

# Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning

Huajie Tan<sup>1,2,\*</sup>, Yuheng Ji<sup>2,3,4,\*</sup>, Xiaoshuai Hao<sup>2,\*†</sup>, Minglan Lin<sup>2</sup>, Pengwei Wang<sup>2,†</sup>, Zhongyuan Wang<sup>2</sup>, Shanghang Zhang<sup>1,2,✉</sup>

<sup>1</sup> State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

<sup>2</sup> Beijing Academy of Artificial Intelligence <sup>3</sup> Institute of Automation, Chinese Academy of Sciences

<sup>4</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences

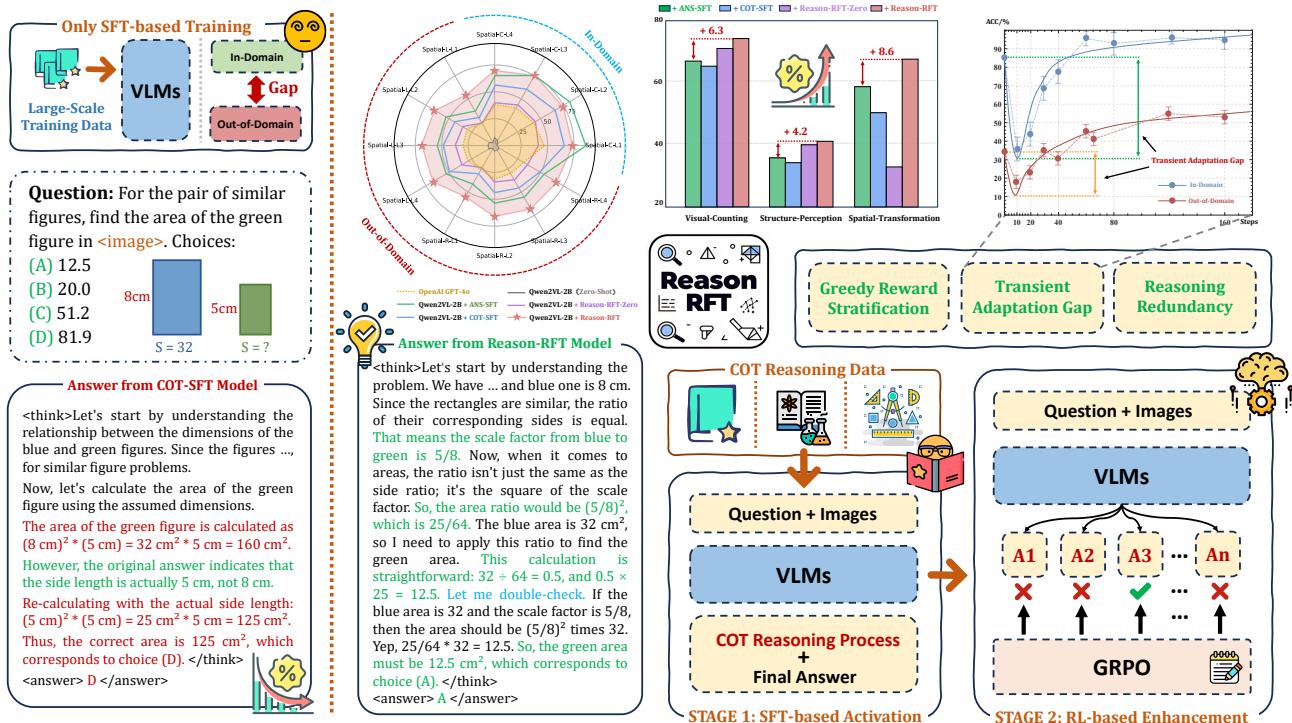


Figure 1. Overview of Reason-RFT. Compared to traditional SFT-based methods, our proposed Reason-RFT framework demonstrates superior generalization in visual reasoning tasks, excelling in reasoning improvement, out-of-domain performance, and data efficiency.

## Abstract

Visual reasoning abilities play a crucial role in understanding complex multimodal data, advancing both domain-specific applications and artificial general intelligence (AGI). Existing methods improve VLM reasoning via Chain-of-Thought (CoT) supervised fine-tuning, using meticulously annotated training data to enhance visual reasoning capabilities. However, this training paradigm may lead to overfitting and cognitive rigidity, restricting the model's ability to transfer visual reasoning skills across domains

and limiting its real-world applicability. To address these limitations, we propose **Reason-RFT**, a novel reinforcement fine-tuning framework that significantly enhances generalization capabilities in visual reasoning tasks. **Reason-RFT** introduces a two-phase training framework for visual reasoning: (1) Supervised Fine-Tuning (SFT) with curated Chain-of-Thought (CoT) data activates the reasoning potential of Vision-Language Models (VLMs), followed by (2) Group Relative Policy Optimization (GRPO)-based reinforcement learning that generates multiple reasoning-response pairs, significantly enhancing generalization in visual reasoning tasks. To evaluate Reason-RFT's visual reasoning capabilities, we reconstructed a comprehensive dataset spanning visual counting, structure percep-

\* Equal contribution.

† Project leaders.

✉ Corresponding author.

*tion, and spatial transformation, serving as a benchmark to systematically assess visual cognition, geometric understanding, and spatial generalization. Experimental results demonstrate Reasoning-RFT’s three key advantages: (1) **Performance Enhancement**: achieving state-of-the-art results across multiple tasks, outperforming most mainstream open-source and proprietary models; (2) **Generalization Superiority**: consistently maintaining robust performance across diverse tasks and domains, outperforming alternative training paradigms; (3) **Data Efficiency**: excelling in few-shot learning scenarios while surpassing full-dataset SFT baselines; **Reason-RFT** introduces a novel paradigm in visual reasoning, significantly advancing multimodal research. Project website: [ReasonRFT](#).*

## 1. Introduction

Visual reasoning is pivotal for understanding complex multimodal data and advancing artificial general intelligence (AGI) [38, 45], making it a central focus in intelligent systems research. Recent advancements in image recognition [27, 48] and scene understanding [9, 64] have enabled transformative applications in healthcare [47, 66], robotics [28, 36], and autonomous driving [19–22]. Consequently, enhancing visual reasoning capabilities has garnered significant attention from both industry and academia for its potential to drive transformative advancements.

Researchers have explored two primary categories of methods to enhance visual reasoning capabilities: (1) neural-symbolic methods [4, 7, 13, 16, 69], which integrate symbolic reasoning with neural networks to improve interpretability and modularity, and (2) Supervised Fine-Tuning (SFT) based on vision-language models (VLMs) [57, 63], which utilize end-to-end training to strengthen reasoning abilities. However, both approaches face significant limitations. Neural-symbolic methods are hindered by high complexity and a strong reliance on program generation, while SFT is constrained by its dependence on high-quality Chain-of-Thought (CoT) annotated data and meticulously designed data mixing strategies, leading to issues such as overfitting, cognitive rigidity, and limited transferability from in-domain to out-of-domain tasks. These challenges reduce their effectiveness in real-world applications.

Recent advances in models such as GPT-o1 [45], DeepSeek-R1 [15], and Kimi-1.5 [56] have demonstrated that large-scale reinforcement learning (RL) during post-training significantly enhances reasoning abilities in mathematics and coding. By providing a dynamic and adaptive alternative to static datasets, RL effectively addresses traditional limitations and improves reasoning, even with minimal data. However, purely RL-based methods face challenges in generalizing across visual reasoning tasks, often limited by the model’s inherent cognitive constraints and the

heterogeneity of multimodal data. Motivated by these observations, this paper investigates advanced RL techniques to enhance the adaptability and transferability of visual reasoning, aiming to overcome generalization challenges and improve performance in real-world applications.

In this paper, we propose **Reason-RFT**, a novel two-phase reinforcement fine-tuning framework designed to enhance generalization in visual reasoning tasks. First, we employ Supervised Fine-Tuning (SFT) with Chain-of-Thought (CoT) reasoning to activate the model’s potential reasoning capabilities, using a high-quality domain-specific visual reasoning dataset tailored to stimulate related reasoning abilities. Subsequently, we further enhance reasoning potential through Group Relative Policy Optimization (GRPO), demonstrating that **Reason-RFT** achieves superior generalization by pushing the model’s reasoning limits. To evaluate its capabilities, we reconstructed a high-quality dataset spanning visual counting, structure perception, and spatial transformation, serving as a benchmark for assessing visual cognition, geometric understanding, and cross-task generalization. Extensive experiments highlight three key advantages of *Reason-RFT*: (1) *Performance Improvement*: It significantly outperforms mainstream VLMs in tasks like visual counting, structure perception, and spatial transformation; (2) *Enhanced Generalization*: It consistently surpasses SFT-only and RL-only paradigms across diverse tasks, validated through multi-dimensional evaluations; (3) *Data Efficiency*: It achieves over 95% of the performance of SFT-only approaches using less than 20% of the data. These results underscore the effectiveness and efficiency of *Reason-RFT*, offering a robust solution for advancing visual reasoning tasks.

Our main contributions are summarized as follows.

- We propose **Reason-RFT**, a novel two-phase reinforcement fine-tuning framework that significantly enhances the visual reasoning capabilities of VLMs by effectively combining the complementary strengths of SFT-based and RL-based methods.
- We systematically analyze SFT-based and RL-based methods in visual reasoning tasks, and highlight the limitations of SFT-based approaches and the superior generalization of RL-based methods in reasoning improvement, out-of-domain performance, and data efficiency.
- We reconstructed a comprehensive dataset that covers three core domains: visual counting, structure perception, and spatial transformation, which serves as a benchmark to systematically assess visual cognition, geometric understanding, and cross-task generalization.
- Extensive experiments validate the effectiveness of **Reason-RFT**, offering valuable insights to advance visual reasoning and introducing a novel paradigm that significantly pushes forward multi-modal research.

## 2. Related Work

**Visual Reasoning** Visual reasoning is a key challenge in advancing artificial general intelligence (AGI), requiring models to perform complex cognitive tasks based on visual perception [18, 38, 42]. It has broad applications, including visual counting [37, 38], geometric problem-solving [12, 30, 42, 51, 70], visual transformation reasoning [23], scientific research [31, 41], and robotic task planning [17, 24, 28]. Traditional methods rely on program generation [16, 29, 53] or neural-symbolic approaches [4, 7, 13, 69], while recent vision-language models (VLMs) leverage large language models (LLMs) for enhanced reasoning. For example, LLaVA-CoT [63] uses multi-stage supervised fine-tuning (SFT) with chain-of-thought (CoT) [61], and Insight-V [10] combines SFT with reinforcement learning (RL). DeepSeek-R1-Zero [14] introduces a rule-based RL method, significantly improving reasoning. Building on DeepSeek-R1 [14], our work compares SFT-based and RL-based paradigms, highlighting the superior performance of R1-based methods in visual reasoning.

**Post-Training** Post-Training is a crucial phase for enhancing the performance of LLMs and VLMs, bridging pre-trained models and their real-world applications [8, 32, 54, 68]. It primarily involves two methodologies: *SFT* [58, 62] and *RL* [43, 46, 52, 65, 71]. SFT adapts pre-trained models to specific tasks using task-oriented datasets, often formatted as instructions. Research like FLAN [60] highlights the importance of diverse instruction-tuning datasets for improving zero-shot performance, while iterative processes, such as Llama 3.1’s six-round strategy [11], integrate rejection sampling, synthetic data, and human annotations. RL aligns models with human preferences or task-specific goals through feedback mechanisms. Reinforcement Learning from Human Feedback (RLHF) [46] refines models using human preference data, as seen in Llama 3.1 [11] and Nemotron-4 [2], which use reward modeling techniques like DPO [49] and RPO [2]. For example, TÜLU 3 [34] employs length-normalized DPO, while DeepSeek-V3 [39] combines rule-based and model-based reward systems. Recently, DeepSeek-R1 [14] achieved significant text reasoning improvements through pure RL [50]. Our work adapts R1 methodologies to VLMs, enhancing visual reasoning, and systematically compares SFT-based and R1-based paradigms, demonstrating the superiority of R1-based methods in visual reasoning tasks.

## 3. Methodology

### 3.1. Problem Definition

Visual reasoning can be formally defined as follows: given a visual input  $I$  (e.g., images or videos) and a corresponding textual description or question  $T$ , the goal is to derive a conclusion or answer  $A$  by analyzing the information in the

visual input. This process can be represented as a mapping:

$$R : (I, T) \rightarrow A,$$

where  $I \in \mathbb{R}^{H \times W \times C}$  is the visual input with  $H$ ,  $W$ , and  $C$  representing height, width, and channels,  $T$  is the textual description or question typically in natural language, and  $A$  is the derived conclusion or answer, often in natural language or structured data. Through this mapping, visual reasoning models integrate visual and textual information to achieve effective reasoning.

### 3.2. Reason-RFT

In this section, we propose **Reason-RFT**, a novel two-phase hybrid training strategy to enhance the reasoning capabilities of VLMs in complex visual reasoning tasks. As shown in Fig. 2, the framework comprises two stages: (1) *SFT-based Visual Reasoning Activation*, which uses SFT with high-quality CoT reasoning data to activate the model’s domain-specific reasoning capabilities, and (2) *RL-based Visual Reasoning Reinforcement*, which employs the GRPO algorithm with rule-based rewards to further push the upper limits of the model’s reasoning potential.

#### 3.2.1. STAGE 1: SFT-based Reasoning Activation

In the initial phase, we employ Supervised Fine-Tuning (SFT) on a structured visual reasoning dataset containing step-by-step reasoning processes. This phase trains the model to decompose complex tasks into logical steps. Each sample is represented as  $(x, q, r, a)$ , where  $x$  denotes the input images,  $q$  is the question,  $r$  is the reasoning steps, and  $a$  is the final answer. The training objective maximizes the likelihood of generating both  $r$  and  $a$  given  $(x, q)$ :

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x, q, r, a) \sim \mathcal{D}} \sum_{t=1}^T \log \pi_\theta(y_t | x, q, y_{<t}), \quad (1)$$

where  $\mathcal{D}$  is the dataset,  $y$  the concatenated sequence of  $r$  and  $a$ , and  $\pi_\theta$  the model’s token distribution. The output model  $\pi_{\text{CoT}}$  serves as the initialization for the next stage, ensuring a robust foundation for reinforcement learning.

#### 3.2.2. STAGE 2: RL-based Reasoning Enhancement

In the second phase, we refine  $\pi_{\text{CoT}}$  using Group Relative Policy Optimization (GRPO), leveraging reinforcement learning for its efficiency and scalability. Unlike Proximal Policy Optimization (PPO), which requires a computationally expensive value network, GRPO calculates relative advantages by comparing rewards within a group of sampled actions, reducing computational overhead and simplifying optimization. This makes GRPO particularly suitable for visual reasoning tasks.

**Sampling Action Groups** For each input state  $s = (x, q)$ , where  $x$  is the visual encoding of the input image

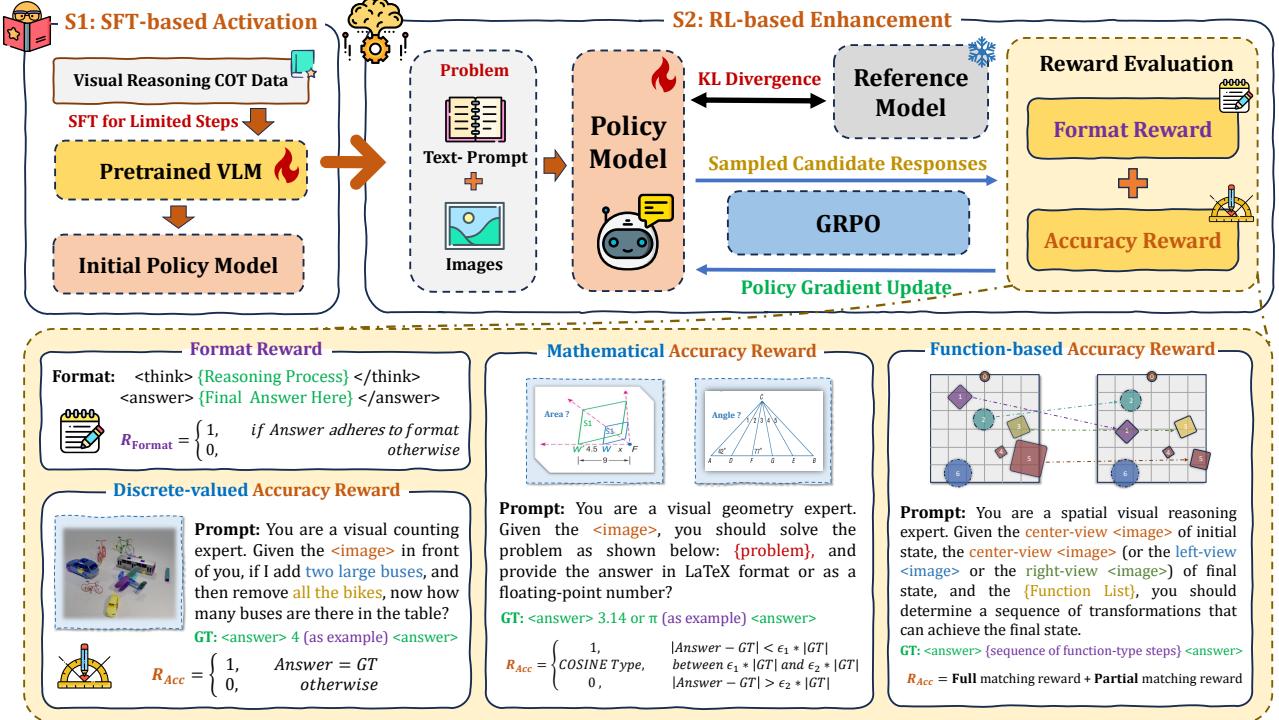


Figure 2. **Framework of Reason-RFT.** Reason-RFT introduces a two-phase training framework for visual reasoning. First, Supervised Fine-Tuning (SFT) with Chain-of-Thought (CoT) reasoning activates the model’s reasoning potential using a high-quality domain-specific visual reasoning dataset in stage 1. Subsequently, in stage 2, Group Relative Policy Optimization (GRPO) enhances reasoning capabilities further, enabling Reason-RFT to achieve superior generalization by pushing the model’s reasoning limits. Specifically, reward evaluation consists of format reward and three different types of accuracy reward.

and  $q$  the textual encoding of the question, GRPO samples a group of actions  $\{a_1, a_2, \dots, a_G\}$  from the current policy  $\pi_\theta$ , initialized from  $\pi_{\text{CoT}}$ . The sampling process is:

$$a_i \sim \pi_\theta(a | x, q), \quad \text{for } i = 1, 2, \dots, G. \quad (2)$$

This strategy ensures diverse responses, promoting exploration and preventing premature convergence.

**Reward Evaluation** Each sampled action  $a_i$  is assigned a reward  $R(a_i)$  based on verifiable criteria, resulting in a reward set  $\{r_1, r_2, \dots, r_G\}$ . For visual reasoning tasks, the reward function  $R(a_i)$  combines two components: Format Reward  $R_{\text{format}}(a_i)$  and Accuracy Reward  $R_{\text{acc}}(a_i)$ . The Format Reward ensures structured response formats, while the Accuracy Reward evaluates correctness, balancing structured reasoning and factual accuracy. The reward function is defined as:

$$R(a_i) = R_{\text{format}}(a_i) + R_{\text{acc}}(a_i). \quad (3)$$

**Policy Update with Relative Advantage** Rewards are normalized within the sampled group to compute relative advantages  $\{A_1, A_2, \dots, A_G\}$ , defined as:

$$A_i = \frac{r_i - \text{mean}\{r_1, r_2, \dots, r_G\}}{\text{std}\{r_1, r_2, \dots, r_G\}}. \quad (4)$$

Based on these advantages, the policy is updated to reinforce actions with positive advantages and reduce the probability of less effective ones. Policy updates are further constrained by minimizing the KL divergence between the updated and reference models, ensuring stable RL learning.

### 3.3. Reward Design for Visual Reasoning Tasks

For the diverse requirements of visual reasoning tasks, including Visual Counting, Structure Perception, and Spatial Transformation, our reward design integrates two essential components: *Format Reward* and *Accuracy Reward*. The Format Reward is uniformly applied across all tasks, ensuring that the model strictly adheres to a structured response format, which significantly enhances interpretability and consistency. For the Accuracy Reward, we carefully tailor the design to the specific characteristics of each task, as shown in Fig. 2, creating task-specific reward mechanisms to evaluate the correctness of the model’s responses.

**Format Reward** This component ensures structured and interpretable responses by requiring the model to adhere to a predefined template: reasoning within `<think>` and `</think>` and the final answer within `<answer>` and `</answer>`. A reward of 1 is given for strict adherence, while deviations result in a reward of 0.

**Accuracy Reward** This component evaluates the correctness of the model’s responses, ensuring alignment with ground truth across diverse visual reasoning tasks. To address task diversity, we design tailored reward mechanisms for discrete-valued, mathematical, and function-based problems. Each mechanism is crafted to handle the unique characteristics of its problem category, enabling precise and fair evaluation. Below, we introduce the three reward forms.

- **Discrete-valued Type** This reward type applies to Visual Counting and partial Structure Perception tasks, where answers are discrete values (*e.g.*, multiple-choice or integer-based responses). The accuracy reward  $R_{\text{acc}}(a_i)$  is defined as:

$$R_{\text{acc}}(a_i) = \begin{cases} 1, & \text{if } a_{\text{pred}} = a_{\text{gt}} \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

where  $a_{\text{pred}}$  is the model’s predicted answer and  $a_{\text{gt}}$  is the ground truth. This binary reward penalizes deviations from the exact ground truth, ensuring precision in tasks requiring unambiguous answers.

- **Mathematical Type** This reward type is designed for Structure Perception tasks involving numerical answers, such as floating-point values or LaTeX-formatted expressions. It uses a tolerance-based evaluation to account for minor numerical deviations. The accuracy reward  $R_{\text{acc}}(a_i)$  is defined as:

$$R_{\text{acc}}(a_i) = \frac{1}{2} \left[ \cos \left( \pi \times \frac{|a_{\text{pred}} - a_{\text{gt}}| - \epsilon_1 \times |a_{\text{gt}}|}{(\epsilon_2 - \epsilon_1) \times |a_{\text{gt}}|} \right) + 1 \right], \quad (6)$$

where  $a_{\text{pred}}$  is the predicted answer,  $a_{\text{gt}}$  is the ground truth,  $\epsilon_1$  is the tolerance threshold for an exact match (*e.g.*, 0.05), and  $\epsilon_2$  is the upper bound for partial rewards (*e.g.*, 0.20). If  $|a_{\text{pred}} - a_{\text{gt}}| < \epsilon_1 \times |a_{\text{gt}}|$ , the reward is 1 (exact match); if  $|a_{\text{pred}} - a_{\text{gt}}| > \epsilon_2 \times |a_{\text{gt}}|$ , the reward is 0 (incorrect). This formulation ensures smooth transitions between full and partial rewards, enabling fair evaluation of numerical accuracy.

- **Function-based Type** This reward type is designed for Spatial Transformation tasks requiring a sequence of transformation functions. The accuracy reward  $R_{\text{acc}}(a_i)$  evaluates the alignment between the predicted sequence  $T_{\text{pred}}$  and the ground truth  $T_{\text{gt}}$ , computed as:

$$R_{\text{acc}}(a_i) = \frac{\text{len}(T_{\text{pred}}^{f+o+v}) + \alpha \cdot \text{len}(T_{\text{pred}}^{f+o/v}) + \beta \cdot \text{len}(T_{\text{pred}}^f)}{\max(\text{len}(T_{\text{pred}}), \text{len}(T_{\text{gt}}))}, \quad (7)$$

where  $T_{\text{pred}}^{f+o+v}$  is the subset of transformation steps with complete matches (*w/* function, object, and value),  $T_{\text{pred}}^{f+o/v}$  are the subsets with partial and only-function matches (*w/* function and object, or *w/* function and value),  $T_{\text{pred}}^f$  is the subset with only-function matches.  $\alpha$  and  $\beta$  are the weighting coefficients for partial matches. This formulation ensures nuanced evaluation, rewarding both exact and partially correct responses while allowing flexible adjustment of partial match contributions.

## 4. Experiment

### 4.1. Experimental Details

**Datasets** In this paper, we comprehensively evaluate the visual reasoning capabilities of our method by leveraging six existing datasets, enhanced through subtask categorization, error-prone data filtering, and dataset restructuring. Detailed protocols for data filtering and restructuring are provided in the supplementary materials. Specifically, we define three task categories as follows.

- **Visual Counting** is a multimodal reasoning task evaluating the integration of linguistic, visual, and mathematical skills by solving arithmetic problems in 3D block-based scenes, which includes four subtask types, with datasets from CLEVR-Math [38] and Super-CLEVR [37]. Specifically, CLEVR-Math provides 35k training and 1k test samples, while Super-CLEVR serves as an out-of-domain test set with 1k samples to assess generalization.
- **Structure Perception** is a visual reasoning task requiring models to perceive and interpret structural information in mathematical geometries, medical imaging, chart layouts, and architectural designs. It includes both choice and non-choice formats, with datasets sourced from Geo170K [12], Math360K [51] (collectively Geo-Math), and Geometry3K [40]. GeoMath provides 4.5k training and 820 test samples, while Geometry3K serves as an out-of-domain test set with 800 samples to evaluate cross-domain adaptability.
- **Spatial Transformation** is a spatial-visual reasoning task requiring models to infer single-step or multi-step transformation actions by analyzing initial and final visual states for 3D scenes from multiple perspectives (*e.g.*, center, left, right). The dataset TRANCE [23] features four distinct difficulty levels, with 60k training and 6k test samples. Out-of-domain test samples are generated by rendering identical data from left and right viewpoints to rigorously assess robustness to perspective variations.

**Evaluation Metrics** We use accuracy-rate (Acc) as the primary metric [67]. For numerical answers, correctness is verified by mathematical equivalence to the ground truth. For multiple-choice questions, we perform a string match. For function-type sequences, we use stepwise multi-level evaluation to assess alignment with the correct solution.

**Implementation Details** We utilize Qwen2-VL-2B and Qwen2-VL-7B [59] as the backbone models for our experiments. Our implementation is built on the open-source frameworks Open-R1 [25] and vLLM [33], ensuring reproducibility and scalability. All experiments were conducted on a cluster of servers, each equipped with 8×A800 GPUs. For further details, see the supplementary materials.

**Training Paradigms and Baselines** To assess the performance and generalization of different training strategies, we compare: (1) SFT-based methods—ANS-SFT, which

Table 1. **Results on three visual reasoning tasks.** The best results among different training paradigms are highlighted in **bold**, while the second-best results are underlined. “ID” denotes in-domain test data, and “OOD” denotes out-of-domain test data.

Method	Visual Counting			Structure Perception			Spatial Transformation			
	Clevr-Math ID	Super-Clevr OOD	Avg	GeoMath ID	Geometry3k OOD	Avg	TRANCE ID	TRANCE-L OOD	TRANCE-R	Avg
<b>Proprietary Models</b>										
GPT-4o-2024-08-06 [26]	68.10	34.31	51.20	50.18	43.49	46.83	42.55	28.67	29.76	35.88
Gemini-1.5-Pro [55]	61.80	37.50	49.65	50.12	48.38	49.45	26.22	18.76	19.88	22.77
<b>Open-Source Models</b>										
Qwen2.5-VL-3B-Instruct [5]	75.90	39.30	57.60	36.75	37.44	37.09	8.57	8.26	8.31	8.42
Phi-3.5-Vision-4B-Instruct [1]	21.40	15.20	18.30	36.83	50.25	43.54	7.42	2.45	4.02	5.33
Llava-OneVision-7B [35]	69.70	29.10	49.40	77.63	43.66	60.64	10.00	8.33	8.74	9.27
Qwen2.5-VL-7B-Instruct [5]	74.60	35.20	54.90	44.00	45.61	44.80	19.63	13.12	13.42	16.45
InternVL-2.5-8B [6]	93.50	35.30	64.40	63.00	47.32	51.60	7.19	6.62	6.63	6.91
Llama-3.2-11B-Vision [44]	10.30	9.50	9.90	13.75	20.85	17.30	8.22	8.40	9.03	8.47
Pixtral-12B [3]	42.60	22.90	32.75	30.38	36.09	33.23	7.35	5.03	5.22	6.42
<b>Qwen2VL-2B-Instruct</b>										
Zero-Shot	82.40	32.00	57.20	25.86	20.63	23.25	3.78	4.60	4.67	4.35
+ ANS-SFT	96.20	39.20	67.70	<b>51.34</b>	22.50	36.92	<b>77.39</b>	<u>49.24</u>	<u>50.33</u>	<u>58.99</u>
+ CoT-SFT	85.50	<u>46.50</u>	66.00	43.05	25.25	34.15	64.37	43.19	42.86	50.14
+ Reason-RFT-Zero	<b>98.40</b>	44.80	<u>71.60</u>	47.68	<u>32.50</u>	<u>40.09</u>	42.13	34.07	33.41	33.74
+ Reason-RFT	<u>96.80</u>	<b>51.20</b>	<b>74.00</b>	<u>49.03</u>	<u>33.13</u>	<b>41.08</b>	<u>74.61</u>	<b>64.05</b>	<b>64.08</b>	<b>67.58</b>
<b>Qwen2VL-7B-Instruct</b>										
Zero-Shot	<u>98.60</u>	42.10	70.35	43.30	43.88	43.59	13.53	12.72	12.78	13.01
+ ANS-SFT	<u>95.00</u>	33.90	64.45	51.34	25.38	38.36	<b>82.19</b>	<u>54.29</u>	<u>54.83</u>	<u>63.77</u>
+ CoT-SFT	87.30	42.40	64.85	50.49	33.00	41.75	<u>81.31</u>	47.90	47.80	59.00
+ Reason-RFT-Zero	<b>99.40</b>	<u>53.00</u>	<u>76.20</u>	<u>55.00</u>	<b>54.75</b>	<b>54.88</b>	67.67	57.20	56.15	56.68
+ Reason-RFT	95.60	<u>51.00</u>	<u>73.30</u>	<b>59.27</b>	<u>49.25</u>	<u>54.26</u>	79.97	<b>59.36</b>	<b>58.61</b>	<b>65.98</b>

Table 2. **More detailed results on Super-Clevr.** “Encountered” refers to the types of questions the model has previously seen during Clevr-Math training, while “UnEncountered” denotes the complicated types that the model has not encountered (*i.e.* questions with mixture of addition and subtraction).

Methods	Encountered	UnEncountered	Total
<b>Qwen2VL-2B-Instruct</b>			
Zero-Shot	42.67	0.00	32.00
+ ANS-SFT	51.07	5.20	39.20
+ CoT-SFT	49.73	<b>36.80</b>	46.50
+ Reason-RFT-Zero	<u>58.00</u>	5.20	44.80
+ Reason-RFT	<b>60.00</b>	28.40	<b>51.20</b>
<b>Qwen2VL-7B-Instruct</b>			
Zero-Shot	54.53	4.80	42.10
+ ANS-SFT	42.53	8.00	33.90
+ CoT-SFT	45.33	<u>33.60</u>	42.40
+ Reason-RFT-Zero	<b>63.60</b>	21.20	<b>53.00</b>
+ Reason-RFT	<u>56.13</u>	<b>35.60</b>	<u>51.00</u>

fine-tunes on answer generation, and CoT-SFT, which uses supervised learning with chain-of-thought (CoT) reasoning; and (2) RL-based methods—Reason-RFT-Zero, which applies reinforcement learning without reasoning activation phase, and Reason-RFT, which uses partial CoT data for reasoning activation before RL training. For comprehensive experiments, we use Qwen2-VL-Instruct [59] as the base model, evaluating both 2B and 7B variants to explore the

impact of model scale. In addition, we also select the most advanced open-source models [1, 3, 5, 6, 35, 44] and the proprietary models [26, 55] available as baselines to evaluate the performance of different training paradigms.

## 4.2. Results on In-Domain Tasks

To evaluate the In-Domain (ID) performance of Reason-RFT against different training methods and baseline models across visual reasoning tasks, we conducted extensive training and validation on 2B/7B models for three tasks. The results in Tab. 1 show: (1) **Visual Counting** RL-based methods outperform all open-source and proprietary baseline models as well as SFT-based methods in both 2B and 7B models, with Reason-RFT-Zero in 7B model achieving the best performance; (2) **Structure Perception** RL-based methods surpass SFT-based approaches in 7B model, while ANS-SFT performs best in 2B model. CoT-SFT shows limited improvement, as enforced reasoning supervision inhibits cognitive enhancement. Additionally, Reason-RFT in 7B model outperforms all proprietary and most open-source models except InternVL-2.5-8B [6] and Llava-OneVision-7B [35]; (3) **Spatial Transformation** SFT-based methods achieve the highest performance, with Reason-RFT achieving comparable performance while totally surpassing all baseline models. However, Reason-RFT-Zero performs poorly, likely due to its difficulty adapting to function-based tasks, whereas Reason-RFT benefits from reason-

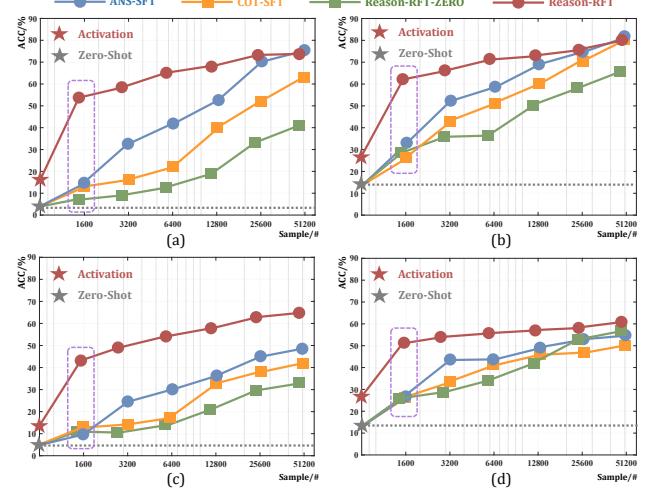
ing activation, reaching a higher performance ceiling. In summary, Reason-RFT achieves performance comparable to SFT-based methods in ID tasks while outperforming all baselines in Visual Counting and Spatial Transformation, and most of baselines in Structure Perception.

### 4.3. Results on Out-of-Domain Generalization

To validate the Out-of-Domain (OOD) performance of Reason-RFT across different training methods and baseline models in visual reasoning tasks, we conducted comprehensive experiments on 2B/7B models for three tasks. Results in Tab. 1 and Tab. 2 show that: **(1) Visual Counting** RL-based methods demonstrate superior generalization over SFT-based methods in both 2B and 7B models, with Reason-RFT surpassing ANS-SFT by 12% (2B) and 17% (7B) while outperforming all open-source and proprietary baselines. Notably, compared with purely RL-based method (*i.e.*, Reason-RFT-Zero), CoT-SFT and Reason-RFT can activate model capabilities on unseen complicated questions, achieving strong results despite its poor zero-shot performance before in-domain training, as shown in Tab. 2. This also highlights the limitations of Reason-RFT-Zero due to cognitive constraints, especially in smaller models (*i.e.* 2B models); **(2) Structure Perception** RL-based methods consistently outperform SFT-based methods, with Reason-RFT achieving the best result in 2B model (8% higher than CoT-SFT) while Reason-RFT-Zero performs strongest in 7B model and Reason-RFT remains competitive (16% higher than CoT-SFT), with SFT-based methods showing limited impact especially in the 7B model; **(3) Spatial Transformation** RL-based methods surpass SFT-based approaches in both 2B and 7B models while significantly outperforming all baseline models, with Reason-RFT (2B) demonstrating remarkable OOD generalization exceeding GPT-4o [26] by 35%. In summary, Reason-RFT surpasses all open-source and proprietary baselines as well as alternative training methods, demonstrating remarkable performance in visual reasoning generalization capabilities.

### 4.4. Assessment on Training Efficiency

To demonstrate the data efficiency of **Reason-RFT** during training, we trained all methods on TRANCE and recorded intermediate and validation results, as shown in Fig. 3. Detailed results are provided in the Appendix. Key findings include: (1) *Reason-RFT exhibits strong data efficiency*. In the ID tasks of the 2B model, Reason-RFT achieves 70% of the performance of Reason-RFT-Zero using only 3% of the training data (1,600 samples), reaching 82.5% with 9% of the data. (2) *This efficiency generalizes to OOD tasks*. In the 7B model, Reason-RFT achieves over 92% of the performance of Reason-RFT-Zero using just 3% of the training data, demonstrating robust generalization capabilities.



**Figure 3. Results of different methods on the Spatial Transformation task across training processes.** (a) Evaluation results for 2B model on ID task, (b) Evaluation results for 7B model on ID task, (c) Evaluation results for 2B model on OOD task, (d) Evaluation results for 7B model on OOD task.

### 4.5. Training Insights

**Transient Adaptation Gap** This phenomenon refers to the temporary performance drop observed during the initial phase of training process with Reason-RFT-Zero. As the model transitions from directly outputting answers to generating structured reasoning processes, it experiences a brief adaptation difficulty, leading to a sharp performance decline followed by gradual recovery. As shown in Fig. 4 (a), which illustrates the training process of Reason-RFT-Zero on the CLEVR-MATH, both ID and OOD test performances exhibit this sharp drop and recovery within the first 100 steps. We further investigate this phenomenon through a case study, shown in Fig. 4 (b), which demonstrates that forcing the model to output its reasoning process can lead to incorrect answers to originally correct question.

**Greedy Reward Stratification** This phenomenon describes the model’s tendency during Reason-RFT-Zero training to prioritize easier rewards (*e.g.*, Format Reward) over harder ones (*e.g.*, Accuracy Reward). As shown in Fig. 5, the model’s Reasoning Token Length initially decreases, then gradually increases before stabilizing. This behavior coincides with the Format Reward reaching its initial peak and the Accuracy Reward entering its rapid growth phase. We infer that the model simplifies its outputs early on to quickly adapt to the structured response format, reducing Reasoning Token Length. Once the Format Reward is maximized, the model shifts focus to improving accuracy, increasing Reasoning Token Length.

**Reasoning Redundancy** This phenomenon refers to the significant difference in Reasoning Token Length be-

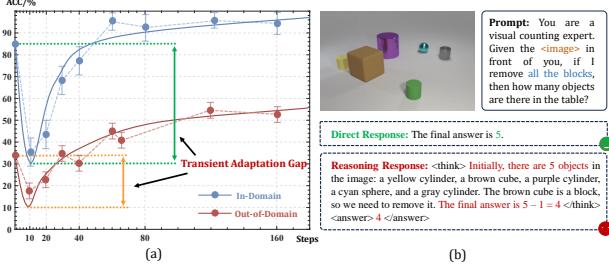


Figure 4. **Illustration of the Transient Adaptation Gap.** (a) shows a sharp drop and recovery in both ID and OOD test performances within the early training steps during training on the CLEVR-MATH task. (b) presents a case study of the prediction result on the 10-th step for further investigation.

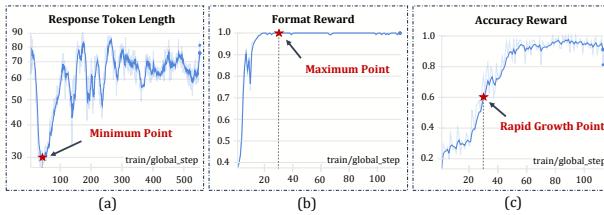


Figure 5. **Analysis of Greedy Reward Stratification.** The model’s Reasoning Token Length initially decreases, then gradually increases before stabilizing during Reason-RFT-Zero training. (a) illustrates the curve of Reasoning Token Length, highlighting the Minimum Point. (b) shows the curve of Format Reward, indicating the Maximum Point. (c) displays the curve of Accuracy Reward, marking the Rapid Growth Point. This trend aligns with the Format Reward reaching its initial peak and the Accuracy Reward entering its rapid growth phase.

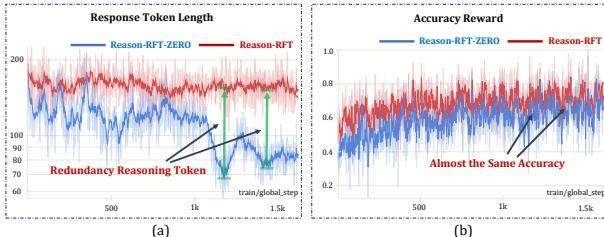


Figure 6. **Analysis of Reasoning Redundancy.** (a) illustrates the curves of Reasoning Token Length for Reason-RFT-Zero and Reason-RFT during training on the Structure Perception task, highlighting the significant difference in Reasoning Token Length between these two paradigms in the later steps of training. (b) shows the curves of Accuracy Reward, indicating that the accuracy rates of both paradigms converge and remain nearly identical.

tween models trained under different paradigms. For example, in the Structure Perception task, Reason-RFT and Reason-RFT-Zero achieve comparable final test accuracy, but Reason-RFT exhibits notably longer Reasoning Token Length, as shown in Fig. 6. This occurs because the CoT data used in Reason-RFT’s Reasoning Activation phase, often distilled from stronger models (*e.g.*, GPT-4o), leads the model to learn longer reasoning chains during CoT-SFT.

Table 3. Results of different format reward strategies on the Spatial Transformation task.

Setting	ID	OOD		AVG
	TRANCE	TRANCE-L	TRANCE-R	
<b>Qwen2VL-2B-Instruct</b>				
Reason-RFT-Zero	<b>42.13</b>	34.07	33.41	33.74
+ visual tokens	42.01	<b>36.05</b>	<b>35.97</b>	<b>38.01</b>
Reason-RFT	<b>74.61</b>	<b>64.05</b>	<b>64.08</b>	<b>69.33</b>
+ visual tokens	71.99	60.13	59.87	65.99
<b>Qwen2VL-7B-Instruct</b>				
Reason-RFT-Zero	67.67	57.2	56.15	62.17
+ visual tokens	<b>70.28</b>	<b>59.52</b>	<b>57.01</b>	<b>64.27</b>
Reason-RFT	<b>79.97</b>	<b>59.36</b>	<b>58.61</b>	<b>69.48</b>
+ visual tokens	79.85	58.71	57.98	69.09

Table 4. Results of different accuracy reward strategies on the Spatial Transformation task.

Setting	$\alpha$	$\beta$	ID	OOD		AVG
			TRANCE	TRANCE-L	TRANCE-R	
<b>Qwen2VL-2B-Instruct</b>						
Baseline	0	0	74.61	64.05	64.08	<b>69.33</b>
(a)	0.50	0.25	<b>79.18</b>	56.36	55.45	67.54
(b)	-0.25	-0.50	73.69	<b>64.41</b>	<b>64.72</b>	69.13
<b>Qwen2VL-7B-Instruct</b>						
Baseline	0	0	79.97	59.36	58.61	69.48
(a)	0.50	0.25	<b>80.89</b>	53.20	52.61	66.90
(b)	-0.25	-0.50	75.03	<b>64.83</b>	<b>63.18</b>	<b>69.52</b>

Table 5. Results of various mixed CoT activation datasets of Reason-RFT on the Structure Perception task.

Setting	CoT Activation Data	ID	OOD		AVG
		GeoMath	Geometry3k		
Baseline	GeoMath-only data	59.27	49.25	54.26	
(a)	Mixed Specific-Domain data	50.61	45.35	48.02	
(b)	Mixed General-Domain data	42.51	40.25	41.38	

During the Reasoning Reinforcement phase, the model retains these lengthy chains due to the lack of penalties or incentives for response length. In contrast, Reason-RFT-Zero, which lacks CoT data, stabilizes at a shorter Reasoning Token Length through exploration. We hypothesize that Reason-RFT’s longer reasoning chains may be unnecessary for the current task difficulty, introducing redundant computational overhead. Experiments limiting Reasoning Token Length during inference for Reason-RFT-trained models show stable performance until a certain threshold, partially confirming Reasoning Redundancy in specific tasks.

#### 4.6. Exploration on Reward Design

**Exploration on Format Reward** In DeepSeek-R1 [15], the Format Reward mandates the use of `<think>` tags for structuring reasoning processes in pure-textual tasks. To enhance visual reasoning tasks, we propose integrating visual observations through captioning, extending the format reward with `<summary>` and `<caption>` tokens along-

side <think>. As shown in Tab. 3, this extension significantly improved the performance of Reason-RFT-Zero, but not Reason-RFT. We attribute this discrepancy to Reason-RFT’s prior exposure to Reasoning Activation, which likely learned the caption-like guidance from a limited CoT reasoning samples, resulting in reduced effectiveness of additional tag guidance. Conversely, Reason-RFT-Zero, being purely RL-based, benefits more from effective tag guidance, highlighting its greater potential for improvement.

**Exploration on Accuracy Reward** We investigate the Accuracy Reward for the challenging Spatial Transformation task, which requires predicting novel transformation sequences. The reward design in Eq. 7 includes coefficients  $\alpha$  and  $\beta$  for partial matches, influencing tolerance for incomplete outputs. We test three configurations: (1)  $\alpha = 0, \beta = 0$  (full matches only), (2)  $\alpha = 0.50, \beta = 0.25$  (full and partial matches), and (3)  $\alpha = -0.25, \beta = -0.50$  (full matches with partial penalties). Experimental results on 2B and 7B models (Tab. 4) revealed: (1) Rewards for partial matches improved in-domain performance but harmed out-of-domain generalization, suggesting “soft rewards” may limit adaptability; (2) Penalties for partial matches enhanced performance on out-of-domain, indicating “hard rewards” are more effective for serialized tasks.

#### 4.7. Exploration on COT Activation Data

To investigate the impact of differently composed CoT activation data on Reason-RFT, we construct two distinct datasets: a mixed domain-specific dataset, which integrates relevant yet distinct data from in-domain tasks, and a mixed general-domain dataset, encompassing a broader range of visual reasoning tasks (*e.g.*, graph topology, visual puzzles). Using these datasets, we perform Reason-RFT training on Structure Perception task, with the results detailed in Tab. 5. From this, two key points emerge: (1) As the proportion of in-domain training data decreases, the model’s performance on specific tasks declines; (2) Models trained on more diverse visual reasoning domain data may also exhibit a reduction in domain-specific performance.

### 5. Conclusion

In this paper, we propose **Reason-RFT**, a novel reinforcement fine-tuning framework that enhances the generalization capabilities of visual reasoning models. By integrating supervised fine-tuning (SFT) with Chain-of-Thought (CoT) reasoning activation data and Group Relative Policy Optimization (GRPO)-based reinforcement learning, **Reason-RFT** effectively mitigates key challenges such as overfitting and cognitive rigidity, thereby improving cross-domain transferability and real-world applicability. To support systematic evaluation, we reconstruct a comprehensive dataset covering visual counting, structure perception, and spatial transformation tasks, establishing a robust benchmark

for assessing model performance across diverse scenarios. Extensive experiments demonstrate the effectiveness of **Reason-RFT**, providing valuable insights for advancing visual reasoning research and introducing a new paradigm in multimodal learning.

### References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 6
- [2] Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, et al. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 2024. 3
- [3] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 6
- [4] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling. In *International Conference on Machine Learning*, pages 279–290. Pmlr, 2020. 2, 3
- [5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6
- [6] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 6
- [7] Minkyu Choi, Harsh Goel, Mohammad Omama, Yunhao Yang, Sahil Shah, and Sandeep Chinchali. Towards neuro-symbolic video understanding. In *European Conference on Computer Vision*, pages 220–236. Springer, 2024. 2, 3
- [8] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025. 3
- [9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2
- [10] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Ex-

- ploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024. 3
- [11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3
- [12] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wan-jun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-l lava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 3, 5, 13
- [13] Artur d’Avila Garcez, Marco Gori, Luis C Lamb, Luciano Serafini, Michael Spranger, and Son N Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088*, 2019. 2, 3
- [14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3
- [15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 8, 13
- [16] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 2, 3
- [17] Peng Hao, Chaofan Zhang, Dingzhe Li, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Tla: Tactile-language-action model for contact-rich manipulation. *arXiv preprint arXiv:2503.08548*, 2025. 3
- [18] Xiaoshuai Hao, Yi Zhu, Srikanth Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multi-modal data augmentation. In *IEEE/CVF winter conference on applications of computer vision*, pages 379–389, 2023. 3
- [19] Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang. Mapdistill: Boosting efficient camera-based hd map construction via camera-lidar fusion model distillation. In *European Conference on Computer Vision*, pages 166–183, 2024. 2
- [20] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, Haimei Zhao, Hui Zhang, Yi Zhou, Qiang Wang, Weiming Li, Lingdong Kong, and Jing Zhang. Is your HD map constructor reliable under sensor corruptions? In *Advances in Neural Information Processing Systems*, 2024.
- [21] Xiaoshuai Hao, Hui Zhang, Yifan Yang, Yi Zhou, Sangil Jung, Seung-In Park, and ByungIn Yoo. Mbfusion: A new multi-modal bev feature fusion method for hd map construction. In *IEEE International Conference on Robotics and Automation*, pages 15922–15928, 2024.
- [22] Xiaoshuai Hao, Yunfeng Diao, Mengchuan Wei, Yifan Yang, Peng Hao, Rong Yin, Hui Zhang, Weiming Li, Shu Zhao, and Yu Liu. Mapfusion: A novel bev feature fusion network for multi-modal map construction. *Information Fusion*, page 103018, 2025. 2
- [23] Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. Transformation driven visual reasoning. In *IEEE/CVF Conference on computer vision and pattern recognition*, pages 6903–6912, 2021. 3, 5, 14
- [24] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023. 3
- [25] Huggingface. open-r1: Fully open reproduction of deepseek-r1. <https://github.com/huggingface/open-r1>, 2025. [Online; accessed: 2025-01-24]. 5, 14
- [26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 7
- [27] Yuheng Ji, Yue Liu, Zhicheng Zhang, Zhao Zhang, Yuting Zhao, Gang Zhou, Xingwei Zhang, Xinwang Liu, and Xiaolong Zheng. Advlora: Adversarial low-rank adaptation of vision-language models. *arXiv preprint arXiv:2404.13425*, 2024. 2
- [28] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:arXiv:2502.21257*, 2025. 2, 3
- [29] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *IEEE international conference on computer vision*, pages 2989–2998, 2017. 3
- [30] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023. 3
- [31] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251, 2016. 3
- [32] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Salman Khan, and Fahad Shahbaz Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025. 3
- [33] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 5, 15
- [34] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. 3

- [35] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6
- [36] Dingzhe Li, Yixiang Jin, Yuhao Sun, Hongze Yu, Jun Shi, Xiaoshuai Hao, Peng Hao, Huaping Liu, Fuchun Sun, Jianwei Zhang, et al. What foundation models can bring for robot learning in manipulation: A survey. *arXiv preprint arXiv:2404.18201*, 2024. 2
- [37] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Koptylewski, Wufei Ma, Benjamin Van Durme, and Alan Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973, 2023. 3, 5, 13
- [38] Adam Dahlgren Lindström and Savitha Sam Abramam. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 2, 3, 5, 13
- [39] 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. 3
- [40] 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. 5, 13
- [41] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, pages 2507–2521, 2022. 3
- [42] 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. 3
- [43] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023. 3
- [44] Meta AI. Llama 3 at connect 2024: Vision for edge and mobile devices, 2024. Accessed: 2025-02-15. 6
- [45] OpenAI. Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/>, 2024. Accessed: 2025-03-02. 2
- [46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, pages 27730–27744, 2022. 3
- [47] Jiazen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025. 2
- [48] Maria MP Petrou and Costas Petrou. *Image processing: the fundamentals*. John Wiley & Sons, 2010. 2
- [49] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, pages 53728–53741, 2023. 3
- [50] 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. 3
- [51] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Mathllava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024. 3, 5, 13
- [52] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 3
- [53] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergrpt: Visual inference via python execution for reasoning. In *IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023. 3
- [54] Yingbo Tang, Shuaike Zhang, Xiaoshuai Hao, Pengwei Wang, Jianlong Wu, Zhongyuan Wang, and Shanghang Zhang. Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. *arXiv preprint arXiv:2503.00778*, 2025. 3
- [55] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 6
- [56] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025. 2
- [57] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-01: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025. 2
- [58] Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *arXiv preprint arXiv:2310.03731*, 2023. 3
- [59] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui

- Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5, 6, 14
- [60] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 3
- [61] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, pages 24824–24837, 2022. 3
- [62] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, pages 24824–24837, 2022. 3
- [63] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 2, 3
- [64] 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. *arXiv preprint arXiv:2412.14171*, 2024. 2
- [65] Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in Neural Information Processing Systems*, pages 110935–110971, 2025. 3
- [66] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *ACM International Conference on Multimedia*, pages 2345–2354, 2020. 2
- [67] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024. 5
- [68] Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. *arXiv preprint arXiv:2502.13451*, 2025. 3
- [69] Mingyu Zhang, Jiting Cai, Mingyu Liu, Yue Xu, Cewu Lu, and Yong-Lu Li. Take a step back: Rethinking the two stages in visual reasoning. In *European Conference on Computer Vision*, pages 124–141. Springer, 2024. 2, 3
- [70] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aoju Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*, 2024. 3
- [71] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 3

## Appendix

This supplementary material provides additional details on the proposed method and experimental results that could not be included in the main manuscript due to page limitations. Specifically, this appendix is organized as follows.

- Sec. A provides more details on the evaluation of reasoning tasks and discusses how we collected, filtered, and reconstructed a high-quality dataset.
- Sec. B outlines the models and training processes, providing more detailed experimental specifics.
- Sec. C presents comprehensive experimental results.
- Sec. D includes more visualization cases.

## A. Details of Evaluation Reasoning Tasks

### A.1. Visual Counting

**Task Definition** Visual Counting is a multi-modal reasoning task that evaluates the integration of language, visual, and mathematical capabilities by solving basic arithmetic problems in visual scenes featuring 3D blocks with diverse attributes like color, size, material, and shape. The task includes four different types: **1) Subtraction** Counting after removing specific object subsets, **2) Addition** Counting after inserting objects with specified quantities and attributes, **3) Adversarial** Designed as trick questions where actions are performed on one object, but the query targets an unaffected object, and **4) Multi-Hop** Counting after sequential steps by addition or subtraction action. This task challenges models to perform attribute-based reasoning in dynamic visual contexts, emphasizing cross-modal understanding and reasoning capabilities. Some examples are shown in Fig. 7.

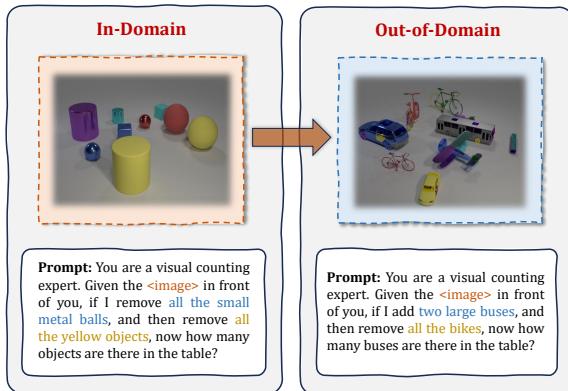


Figure 7. The sample of Visual Counting.

**Dataset Preparation** For in-domain dataset, we refined the original dataset from CLEVR-Math [38] by filtering out low-quality or incorrect samples using GPT-4o, resulting in a clean dataset comprising 35K training samples and 1K test samples. These samples are categorized into four specific types: subtraction, addition, adversarial, and

multihop-subtraction. To evaluate out-of-domain generalization, we extended CLEVR-Math by enhancing the diversity of objects through the incorporation of 3D assets from Super-CLEVR [37], which leads to the creation of Super-CLEVR-Math, an advanced benchmark with 1K test samples designed to assess model generalization under increased complexity. These test samples are also divided into four task types: addition, subtraction, addition-subtraction, and subtraction-multihop. Notably, the addition-subtraction type introduces a novel category consisting both addition and subtraction, which is not present in CLEVR-Math, further elevating the benchmark’s challenge and diversity.

**Reward Design** Following the reward methodology of DeepSeek-R1 [15], we define two distinct reward functions: Format Reward and Accuracy Reward. The Format Reward is assigned a value of 1 if the response adheres to the predefined template structure, specifically in the form of <think>...</think><answer>...</answer>; otherwise, it is assigned a value of 0. The Accuracy Reward is assigned a value of 1 if the numerical counting result in the response is correct; otherwise, it is assigned a value of 0. This dual-reward mechanism ensures both structural compliance and numerical accuracy in model responses.

### A.2. Structure Perception

**Task Definition** Structure Perception represents a complex class of visual mathematical reasoning tasks, which focuses on assessing the model’s capacity to determine geometric structure relationships and perform calculations involving angles, lengths, areas, and other geometric properties. The task includes problems such as identifying congruent or similar shapes, calculating perimeters and areas, determining angles between lines or shapes, and solving problems related to geometric transformations (e.g., rotations, translations, and reflections). By combining mathematical rigor with visual reasoning, this task challenges models to demonstrate a deep understanding of geometric principles in both abstract and real-world scenarios. Some examples are shown in Fig. 8.

**Dataset Preparation** For the in-domain dataset, we utilized GeoMath-8K, a dataset specifically designed for geometric problem-solving, which is constructed based on Math360K [51] and Geo170K [12]. To ensure data quality, we employed GPT-4o to filter out incorrect samples and removed those with answers that were neither numerical nor included in the provided options, thereby streamlining the validation process during training and testing. This refinement process resulted in a curated dataset consisting of 4.5K training samples and 820 test samples. For out-of-domain evaluation, we selected 800 samples from Geometry3K [40] (including 400 multiple-choice and 400 open-ended questions) to comprehensively assess the model’s generalization capabilities on geometry reasoning.

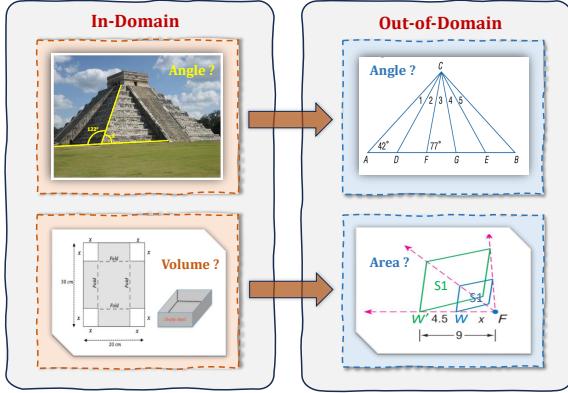


Figure 8. The sample of Structure Perception.

**Reward Design** We maintain the same Format Reward as used in the Visual Counting task above. The Accuracy Reward is extended to support the evaluation of both multiple-choice questions and mathematical expressions, ensuring comprehensive assessment across various problem types. Specifically, mathematical reward type is designed for Structure Perception tasks involving numerical answers, such as floating-point values or LaTeX-formatted expressions. It uses a tolerance-based evaluation to account for minor numerical deviations. The accuracy reward  $R_{\text{acc}}(a_i)$  is defined as:

$$R_{\text{acc}}(a_i) = \frac{1}{2} \left[ \cos \left( \pi \times \frac{|a_{\text{pred}} - a_{\text{gt}}| - \epsilon_1 \times |a_{\text{gt}}|}{(\epsilon_2 - \epsilon_1) \times |a_{\text{gt}}|} \right) + 1 \right], \quad (8)$$

where  $a_{\text{pred}}$  is the predicted answer,  $a_{\text{gt}}$  is the ground truth,  $\epsilon_1$  is the tolerance threshold for an exact match (e.g., 0.05), and  $\epsilon_2$  is the upper bound for partial rewards (e.g., 0.20). If  $|a_{\text{pred}} - a_{\text{gt}}| < \epsilon_1 \times |a_{\text{gt}}|$ , the reward is 1 (exact match); if  $|a_{\text{pred}} - a_{\text{gt}}| > \epsilon_2 \times |a_{\text{gt}}|$ , the reward is 0 (incorrect). This formulation ensures smooth transitions between full and partial rewards, enabling fair numerical evaluation.

### A.3. Spatial Transformation

**Task Definition** Spatial Transformation is a spatial-visual reasoning task designed to infer single-step or multi-step transformation actions by analyzing the initial and final visual states from multiple perspectives (e.g., center, left, right). The task utilizes transformation functions, including `change_size`, `change_color`, `change_material`, `change_shape`, and `change_position`, to modify object properties such as size, color, material, shape, and position using predefined values. This task evaluates the model’s ability to reason about spatial relationships and object transformations across diverse viewpoints in dynamic visual scenarios. Some examples are shown in Fig. 9.

**Dataset Preparation** We generated 100K samples using the environment and configuration from Trance [23], with each sample comprising initial object attributes, front-view image of initial state, and images of final state captured

from front, left, and right perspectives. To ensure high data quality, we implemented a rigorous filtering process: (1) removing samples containing occluded or invisible objects in either the initial or final states, (2) eliminating redundant actions within the transformation sequences, and (3) consolidating multi-step displacement actions, which collectively ensure the uniqueness and correctness of the solutions. The refined dataset consists of 60K training samples and 6K test samples. For the training set, we constructed the Trans-Center-60K dataset using the Center-Center configuration, which pairs front-view initial and final state images. For in-domain evaluation, we derived the Trans-Center-6K dataset from the 6K test samples under the same Center-Center configuration. To evaluate out-of-domain generalization, we constructed two additional datasets: Trans-Left-6K and Trans-Right-6K, leveraging the Center-Left and Center-Right configurations to systematically assess the model’s generalization capabilities in spatial reasoning under diverse viewpoint conditions.

**Reward Design** For the Format Reward, we adopted the same formulation as used in the Visual Counting task. As for the Accuracy Reward, a specialized design was developed to evaluate the sequence of transformation functions. Function-based type is designed for Spatial Transformation tasks requiring a sequence of transformation functions. The accuracy reward  $R_{\text{acc}}(a_i)$  evaluates the alignment between the predicted sequence  $T_{\text{pred}}$  and the ground truth  $T_{\text{gt}}$ , computed as:

$$R_{\text{acc}}(a_i) = \frac{\text{len}(T_{\text{pred}}^{f+o+v}) + \alpha \cdot \text{len}(T_{\text{pred}}^{f+o/v}) + \beta \cdot \text{len}(T_{\text{pred}}^f)}{\max(\text{len}(T_{\text{pred}}), \text{len}(T_{\text{gt}}))}, \quad (9)$$

where  $T_{\text{pred}}^{f+o+v}$  is the subset of transformation steps with complete matches (w/ function, object, and value),  $T_{\text{pred}}^{f+o/v}$  are the subsets with partial and only-function matches (w/ function and object, or w/ function and value),  $T_{\text{pred}}^f$  is the subset with only-function matches.  $\alpha$  and  $\beta$  are the weighting coefficients for partial matches. This formulation ensures nuanced evaluation, rewarding both exact and partially correct responses while allowing flexible adjustment of partial match contributions.

**System Prompts** For the Spatial Transformation task, we designed two versions of the system prompt. The first version specifies the answer output format using the `<think>` and `<answer>` tags, while the second version includes additional outputs `<summary>` and `<caption>` for experiments on exploration of format reward design in main paper. These two versions are illustrated in Fig. 10 and Fig. 11, respectively.

## B. Details of Models and Training

We utilize Qwen2-VL-2B and Qwen2-VL-7B [59] as the backbone models for our experiments. Our implementation is built on the open-source frameworks Open-R1 [25] and

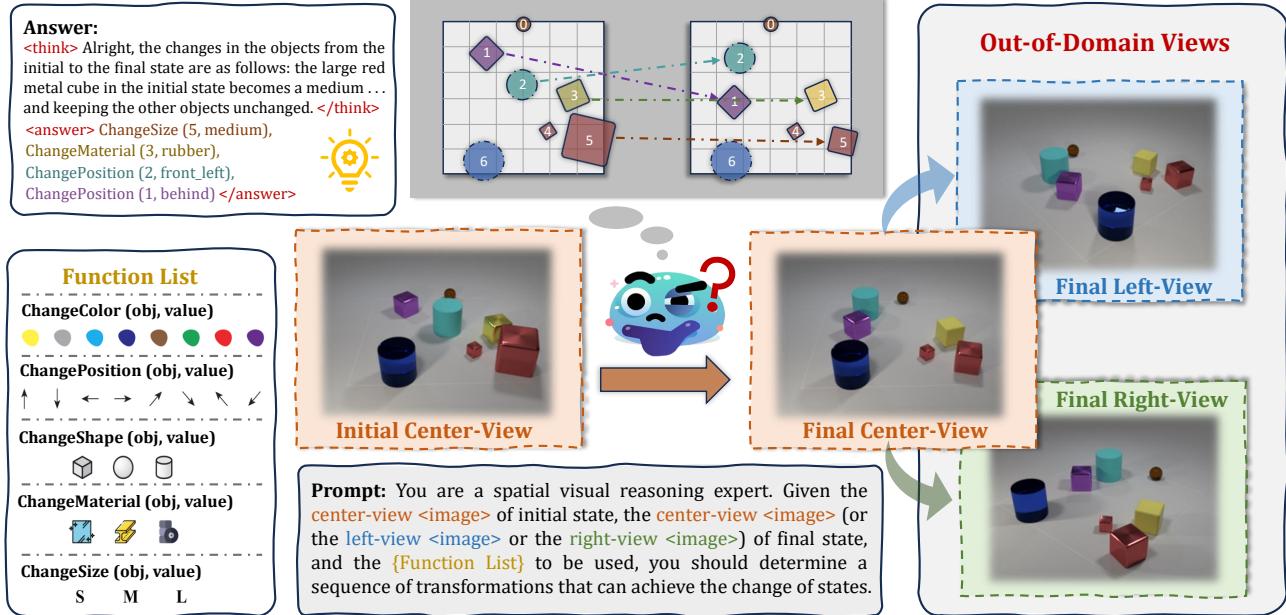


Figure 9. The sample of Spatial Transformation.

vLLM [33], ensuring reproducibility and scalability. All experiments were conducted on a cluster of servers, each equipped with  $8 \times$ A800 GPUs. For the Visual Counting task and Spatial Transformation task, we trained the models for 1 epoch each on their respective training datasets, ensuring sufficient exposure to task-specific patterns while avoiding overfitting. For the Structure Perception task, due to its GeoMath training dataset consisting of a relatively small number of training samples (a total of 4,500), we extended the training duration to 6 epochs to allow the models to fully capture the underlying structural and geometric relationships. In the Reason-RFT training pipeline, all models underwent an initial CoT activation stage with 1,600 samples before proceeding to the RL phase.

## C. More Experiment Results

**Performance at Different Training Samples** Fig. 12 and Fig. 13 illustrate the In-Domain (ID) and Out-of-Domain (OOD) performance of all methods across three visual reasoning tasks, evaluated at various training sample sizes. This analysis helps us understand how each method scales with training data. More detail evaluation results for each subset of three tasks are in Tab. 6 - Tab. 13.

We systematically varied the number of training samples, from minimal to substantial, allowing us to identify performance thresholds and data efficiency for each method in both ID and OOD contexts. Key findings from this analysis include: Data Efficiency of Reason-RFT: *Reason-RFT* demonstrates exceptional data efficiency, achieving approx-

imately 70% of the performance of *Reason-RFT-Zero* with only 3% of the training data (1,600 samples), and 82.5% with just 9Robust Generalization to OOD Tasks: In the 7B model, *Reason-RFT* achieves over 92% of *Reason-RFT-Zero*'s performance using just 3% of the training data, showcasing its strong generalization capabilities. Comparison Across Methods: *Reason-RFT* consistently outperforms other methods, particularly in data-constrained scenarios, indicating its suitability for applications with limited data availability. Performance Saturation: As training sample size increases, some methods experience performance plateaus, suggesting that beyond a certain point, additional data yields diminishing returns.

In conclusion, the evaluation of performance across different training samples not only highlights the strengths of *Reason-RFT* in terms of data efficiency and generalization but also provides critical insights into the performance dynamics of various methods. These findings are essential for practitioners aiming to maximize performance while effectively managing training resources.

## D. Visualization

In this section, we present additional visualization results, see Fig. 14 - Fig. 22. *Reason-RFT* demonstrates superior performance over CoT-SFT in terms of logical consistency, reasoning quality, and correctness. CoT-SFT's flaws stem from incorrect assumptions and misinterpretations, highlighting the importance of accurate problem interpretation and reasoning in visual reasoning tasks.

## System Prompt for Spatial Transformation Task

"Your need to complete the spatial visual reasoning task according to the following rules.

### ### Task Description:

Given the image of the initial state, the image of the final state, and the attributes of the initial objects, you should determine a transformation that can achieve the change of states.

The \*\*attributes of the initial objects\*\* are provided as a list of tuples in the following format:

\*\*('object\_id', 'shape', 'size', 'color', 'material')\*\*

Each tuple represents an object and its properties in the initial state.

The transformation should be a sequence of functions with a length ranging from 1 to 4, where each function is represented as \*\*'func(object\_id, value)'\*\*.

### ### Available functions and values:

1. \*\*'change\_size(object\_id, value)'\*\* - Changes the object to a new size relative to its initial size.

- Possible values: `['small', 'medium', 'large']`

2. \*\*'change\_color(object\_id, value)'\*\* - Changes the object to a new color relative to its initial color.

- Possible values: `['yellow', 'gray', 'cyan', 'blue', 'brown', 'green', 'red', 'purple']`

3. \*\*'change\_material(object\_id, value)'\*\* - Changes the object to a new material relative to its initial material.

- Possible values: `['glass', 'metal', 'rubber']`

4. \*\*'change\_shape(object\_id, value)'\*\* - Changes the object to a new shape relative to its initial shape.

- Possible values: `['cube', 'sphere', 'cylinder']`

5. \*\*'change\_position(object\_id, value)'\*\* - Moves the object to a new position relative to its initial location.

- Possible values: `['front', 'behind', 'left', 'right', 'front\_left', 'front\_right', 'behind\_left', 'behind\_right']`

- 'front' means moving forward along the object's initial direction.

- 'behind' means moving backward along the object's initial direction.

- 'left' means moving to the left of the object's initial orientation.

- 'right' means moving to the right of the object's initial orientation.

- 'front\_left' means moving diagonally toward the front and left of the initial location.

- 'front\_right' means moving diagonally toward the front and right of the initial location.

- 'behind\_left' means moving diagonally toward the behind and left of the initial location.

- 'behind\_right' means moving diagonally toward the behind and right of the initial location.

### ### Output Format

You should first thinks about the reasoning process internally and then provides the user with the answer. The \*\*reasoning process\*\* and \*\*answer\*\* are enclosed within specific tags:

- \*\*Reasoning process\*\*: Provide a chain-of-thought, logical explanation of the problem. This should outline step-by-step reasoning, enclosed within `<think>...</think>`

- \*\*Final answer (sequence of functions only)\*\*: Enclosed within `<answer>...</answer>`

Now, it's your turn!

{Question} Output the thinking process in <think> </think> and final answer in <answer> </answer> tags.

""

Figure 10. The system prompt used in Spatial Transformation task.

## System Prompt for Spatial Transformation Task (Add <summary> <caption> in FORMAT)

"Your need to complete the spatial visual reasoning task according to the following rules.

### ### Task Description:

Given the image of the initial state, the image of the final state, and the attributes of the initial objects, you should determine a transformation that can achieve the change of states.

The \*\*attributes of the initial objects\*\* are provided as a list of tuples in the following format:

\*\*('object\_id', 'shape', 'size', 'color', 'material')\*\*

Each tuple represents an object and its properties in the initial state.

The transformation should be a sequence of functions with a length ranging from 1 to 4, where each function is represented as \*\*'func(object\_id, value)'\*\*.

### ### Available functions and values:

1. \*\*'change\_size(object\_id, value)'\*\* - Changes the object to a new size relative to its initial size.  
- Possible values: `['small', 'medium', 'large']`

2. \*\*'change\_color(object\_id, value)'\*\* - Changes the object to a new color relative to its initial color.  
- Possible values: `['yellow', 'gray', 'cyan', 'blue', 'brown', 'green', 'red', 'purple']`

3. \*\*'change\_material(object\_id, value)'\*\* - Changes the object to a new material relative to its initial material.  
- Possible values: `['glass', 'metal', 'rubber']`

4. \*\*'change\_shape(object\_id, value)'\*\* - Changes the object to a new shape relative to its initial shape.  
- Possible values: `['cube', 'sphere', 'cylinder']`

5. \*\*'change\_position(object\_id, value)'\*\* - Moves the object to a new position relative to its initial location.  
- Possible values: `['front', 'behind', 'left', 'right', 'front\_left', 'front\_right', 'behind\_left', 'behind\_right']`  
- 'front' means moving forward along the object's initial direction.  
- 'behind' means moving backward along the object's initial direction.  
- 'left' means moving to the left of the object's initial orientation.  
- 'right' means moving to the right of the object's initial orientation.  
- 'front\_left' means moving diagonally toward the front and left of the initial location.  
- 'front\_right' means moving diagonally toward the front and right of the initial location.  
- 'behind\_left' means moving diagonally toward the behind and left of the initial location.  
- 'behind\_right' means moving diagonally toward the behind and right of the initial location.

### ### Output Format

You should first thinks about the reasoning process internally and then provides the user with the answer. The \*\*reasoning process\*\* and \*\*answer\*\* are enclosed within specific tags:

- **Summary process**: Summary how you will approach the problem and explain the steps you will take to reach the answer, enclosed within `<summary>...</summary>`

- **Caption process**: Provide a detailed description of the image, particularly emphasizing the aspects related to the question, enclosed within `<caption>...</caption>`

- **Reasoning process**: Provide a chain-of-thought, logical explanation of the problem. This should outline step-by-step reasoning, enclosed within `<think>...</think>`

- **Final answer (sequence of functions only)**: Enclosed within `<answer>...</answer>`

Now, it's your turn!

**[Question]** Output the summary process in <summary> </summary>, caption process in <caption>...</caption>, thinking process in <think> </think> and final answer in <answer> </answer> tags.  
"

Figure 11. The system prompt used in Spatial Transformation task w/ <summary> and <caption> tags in format.

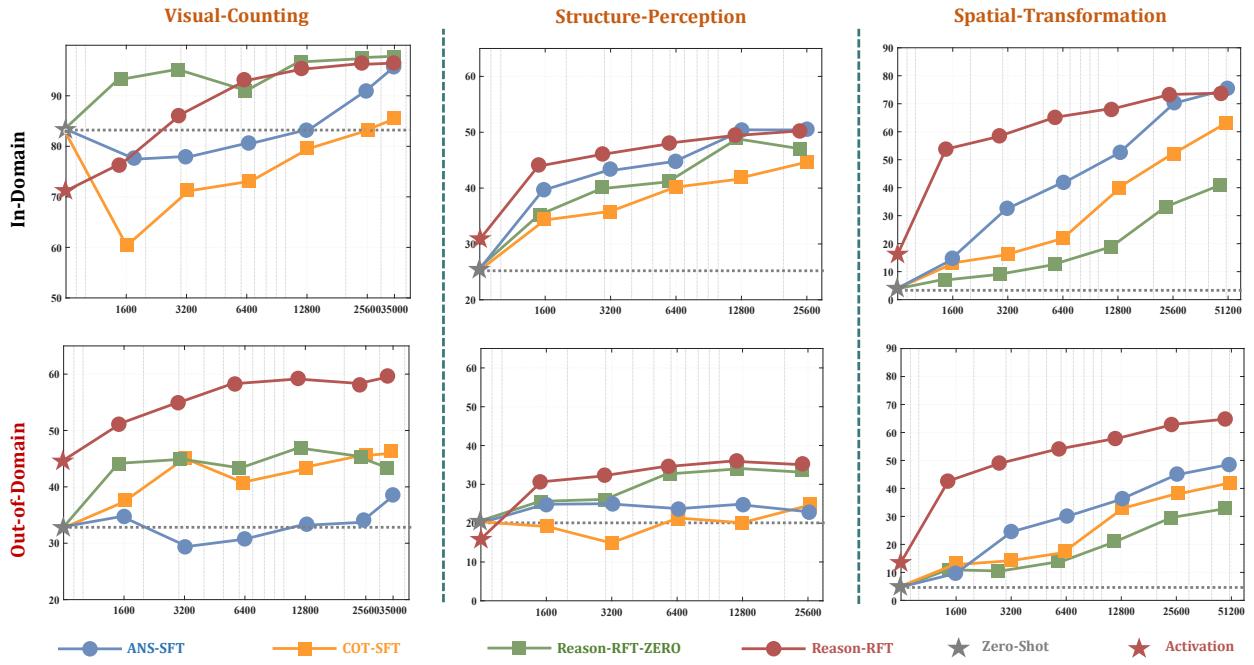


Figure 12. Results of all methods on Qwen2VL-2B-Instruct, ID and OOD performance at different training checkpoints.

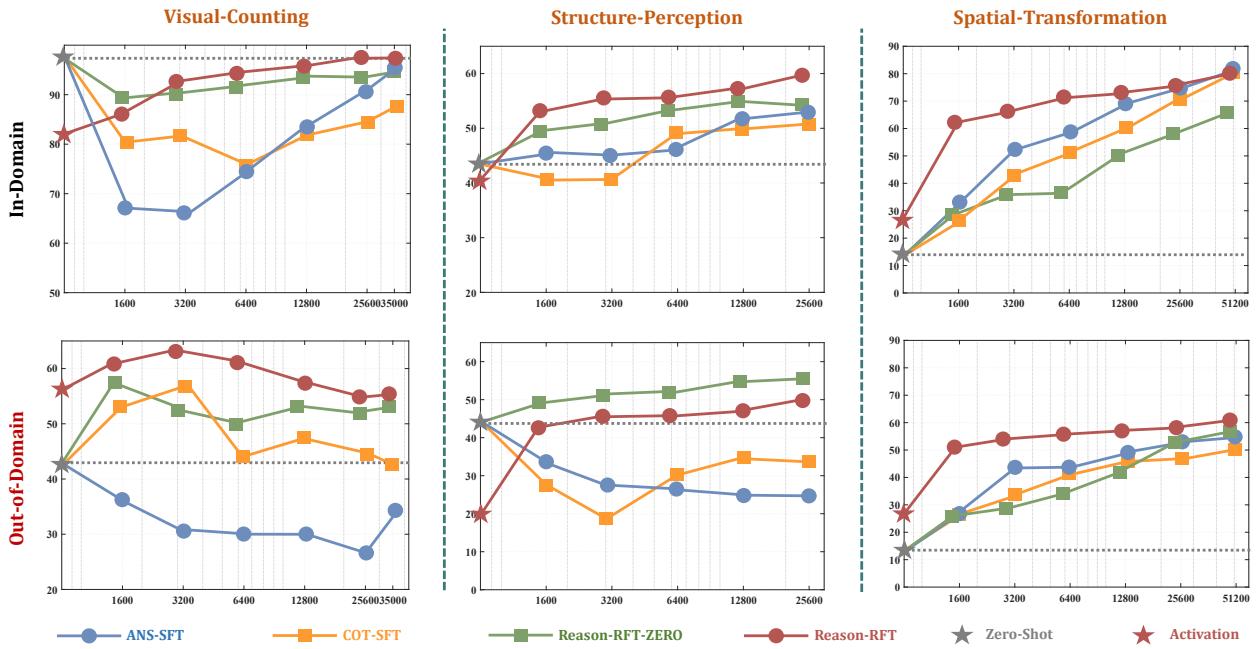


Figure 13. Results of all methods on Qwen2VL-7B-Instruct, ID and OOD performance at different training checkpoints.

Methods	Steps	Visual Counting				
		Clevr-Math (ID)				
		adversarial	sub-multi	addition	subtraction	AVG
<b>Zero-Shot</b>	-	93.60	84.00	55.60	96.40	82.40
ANS-SFT	100	83.60	56.40	91.20	81.60	78.20
	200	69.20	67.60	91.60	82.00	77.60
	400	81.60	65.60	90.80	84.80	80.70
	600	72.40	73.20	92.40	89.20	81.80
	800	78.40	77.20	82.80	90.40	82.20
	1200	85.60	78.00	91.60	95.60	87.70
	1600	92.80	82.40	94.80	96.80	91.70
	2187	95.20	92.80	97.60	99.20	96.20
CoT-SFT	100	49.20	40.00	82.00	69.20	60.10
	200	65.20	55.60	88.00	76.40	71.30
	400	66.00	57.20	90.00	79.60	73.20
	600	67.20	59.20	87.20	82.80	74.10
	800	77.60	61.60	92.40	85.20	79.20
	1200	76.80	70.00	91.20	93.60	82.90
	1600	80.80	66.80	91.60	92.00	82.80
	2187	83.20	71.20	93.20	94.40	85.50
Reason-RFT-Zero	100	92.80	88.80	94.40	96.00	93.00
	200	95.60	91.60	95.60	97.60	95.10
	400	92.00	87.60	84.00	96.40	90.00
	600	94.40	92.80	93.60	96.00	94.20
	800	96.40	96.40	96.00	98.80	96.90
	1200	98.40	95.60	100.00	99.60	98.40
	1600	96.40	94.80	98.80	99.60	97.40
	2500	98.40	95.60	99.60	100.00	98.40
Reason-RFT	100	89.60	73.20	93.60	95.60	88.00
	200	89.20	78.00	95.20	96.40	89.70
	400	92.80	82.40	95.20	97.60	92.00
	600	94.80	86.00	96.80	97.20	93.70
	800	96.80	88.40	96.80	98.80	95.20
	1200	94.80	86.00	96.40	98.80	94.00
	1600	94.40	91.60	97.20	99.60	95.70
	2500	98.40	92.80	96.80	99.20	96.80

Table 6. Complete experimental results of Qwen2VL-2B-Instruct on the Clevr-Math test set after training on Clevr-Math. “sub-multi” donates the subtraction-multihop task.

Methods	Steps	Visual Counting						
		Super-Clevr (OOD)						
		addition	subtraction	add-sub	sub-multi	Avg	Encountered	UnEncountered
<b>Zero-Shot</b>	-	10.40	54.40	0.00	63.20	32.00	42.67	0.00
ANS-SFT	100	51.20	37.60	11.60	39.20	34.90	42.67	11.60
	200	38.40	55.60	8.40	15.60	29.50	36.53	8.40
	400	40.80	45.20	5.60	35.20	31.70	40.40	5.60
	600	41.20	61.60	8.00	35.60	36.60	46.13	8.00
	800	49.20	50.40	7.20	26.00	33.20	41.87	7.20
	1200	44.00	53.20	5.60	38.80	35.40	45.33	5.60
	1600	48.80	53.60	6.00	26.00	33.60	42.80	6.00
	2187	49.60	62.00	5.20	41.60	39.60	51.07	5.20
	100	47.20	50.00	28.80	25.60	37.90	40.93	28.80
CoT-SFT	200	56.00	52.40	38.00	34.00	45.10	47.47	38.00
	400	55.20	57.20	22.40	30.40	41.30	47.60	22.40
	600	58.40	55.20	24.00	35.60	43.30	49.73	24.00
	800	57.60	47.60	26.80	41.60	43.40	48.93	26.80
	1200	58.00	54.40	35.60	32.40	45.10	48.27	35.60
	1600	53.20	58.40	33.20	40.40	46.30	50.67	33.20
	2187	53.60	58.80	36.80	36.80	46.50	49.73	36.80
	100	46.00	65.20	6.80	58.80	44.20	56.67	6.80
	200	48.80	66.00	9.20	57.60	45.40	57.47	9.20
Reason-RFT-Zero	400	42.00	71.20	8.40	50.80	43.10	54.67	8.40
	600	47.20	65.20	7.60	47.60	41.90	53.33	7.60
	800	56.40	69.20	6.80	55.20	46.90	60.27	6.80
	1200	52.00	73.60	7.20	59.20	48.00	61.60	7.20
	1600	51.60	71.60	6.40	54.80	46.10	59.33	6.40
	2500	49.60	71.20	5.20	53.20	44.80	58.00	5.20
	100	59.20	57.60	38.00	41.60	49.10	52.80	38.00
	200	59.60	64.40	39.20	42.00	51.30	55.33	39.20
	400	61.60	64.00	39.20	37.20	50.50	54.27	39.20
Reason-RFT	600	66.80	67.20	32.00	46.00	53.00	60.00	32.00
	800	66.00	65.60	34.00	39.20	51.20	56.93	34.00
	1200	67.20	65.20	33.60	40.80	51.70	57.73	33.60
	1600	63.60	66.00	33.20	44.80	51.90	58.13	33.20
	2500	68.00	67.20	28.40	44.80	52.10	60.00	28.40

Table 7. Complete experimental results of Qwen2VL-2B-Instruct on the Super-Clevr test set after training on Clevr-Math. “add-sub” donates the addition-subtraction task, while “sub-multi” donates the subtraction-multiphop task.

Methods	Steps	Visual Counting				
		Clevr-Math (ID)				
		adversarial	sub-multi	addition	subtraction	AVG
<b>Zero-Shot</b>	-	99.60	98.40	97.60	98.80	98.60
ANS-SFT	100	69.20	54.00	81.20	69.20	68.40
	200	61.20	50.00	82.40	75.60	67.30
	400	69.20	63.60	89.20	77.60	74.90
	600	70.40	54.00	90.40	81.20	74.00
	800	80.00	74.00	91.20	89.20	83.60
	1200	86.80	79.20	94.40	91.20	87.90
	1600	90.40	84.40	95.20	92.00	90.50
	2187	96.80	89.20	96.80	97.20	95.00
CoT-SFT	100	81.60	63.60	91.20	83.60	80.00
	200	80.00	64.00	92.00	88.80	81.20
	400	72.40	66.00	88.80	79.60	76.70
	600	77.60	66.00	94.40	89.20	81.80
	800	78.40	65.20	94.00	87.20	81.20
	1200	79.60	76.80	92.40	88.00	84.20
	1600	86.40	78.00	92.80	93.20	87.60
	2187	87.20	78.80	93.60	89.60	87.30
Reason-RFT-Zero	100	98.00	94.40	98.80	99.60	97.70
	200	99.60	93.20	99.20	100.00	98.00
	400	99.60	95.20	99.60	98.80	98.30
	600	98.00	98.40	100.00	99.60	99.00
	800	99.60	98.40	99.60	98.80	99.10
	1200	100.00	98.00	99.60	99.20	99.20
	1600	99.60	97.60	100.00	99.20	99.10
	2500	99.60	98.40	100.00	99.60	99.40
Reason-RFT	100	88.80	79.20	95.60	94.40	89.50
	200	92.00	80.00	96.40	95.20	90.90
	400	94.40	84.40	96.00	95.60	92.60
	600	92.80	84.00	96.40	97.60	92.70
	800	92.80	85.20	96.80	96.40	92.80
	1200	94.80	89.60	97.20	97.60	94.80
	1600	94.80	86.40	97.60	97.20	94.00
	2500	96.80	88.40	99.20	98.00	95.60

Table 8. Complete experimental results of Qwen2VL-7B-Instruct on the Clevr-Math test set after training on Clevr-Math. “sub-multi” donates the subtraction-multihop task.

Methods	Steps	Visual Counting						
		Super-Clevr (OOD)						
		addition	subtraction	add-sub	sub-multi	Avg	Encountered	UnEncountered
<b>Zero-Shot</b>	-	46.80	75.20	4.80	41.60	42.10	54.53	4.80
ANS-SFT	100	57.60	41.20	5.60	46.40	37.70	48.40	5.60
	200	42.00	38.80	8.00	33.60	30.60	38.13	8.00
	400	37.20	46.40	5.20	31.60	30.10	38.40	5.20
	600	32.00	44.80	12.40	19.20	27.10	32.00	12.40
	800	38.80	38.00	6.80	37.20	30.20	38.00	6.80
	1200	42.00	42.80	12.80	32.00	32.40	38.93	12.80
	1600	36.40	48.40	11.20	17.20	28.30	34.00	11.20
	2187	39.60	58.80	8.00	29.20	33.90	42.53	8.00
CoT-SFT	100	60.00	63.60	44.00	41.60	52.30	55.07	44.00
	200	67.60	66.40	48.00	46.80	57.20	60.27	48.00
	400	55.20	60.40	19.60	42.00	44.30	52.53	19.60
	600	64.80	61.20	35.20	43.20	51.10	56.40	35.20
	800	60.00	53.60	37.60	42.40	48.40	52.00	37.60
	1200	51.20	56.00	35.20	39.60	45.50	48.93	35.20
	1600	53.20	56.40	34.40	35.20	44.80	48.27	34.40
	2187	51.60	51.60	33.60	32.80	42.40	45.33	33.60
Reason-RFT-Zero	100	58.80	82.80	24.00	62.40	57.00	68.00	24.00
	200	56.00	83.20	18.80	50.00	52.00	63.07	18.80
	400	62.40	79.60	22.80	37.60	50.60	59.87	22.80
	600	61.20	85.20	17.20	49.20	53.20	65.20	17.20
	800	52.80	86.80	20.40	52.00	53.00	63.87	20.40
	1200	53.60	83.20	19.20	46.80	50.70	61.20	19.20
	1600	61.20	84.80	18.40	43.20	51.90	63.07	18.40
	2500	59.20	86.40	21.20	45.20	53.00	63.60	21.20
Reason-RFT	100	53.60	56.80	33.20	39.60	45.80	50.00	33.20
	200	52.00	61.20	31.60	44.00	47.20	52.40	31.60
	400	56.00	59.60	30.80	45.20	47.90	53.60	30.80
	600	56.00	64.00	31.60	50.00	50.40	56.67	31.60
	800	56.00	60.00	28.00	41.60	46.40	52.53	28.00
	1200	66.00	65.60	38.00	50.40	55.00	60.67	38.00
	1600	64.40	59.60	32.40	48.80	51.30	57.60	32.40
	2500	62.80	60.80	35.60	44.80	51.00	56.13	35.60

Table 9. Complete experimental results of Qwen2VL-7B-Instruct on the Super-Clevr test set after training on Clevr-Math. “add-sub” donates the addition-subtraction task, while “sub-multi” donates the subtraction-multihop task.

Methods	Steps	Structure Perception					
		Geometry3k (OOD)			GeoMath (ID)		
		CHOICE	NON-CHOICE	AVG	CHOICE	NON-CHOICE	AVG
<b>Zero-Shot</b>	-	40.25	1.00	20.63	35.57	20.31	25.86
ANS-SFT	100	35.25	16.25	25.75	58.72	29.89	40.37
	200	33.25	17.50	25.38	56.38	35.44	43.05
	400	30.75	17.00	23.88	64.77	35.06	45.86
	600	-	-	-	73.83	38.12	51.10
	800	32.75	16.00	24.38	72.15	36.40	49.39
	1200	-	-	-	73.83	35.44	49.39
	1600	29.00	16.00	22.50	74.83	37.36	50.98
	1686	28.75	16.25	22.50	74.83	37.93	51.34
	100	16.50	21.50	19.00	31.54	34.10	33.17
CoT-SFT	200	7.50	23.50	15.50	32.89	35.25	34.39
	400	21.50	21.25	21.38	41.61	40.04	40.61
	600	-	-	-	43.62	36.59	39.14
	800	16.50	23.50	20.00	45.97	39.27	41.70
	1200	-	-	-	53.02	40.04	44.76
	1600	24.25	24.00	24.13	53.69	37.93	43.66
	1686	26.75	23.75	25.25	51.34	38.31	43.05
	100	32.25	17.75	25.00	41.61	31.23	35.00
Reason-RFT-Zero	200	33.00	18.50	25.75	48.99	35.06	40.12
	400	41.50	23.50	32.50	52.68	34.87	41.34
	600	37.00	22.75	29.88	60.74	37.55	45.98
	800	42.25	25.00	33.63	62.42	40.42	48.42
	1200	43.00	23.75	33.38	61.07	39.66	47.44
	1600	42.75	22.25	32.50	63.09	38.31	47.32
	1610	43.25	21.75	32.50	63.09	38.89	47.68
	100	37.50	23.25	30.38	50.34	41.00	44.39
	200	33.50	29.25	31.38	56.71	40.04	46.10
Reason-RFT	400	38.25	28.75	33.50	56.38	39.27	45.49
	600	40.50	27.25	33.88	61.41	41.19	48.54
	800	41.25	29.50	35.38	58.05	41.19	47.32
	1200	40.25	31.00	35.63	61.74	42.34	49.39
	1600	38.00	29.25	33.63	62.08	43.10	50.00
	1610	36.75	29.50	33.13	60.74	42.34	49.03

Table 10. Complete experimental results of Qwen2VL-2B-Instruct on the Structure Perception task after training on GeoMath.

Methods	Steps	Structure Perception					
		Geometry3k (OOD)			GeoMath (ID)		
		CHOICE	NON-CHOICE	AVG	CHOICE	NON-CHOICE	AVG
<b>Zero-Shot</b>	-	45.25	23.00	34.13	61.07	38.12	46.46
ANS-SFT	100	38.50	18.25	28.38	64.77	34.87	45.74
	200	32.50	22.75	27.63	69.46	35.25	47.68
	400	-	-	-	72.48	40.42	52.07
	600	32.25	18.00	25.13	73.49	39.27	51.71
	800	-	-	-	75.50	37.93	51.58
	1200	32.50	18.50	25.50	75.84	37.74	51.59
	1600	32.50	18.25	25.38	75.84	37.36	51.34
	1686	18.25	38.75	28.50	38.59	42.72	41.22
	100	6.50	32.00	19.25	38.26	43.10	41.34
CoT-SFT	200	27.00	34.50	30.75	56.71	44.64	49.03
	400	-	-	-	52.68	44.06	47.19
	600	35.50	36.25	35.88	63.09	43.49	50.61
	800	-	-	-	63.42	42.91	50.36
	1200	29.50	37.50	33.50	64.09	44.06	51.34
	1600	29.25	36.75	33.00	61.74	44.06	50.49
	1686	58.50	41.75	50.13	56.71	45.98	49.88
	100	59.00	44.25	51.63	63.42	45.21	51.83
Reason-RFT-Zero	200	62.00	43.00	52.50	70.47	45.40	54.51
	400	-	-	-	70.13	46.74	55.24
	600	64.75	45.25	55.00	70.47	49.23	56.95
	800	-	-	-	66.11	46.17	53.42
	1200	69.00	43.25	56.13	71.14	45.59	54.88
	1600	66.25	43.25	54.75	69.80	46.55	55.00
	1610	46.75	37.50	42.13	67.79	45.79	53.79
	100	53.00	37.00	45.00	72.82	46.93	56.34
	200	52.75	37.25	45.00	71.14	46.55	55.49
Reason-RFT	400	51.50	37.00	44.25	73.49	48.28	57.44
	600	56.75	37.25	47.00	77.52	46.17	57.56
	800	59.00	40.00	49.50	79.87	48.08	59.63
	1200	56.00	39.50	47.75	74.50	49.62	58.66
	1600	59.00	39.50	49.25	78.52	48.28	59.27
	1610	59.00	39.50	49.25	78.52	48.28	59.27

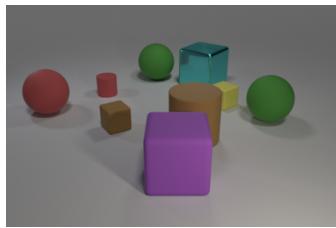
Table 11. Complete experimental results of Qwen2VL-7B-Instruct on the Structure Perception task after training on GeoMath.

Method	Steps	Spatial Transformation										TRANCE-R (OOD)					
		TRANCE (ID)				TRANCE-L (OOD)						TRANCE-R (OOD)				Avg	
		Level-1	Level-2	Level-3	Level-4	Avg	Level-1	Level-2	Level-3	Level-4	Avg	Level-1	Level-2	Level-3	Level-4	Avg	
GPT-4o Zero-Shot	/	47.28	42.96	40.87	39.08	42.55	23.16	30.56	30.73	30.22	28.67	24.38	31.74	31.13	31.77	29.76	
	/	2.10	3.27	4.08	5.68	3.78	2.02	4.73	5.57	6.08	4.60	2.39	4.73	5.59	5.95	4.67	
ANS-SFT	100	15.90	19.12	14.67	13.12	15.70	10.60	11.33	10.01	9.07	10.25	11.08	12.17	10.17	10.07	10.87	
	200	23.97	29.56	33.98	33.95	30.37	13.25	26.76	31.26	32.58	25.96	13.53	26.54	30.53	31.20	25.45	
	400	44.95	42.58	40.75	33.65	40.48	26.03	35.98	34.26	29.38	31.41	24.06	35.62	35.46	31.73	31.72	
	800	62.10	56.55	53.01	47.55	54.80	24.70	42.27	43.05	42.65	38.17	26.09	38.98	42.95	42.55	37.64	
	1600	80.70	75.68	68.34	64.60	72.33	34.38	49.00	53.83	53.10	47.58	35.84	49.28	51.23	52.50	47.21	
	3200	82.85	80.30	78.00	71.60	78.19	36.22	52.61	55.23	54.27	49.58	38.51	52.78	54.47	53.98	49.94	
	final	82.70	79.93	76.70	70.22	77.39	36.00	52.82	54.59	53.55	49.24	39.63	53.75	54.33	53.60	50.33	
COT-SFT	100	6.99	14.90	15.99	20.36	14.56	10.32	13.74	11.82	14.69	12.64	6.97	13.38	12.84	13.41	11.65	
	200	15.45	19.12	14.53	16.46	16.39	12.90	17.51	14.22	15.82	15.11	11.23	17.71	13.87	16.32	14.78	
	400	25.98	26.74	19.94	16.02	22.17	15.73	21.19	17.73	15.96	17.65	16.13	20.55	17.91	15.97	17.64	
	800	43.85	43.19	41.77	37.84	41.66	22.88	38.29	37.54	35.29	33.50	22.74	35.81	37.60	36.23	33.10	
	1600	52.82	61.06	54.38	45.85	53.53	28.65	43.95	40.99	40.13	38.43	29.00	41.23	40.78	39.26	37.57	
	3200	61.40	69.15	65.32	62.28	64.54	28.67	45.97	50.06	52.10	44.20	31.19	45.87	45.92	51.35	43.58	
	final	67.47	67.52	62.78	64.37	28.87	44.41	49.16	50.30	43.19	30.20	44.77	47.15	49.33	42.86		
Reason-RFT-Zero	100	8.44	17.96	20.69	26.22	18.33	8.53	17.42	21.16	25.05	18.04	8.08	18.12	21.09	25.70	18.25	
	200	9.59	18.76	22.97	28.73	20.01	9.49	20.08	23.19	27.00	19.94	9.72	18.97	22.82	28.50	20.00	
	400	12.35	21.47	27.01	26.25	21.77	11.10	21.47	25.73	25.30	20.90	10.54	21.19	25.44	25.60	20.69	
	800	18.47	32.08	32.77	27.85	27.79	15.40	29.12	30.93	27.38	25.71	15.52	27.75	31.50	27.88	25.66	
	1600	36.78	40.20	37.78	34.51	37.32	19.96	33.03	35.84	34.49	30.83	20.39	32.85	33.87	33.90	30.25	
	3200	43.72	46.89	44.07	40.50	43.80	18.67	34.11	37.85	39.69	32.58	18.08	34.01	37.27	40.27	32.41	
	final	46.21	45.01	44.53	42.11	44.47	18.33	34.57	37.94	40.57	32.85	18.28	33.45	37.44	40.34	32.38	
Reason-RFT	100	53.52	55.47	58.91	53.35	55.31	31.84	47.02	50.62	50.39	44.97	31.29	46.08	48.61	49.85	43.96	
	200	54.97	59.77	63.67	59.46	59.47	35.72	49.28	54.85	54.67	48.63	36.74	52.27	53.68	54.16	49.21	
	400	63.80	66.97	68.47	64.70	65.99	39.74	55.94	61.27	57.94	53.72	41.10	56.35	59.16	57.22	53.46	
	800	64.33	68.13	66.88	63.15	65.62	47.40	61.64	63.00	58.45	57.62	46.76	60.60	61.40	59.91	57.17	
	1600	76.47	73.42	74.05	69.16	73.28	52.68	62.22	66.56	64.71	61.54	53.47	65.17	64.98	63.52	61.79	
	3200	72.88	74.85	75.77	72.45	73.99	52.58	63.60	68.51	66.01	62.68	52.67	65.60	67.60	65.66	62.88	
	final	74.10	74.52	76.68	73.12	74.61	53.49	65.72	69.64	67.34	64.05	54.95	66.25	68.32	66.80	64.08	

Table 12. Complete experimental results of Qwen2VIL-2B-Instruct on the Spatial Transformation task after training on TRANCE.

Method	Steps	Spatial Transformation										TRANCE-R (OOD)					
		TRANCE (ID)					TRANCE-L (OOD)					TRANCE-R (OOD)					
		Level-1	Level-2	Level-3	Level-4	Avg	Level-1	Level-2	Level-3	Level-4	Avg	Level-1	Level-2	Level-3	Level-4	Avg	
GPT-4o Zero-Shot	/	47.28	42.96	40.87	39.08	42.55	23.16	30.56	30.73	30.22	28.67	24.38	31.74	31.13	31.77	29.76	
	/	16.25	16.42	10.96	10.48	13.53	11.71	16.80	11.50	10.85	12.72	13.30	16.08	10.55	11.18	12.78	
ANS-SFT	100	40.30	37.05	30.67	28.35	34.09	32.07	31.12	26.00	26.80	29.00	26.38	29.71	27.74	26.48	27.58	
	200	65.18	53.33	49.43	45.15	53.27	33.29	45.14	45.61	45.52	42.39	35.17	43.43	45.97	43.02	41.90	
COT-SFT	400	65.33	59.35	57.17	50.77	58.16	32.40	44.13	47.69	46.23	42.61	32.10	44.88	47.04	45.25	42.32	
	800	78.90	70.67	63.97	62.10	68.91	34.08	50.62	51.99	52.95	47.41	34.22	50.40	50.62	52.88	47.03	
Reason-RFT-Zero	1600	78.50	76.12	73.80	66.25	73.67	38.85	52.97	57.93	56.05	51.45	37.77	53.57	56.45	55.92	50.93	
	3200	83.80	83.23	82.83	78.17	82.01	40.10	56.02	61.02	59.90	54.26	40.78	55.06	61.67	60.98	54.62	
Reason-RFT	final	83.70	84.10	82.50	78.45	82.19	39.67	55.58	61.84	60.05	54.29	42.64	54.84	61.44	60.38	54.83	
	100	20.58	28.98	25.97	30.00	26.38	21.89	29.94	29.18	29.33	27.59	18.49	28.43	30.43	30.36	26.93	
Reason-RFT-Zero	200	41.80	44.08	46.02	42.16	43.52	25.31	36.28	36.81	38.86	34.32	21.19	34.55	37.06	37.54	32.59	
	400	45.39	51.32	58.20	52.42	51.83	32.53	44.61	47.73	46.10	42.74	31.87	38.96	44.62	45.90	40.34	
Reason-RFT	800	54.87	61.97	62.20	59.93	59.74	30.19	46.01	50.01	52.88	44.77	29.91	45.02	49.60	52.83	44.34	
	1600	71.27	71.14	72.82	69.93	71.29	28.82	46.43	51.01	58.94	46.30	29.08	45.25	52.24	58.31	46.22	
Reason-RFT	3200	84.13	80.62	79.99	78.42	80.79	29.93	47.63	56.49	62.54	49.15	30.46	47.85	54.83	61.02	48.54	
	final	86.50	79.43	80.54	78.77	81.31	28.07	47.54	54.42	61.58	47.90	29.69	45.32	54.69	61.48	47.80	
Reason-RFT-Zero	100	23.59	31.62	33.22	31.27	29.93	15.88	26.86	28.13	30.19	25.27	15.21	27.29	27.54	29.88	24.98	
	200	35.06	39.45	36.80	34.77	36.52	20.39	30.22	31.15	31.20	28.24	18.10	29.27	30.57	30.81	27.19	
Reason-RFT	400	25.28	40.78	41.70	35.35	35.78	20.20	39.28	35.43	33.44	32.09	21.72	39.47	37.63	32.79	32.90	
	800	50.18	51.55	50.43	46.06	49.56	35.44	46.15	45.65	39.55	41.70	33.90	45.89	46.48	41.57	41.96	
Reason-RFT	1600	59.60	61.90	57.30	55.36	58.54	43.95	55.03	52.96	50.60	50.64	41.28	56.46	51.64	49.08	49.62	
	3200	62.50	68.53	68.79	66.22	66.51	42.54	58.05	58.97	60.10	54.92	42.56	56.93	60.02	60.21	54.93	
Reason-RFT	final	65.63	68.30	69.45	67.30	67.67	46.61	58.22	61.69	62.26	57.20	45.53	58.40	61.81	58.85	56.15	
	100	60.07	62.95	68.53	62.78	63.58	37.09	52.54	60.21	54.99	51.21	35.26	52.56	57.30	55.42	50.14	
Reason-RFT	200	67.69	66.98	69.53	66.00	67.55	36.26	53.94	60.20	60.05	52.61	34.33	54.21	58.25	60.79	51.90	
	400	74.72	71.31	73.62	69.14	72.20	36.11	56.00	62.25	63.01	54.34	36.95	55.72	60.82	64.88	54.59	
Reason-RFT	800	72.57	71.04	75.58	73.26	73.11	38.93	56.86	64.10	63.69	55.90	35.06	54.39	60.82	64.88	53.79	
	1600	79.18	76.70	76.59	73.26	76.43	37.25	53.44	66.38	66.76	55.96	35.01	55.39	66.02	66.67	55.77	
Reason-RFT	3200	79.70	79.22	80.13	76.61	78.92	37.63	55.73	65.24	68.02	56.66	37.18	56.34	63.90	68.89	56.58	
	final	80.48	79.15	81.97	78.26	79.97	43.68	61.04	64.66	68.05	59.36	45.45	56.34	63.33	69.33	58.61	

Table 13. Complete experimental results of Qwen2VLL-7B-Instruct on the Spatial Transformation task after training on TRANCE.



**Question:** Subtract all spheres. How many objects are left?

#### COT-SFT:

<think>Let's start by identifying all the objects in the image. We can see various shapes: spheres, cubes, and cylinders. First, let's count the total number of objects in the image: 1. Red sphere 2. Green sphere 3. Another green sphere 4. Purple cube 5. Brown cube 6. Yellow cube 7. Brown cylinder 8. Blue transparent cube.  
**(Miss one object)** Now, let's identify and count the spheres: - Red sphere - Green sphere - Another green sphere. There are 3 spheres in total. Next, we subtract the number of spheres from the total number of objects:  
**Total objects = 8, Spheres = 3.** Objects left after subtracting spheres = **Total objects - Spheres = 8 - 3 = 5.** Oh, I see, after removing all the spheres, we are left with 5 objects.</think>

<answer>5</answer>

#### Reason-RFT:

<think>Let's start by identifying all the objects in the image. I see a total of 9 objects. The question asks us to subtract all the spheres.  
**There are 3 spheres in the image:** a red sphere, a green sphere, and a green sphere. So, we need to subtract these 3 spheres from the total of 9 objects. **9 (total objects) - 3 (spheres) = 6.** Therefore, after subtracting the spheres, there are 6 objects left.</think>

<answer>6</answer>



**Question:** Add 7 small cars. How many cars are left?

#### COT-SFT:

<think>Let's start by identifying the number of cars in the image. I see a blue SUV, a red airplane, a purple dirt bike, and a blue small car. That makes a total of 4 cars initially. Now, according to the question, we need to add 7 small cars to the existing ones. So, let's calculate the total number of cars after adding these 7 small cars. Initial number of cars = 4. Number of small cars to add = 7. Total number of cars = Initial number of cars + Number of small cars to add. **Total number of cars = 4 + 7 = 11.** Therefore, after adding 7 small cars, there are 11 cars in total.</think>

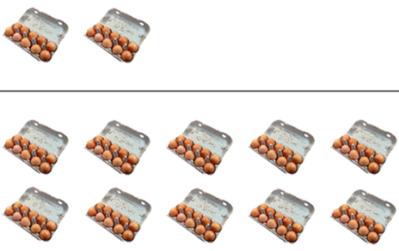
<answer>11</answer>

#### Reason-RFT:

<think>Let's start by identifying the cars in the image. I see two cars: a blue SUV and a blue minivan. The question asks to add 7 small cars to the existing cars. So, we begin with 2 cars and add 7 more.  
**2 (existing cars) + 7 (additional cars) = 9 cars in total.** Therefore, after adding 7 small cars, there are 9 cars in total.</think>

<answer>9</answer>

Figure 14. Visualization examples of Reason-RFT and CoT-SFT on Visual Counting task.



#### COT-SFT:

<think>First, let's count the number of egg trays in the first group. There are 3 trays in the first group. **Each tray holds 12 eggs, so the total number of eggs in the first group is  $3 \text{ trays} * 12 \text{ eggs/tray} = 36 \text{ eggs}$ .** Next, let's count the number of egg trays in the second group. There are 10 trays in the second group. Each tray holds 12 eggs, so the total number of eggs in the second group is  $10 \text{ trays} * 12 \text{ eggs/tray} = 120 \text{ eggs}$ . **Now, let's find the ratio of the total number of eggs in the second group to the first group:  $120 \text{ eggs} / 36 \text{ eggs} = 10$ .** Therefore, the second group has 10 times more eggs than the first group.</think>

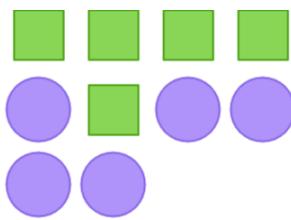
<answer>10</answer>

**Question:** How many times larger is the total number of eggs in the second group compared to the first group?

#### Reason-RFT:

<think>First, let's count the number of egg trays in the first group. There are 2 trays in the first group. **Each tray contains 10 eggs, so the total number of eggs in the first group is  $2 \text{ trays} * 10 \text{ eggs/tray} = 20 \text{ eggs}$ .** Next, let's count the number of egg trays in the second group. There are 10 trays in the second group. Each tray contains 10 eggs, so the total number of eggs in the second group is  $10 \text{ trays} * 10 \text{ eggs/tray} = 100 \text{ eggs}$ . Now, let's compare the total number of eggs in the second group to the first group. The second group has 100 eggs, and the first group has 20 eggs. To find how many times larger the second group is compared to the first group, **we divide the total number of eggs in the second group by the total number of eggs in the first group:  $100 \text{ eggs} / 20 \text{ eggs} = 5$ .**</think>

<answer>5</answer>



#### COT-SFT:

<think>First, let's count the number of green squares. **There are 4 green squares in the image.** Next, let's count the number of purple circles. There are 5 purple circles in the image. To find the difference, we subtract the number of green squares from the number of purple circles:  **$5 \text{ (purple circles)} - 4 \text{ (green squares)} = 1$** .</think>

<answer>1</answer>

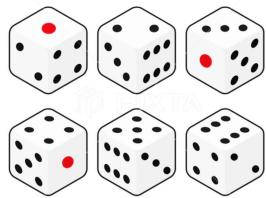
**Question:** How many more purple circles are there than green squares?

#### Reason-RFT:

<think>First, let's count the number of purple circles. I see 1, 2, 3, 4, 5 purple circles. Next, let's count the number of green squares. I see 1, 2, 3, 4, 5 green squares. Now, to find out how many more purple circles there are than green squares, we subtract the number of green squares from the number of purple circles:  **$5 \text{ (purple circles)} - 5 \text{ (green squares)} = 0$** . Oh, I see, there are no more purple circles than green squares.</think>

<answer>0</answer>

Figure 15. Visualization examples of Reason-RFT and CoT-SFT on Structure Perception task.



#### COT-SFT:

<think>Let me think about the properties of a standard six-sided die. On a six-sided die, each number from 1 to 6 is opposite to another number. **The sum of the numbers on opposite sides is always 7. Number 4 is opposite to number 2 because  $4 + 2 = 6$ , and the sum of 6 and 1 is 7. Therefore, the number directly opposite to number 4 is 2.**</think>

<answer>**B**</answer>

**Question:** Based on the image, determine which number is directly opposite to number 4 on the die depicted. Use reasoning to arrive at the correct answer from the choices provided.  
Choices: A. 1    B. 2    C. 3    D. 6

#### Reason-RFT:

<think>Let's analyze the image. The numbers shown are 2, 3, and 4. To find the number opposite to 4, we need to consider the properties of a standard six-sided die. **On a die, opposite numbers add up to 7. Since 4 is opposite to 3, we can conclude that the number opposite to 4 is 3.**</think>

<answer>**C**</answer>



#### COT-SFT:

<think>To find the total surface area of a rectangular prism, we need to calculate the sum of the areas of all six faces. **The formula for the surface area is  $2 * (\text{length} * \text{width} + \text{width} * \text{height} + \text{height} * \text{length})$ .** Let's calculate it step-by-step. 1. Calculate the area of the face with dimensions 3 units by 4 units:  $3 * 4 = 12$  square units. 2. Calculate the area of the face with dimensions 4 units by 5 units:  $4 * 5 = 20$  square units. 3. Calculate the area of the face with dimensions 5 units by 3 units:  $5 * 3 = 15$  square units. Now, add these areas together:  $12 + 20 + 15 = 47$  square units.

**Therefore, the total surface area of the rectangular prism is 47 square units.**</think>

<answer>**47**</answer>

**Question:** What is the total surface area of this rectangular prism if its dimensions are 3 units by 4 units by 5 units?

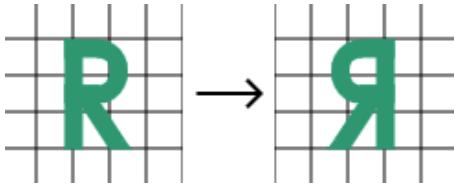
#### Reason-RFT:

<think>To find the total surface area of a rectangular prism, we need to calculate the sum of the areas of all six faces. **The formula for the total surface area is  $2 * (\text{length} * \text{width} + \text{width} * \text{height} + \text{height} * \text{length})$ .** Let's calculate it step by step. First, calculate the area of the face with dimensions 3 units and 4 units:  $3 * 4 = 12$  square units. Next, calculate the area of the face with dimensions 4 units and 5 units:  $4 * 5 = 20$  square units. Then, calculate the area of the face with dimensions 5 units and 3 units:  $5 * 3 = 15$  square units. **Now, add these areas together:  $12 + 20 + 15 = 47$  square units.**

Finally, multiply by 2 to account for both pairs of each face:  $2 * 47 = 94$  square units.</think>

<answer>**94**</answer>

Figure 16. Visualization examples of Reason-RFT and CoT-SFT on Structure Perception task.



#### COT-SFT:

<think>Let me think about the transformation applied to the letter on the left. The letter on the left is a grid of squares, and the letter on the right is also a grid of squares. **The orientation and shape of the letter remain the same**, but the position of the letter has changed. Oh, I see, **this is a slide transformation because the letter has been moved horizontally.** </think>

<answer>A</answer>

**Question:** What the transformation has been applied to the letter on the left to obtain the letter on the right, and which option correctly describes this transformation?  
Choices: (A) slide (B) flip (C) turn

#### Reason-RFT:

<think>Let's analyze the image. The letter on the left is the original image, and the letter on the right is the transformed image. **The transformation involves flipping the letter horizontally.** This means the letter has been **mirrored along a vertical axis**. Therefore, the correct option describing **this transformation is a flip.** </think>

<answer>B</answer>

Kilometers walked	
Day	Kilometers
Monday	4
Tuesday	3
Wednesday	7
Thursday	4
Friday	7
Saturday	3
Sunday	7

#### COT-SFT:

<think>First, let's list the kilometers Josiah walked each day: 4, 3, 7, 4, 7, 3, 7. Now, let's count how many times each number appears: **4 appears 3 times**, **3 appears 2 times**, and **7 appears 2 times**. Oh, I see, the number that appears the most frequently is the mode. **Since 4 appears 3 times, which is more than any other number, the mode is 4.** </think>

<answer>4</answer>

**Question:** Josiah kept track of how many kilometers he walked during the past 7 days. Can you determine which number of kilometers he walked most frequently?

#### Reason-RFT:

<think>First, let's list the kilometers Josiah walked each day: 4, 3, 7, 4, 7, 3, 7. Now, let's count how many times each distance appears: - **4 km appears 2 times**. - **3 km appears 2 times**. - **7 km appears 3 times**. Oh, I see, **the highest frequency is 3, which corresponds to 7 km.** </think>

<answer>7</answer>

Figure 17. Visualization examples of Reason-RFT and CoT-SFT on Structure Perception task.



#### COT-SFT:

<think>Let me think about the main structural elements visible in the image. First, I see the two large white towers, which are the main supports for the suspension bridge. These towers are tall and robust, providing the necessary height and strength to maintain the tension in the cables. So, there are two main structural elements: the towers. Oh, I see, the towers are crucial for supporting the weight of the bridge and ensuring its stability. Without them, the bridge would not be able to span the valley effectively.</think>

<answer>2</answer>

**Question:** How many main structural elements can be identified in the suspension bridge, and what are their roles in supporting the bridge?

#### Reason-RFT:

<think>Let me think about the main structural elements visible in the image. First, I notice the two large towers supporting the bridge. These towers are crucial for providing the necessary height and strength to maintain the bridge's structure. Next, I see the main span of the bridge, which is the horizontal part connecting the towers. This span is supported by the towers and is essential for carrying the weight of the bridge and any traffic it carries. Finally, there are additional support structures visible, such as the curved steel cables, which help distribute the weight evenly and maintain the bridge's stability. So, the main structural elements are the towers, the main span, and the curved steel cables. Counting these elements, I see two towers, one main span, and multiple curved steel cables. Therefore, the main structural elements are three: the towers, the main span, and the curved steel cables.</think>

<answer>3</answer>



**Question:** How many distinct colors of lights are used in the Christmas tree decoration?

#### COT-SFT:

<think>Let me think about the colors described in the image. The Christmas tree is decorated with lights that are described as being red, blue, and white. I need to count these distinct colors. First, I identify the color red, which is mentioned directly. Next, I see the color blue, which is also mentioned. Finally, I observe the color white, which is described as shining brightly. So, there are three distinct colors of lights used in the decoration: red, blue, and white.</think>

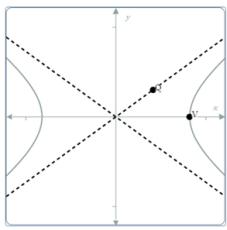
<answer>3</answer>

#### Reason-RFT:

<think>Let me think. The image shows a Christmas tree decorated with lights. The description mentions that the lights are of different colors. I need to identify these colors. The description specifically mentions red and white lights. So, there are two distinct colors of lights used in the decoration.</think>

<answer>2</answer>

Figure 18. Visualization examples of Reason-RFT and CoT-SFT on Structure Perception task.



### COT-SFT:

<think>Let's think about the properties of a hyperbola. The slope of the asymptote of a hyperbola is given by the equation  $m = \frac{a}{b}$ , where  $a$  and  $b$  are the coefficients from the equation of the hyperbola  $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ . The distance from the center to the vertex is  $c$ , which is related to the semi-major and semi-minor axes. However, to find the equation of the hyperbola, we need more information than just the slope of the asymptote and the distance from the center to the vertex. We also need the values of  $a$  and  $b$ , which are not directly provided in the image. Therefore, the information given is insufficient to determine the equation of the hyperbola.</think>

<answer>B</answer>

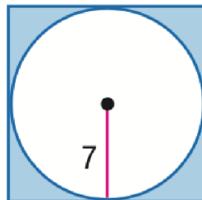
**Question:** Given the hyperbola shown in the image, can we determine the equation of the hyperbola by analyzing the relationship between the slope of the asymptote and the distance from the center to the vertex?

Choices: A. Yes B. No

### Reason-RFT:

<think>Let's analyze the image. The hyperbola is shown with two asymptotes intersecting at the center. To find the equation of the hyperbola, we need to consider the relationship between the slope of the asymptote and the distance from the center to the vertex. The equation of a hyperbola with its center at the origin and aligned with the coordinate axes is given by  $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ , where  $a$  and  $b$  are the distances from the center to the vertices and asymptotes, respectively. The center to the vertices is given, which suggests that  $a^2 + b^2$  is given. The slope of the asymptotes is given by  $\pm \frac{b}{a}$ . In this case, the slope is 3, which implies that  $\frac{b}{a} = 3$ . Therefore, the equation of the hyperbola can be determined by these relationships.</think>

<answer>A</answer>



### COT-SFT:

<think>To find the circumference of the circle, I need to know the diameter of the circle. The diameter is the longest straight line that can be drawn across a circle, passing through the center. Once I have the diameter, I can use the formula for the circumference of a circle, which is  $C = \pi d$ , where  $d$  is the diameter. Oh, I see, the diameter of the circle is given as 7. So, I can calculate the circumference as  $C = \pi * 7$ . Using the value of  $\pi$  as approximately 3.14159, I can calculate the circumference as  $7 * 3.14159 = 21.99633$ . Rounding this to one decimal place, the circumference is approximately 22.</think>

<answer>22</answer>

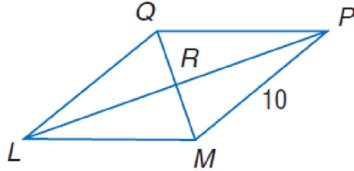
**Question:** What is the circumference of the circle?

### Reason-RFT:

<think>First, let's determine the radius of the circle. The problem states that the radius is 7. The formula for the circumference of a circle is  $C = 2\pi r$ , where  $r$  is the radius. Substituting the given radius,  $r = 7$ , into the formula, we have  $C = 2\pi(7)$ . Calculating this gives  $C = 14\pi$ . Therefore, the circumference of the circle is  $14\pi$ .</think>

<answer>14\pi</answer>

Figure 19. Visualization examples of Reason-RFT and CoT-SFT on Structure Perception task.



**Question:** What is the length of side QP in the rhombus?

**COT-SFT:**

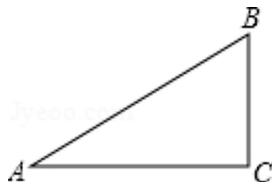
<think>Let's think about this step-by-step. The rhombus LMPQ is a type of quadrilateral where all sides are equal in length. We are given that side LM is 8 units long. Since all sides of a rhombus are equal, the length of side QP must also be equal to the length of side LM. Therefore,  $QP = LM = 8$  units.</think>

<answer>8</answer>

**Reason-RFT:**

<think>Let's analyze the rhombus LMPQ. We know that all sides of a rhombus are equal in length. The side QP is one of these equal sides. Since the problem states that the side MP is 10, we can confirm that this is indeed the length of side QP in the rhombus LMPQ. </think>

<answer>10</answer>



**Question:** Consider the right triangle ACB, where angle C is 90 degrees, angle A is 37 degrees, and the length of AC is 4 units. What is the approximate length of BC, expressed as 'k' in terms of the given angles and length? ( $\sin 37.0 \approx 0.6$ ,  $\cos 37.0 \approx 0.8$ ,  $\tan 37.0 \approx 0.75$ )  
Choices: A: 2.4 B: 3.0 C: 3.2 D: 5.0

**COT-SFT:**

<think>Let's think through this step-by-step. We know that angle A is 37 degrees, and we have the tangent function, which is the ratio of the opposite side (AB) to the adjacent side (AC). Given that  $\tan(37^\circ) = 0.75$ , we can write the equation as  $AB = AC / \tan(37^\circ)$ . Substituting the given value,  $AB = 4 / 0.75$ . Calculating this gives  $AB = 5.33$ . Now, we need to find the hypotenuse BC using the Pythagorean theorem, which states that in a right triangle, the square of the hypotenuse is equal to the sum of the squares of the other two sides. So,  $BC = \sqrt{AB^2 + AC^2}$ . Substituting the known values,  $BC = \sqrt{5.33^2 + 4^2}$ . Calculating this gives  $BC = 7.34$ . Therefore, the approximate length of BC is 7.34, which is not an exact match with the options. However, the closest approximation from the choices given is 3.0, which is not correct based on our calculations. Let's re-evaluate the choices and the reasoning provided.</think>

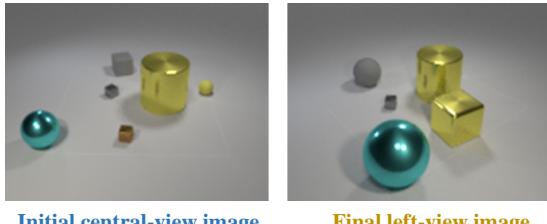
<answer>Based on the reasoning and calculations, the closest approximation from the choices given is 3.0, which is not correct based on our calculations</answer>

**Reason-RFT:**

<think>Let's start by finding the length of BC using the tangent of angle A. We know that  $\tan(\text{angle A}) = \text{opposite} / \text{adjacent} = BC/AC$ . Given that angle A is 37 degrees, we can write:  $\tan(37^\circ) = BC/4$ . We also know that  $\tan(37^\circ)$  is approximately 0.75. So,  $BC = 4 * 0.75 = 3$ . Therefore, the length of BC is approximately 3.0.</think>

<answer>B</answer>

Figure 20. Visualization examples of Reason-RFT and CoT-SFT on Structure Perception task.



Initial central-view image

Final left-view image

**Question:** The first image shows the central view of initial state, the second image shows the left view of final state, and the attributes of the initial objects are listed below:

(0, cube, medium, gray, rubber) (1, sphere, medium, cyan, metal)

(2, cube, small, gray, metal) (3, cube, small, brown, metal)

(4, cylinder, large, yellow, metal) (5, sphere, small, yellow, rubber)

What is the transformation process that can achieve the change from initial state to final state?

#### COT-SFT:

<think>Let's analyze the initial and final states of the objects in the images. We have the attributes of the initial objects, and we need to determine the transformations that lead to the final state.

1. \*\*Object 0\*\*: Initially, it's a medium gray rubber cube. In the final state, it appears to be a sphere. So, we need to change shape.  
- Transformation: `change\_shape(0, sphere)`
2. \*\*Object 1\*\*: Initially, it's a medium cyan metal sphere. In the final state, it seems to have moved to the right and increased in size.  
- Transformation: `change\_position(1, right)`, `change\_size(1, large)`
3. \*\*Object 2\*\*: Initially, it's a small gray metal cube. In the final state, it appears unchanged.
4. \*\*Object 3\*\*: Initially, it's a small brown metal cube. In the final state, it appears unchanged.
5. \*\*Object 4\*\*: Initially, it's a large yellow metal cylinder. In the final state, it appears unchanged.
6. \*\*Object 5\*\*: Initially, it's a small yellow rubber sphere. In the final state, it appears unchanged.

Now, let's compile the transformations needed to achieve the final state from the initial state:

- `change\_shape(0, sphere)`
- `change\_position(1, right)`
- `change\_size(1, large)`

These transformations match the original answer provided.</think>

<answer>change\_shape(0, sphere), change\_position(1, right), change\_size(1, large)</answer>

#### Reason-RFT:

<think>Let's analyze the initial and final states to determine the transformations needed.

1. \*\*Object 0\*\*: Initially a medium gray rubber cube. In the final state, it appears to be a medium gray rubber sphere.  
- Transformation: `change\_shape(0, sphere)`
2. \*\*Object 1\*\*: Initially a medium cyan metal sphere. In the final state, it appears to be a large cyan metal sphere.  
- Transformation: `change\_size(1, large)`
3. \*\*Object 2\*\*: Initially a small gray metal cube. In the final state, it appears to be a small gray metal cube.  
- Transformation: None needed as it remains unchanged.
4. \*\*Object 3\*\*: Initially a small brown metal cube. In the final state, it appears to be a large yellow metal cube.  
- Transformation: `change\_color(3, yellow)` and `change\_size(3, large)`
5. \*\*Object 4\*\*: Initially a large yellow metal cylinder. In the final state, it appears to be a large yellow metal cylinder.  
- Transformation: None needed as it remains unchanged.
6. \*\*Object 5\*\*: Initially a small yellow rubber sphere. In the final state, it appears to be a small yellow rubber sphere.  
- Transformation: None needed as it remains unchanged.

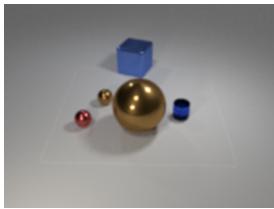
By analyzing the changes, we can determine the sequence of transformations needed to achieve the final state from the initial state. Therefore, the transformations are:

- `change\_color(3, yellow)`
- `change\_size(1, large)`
- `change\_size(3, large)`
- `change\_shape(0, sphere)`

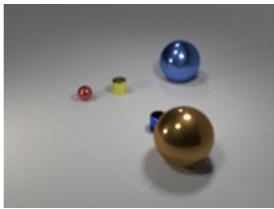
These transformations match the original answer provided. </think>

<answer>change\_color(3, yellow), change\_size(1, large), change\_size(3, large), change\_shape(0, sphere)</answer>

Figure 21. Visualization examples of Reason-RFT and CoT-SFT on Spatial Transformation task.



Initial central-view image



Final right-view image

**Question:** The first image shows the central view of initial state, the second image shows the right view of final state, and the attributes of the initial objects are listed below:

- (0, sphere, small, red, metal) (1, sphere, large, brown, metal)
- (2, cube, large, blue, metal) (3, sphere, small, brown, metal)
- (4, cylinder, small, blue, glass)

What is the transformation process that can achieve the change from initial state to final state?

#### COT-SFT:

<think>Let's analyze the initial and final states of the objects in the images. We have the initial attributes of the objects, and we need to determine the transformations that occurred.

1. \*\*Object 0\*\*: Initially a small red metal sphere. In the final image, it appears to have moved to the front left. So, we apply `change\_position(0, front\_left)`.
2. \*\*Object 1\*\*: Initially a large brown metal sphere. In the final image, it seems to have moved behind the other objects. So, we apply `change\_position(1, behind)`.
3. \*\*Object 2\*\*: Initially a large blue metal cube. In the final image, it appears to have changed shape to a sphere. So, we apply `change\_shape(2, sphere)`.
4. \*\*Object 3\*\*: Initially a small brown metal sphere. In the final image, it appears to have changed color to yellow. So, we apply `change\_color(3, yellow)`.
5. \*\*Object 4\*\*: Initially a small blue glass cylinder. There doesn't seem to be any change in its attributes or position in final image.

Now, let's compile the transformations in the order they were identified:

- `change\_color(3, yellow)`
- `change\_position(0, front\_left)`
- `change\_shape(2, sphere)`
- `change\_position(1, behind)`

These transformations match the original answer provided.</think>

<answer>change\_color(3, yellow), change\_position(0, front\_left), change\_shape(2, sphere), change\_position(1, behind)</answer>

#### Reason-RFT:

<think>Let's analyze the initial and final states to determine the transformations needed.

1. \*\*Object 3 (sphere, small, brown, metal):\*\*  
  - In the initial state, it is a small brown metal sphere.
  - In the final state, it appears to be a small yellow cylinder.
  - Therefore, the transformations needed are: - Change its color from brown to yellow: `change\_color(3, yellow)` - Change its shape from a sphere to a cylinder: `change\_shape(3, cylinder)`
2. \*\*Object 1 (sphere, large, brown, metal):\*\*  
  - In the initial state, it is a large brown metal sphere.
  - In the final state, it has moved to the right.
  - Therefore, the transformation needed is: - Change its position to the right: `change\_position(1, right)`
3. \*\*Object 2 (cube, large, blue, metal):\*\*  
  - In the initial state, it is a large blue metal cube.
  - In the final state, it appears to be a large blue metal sphere.
  - Therefore, the transformation needed is: - Change its shape from a cube to a sphere: `change\_shape(2, sphere)`

By applying these transformations, we can achieve the change from the initial state to the final state. The transformations are:

- `change\_color(3, yellow)`
- `change\_shape(3, cylinder)`
- `change\_position(1, right)`
- `change\_shape(2, sphere)`

These transformations match the original answer provided.</think>

<answer>change\_color(3, yellow), change\_shape(3, cylinder), change\_position(1, right), change\_shape(2, sphere)</answer>

Figure 22. Visualization examples of Reason-RFT and CoT-SFT on Spatial Transformation task.