-augmented responses per prompt using Qwen2.5-VL-7B (Bai et al., 2025) and compute the *response accuracy* as the proportion of correct answers among these responses. Prompts yielding 0% or 100% accuracy are excluded from the final dataset, as these extremes represent either overly challenging or trivial cases that fail to provide meaningful gradient

signals during RLVR training. For the remaining prompts, we define their *difficulty* as  $1 - \text{accuracy}$ , where higher values correspond to greater task difficulty.

## 2.2 Spatial Reasoning Data

To improve the model’s perceptual understanding of real-world images and videos, we design a spatial reasoning dataset based on controlled spatial simulations. Notably, the video simulation data is specifically crafted to capture dynamic, video-dependent information such as room size and object appearance order.

We introduce a comprehensive **spatial data synthesis pipeline**, which ensures the creation of high-quality, semantically meaningful data tailored for training and evaluating spatial reasoning capabilities. Furthermore, we design a **spatial data augmentation pipeline** that enhances the diversity of questions, options, and instructions through secondary augmentation.

### 2.2.1 Spatial Data Synthesis Pipeline

**Automated Data Annotation.** Our annotation process first performs preprocessing steps on real-world images, including depth estimation and instance segmentation, to obtain fundamental modalities such as depth maps, segmentation masks, surface normal maps, and camera parameters. Based on these, the module further derives 3D point clouds, 3D bounding boxes, and object labels, which are key informations for building spatial-related QA pairs. For simulated data, the annotation process prioritizes the use of built-in annotations, including point cloud-based instance segmentation, 3D bounding boxes, and object classification. These built-in labels in simulated data provide more accurate ground truth compared to real-world data.

**Question-Answer Generation.** To endow the model with more comprehensive spatial understanding and reasoning abilities, we designed 10 categories of tasks. The image-based tasks include: ObjectCount, SpatialRelation, RelativeSize, AbsoluteSize, AbsoluteDistance, RelativeDistance, and RelativeDepth. The video-based tasks focus on dynamic spatial-temporal understanding and include: RoomSize, AppearanceOrder, and RelativeDirection.

For each question type, we design a corresponding QA generation algorithm that leverages the pre-obtained annotations to produce semantically relevant questions and answers. We also implement question-specific filtering mechanisms to eliminate ambiguous or invalid object-relation pairs, ensuring clarity and well-defined reasoning tasks.

Take the image-based RelativeDistance task as an example, we applied three filtering criteria to ensure unambiguous spatial comparisons: 1) Objects with multiple instances were removed to avoid localization ambiguity; 2) Object with planar surfaces such as *Floor* and *Ceiling* were excluded due to potential perceptual ambiguity; 3) Among the remaining candidates, objects with similar distances to the reference object were filtered out, and only triplets  $(A, B, C)$  with clearly distinct distances were selected to construct the question: “*Which is closer to A, B or C?*”.

As another example, consider the video-based task AppearanceOrder. We first record object names and detection results frame-by-frame. Then, the following filters were applied: 1) Objects with excessively small pixel areas were discarded to reduce perceptual ambiguity; 2) Objects appearing simultaneously in the same frame or across adjacent frames were filtered out to prevent choice ambiguity. Finally, 3–4 objects were sampled from the remaining set, and temporal deduplication was performed to generate well-structured questions and answer options based on their appearance order.

**Data Quality Validation.** We adopt a hybrid quality control strategy combining model-based assessments with human verification. The model-based evaluation, performed by the Qwen2.5-VL-32B-Instruct (Bai et al., 2025) MLLM, encompassed two distinct types of assessments:

- ◆ **Difficulty Assessment.** We fed the visual input, question, and answer into the MLLM. The model performed multiple reasoning processes, and the accuracy across responses was computed to derive a difficulty score, defined as  $1 - \text{accuracy}$ .
- ◆ **Comprehensive Assessment.** The visual input, question, and answer were jointly input into the MLLM, which independently evaluated three aspects — *the correctness of the answer*, *the completeness of the objects involved in the question*, and *the discriminability of the answer choices* — and generated corresponding labels for each.

After obtaining all the above labels, we conducted manual verification to ensure their accuracy and performed preliminary data filtering based on these labels. Additionally, we carried out a second round of human evaluation on a small random sample to further ensure data quality.

## 2.2.2 Spatial Data Augmentation Pipeline

Each synthesized spatial data consists of VisualInput, Question, TargetAnswer, and CandidateOptions. We perform the following key data augmentation strategies:

- ◆ **Question-Type Augmentation.** To balance the distribution of question types, we transform original multiple-choice questions into fill-in-the-blank and true/false formats based on the TargetAnswer and CandidateOptions. For instance, a portion of ObjectCount questions are converted into numerical fill-in-the-blank tasks, while some SpatialRelation questions are reformulated as true/false questions. In the latter case, a question originally asking for "*left*" or "*right*" becomes a binary judgment such as: "*Is object A to the left of object B?*"
- ◆ **Instruction Augmentation.** To preserve instruction-following performance, we augment the Question by enriching prompts while ensuring consistency with the TargetAnswer. Specifically, we embed formatting requirements—such as units, response types, or structured formats—directly into the questions. This enables the accuracy-based reward to also encourage proper adherence to instructions. For example, a question like "*What is the distance between A and B?*" may be expanded to include "*measured in centimeters*" or "*measured in meters*", each paired with the corresponding correct answer.
- ◆ **Distribution Augmentation.** To address the bias introduced by imbalanced TargetAnswer distributions in the original data, we perform augmentation on the CandidateOptions. Specifically, we introduce a random offset to the option letters in both TargetAnswer and CandidateOptions, which helps reduce the model's reliance on specific option positions and encourages more robust and generalizable learning.

## 2.3 Overview of Data Configuration

The overview of data configurations for M2-Reasoning are depicted in Figure 2. The cold-start stage (Figure 2(a)) employs a large-scale dataset of 3.3 million image-text data and 2.9 million text-only samples. Within this stage, the "General Reasoning" data is sourced from Section 2.1 for the image-text modality and from Ling-Lite (Team, 2025) for the text-only modality. Additionally, to mitigate catastrophic forgetting and preserve the model's foundational capabilities, we incorporate a substantial amount of "Non-Thinking" data from the Instruction-Tuning-Stage-3 of M2-Omni (Guo

et al., 2025c). The subsequent RLVR stage (Figure 2(b)) utilizes a smaller, curated dataset to hone the model’s advanced reasoning abilities. This dataset is heavily focused on complex tasks such as mathematics, science, and fine-grained spatial reasoning, spanning both image and video modalities.



**Figure 2** Overview of the data configurations during cold-start and RLVR.

### 3 Approach

As depicted in Figure 3, M2-Reasoning is a unified model capable of supporting general and spatial reasoning tasks. The development of M2-Reasoning comprises two core components: (1) a dynamic reasoning incentive framework that harmonizes multi-task optimization, and (2) task-specific reward formulation that preserve domain specialization. The following sections will explore how M2-Reasoning achieves the superior ability across all tasks.

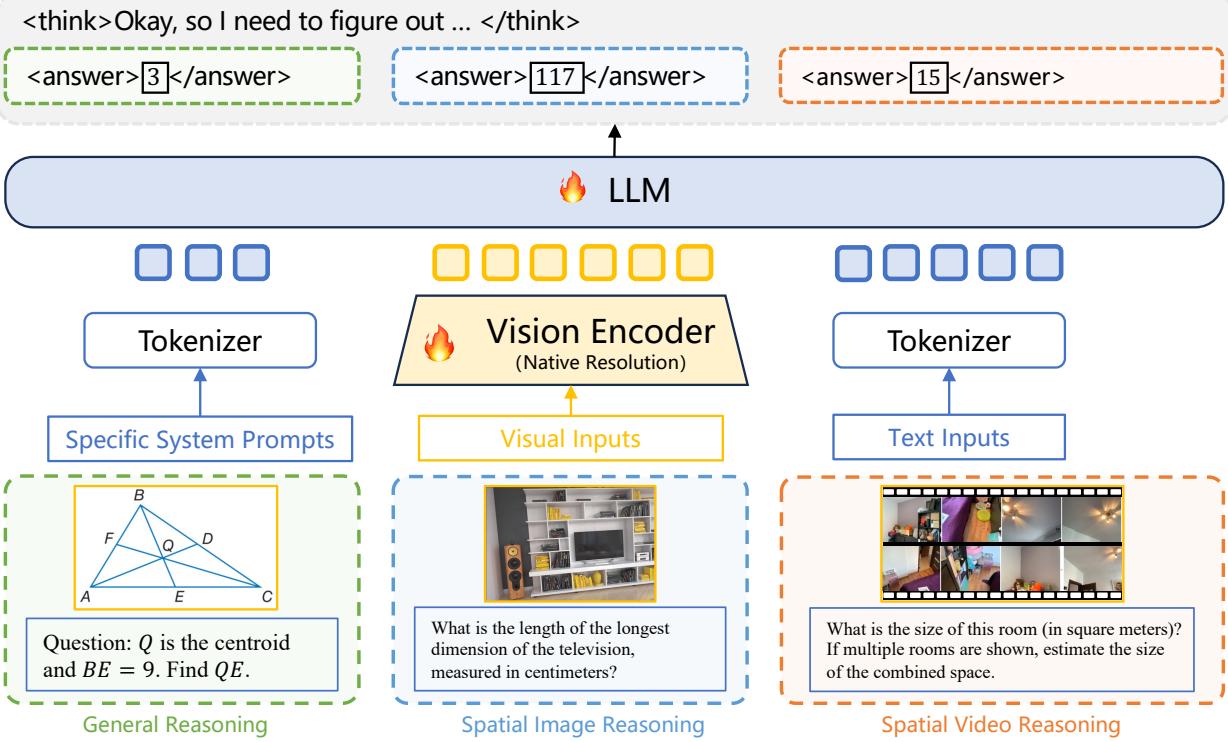
#### 3.1 Dynamic Reasoning Incentivizing

##### 3.1.1 Two-Stage Training Strategy

M2-Reasoning initializes by adapting the M2-Omni (Guo et al., 2025c) framework. We leverage its established training process and data, but crucially replace the Llama-3 (Grattafiori et al., 2024) LLM with Qwen2.5-7B-Instruct (Team, 2024) for its enhanced language modeling capabilities.

To advance multimodal reasoning performance, we propose a two-stage training framework.

**Stage 1: Cold-start.** In this stage, supervised fine-tuning (SFT) is conducted on cold-start data to activate the model’s latent reasoning capabilities and standardize its output format. This step helps mitigate instability during the early phases of RLVR training. Our preliminary experiments align with findings from previous studies (Huang et al., 2025; Yang et al., 2025d; Chen et al., 2025), which demonstrate that initializing the training process with high-quality and semantically



**Figure 3 The M2-Reasoning’s model architecture** is built upon the Qwen2.5-7B language model, incorporating a native-resolution vision encoder. Notably, the figure omits the MLP projector typically used to connect the vision encoder and the language model.

coherent trajectory data can enhance the upper bound of the model’s reasoning performance. Upon completing this stage, we obtained the cold-start initialized model, denoted as M2-Reasoning-CI.

**Stage 2: Dynamic Multi-task RLVR.** In the second stage, Reinforcement Learning with Verifiable Rewards (RLVR) is applied to data with verifiable answers, encouraging the model to adopt the correct reasoning process and improve its generalization across a diverse range of multimodal tasks. We adopt GRPO for our reasoning tasks due to its demonstrated effectiveness and efficiency in textual mathematical reasoning (Shao et al., 2024; Guo et al., 2025a). Furthermore, we optimize the GRPO objective function by introducing dynamic hyper-parameter adjustment, which leads to improved model inference performance.

Mathematically, let  $q$  denote a query and  $\{o_i\}_{i=1}^G$  represent a set of  $G$  completions generated by the old policy model  $\pi_{old}$ . Each completion  $o_i$  is assigned a reward  $R_i$  based on the task type of  $q$ , and the corresponding advantage estimate is formulated as:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)} \quad (1)$$

The reward  $R_i$  for each completion  $o_i$  consists of two components: the accuracy reward  $R_i^{acc}$ , which measures how correct the generated answer is, and the format reward  $R_i^{fmt}$ , which evaluates whether the response adheres to the required output structure. The **accuracy reward**  $R_i^{acc}$  is computed in a rule-based manner, with the specific calculation method determined by the task type (general or spatial). A detailed description of the computation can be found in Section 3.2. The **format reward**  $R_i^{fmt}$ , on the other hand, checks whether the response includes valid thinking and answer blocks, i.e.

`<think></think>` and `<answer></answer>`. It is defined as:

$$R_i^{\text{fmt}} = \mathbb{I}(\text{Valid thinking and answer blocks in } o_i) \quad (2)$$

where the indicator function  $\mathbb{I}$  returns 1 if the formatting is valid and 0 otherwise. The final reward can be computed as:

$$R_i = R_i^{\text{acc}} + R_i^{\text{fmt}}. \quad (3)$$

Our variant of GRPO maximizes the following objective function:

$$\begin{aligned} \mathcal{J}_{GRPO} = & \mathbb{E}_{q, \{o_i\} \sim \pi_{\theta_{old}}} \\ & \left[ \frac{\hat{\alpha}_{\{o_i\}}}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \left( r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_{i,t} \right) - \hat{\beta} \mathbb{D}_{KL}(\pi_\theta \| \pi_{\text{ref}}) \right] \end{aligned} \quad (4)$$

where  $r_{i,t}(\theta)$  represents the importance sampling term and is computed as:

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})} \quad (5)$$

The highlighted terms in Eq. 4 represent our key modification—the introduction of dynamic hyper-parameter adjustment—distinguishing our method from the vanilla GRPO. These extensions are further elaborated in Section 3.1.3.

### 3.1.2 Curriculum Sampling

Motivated by the progressive nature of human learning, we design our training framework to emulate the way humans acquire knowledge—starting with simple concepts and gradually advancing to more complex and challenging tasks. This staged learning process mirrors a curriculum-like structure, where each level builds upon the previous one to foster deeper understanding and reasoning.

Several recent studies (Team et al., 2025a,c; Guo et al., 2025d) have highlighted the effectiveness of curriculum-based sampling strategies in improving model generalization and learning efficiency. Following this principle, we *offline* pre-score and filter the training data based on task difficulty prior to the RLVR stage. The selected samples are then organized in increasing order of difficulty, enabling a step-wise, difficulty-ascending training paradigm that supports stable and effective learning.

### 3.1.3 Step-wise Dynamic Optimization

Our joint training framework is designed to overcome two primary obstacles to efficient learning. First, we address the pronounced data heterogeneity among tasks with differing modalities and complexities. To harmonize the training process, we employ a balanced data sampling mechanism that prevents any single task from dominating the learning process. Second, we recognize that a one-size-fits-all training approach is suboptimal for handling data of varying difficulty. To address this, we introduce dynamic hyper-parameter adjustment to focus the model’s effort on more instructive samples, thereby optimizing the learning process.

**Data Sampling Balance.** Following (Guo et al., 2025c), we adopt two principles for data sampling during both the cold-start and RLVR stages: (1) each training batch contains data from only one

specific task, and (2) training steps are uniformly distributed across all tasks throughout the training pipeline. These strategies ensure balanced computational resource allocation while preserving task-specific learning characteristics.

**Dynamic Hyper-parameter Adjustment.** During the RLVR stage, we introduce several modifications to the standard GRPO algorithm to improve performance, including dynamic advantage weighting and a cosine-annealed KL penalty, as described in Section 3.1.1.

Specifically, for advantage weighting, we adopt the insight from (Guo et al., 2025d) that data of moderate difficulty, which corresponding to an *online* accuracy reward of approximately 0.5, contains more informative signals. Such samples are therefore assigned higher weights. This leads to the following weight calculation:

$$\hat{\alpha}_{\{o_i\}_{i=1}^G} = \sigma \text{mean}(\{R_i^{\text{acc}}\}_{i=1}^G) (1 - \text{mean}(\{R_i^{\text{acc}}\}_{i=1}^G)) \quad (6)$$

where the scaling factor  $\sigma$  is set to 7.2. It is crucial to note that our dynamic weighting and curriculum sampling are two orthogonal components: curriculum sampling provides a static, pre-determined data ordering based on intrinsic difficulty, while dynamic weighting modulates sample importance in real-time, adapting to the model’s evolving capabilities.

Second, we employ a cosine annealing strategy for the KL penalty coefficient (Xia et al., 2025). Combined with our curriculum sampling approach, this helps stabilize training on easier data while encouraging longer and more meaningful outputs for more challenging examples. The updated coefficient  $\hat{\beta}$  is computed as:

$$\hat{\beta} = \frac{\beta}{2} \left( 1 + \cos \left( \pi \frac{T_{\text{cur}}}{T_{\max}} \right) \right) \quad (7)$$

where  $T_{\text{cur}}$  and  $T_{\max}$  denote the current and maximum training steps, respectively.

### 3.2 Task-specific Accuracy Reward Formulation

In this section, we provide a comprehensive explanation of the accuracy reward computation methods across various reasoning tasks, detailing the evaluation criteria and implementation strategies for each task type.

**General Reasoning.** The general reasoning data comprises two question types: multiple-choice questions and fill-in-the-blank questions. To effectively evaluate model outputs across these diverse formats, we design a rule-based reward mechanism that accounts for four distinct response types — exact option letter matching, numeric digit matching, extract string matching, and formal mathematical expression verification. Toward this end, we integrate the Math-Verify<sup>1</sup> package and customized string matching algorithm with case and punctuation insensitivity to rigorously assess the accuracy of generated responses.

Accordingly, the accuracy reward for the general reasoning task is defined as:

$$R_i^{\text{gen}} = \mathbb{I}(o_i, \text{ground truth}) \quad (8)$$

where  $\mathbb{I}$  denotes an indicator function that returns 1 if  $o_i$  matches the ground truth, and 0 otherwise.

**Spatial Reasoning.** In spatial reasoning tasks, certain fill-in-the-blank questions demand specific numerical answers—such as size of objects and distances between objects—where an exact match may not always be appropriate. Current MLLMs often struggle to percept absolute spatial metrics,

---

<sup>1</sup><https://github.com/huggingface/Math-Verify>

frequently producing rough and inaccurate estimates, particularly in the early stages of training. When binary rewards are assigned solely based on exact matches, the model faces significant challenges in generating responses that qualify as positive samples. This limitation hampers gradient accumulation impeding the learning process. Consequently, a more nuanced reward function is essential to better assess model outputs and facilitate meaningful learning, as it can provide a smooth reward even when the model’s initial predictions deviate significantly from the actual values, consistently encouraging it to optimize in the correct direction.

We designed a reward calculation mechanism, which we term Exponential Decay Numeric Matching (EDNM), based on smooth continuous absolute values. The core of this mechanism is a continuous reward function that penalizes predictions based on their normalized relative error. The reward  $R^{\text{EDNM}}(x)$  is calculated as follows:

$$R^{\text{EDNM}}(x) = \gamma \cdot \exp\left(-\lambda \cdot \frac{|x - x_{\text{gt}}|}{|x_{\text{gt}}| + \epsilon}\right) \quad (9)$$

where  $\gamma$  and  $\lambda$  are tunable hyper-parameters. Specifically,  $\gamma$  serves as a scaling factor defining the maximum reward, while  $\lambda$  controls the decay rate, determining the function’s sensitivity to prediction errors. Figure 4 illustrates this behavior, showing the reward for various predictions under different  $\lambda$  settings, given a ground truth of  $x_{\text{gt}} = 300$ . For our experiments, we set the hyper-parameters to  $\gamma = 1$  and  $\lambda = 2$ .

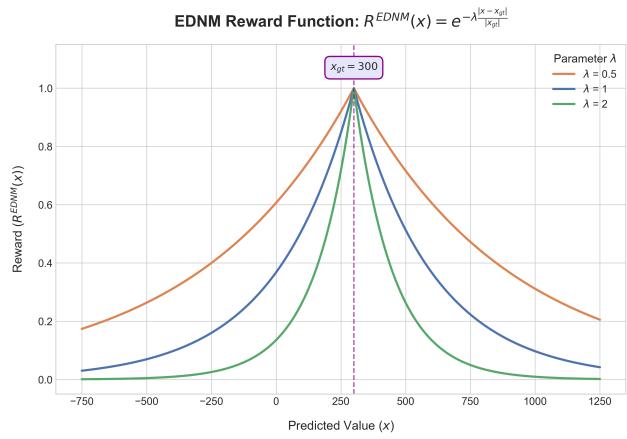
## 4 Experiments

We conduct a comprehensive evaluation of our models across two key domains: general and spatial reasoning. Our evaluation utilizes a diverse set of public benchmarks, grouped by the primary capability they measure:

- ◆ **General Reasoning (Mathematical & Logical, Section 4.1):** To evaluate this capability, we employ six benchmarks: MathVista (Lu et al., 2023), MathVision (Wang et al., 2024a), MathVerse (Zhang et al., 2024a), DynaMath (Zou et al., 2024), WeMath (Qiao et al., 2024), and LogicVista (Xiao et al., 2024).
- ◆ **Spatial Reasoning (Section 4.2):** We assess this skill using 2 benchmarks: CV-Bench (Tong et al., 2024) and VSI-Bench (Yang et al., 2025c).

### 4.1 General Reasoning Evaluation

As presented in Table 1, we evaluate our M2-Reasoning-7B against other leading base-scale MLLMs on general multimodal reasoning benchmarks. Our model achieves a new state-of-the-art average score of 45.0, surpassing both top general models like InternVL3-8B (41.4) and reasoning models such as WeThink-VL-7B (44.3). Notably, M2-Reasoning-7B obtains the best results on MathVista (75.0) and DynaMath (26.8), and secures competitive second-place rankings on MathVision (31.5) and LogicVista (50.0). This performance, representing a significant **+9.5** point improvement over



**Figure 4** Visualization of the EDNM reward function for different values of the hyperparameter  $\lambda$ .

**Table 1 Performance of MLLMs on general multimodal reasoning benchmarks.** Best and second-best results of base-scale MLLMs are in **bold** and underlined, respectively.  $\Delta$  denotes the performance improvement over the corresponding base model. All scores are from the OpenCompass leaderboard (Duan et al., 2024).

Models	Math-Vista	Math-Vision	Math-Verso	Dyna-Math	WeMath	Logic-Vista	Avg. ( $\Delta$ )
<b>Base-Scale General Models</b>							
InternVL3-8B (Zhu et al., 2025)	70.5	30.0	38.5	<u>25.7</u>	39.5	44.5	41.4
InternVL3-9B (Zhu et al., 2025)	69.0	29.3	37.9	25.1	34.8	49.0	40.8
Qwen2.5-VL-7B (Bai et al., 2025)	68.1	25.4	41.1	21.8	36.2	47.9	40.1
MUG-U-7B (Shopee, 2025)	<b>74.8</b>	26.1	35.4	17.2	26.5	39.8	36.6
SAIL-VL-1.6-8B (Dong et al., 2025)	74.2	23.2	33.4	14.0	29.6	41.4	36.0
<b>Base-Scale Reasoning Models</b>							
WeThink-VL-7B (Yang et al., 2025b)	71.6	26.0	44.2	24.8	<b>48.0</b>	<b>51.2</b>	<u>44.3</u> (+4.2)
Taichu-VLR-7B (WAIR, 2025)	72.3	27.1	<u>46.7</u>	23.0	<u>44.0</u>	48.3	43.6
VLAA-Thinker-7B (Chen et al., 2025)	68.0	26.4	<b>48.2</b>	22.4	41.5	48.5	42.5 (+2.4)
URSA-8B-PS-GRPO (Luo et al., 2025)	67.8	<b>31.8</b>	41.5	22.4	38.3	44.7	41.1 (+8.2)
Ovis2-8B (Lu et al., 2024)	71.8	25.9	42.3	20.4	27.2	39.4	37.8
<b>Our Models</b>							
Base Model	70.2	25.9	30.5	20.2	27.2	37.8	35.5
M2-Reasoning-CI-7B	71.7	29.2	42.1	25.0	42.8	46.8	42.9 (+7.4)
<b>M2-Reasoning-7B</b>	<b>75.0</b>	<u>31.5</u>	44.7	<b>26.8</b>	41.8	<u>50.0</u>	<b>45.0</b> (+9.5)

our base model, underscores the effectiveness of our approach in enhancing multimodal reasoning capabilities.

**Table 2 Performance of MLLMs on CV-Bench.** Best and second-best scores of base-scale MLLMs for each metric are shown in **bold** and underlined, respectively.  $\dagger$  Scores adapted from (Ray et al., 2025); \* scores reproduced by us.

Models	Count	Relation	Depth	Distance	Avg.
<b>Large-Scale Models</b>					
GPT-4O $\dagger$ (Achiam et al., 2023)	65.9	85.7	87.8	78.2	78.9
Gemini-1.5-pro $\dagger$ (Team et al., 2023)	70.4	85.2	82.4	72.8	77.4
<b>Base-Scale Models</b>					
InternVL3-8B* (Zhu et al., 2025)	<b>74.0</b>	<u>90.6</u>	<u>84.3</u>	<u>81.0</u>	<u>82.0</u>
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	65.2	86.6	70.6	79.8	75.0
LLava-NEXT-Video-7B $\dagger$ (Zhang et al., 2024b)	59.3	77.0	71.3	54.7	65.2
<b>Our Models</b>					
M2-Reasoning-7B	<u>66.6</u>	<b>92.8</b>	<b>89.3</b>	<b>84.3</b>	<b>82.3</b>

## 4.2 Spatial Reasoning Evaluation

As shown in Table 2, our model, M2-Reasoning-7B, achieves state-of-the-art performance on CV-Bench with an average score of 82.3, slightly surpassing the previous leading model, InternVL3-8B (82.0). Our model demonstrates exceptional strength in understanding complex spatial configurations, securing the top scores in Relation (92.8), Depth (89.3), and Distance (84.3). While InternVL3-8B maintains an edge in object counting, M2-Reasoning-7B’s superior ability to process and reason about relational and depth-related cues highlights its advanced spatial understanding capabilities.

On the more challenging VSI-Bench for nuanced video spatial imagination, M2-Reasoning-7B demonstrates highly competitive performance (Table 3). Its average score of 42.3 surpasses the strong InternVL3-8B baseline (42.1) and is second only to Gemini-1.5-pro (45.4). Notably, our

**Table 3 Performance of MLLMs on VSI-Bench.** Best and second-best scores of base-scale MLLMs are shown in **bold** and underlined. Abbreviations stand for: OC (Object Count), AD (Absolute Distance), OS (Object Size), RS (Room Size), RDs (Relative Distance), RDr (Relative Direction), RP (Route Plan), and AO (Appearance Order). <sup>†</sup>Scores adapted from the InternVL3 technical report (Zhu et al., 2025).

Models	OC	AD	OS	RS	RDs	RDr	RP	AO	Avg.
<b>Large-Scale Models</b>									
Gemini-1.5-pro <sup>†</sup> (Team et al., 2023)									
Gemini-1.5-pro <sup>†</sup> (Team et al., 2023)	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6	45.4
GPT-4O <sup>†</sup> (Achiam et al., 2023)	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5	34.0
<b>Base-Scale Models</b>									
InternVL3-8B <sup>†</sup> (Zhu et al., 2025)	<b>68.1</b>	<b>39.0</b>	48.4	33.6	<b>48.3</b>	36.4	27.3	<b>35.4</b>	<u>42.1</u>
Video-R1-7B (Feng et al., 2025)	-	-	-	-	-	-	-	-	37.1
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	37.7	20.1	<u>49.7</u>	<u>37.4</u>	38.5	40.4	<u>31.4</u>	<u>32.0</u>	35.9
LLava-NeXT-Video-7B <sup>†</sup> (Zhang et al., 2024b)	<u>48.5</u>	14.0	47.8	24.2	<u>43.5</u>	<u>42.4</u>	<b>34.0</b>	30.6	35.6
<b>Our Models</b>									
M2-Reasoning-7B	41.0	<u>34.0</u>	<b>60.9</b>	<b>55.4</b>	40.7	<b>47.3</b>	29.9	28.8	<b>42.3</b>

model’s key advantages lie in fine-grained analysis, where it establishes new state-of-the-art records for inferring Room Size (RS) at 55.4 and determining Relative Direction (RDr) at 47.3. Furthermore, it secures the second-best performance in estimating Object Size (OS) and Absolute Distance (AD). This highlights a sophisticated capability to infer complex spatial attributes like scale and orientation from dynamic scenes, showcasing its particular strengths in these nuanced reasoning tasks.

## 5 Conclusion

In this paper, we introduce M2-Reasoning-7B, a model engineered to master both general and spatial tasks to address the critical deficiency of MLLMs in dynamic spatial reasoning. Our success hinges on the synergy between two core innovations: a sophisticated data pipeline that generates and curates high-quality data with verifiable reasoning trajectories, and a dynamic training strategy that resolves learning conflicts through step-wise optimization and task-specific rewards. This powerful integration enables M2-Reasoning-7B to establish a new state-of-the-art across eight diverse benchmarks. However, we acknowledge several limitations that warrant further investigation:

- ◆ **Constrained Reasoning Depth:** Compared to specialized language-only reasoning models like DeepSeek-R1, M2-Reasoning-7B tends to produce shorter reasoning chains. This limits its ability to deconstruct and solve highly complex, multi-step problems that require deeper logical deliberation.
- ◆ **Pathological Repetition:** In some instances, the model exhibits pathological repetition, where it gets trapped in a loop, redundantly generating the same phrases or reasoning steps. This behavior indicates a potential instability in the generation process that can derail coherent thought progression.
- ◆ **Suboptimal Visual Perception:** Despite its strong reasoning capabilities, the model’s visual perception is not infallible. We observed occasional weaknesses in perceiving fine-grained details, leading to imprecise descriptions or, in some cases, visual hallucinations where the model fabricates non-existent objects or attributes.

We will continue to explore and address these challenges in our future work to further enhance the model’s robustness and reasoning capabilities.

## 6 Contributors

Authors are listed **alphabetically by the first name**.

Fudong Wang

Lixiang Ru

Yifan Mao

Jiajia Liu

Qingpei Guo

Yuting Xiao

Jingdong Chen

Ruobing Zheng

Ziping Ma

Jun Zhou

Tianqi Li

Kaixiang Ji

Yi Yuan

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Inclusion AI. Ming-omni: A unified multimodal model for perception and generation, 2025. <https://arxiv.org/abs/2506.09344>.

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.

Zheng Cai, Maosong Cao, Haojong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenna Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024.

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.

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*, 2025.

Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. Scalable vision language model training via high quality data curation. *arXiv preprint arXiv:2501.05952*, 2025.

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.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-rl: Reinforcing video reasoning in llms. *arXiv preprint arXiv:2503.21776*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, Shixiong Zhao, Shuai Peng, Shuangye Li, Sihang Yuan, Sijin Wu, Tianheng Cheng, Weiwei Liu, Wenqian Wang, Xianhan Zeng, Xiao Liu, Xiaobo Qin, Xiaohan Ding, Xiaojun Xiao, Xiaoying Zhang, Xuanwei Zhang, Xuehan Xiong, Yanghua Peng, Yangrui Chen, Yanwei Li, Yanxu Hu, Yi Lin, Yiyuan Hu, Yiyuan Zhang, Youbin Wu, Yu Li, Yudong Liu, Yue Ling, Yujia Qin, Zanbo Wang, Zhiwu He, Aoxue Zhang, Bairen Yi, Bencheng Liao, Can Huang, Can Zhang, Chaorui Deng, Chaoyi Deng, Cheng Lin, Cheng Yuan, Chenggang Li,

Chenhui Gou, Chenwei Lou, Chengzhi Wei, Chundian Liu, Chunyuan Li, Deyao Zhu, Donghong Zhong, Feng Li, Feng Zhang, Gang Wu, Guodong Li, Guohong Xiao, Haibin Lin, Haihua Yang, Haoming Wang, Heng Ji, Hongxiang Hao, Hui Shen, Huixia Li, Jiahao Li, Jialong Wu, Jianhua Zhu, Jianpeng Jiao, Jiashi Feng, Jiaze Chen, Jianhui Duan, Jihao Liu, Jin Zeng, Jingqun Tang, Jingyu Sun, Joya Chen, Jun Long, Junda Feng, Junfeng Zhan, Junjie Fang, Junting Lu, Kai Hua, Kai Liu, Kai Shen, Kaiyuan Zhang, Ke Shen, Ke Wang, Keyu Pan, Kun Zhang, Kunchang Li, Lanxin Li, Lei Li, Lei Shi, Li Han, Liang Xiang, Liangqiang Chen, Lin Chen, Lin Li, Lin Yan, Liying Chi, Longxiang Liu, Mengfei Du, Mingxuan Wang, Ningxin Pan, Peibin Chen, Pengfei Chen, Pengfei Wu, Qingqing Yuan, Qingyao Shuai, Qiuyan Tao, Renjie Zheng, Renrui Zhang, Ru Zhang, Rui Wang, Rui Yang, Rui Zhao, Shaoqiang Xu, Shihao Liang, Shipeng Yan, Shu Zhong, Shuaishuai Cao, Shuangzhi Wu, Shufan Liu, Shuhan Chang, Songhua Cai, Tenglong Ao, Tianhao Yang, Tingting Zhang, Wanjun Zhong, Wei Jia, Wei Weng, Weihao Yu, Wenhai Huang, Wenjia Zhu, Wenli Yang, Wenzhi Wang, Xiang Long, XiangRui Yin, Xiao Li, Xiaolei Zhu, Xiaoying Jia, Xijin Zhang, Xin Liu, Xincheng Zhang, Xinyu Yang, Xiongcai Luo, Xiuli Chen, Xuantong Zhong, Xuefeng Xiao, Xujing Li, Yan Wu, Yawei Wen, Yifan Du, Yihao Zhang, Yining Ye, Yonghui Wu, Yu Liu, Yu Yue, Yufeng Zhou, Yufeng Yuan, Yuhang Xu, Yuhong Yang, Yun Zhang, Yunhao Fang, Yuntao Li, Yurui Ren, Yuwen Xiong, Zehua Hong, Zehua Wang, Zewei Sun, Zeyu Wang, Zhao Cai, Zhaoyue Zha, Zhecheng An, Zehui Zhao, Zhengzhuo Xu, Zhipeng Chen, Zhiyong Wu, Zhuofan Zheng, Zihao Wang, Zilong Huang, Ziyu Zhu, and Zuquan Song. Seed1.5-vl technical report, 2025b. <https://arxiv.org/abs/2505.07062>.

Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Tai-Wei Chang, Jingdong Chen, et al. M2-omni: Advancing omni-mllm for comprehensive modality support with competitive performance. *arXiv preprint arXiv:2502.18778*, 2025c.

Zirun Guo, Minjie Hong, and Tao Jin. Observe-r1: Unlocking reasoning abilities of mllms with dynamic progressive reinforcement learning. *arXiv preprint arXiv:2505.12432*, 2025d.

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.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. <https://arxiv.org/abs/2408.03326>.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024b. <https://arxiv.org/abs/2407.07895>.

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.

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.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

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.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursu: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*, 2025.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhui Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

OpenAI. Gpt-4o system card, August 2024. <https://openai.com/index/gpt-4o-system-card/>.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoqian Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models, 2025. <https://arxiv.org/abs/2412.07755>.

ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, et al. Seed1. 5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*, 2025.

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.

Shopee. Mug-u. <https://github.com/Shopee-MUG/MUG-U>, 2025. Accessed: July 11, 2025.

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.

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*, 2025a.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinhao Li, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yuhao Dong, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. Kimi-vl technical report, 2025b. <https://arxiv.org/abs/2504.07491>.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025c.

Ling Team. Every flop counts: Scaling a 300b mixture-of-experts ling llm without premium gpus. *arXiv preprint arXiv:2503.05139*, 2025.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. <https://qwenlm.github.io/blog/qwen2.5/>.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

WAIR. Taichu-vlr. <https://platform.wair.ac.cn/taichu#/modellist>, 2025. Accessed: July 11, 2025.

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, 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

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

Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.

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

LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. <https://arxiv.org/abs/2506.03569>.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jie Yang, Feipeng Ma, Zitian Wang, Dacheng Yin, Kang Rong, Fengyun Rao, and Ruimao Zhang. Wethink: Toward general-purpose vision-language reasoning via reinforcement learning. *arXiv preprint arXiv:2506.07905*, 2025b.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025c.

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*, 2025d.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

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, 2024a.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b. <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>.

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, 2025. <https://arxiv.org/abs/2503.05132>.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

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.

## A Prompts Used for High-Quality Data Construction Pipeline

### Prompt for scoring synthesized general CoT data

You are given a written mathematical solution. Please rate the quality of its reasoning process on a scale from 1 (lowest) to 5 (highest) based on the following three criteria:

1. **Optimal Structural Organization:** Is the reasoning process clearly and logically structured? Does it break down complex steps with sufficient detail, while keeping straightforward steps concise? High-quality solutions adapt the level of detail (granularity) to the complexity of each step, ensuring that critical transitions receive appropriate elaboration.
2. **Effective Cognitive Scaffolding:** Does the solution guide the reader's understanding by gradually building up concepts and insights? Are key ideas introduced at the right time, and does the explanation bridge conceptual gaps thoughtfully, making the solution easier to follow and learn from?
3. **Rigorous Verification:** Does the reasoning include frequent verification of intermediate results? Are assumptions explicitly checked, and is logical consistency maintained throughout the solution? A high-quality answer thoroughly validates its correctness at each stage.

#### Scoring Rubric:

- **5 — Excellent:** Exemplary reasoning with adaptive step granularity, strong pedagogical clarity, and thorough verification.
- **4 — Good:** Clear and mostly well-paced explanation with minor issues in elaboration or verification.
- **3 — Fair:** Reasonable logic but with noticeable issues in structure, clarity, or verification.
- **2 — Poor:** Disorganized or incomplete reasoning with major gaps in explanation or checks.
- **1 — Very Poor:** Lacks coherent structure, explanation, and verification; difficult to understand or trust.

The given reasoning process is: REASONING

You MUST output your score with one single number.

### Prompt for scoring the difficulty of RLVR data

### For multiple-choice questions:

According to the following question, please first conduct step by step reasoning, then answer the question and put the correct option letter, e.g., A, B, C, D, within \boxed{ }.

### For fill-in-the-blank questions:

According to the following question, please first conduct step by step reasoning, then answer the question and put the final answer within \boxed{ }.

## Prompt for validating the spatial reasoning data

You are a logic consistency evaluator. Evaluate the Input by the Evaluation Rules and output the final response with Output Format.

### Input:

- Question: QUESTION
- Choice List: ANSWER\_CHOICES
- Answer: CORRECT\_ANSWER

### Evaluation Rules:

1. Thoroughly analyze the provided image, question, answer, and the choice list before scoring.
2. Evaluate the data quality based on the following three dimensions, providing a score of 1 (Good) or 0 (Bad) for each.
  - ObjectS (Object Score): Assess if the object(s) relevant to the question are clearly visible and identifiable in the image.
    - Score 1 if the objects are clear and unambiguous.
    - Score 0 if the objects are significantly occluded, blurry, overlapping, too small, cut off at the image edge, or otherwise difficult to discern in a way that impedes answering the question.
  - AnswerS (Answer Score): Verify if the provided answer is correct and unambiguous based on the image and the question.
    - Score 1 if the answer is factually correct and logically sound.
    - Score 0 if the answer is ambiguous, factually incorrect, inconsistent with the image, or cannot be confirmed from the provided information.
  - OptionS (Options Score): Examine the quality and validity of the Choice List.
    - Score 1 if all choices are meaningful, distinct, and plausible distractors that don't confuse the question's intent.
    - Score 0 if choices are redundant, highly similar, nonsensical, or clearly irrelevant to the question or image.

### Output Format:

1. Output MUST be valid JSON format which can be loaded by json.loads() and contains the reasoning step.

2. Example:

```
{  
    "Reason\_Step": <reasoning step by step>,  
    "ObjectS": 1,  
    "AnswerS": 1,  
    "OptionS": 0  
}
```

3. Please reason step by step, and put your final JSON format output within \boxed{ }.

## B Experimental Implementation Details

### B.1 System Prompts

To guide the reasoning process during training, we employ the following system prompt in both the cold-start and RLVR stages:

#### System Prompt

You are a helpful assistant. When the user asks a question, your response must include two parts: first, the reasoning process enclosed in `<think>...</think>` tags, then the final answer enclosed in `<answer>...</answer>` tags. The critical answer or key result should be placed within `\boxed{}`.

### B.2 Training Hyper-parameters for RLVR

Our training framework is built upon VisualThinker-R1-Zero ([Zhou et al., 2025](#)), and all RLVR hyper-parameters are presented in Table 4.

**Table 4** Training Hyper-parameters for RLVR

	General	Spatial <sub>image</sub>	Spatial <sub>video</sub>
Learning Rate		1e-6	
Initial KL Coefficient		0.01	
Completion Max Length		1024	
Temperature		1.0	
Batch Size		64	
Epochs	1	2	3
Number of Generations( $G$ )	16	16	4

### B.3 Benchmarks

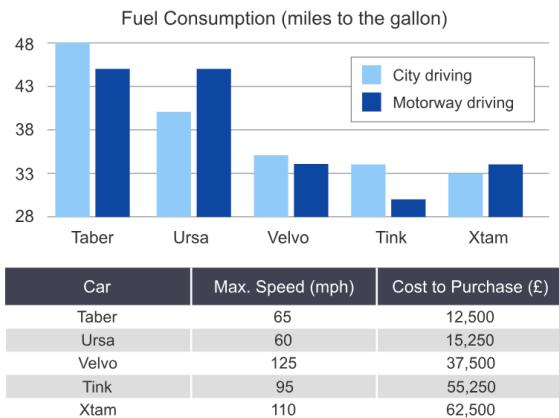
As illustrated in Section 4.1, we conduct evaluation across six mathematical and logical reasoning benchmarks. Below are the details:

- ◊ **MathVista:** We use the Test Mini split of the MathVista dataset.
- ◊ **MathVision:** We use the full test set of the MathVision dataset.
- ◊ **MathVerse:** We use the Test Mini split of MathVerse in "Vision Only" mode.
- ◊ **DynaMath:** We use the full test set of DynaMath, and report the worst-case accuracy as the main metric.
- ◊ **WeMath:** We use the Test Mini split of WeMath, and report the "Score (Strict)" as the main metric.
- ◊ **LogicVista:** We use the full test set of LogicVista.

## C Visualization Results

The following examples showcases visualization results of M2-Reasoning on general and spatial reasoning tasks. These examples demonstrate that M2-Reasoning-7B possesses strong logical and analytical capabilities, enabling it to arrive at the correct answers.

### General Reasoning



**Question:** If a driver travels an average of 4,250 miles per month driving only along motorways in an Xtam car, what is the predicted annual consumption of fuel (in gallons)? Select from A, B, C, D and E.

**Options:**

- (A) Cannot say
- (B) 1,500
- (C) 125
- (D) 150
- (E) 1,250

**GT Answer:** B

---

### M2-Reasoning-7B:

<think>

To determine the predicted annual consumption of fuel for an Xtam car driven only along motorways, we need to follow these steps:

1. Identify the fuel consumption for the Xtam car on the motorway from the bar chart.
2. Calculate the monthly fuel consumption.
3. Multiply the monthly fuel consumption by 12 to get the annual consumption.

From the bar chart:

- The fuel consumption for the Xtam car on the motorway is 33.5 miles per gallon.

Next, we need to calculate the monthly fuel consumption:

- The driver travels 4,250 miles per month.

- Fuel consumption = Total miles driven / Miles per gallon
- Monthly fuel consumption = 4,250 miles / 33.5 miles per gallon  $\approx$  126.76 gallons

Finally, we calculate the annual fuel consumption:

- Annual fuel consumption = Monthly fuel consumption \* 12
- Annual fuel consumption  $\approx$  126.76 gallons/month \* 12 months  $\approx$  1,521.12 gallons

The closest option to 1,521.12 gallons is 1,500 gallons.

Therefore, the predicted annual consumption of fuel is approximately 1,500 gallons.

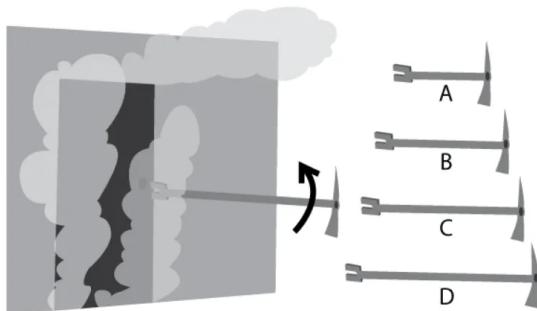
</think>

<answer>

B

</answer>

## General Reasoning



**Question:** Firefighters use a halligan to forcibly enter locked doors during fire rescue operations. Which halligan would require the least effort to forcibly open the door? Select from A, B, C, D, and E.

**Options:**

- (A) A
- (B) B
- (C) C
- (D) D
- (E) They would all require the same effort

**GT Answer:** D

### M2-Reasoning-7B:

<think>

To determine which Halligan tool would require the least effort to forcibly open the door, we need to consider the mechanical advantage provided by each tool. The Halligan tool is a lever, and the effort required to lift the tool is inversely proportional to the distance from the fulcrum to the point where the force is applied.

- Tool A has a short handle and a long lever arm.
- Tool B has a slightly longer handle and a shorter lever arm.
- Tool C has a medium-length handle and a medium-length lever arm.
- Tool D has the longest handle and the shortest lever arm.

The longer the handle, the greater the mechanical advantage, meaning less effort is required to lift the tool. Therefore, the tool with the longest handle and the shortest lever arm would require the least effort.

From the options given, Tool D has the longest handle and the shortest lever arm, providing the greatest mechanical advantage.

Thus, the tool that would require the least effort to forcibly open the door is Tool D.

</think>

<answer>

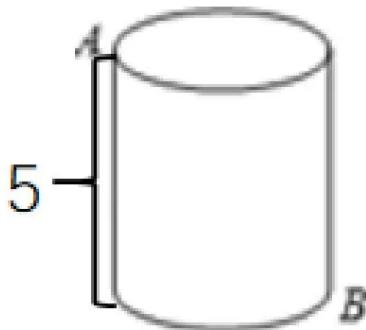
D

</answer>

## General Reasoning

As shown in the figure, a cylinder with a bottom circumference of 24.0, the shortest route that an ant passes along the surface from point A to point B is ()

Choices: A:12m B:15m C:13m D:14m



**Question:** According to the question shown in the image, please directly answer the question and provide the correct option letter, e.g., A, B, C, D.

**GT Answer:** C

---

### M2-Reasoning-7B:

<think>

To solve this problem, we need to determine the shortest path an ant would take on the surface of the cylinder from point A to point B. The key is to unroll the cylinder into a flat surface, which will help us visualize the path more easily.

1. **Circumference of the Cylinder:** The circumference is given as 24.0 meters. This means if we unroll the cylinder, the length of the bottom edge of the rectangle will be 24.0 meters.

2. **Height of the Cylinder:** The height of the cylinder is given as 5 meters.

3. **Unrolling the Cylinder:** When we unroll the cylinder, it forms a rectangle where one side is the height of the cylinder (5 meters) and the other side is the circumference of the base (24.0 meters).

4. **Points A and B:** Point A is at the bottom left corner of the rectangle, and point B is at the top right corner of the rectangle.

5. **Shortest Path:** The shortest path from A to B on the unrolled rectangle is the straight-line distance between these two points. This forms a right triangle where one leg is the height of the cylinder (5 meters) and the other leg is half the circumference (since the ant can go either directly around the cylinder or the other way, but the shortest path is the direct line across the rectangle).

The distance can be calculated using the Pythagorean theorem:

$$\text{Distance} = \sqrt{(24.0)^2 + (5)^2} = \sqrt{576 + 25} = \sqrt{601} \approx 24.5 \text{ meters}$$

However, the options provided are much smaller, suggesting we might need to consider the direct path around the cylinder or a simpler calculation. Given the options, the correct path is likely the direct path across the rectangle, which is the hypotenuse of a right triangle with legs 5 meters and 12 meters (half the circumference).

$$\text{Distance} = \sqrt{(12)^2 + (5)^2} = \sqrt{144 + 25} = \sqrt{169} = 13 \text{ meters}$$

Thus, the shortest path is 13 meters.

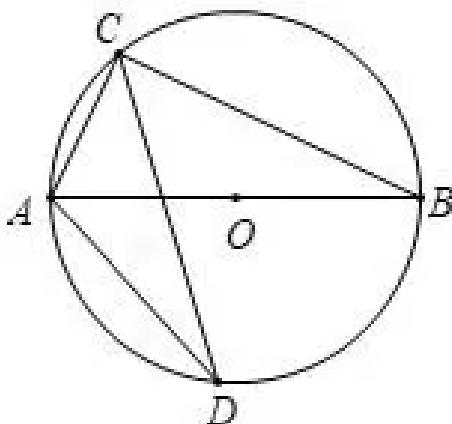
</think>

<answer>

C

</answer>

## General Reasoning



**Question:** As shown in the figure, AB is the diameter of  $\odot O$ , CD is the chord of  $\odot O$ ,  $\angle ADC = 26.0$ , then the degree of  $\angle CAB$  is ()

**Options:**

- (A)  $26^\circ$   
 (B)  $74^\circ$   
 (C)  $64^\circ$   
 (D)  $54^\circ$

**GT Answer:** C

### M2-Reasoning-7B:

<think>

To solve the problem, we need to use some properties of circles and angles.

1. Since AB is the diameter of the circle, the angle subtended by the diameter at any point on the circle is a right angle. This means  $\angle ACB = 90^\circ$ .
2. We are given that  $\angle ADC = 26^\circ$ . Since  $\angle ADC$  and  $\angle ABC$  are both inscribed angles that subtend the same arc AC, they are equal. Therefore,  $\angle ABC = 26^\circ$ .
3. In triangle ABC, we know that the sum of the angles in a triangle is  $180^\circ$ . So, we can write:  $\angle CAB + \angle ABC + \angle ACB = 180^\circ$

Substituting the known values:

$$\angle CAB + 26^\circ + 90^\circ = 180^\circ$$

Simplifying this, we get:

$$\angle CAB + 116^\circ = 180^\circ$$

$$\angle CAB = 180^\circ - 116^\circ = 64^\circ$$

Therefore, the measure of  $\angle CAB$  is  $64^\circ$ .

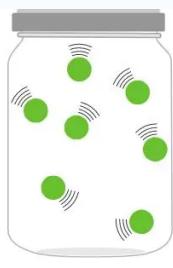
</think>

<answer>

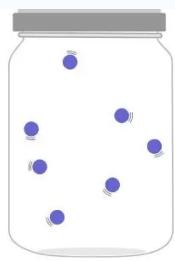
C)  $64^\circ$

</answer>

### General Reasoning



Sample A



Sample B

Mass of each particle: 40 u

Mass of each particle: 32 u

Average particle speed: 950 m/s

Average particle speed: 750 m/s

**Question:** The diagrams below show two pure samples of gas in identical closed, rigid containers. Each colored ball represents one gas particle. Both samples have the same number of particles. Compare the average kinetic energies of the particles in each sample. Which sample has the higher temperature?

**Options:**

- (A) neither; the samples have the same temperature
- (B) sample A
- (C) sample B

**GT Answer:** B**M2-Reasoning-7B:**

&lt;think&gt;

To determine which sample has the higher temperature, we need to compare the average kinetic energies of the particles in each sample. The average kinetic energy of a gas particle is given by the formula:

$$\text{Kinetic Energy} = \frac{1}{2}mv^2$$

where  $m$  is the mass of the particle and  $v$  is its speed.

The temperature of a gas is directly proportional to the average kinetic energy of its particles. Therefore, the sample with the higher average kinetic energy will have the higher temperature.

Let's calculate the average kinetic energy for each sample:

For Sample A:

- Mass of each particle,  $m_A = 40\text{u}$
- Average particle speed,  $v_A = 950\text{m/s}$

$$\text{Kinetic Energy for Sample A: } KE_A = \frac{1}{2} \times 40 \times (950)^2$$

For Sample B:

- Mass of each particle,  $m_B = 32\text{u}$
- Average particle speed,  $v_B = 750\text{m/s}$

Kinetic Energy for Sample B:

$$KE_B = \frac{1}{2} \times 32 \times (750)^2$$

Now, let's compare the two kinetic energies:

$$KE_A = \frac{1}{2} \times 40 \times (950)^2$$

$$KE_B = \frac{1}{2} \times 32 \times (750)^2$$

Since the mass and speed of the particles in Sample A are both higher than those in Sample B, the kinetic energy of the particles in Sample A will be higher.

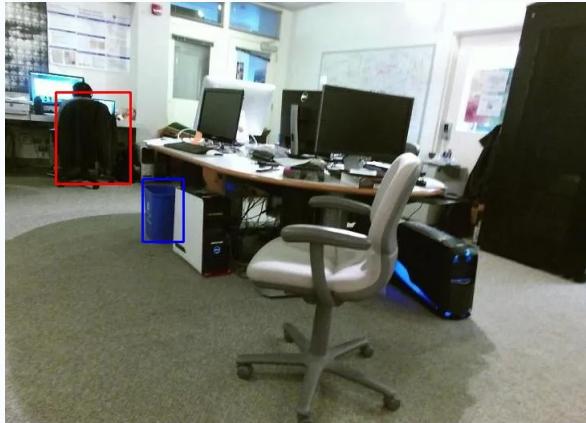
Therefore, Sample A has a higher average kinetic energy, and thus a higher temperature.

&lt;/think&gt;

&lt;answer&gt;

*B) sampleA*  
  </answer>

## Spatial Reasoning



**Question:** Which object is closer to the camera taking this photo, the chair (highlighted by a red box) or the bin (highlighted by a blue box)?

**Options:**

- (A) chair
- (B) bin

**GT Answer:** B

---

### M2-Reasoning-7B:

<think>

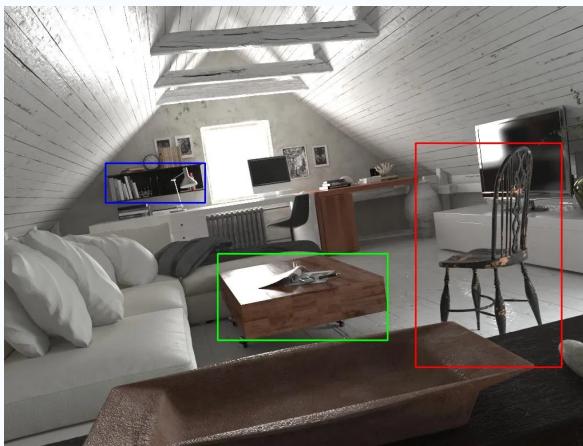
To determine which object is closer to the camera, we need to analyze the relative positions and sizes of the objects in the image. The chair highlighted by the red box is positioned in the background, far from the camera, as it is located at the far end of the room. The bin highlighted by the blue box is closer to the foreground, as it is situated near the center of the image and appears larger in the frame, indicating it is nearer to the camera.

Given the spatial relationships and the perspective, the bin is closer to the camera than the chair.

</think>

<answer>  
  (B)bin  
  </answer>

## Spatial Reasoning



**Question:** Estimate the real-world distances between objects in this image. Which object is closer to the chair (highlighted by a red box), the bookcase (highlighted by a blue box) or the table (highlighted by a green box)?

**Options:**

- (A) bookcase
- (B) table

**GT Answer:** B

---

### M2-Reasoning-7B:

<think>

To determine which object is closer to the chair highlighted by the red box, we need to consider the spatial relationships between the objects in the image.

1. The chair is located on the right side of the image.
2. The bookcase is on the left side of the image, behind the couch and the desk.
3. The table is in the center of the image, closer to the chair than the bookcase.

Given the positions:

- The table is directly in front of the chair and appears to be closer to it.
- The bookcase is further away, on the opposite side of the room.

Therefore, the table is closer to the chair than the bookcase.

</think>

<answer>

(B)table

</answer>

## Spatial Reasoning



**Question:** Considering the relative positions of the person and the dog in the image provided, where is the person located with respect to the dog? Select from the following choices.

**Options:**

- (A) left
- (B) right

**GT Answer:** A

---

### M2-Reasoning-7B:

<think>

The image shows a dog wearing a cowboy hat sitting in the back seat of a car. To the left of the dog, there is a person wearing a plaid shirt. The dog is positioned more towards the right side of the image, and the person is to the left of the dog. Therefore, the person is located to the left of the dog.

</think>

<answer>

- (A) left

</answer>

## Spatial Reasoning



**Question:** These are frames of a video. How many chair(s) are in this room? Please answer the question using a single word or phrase.

**GT Answer:** 3

---

### M2-Reasoning-7B:

<think>

To determine the number of chairs in the room, I will carefully analyze the images provided, focusing on identifying distinct chair shapes and structures. I will note any visible chairs and ensure no duplicates are counted.

The images show a modern kitchen and a dining area. In the kitchen, there is a visible chair near the dining table. Another chair is seen in the dining area, and a third chair is partially visible in the background. No other chairs are evident in the other sections of the room.

Counting the chairs: 1 (near the dining table) + 1 (in the dining area) + 1 (partially visible) = 3 chairs.

</think>

<answer>

</answer>

## Spatial Reasoning



**Question:** These are frames of a video. What will be the first-time appearance order of the following categories in the video: blanket, sofa, trash can, microwave?

**Options:**

- A. blanket, sofa, trash can, microwave
- B. sofa, microwave, blanket, trash can
- C. sofa, blanket, trash can, microwave
- D. sofa, microwave, trash can, blanket

Answer with the option's letter from the given choices directly.

**GT Answer: C**

---

### M2-Reasoning-7B:

<think>

To determine the first-time appearance order of the categories in the video, I will analyze the sequence of frames provided and identify the order in which each category (blanket, sofa, trash can, microwave) appears.

1. The first frame shows a living room with a sofa, a table, and a potted plant. The sofa is visible.
2. The second frame shows a dining area with a table and chairs, but no sofa, blanket, or microwave.
3. The third frame shows a view of the living room again, with the sofa and a blanket on it.
4. The fourth frame shows a view of the living room, with the sofa and a blanket.
5. The fifth frame shows a view of the living room, with the sofa and a blanket.
6. The sixth frame shows a view of the living room, with the sofa and a blanket.
7. The seventh frame shows a view of the living room, with the sofa and a blanket.
8. The eighth frame shows a view of the kitchen, with a trash can and a microwave.

From the frames, the order of appearance is:

- Sofa appears first.
- Blanket appears next.
- Trash can appears next.
- Microwave appears last.

Thus, the correct order is: sofa, blanket, trash can, microwave.

</think>

<answer>

C

</answer>