
MIRAGE: Assessing Hallucination in Multimodal Reasoning Chains of MLLM

Bowen Dong^{1,2} Minheng Ni^{1,2} Zitong Huang¹ Guanglei Yang^{1✉} Wangmeng Zuo¹ Lei Zhang²

¹Harbin Institute of Technology ²The Hong Kong Polytechnic University
cndongsky@gmail.com kodenii@outlook.com cslzhang@comp.polyu.edu.hk yangguanglei@hit.edu.cn

Abstract

Multimodal hallucination in multimodal large language models (MLLMs) restricts the correctness of MLLMs. However, multimodal hallucinations are multi-sourced and arise from diverse causes. Existing benchmarks fail to adequately distinguish between perception-induced hallucinations and reasoning-induced hallucinations. This failure constitutes a significant issue and hinders the diagnosis of multimodal reasoning failures within MLLMs. To address this, we propose the MIRAGE benchmark, which isolates reasoning hallucinations by constructing questions where input images are correctly perceived by MLLMs yet reasoning errors persist. MIRAGE introduces multi-granular evaluation metrics: accuracy, factuality, and LLMs hallucination score for hallucination quantification. Our analysis reveals that (1) the model scale, data scale, and training stages significantly affect the degree of logical, fabrication, and factual hallucinations; (2) current MLLMs show no effective improvement on spatial hallucinations caused by misinterpreted spatial relationships, indicating their limited visual reasoning capabilities; and (3) question types correlate with distinct hallucination patterns, highlighting targeted challenges and potential mitigation strategies. To address these challenges, we propose Logos, a method that combines curriculum reinforcement fine-tuning to encourage models to generate logic-consistent reasoning chains by stepwise reducing learning difficulty, and collaborative hint inference to reduce reasoning complexity. Logos establishes a baseline on MIRAGE, and reduces the logical hallucinations in original base models. Link: <https://bit.ly/25mirage>.

1 Introduction

Multimodal large language models (MLLMs) [23, 56, 65] achieve advancements in multimodal perception [38, 5, 61], as evidenced by standard benchmarks [60, 14, 45, 42, 44]. Recent studies further enhance their reasoning capacities through post-training [57, 12, 24, 62, 86]. However, two critical challenges remain, *i.e.*, erroneous visual perception that fabricates non-existent content, and defective logical reasoning yielding inconsistent conclusions. These multi-source hallucinations (stemming from distinct perceptual and cognitive origins) fundamentally limit the practical utility.

To quantitatively measure hallucination in MLLMs, several multimodal benchmarks have been applied to detect and measure multimodal hallucination in object recognition [14, 34, 59, 20, 30] or academic reasoning [41] aspects. Existing benchmarks [14, 34, 59, 20, 30, 41] attempt to measure hallucinations via object recognition or academic tasks. However, two critical gaps remain. First, current evaluations fail to distinguish between different types of hallucinations, *i.e.*, perception-induced hallucinations caused by inaccurate visual understanding and reasoning-induced hallucinations stemming from logical flaws, making it difficult to pinpoint errors. Second, most benchmarks focus on validating the content of answers or intermediate steps, while lacking fine-grained evaluation of the reasoning process in terms of perception and logic, thereby hindering the ability to trace error

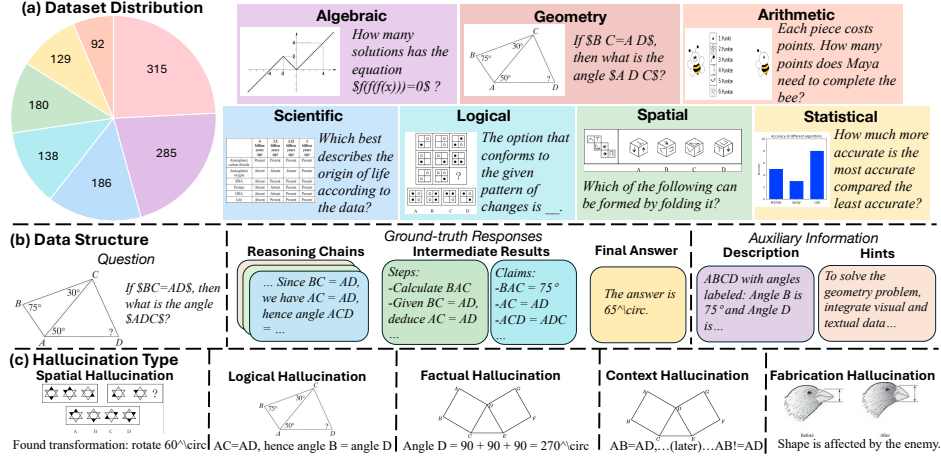


Figure 1: Data distribution and structure of MIRAGE benchmark. (a) shows the classes and amounts of questions. (b) shows the data structure of each question in MIRAGE. And (c) shows the multimodal reasoning hallucination types we explored, which are clearly defined in Appendix A.

propagation patterns. This absence of hierarchical analysis spanning answer-level outputs, step-level intermediate results, and thought-level reasoning logic prevents systematic diagnosis of reasoning failures. Addressing these gaps is essential for building trustworthy MLLMs.

To address these challenges, we propose MIRAGE, a diagnostic benchmark specifically designed to isolate reasoning-induced hallucinations in MLLMs. As shown in Fig. 1, MIRAGE contains 1,329 questions where MLLMs demonstrate accurate visual perception but exhibit defective reasoning. Each question provides three-tier annotations: final answers, intermediate reasoning steps and claims, and ground-truth reasoning chains, enabling precise tracking of hallucination propagation in multimodal reasoning processes. To comprehensively assess reasoning hallucination in MLLMs, MIRAGE proposes three evaluation metrics, *i.e.*, accuracy assessment measuring overall answer correctness, factuality assessment verifying the correctness in intermediate steps and claims, and LLMs Hallucination Score to assess hallucination from the whole reasoning chain level. By evaluating MLLMs on MIRAGE from different levels, we aim to answer three critical research questions. First, how reasoning hallucinations compromise MLLM robustness and correlate with answer accuracy. Second, whether specific question types induce distinct hallucination patterns unique to multimodal reasoning. Finally, the efficacy of current mitigation methods against reasoning-specific hallucinations.

We conduct extensive experiments on MIRAGE, leading to several key insights. First, the model scale, data scale, and training stages of MLLMs significantly influence the severity of logical, fabrication, and factual hallucinations. Second, these factors offer limited improvements in addressing spatial hallucinations, which are primarily caused by misinterpretations of spatial relationships—highlighting the limited visual reasoning capabilities of current MLLMs and their inability to benefit from straight-forward scaling. Third, we observe strong correlations between question types and specific patterns of reasoning hallucination, underscoring critical challenges and suggesting targeted mitigation strategies. These findings offer valuable guidance for the future development of more reliable and reasoning-aware MLLMs.

Building on the insight that increasing the probability of logic-consistent reasoning chains reduces specific logical hallucinations, we propose Logos, which integrates curriculum reinforcement fine-tuning (CRFT) for training and collaborative hint inference for testing. During training, CRFT with online reward filtration (ORF) gradually increases question difficulty while dynamically selecting high-reward samples, guiding the model toward accurate and logic-consistent reasoning. At the testing stage, collaborative hint inference provides topic- and question-specific hints from LLMs, reducing reasoning complexity for optimized models. Experiments demonstrate that Logos significantly reduces reasoning hallucinations and achieves strong performance on both MIRAGE and standard benchmarks [41]. In conclusion, the contribution of this paper can be summarized as follows:

- We propose MIRAGE, the first benchmark for evaluating multimodal reasoning hallucinations in MLLMs. It isolates reasoning hallucinations with tasks where inputs are correctly

Table 1: Comparison of MIRAGE with existing benchmarks. “MCQ” means multiple-choice questions, “A” means answers, “D” means multimodal input descriptions, “R” means full reasoning chains, and “S” means intermediate results. * means multiple reasoning chains.

| Benchmarks | Dataset Properties | | | | Hallucination Assessment | | Usage |
|----------------------|--------------------|--------------|-----------------|--------------------|--------------------------|--------|---------------------|
| | Taxonomy | Scale | Annotation | Intermediate | Steps | Chains | |
| POPE [34] | Object | 18K | A | - | ✗ | ✗ | Object Hallu |
| MMVP [59] | Object | 300 | A | - | ✗ | ✗ | MCQ |
| HallusionBench [20] | Object | 1,129 | A | - | ✗ | ✗ | Illusion |
| MME [14] | Various | 2,374 | A | - | ✗ | ✗ | General VQA |
| SEEDBench [30] | Obj+Act | 19K | A | - | ✗ | ✗ | MCQ |
| MathVista [41] | Math | 1,000 | A | - | ✗ | ✗ | Math Reasoning |
| OmniBench [35] | Various | 1,142 | A/D | - | ✗ | ✗ | MCQ |
| MME-CoT [27] | Various | 1,130 | A/D/S | Steps | ✓ | ✗ | General CoT |
| MIRAGE (Ours) | Various | 1,329 | A/D/R*/S | Steps+Hints | ✓ | ✓ | Reason Hallu |

perceived but reasoning errors persist, and introduces multi-level metrics for comprehensive assessment: accuracy, factuality, and LLMs hallucination score.

- Our findings reveal that the model scale, data scale, and training stages of MLLMs, and highlight critical challenges and mitigation for specific hallucination types. These findings will provide insights for future MLLM development.
- We propose Logos, a baseline method of MIRAGE to encourage model for logic-consistent reasoning via curriculum reinforcement fine-tuning and collaborative hint inference. Logos reduces the multimodal logical hallucination and improves the answer accuracy.

2 Related Work

Reasoning Multimodal Large Language Models. Reasoning MLLMs can be roughly divided into three groups. First is the prompt-based reasoning method [62, 83, 47] to guide MLLMs by in-context learning [11]. Second is the plan-based method [69, 85, 78], which uses searching methods [55, 82] to explore optimal reasoning chains. And last is learning-based method by supervised fine-tuning (SFT) [71] or reinforcement learning (RL) [53, 39, 46]. RL methods generalize better by optimizing with their high-reward predictions instead of fixed ground-truths. Hence, we build Logos on RL for hallucination mitigation, uniquely focusing on dynamic training difficulty adjustment.

Multimodal Hallucination Evaluation. Existing MLLMs still suffer from multimodal hallucination, where generated text either contradicts the visual input or deviates from correct logical reasoning. To assess the hallucination and its effect in MLLMs, recent works measure the accuracy degradation among object perception [34], illusion [20, 64], mathematic [41, 60, 80], IQ test [4, 7, 17], and general multimodal abilities [30, 14]. While existing benchmarks have advanced multimodal evaluation, they often conflate perception-induced hallucinations with reasoning-induced ones, making it challenging to diagnose reasoning failures. In contrast, MIRAGE focuses on reasoning hallucinations by isolating reasoning errors from correctly perceived inputs, providing multi-level metrics for assessment.

Multimodal Hallucination Mitigation. Hallucination evaluations provide critical insights for future mitigation strategies, which broadly fall into training-free and training-based approaches. First is training-free methods [86, 16, 73]. They use prompting [66], question decomposition [86], reasoning chain reflection [16, 25], and ensembling [63] to mitigate perception-induced hallucination. Second is training-based mitigation methods [43, 71, 12, 70, 79, 75]. These methods mitigate hallucination by data curation [75, 37], SFT [51, 3, 26, 6], and RL [84, 88, 77]. Previous hallucination mitigation methods mainly target perception errors in MLLMs, with only a few early efforts addressing multimodal reasoning hallucinations [86, 43]. Our proposed Logos focuses on reasoning hallucinations and shows that combining CRFT with CHI effectively reduces hallucinations beyond prior methods.

3 MIRAGE Dataset

3.1 Data Construction

To evaluate reasoning hallucinations, as in Table 1, we present MIRAGE, emphasizing tasks with accurate perception but challenging reasoning. MIRAGE offers multi-level annotations and rich auxiliary data for error diagnosis. As in Fig. 2, the construction involves data collection and curation.

Figure 2: The construction and evaluation of MIRAGE. (a) shows the construction of MIRAGE. And (b) shows multi-granular evaluation metrics: accuracy, factuality, and LLMs hallucination score.

Data Collection. To systematically evaluate multimodal reasoning capacities across diverse cognitive dimensions, the MIRAGE is constructed through rigorous selection of seven distinct taxonomies, including geometry, algebraic, arithmetic, scientific, spatial reasoning, and statistical reasoning. Based on above taxonomies, we collect the original benchmark data from both publicly available datasets and questions from Internet. Finally, the size of original dataset is roughly 18K.

Data Curation. To ensure MIRAGE isolates reasoning hallucinations and comprehensively evaluate each topic, we apply a two-step curation process, *i.e.*, difficulty curation and balance curation. For difficulty curation, we use three open-source MLLMs [2, 5, 18] to generate image descriptions, retaining only questions where these descriptions are consistently accurate (verified by a secondary LLM [36]) but lead to frequent reasoning errors, aligning with our benchmark focus. For balance curation, we sample the resulting data to ensure a balanced distribution [10, 28, 33] across seven topics, maintaining a small imbalanced rate, and resulting in a final dataset of 1,329 questions.

3.2 Data Annotation and Verification

Reasoning Chain Annotation. To address the lack of ground-truth reasoning chains, we propose a cost-effective automated annotation framework that optimizes both computational efficiency and output quality. Our approach follows a two-stage refinement process. Firstly, we generate initial reasoning chains using the lightweight O3-mini [24]. And then we refine these chains with a strong LLM [21], guided by known ground-truths. We will discuss the annotation cost in appendix.

Collaborative Verification. Next, we conduct annotation verification to ensure the correctness. To improve both verification speed and accuracy, we introduce a human-AI collaborative verification framework. Specifically, each question is independently assessed by a human expert and an MLLM [65] for potential hallucinations in the reasoning chain. If both assessments are accurate, the reasoning chain is retained as the ground-truth. In cases of discrepancies, the human expert either guides the MLLM to correct the reasoning chains or manually provides reasoning steps if the MLLM remains inaccurate. Finally, all newly annotated chains undergo cross-checking by other experts.

Step and Claim Extraction. Finally, given the verified reasoning chain and final answer for each question, we use a state-of-the-art LLM [36] to extract critical intermediate steps and claims via in-context learning. Specifically, for each ground-truth reasoning chain \hat{y} , we use hand-crafted few-shot prompts to guide the LLM in selecting \hat{K}_s representative reasoning steps $\hat{S} = \hat{s}_1, \dots, \hat{s}_{K_s}$ and \hat{K}_c intermediate claims $\hat{C} = \hat{c}_1, \dots, \hat{c}_{K_c}$. To ensure reliability, we limit $1 \leq K_s \leq 10$ and $1 \leq K_c \leq 10$, preventing over-detailed and unreliable outputs. The final intermediate steps and claims are then parsed using regular expressions for fine-grained reasoning hallucination evaluation.

Auxiliary Information Annotation. MIRAGE also uses MLLM [23] to annotate image descriptions and hints. This information is verified by experts and can help researchers to diagnose hallucinations.

4 MIRAGE Benchmark Evaluation

4.1 Accuracy Assessment

The accuracy is a fundamental metric since incorrect **final answers** often indicate reasoning chain hallucinations [86, 50]. To accommodate different question types, MIRAGE parses the final predictions and matches parsed answers \mathbf{A}_{pred} with ground-truths \mathbf{A}_{gt} for both multiple-choice and deterministic free-form answers. For questions with approximate answers (*e.g.*, statistical questions without precise annotations on charts), MIRAGE calculates the relative error between predictions and ground-truths, considering answers correct if the error falls below a threshold.

4.2 Factuality Assessment

Step and Claim Factuality Evaluation. For each predicted reasoning chain y , MIRAGE follows the extraction pipeline in Sec. 3.2 to extract **intermediate** steps $\mathbb{S} = \{s_1, \dots, s_{K_s}\}$ and claims $\mathbb{C} = \{c_1, \dots, c_{K_c}\}$. With given corresponding ground-truth intermediate steps $\hat{\mathbb{S}}$ and claims $\hat{\mathbb{C}}$, we utilize an LLM [36] and use $\{\mathbb{S}, \hat{\mathbb{S}}\}$ as input, and guide the LLM to detect whether a predicted step s_i

is covered in $\hat{\mathbb{S}}$ and whether a ground-truth step \hat{s}_i is stated in \mathbb{S} , and then predicts the binary matching results $\mathbf{M}_{s,\text{pred}}$ and $\mathbf{M}_{s,\text{gt}}$. By this formulation, MIRAGE can efficiently match free-form steps and claims for flexible factuality evaluation. Finally, we calculate the step factuality score F_{step} by:

$$F_{\text{step}} = \frac{2 \times \text{Precision}_s \times \text{Recall}_s}{\text{Precision}_s + \text{Recall}_s}, \quad (1)$$

where $\text{Precision} = \frac{|\mathbf{M}_{s,\text{pred}=1}|}{|\mathbf{M}_{s,\text{pred}}|}$ means correctly answered steps and $\text{Recall} = \frac{|\mathbf{M}_{s,\text{gt}=1}|}{|\mathbf{M}_{s,\text{gt}}|}$ means correctly matched ground-truth steps. Similarly, the claim factuality score F_{claim} is defined by:

$$F_{\text{claim}} = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \quad (2)$$

Hallucination Type Detection. Besides, to qualitatively assess which kind of reasoning hallucination does an MLLM suffer in a specific question, we first clearly define five reasoning hallucination types, including *logical*, *factual*, *spatial*, *context*, and *fabrication* hallucinations. The detailed definitions of each type are provided in App. A. Then, we follow LLM-as-a-Judge [15, 19] and introduce an LLM-based hallucination detector. Specifically, rather than compare two plain reasoning chains directly, MIRAGE detects the reasoning hallucination by examining extracted intermediate steps $\{\mathbb{S}, \hat{\mathbb{S}}\}$ and claims $\{\mathbb{C}, \hat{\mathbb{C}}\}$, then predict the hallucination detection results by in-context learning.

4.3 LLMs Hallucination Score (LHS) Assessment

Finally, we also assess the hallucination from **the whole reasoning chains**. While entropy-based uncertainty estimation methods [13, 81, 87, 58] can identify unreliable reasoning or information-deficient chains, they still face two limitations. First is token-level likelihood dependencies. Existing methods rely on token-level likelihood to quantify uncertainty, but it is inaccessible in black-box MLLMs [56, 23, 65]. And second is high computational cost. Accurate uncertainty assessment typically requires sampling numerous responses per query, escalating evaluation overhead. Therefore, inspired by LLM as judges [32, 15, 19, 31], we propose LLMs Hallucination Score (LHS) to simulate uncertainty estimation via multi-LLMs and multi-reference ensemble. Specifically, we first define multi-dimension scoring rules to measure the hallucination in the whole reasoning chain rather than extracted steps, including factual accuracy, logical consistency, reasoning completeness, conceptual accuracy, and strategy appropriateness. Above dimensions can be used to simulate the uncertainty in responses and formulate the scoring template $\mathbf{h}_{\text{score}}$. Our aim is to predict LHS by M (e.g., 3) LLM judges. To improve the confidence of LHS, MIRAGE leverage an LLM [36] to rewrite $\hat{\mathbf{r}}$ by $N - 1$ variants, thus formulating $N = 3$ reference chains $\{\mathbf{r}_{\text{ref}}^1, \dots, \mathbf{r}_{\text{ref}}^N\}$. Then, for each response $\mathbf{r}_{\text{ref}}^i$ with corresponding ground-truth $\hat{\mathbf{r}}$, both responses are integrated into the template $\mathbf{h}_{\text{score}}$ and then generate the judgement scores $\{s_1^{i,j}, \dots, s_5^{i,j}\}$ by each j -th LLM judge. Finally, the LHS \bar{s} of response r is:

$$\bar{s} = \frac{1}{M} \sum_{j=1}^M \frac{1}{N} \sum_{i=1}^N \text{mean}(\{s_1^{i,j}, \dots, s_5^{i,j}\}). \quad (3)$$

By accumulating responses in MIRAGE, one can calculate the mean and standard deviation of \bar{s} for a specific reasoning MLLM. Generally, lower mean indicates higher uncertainty (i.e., hallucination rate), and lower standard deviation means higher confidence of the LHS. We further conduct consistency checks on LHS using human evaluators. We randomly sample 100 responses from Gemini-2-flash and Qwen2.5-VL-7B, comparing the human evaluation from three experts. The average difference rate is 7.5%, showing the reliability of LHS for measuring reasoning hallucinations.

5 Logos: A Baseline Method of MIRAGE

5.1 Revisit Multimodal Reinforcement Fine-Tuning

As shown in Sec. 6.1, reasoning chains with correct answers generally have lower hallucination rates (e.g., logical hallucination). This suggests that reducing hallucinations in MLLMs can be approached by increasing the generation probability of logic-consistent and correct reasoning chains, aligning inherently with Group Relative Policy Optimization (GRPO). To address this, we propose the baseline method Logos for MIRAGE, leveraging GRPO to optimize MLLM (the policy model π

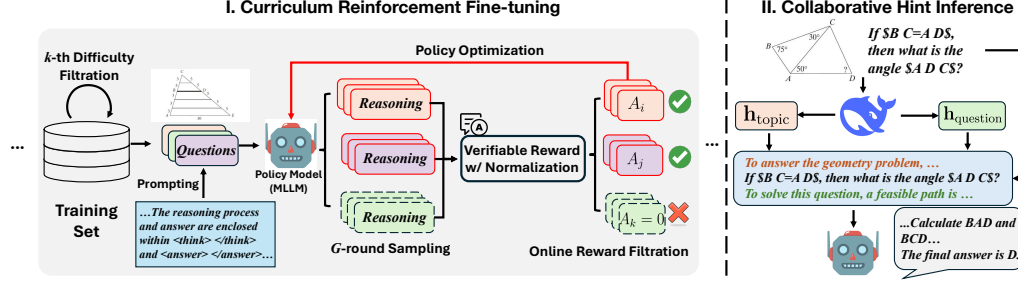


Figure 3: Overview of our baseline Logos. During training, Logos adopts curriculum reinforcement fine-tuning (CRFT) with online reward filtration (ORF) to progressively increase data difficulty and filter low-impact samples. During testing, Logos introduces Collaborative Hint Inference, leveraging LLM-guided hints to simplify the reasoning process. Logos effectively reduces logical hallucination.

with parameter θ). Specifically, we leverage in-context learning [62] to guide π to generate formatted response with “<think>...</think>” and “<answer>...</answer>” blocks, where the former contains the reasoning chain and the latter includes the final answer. Rather than using a separate value model to calculate advantages of responses, GRPO directly samples G different responses $\{\mathbf{r}_1, \dots, \mathbf{r}_G\}$ with given multimodal question \mathbf{x} . To measure the relative advantages $\{A_1, \dots, A_G\}$, we define the reward function \mathcal{R} as format reward \mathcal{R}_{fmt} and accuracy reward \mathcal{R}_{acc} , where the former is a binary function to judge whether the i -th response \mathbf{r}_i follows response format, and the latter is a binary function to judge the correctness of final answer. Then the reward of \mathbf{r}_i is defined by $r_i = \mathcal{R}_{\text{fmt}}(\mathbf{r}_i) + \mathcal{R}_{\text{acc}}(\mathbf{r}_i)$. And the advantage A_i is:

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

Finally, we optimize π via minimizing GRPO loss $\mathcal{L}_{\text{GRPO}}$ with above advantages as follows:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{\{\mathbf{r}_i\}_1^G \sim \pi(\mathbf{x})} \frac{1}{G} \sum_{i=1}^G \frac{1}{|\mathbf{r}_i|} \sum_{t=1}^{|\mathbf{r}_i|} \min(\tilde{r}_{i,t}(\theta) A_i, \text{clip}(\tilde{r}_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon) A_i), \quad (5)$$

$$\text{where } \tilde{r}_{i,t} = \frac{\pi(\mathbf{r}_{i,t} | \mathbf{x}, \mathbf{r}_{i,1:t-1})}{\pi_{\text{old}}(\mathbf{r}_{i,t} | \mathbf{x}, \mathbf{r}_{i,1:t-1})}.$$

We remove the KL-divergence term since reasoning models have a non-negligible distribution gap with base models [49]. We will investigate the effect in the App. D.

5.2 Curriculum Reinforcement Fine-Tuning

Note that if responses from a specific training sample are all correct or all incorrect, the advantage of each response is 0, which is harmful for GRPO optimization. To reduce the difficulty and improve the training efficiency, we propose curriculum reinforcement fine-tuning (CRFT). Specifically, before the optimization, we first leverage π to conduct G -round sampling, and calculate average accuracy reward $\bar{r}_{\text{acc}} = \text{mean}(\{\mathcal{R}(\mathbf{r}_1), \dots, \mathcal{R}(\mathbf{r}_G)\})$. During the first stage, we keep questions with $\bar{r}_{\text{acc}} > 0$ to ensure that π can sample at least one reasoning chain with correct answer and logic-consistent reasoning during training, thus making the advantages non-zero for smooth optimization. Then, during each k -round ($k > 1$) curriculum training, we repeat G -round sampling and keep questions with $\bar{r}_{\text{acc}} < 0.5$ to ensure π can face more difficult questions during further CRFT. Our experimental results in Sec . 6.2 illustrate the efficiency of CRFT.

5.3 Online Reward Filtration

While CRFT effectively controls data difficulty, it may still encounter training samples where all generated responses receive identical rewards, disrupting the optimization process. To address this without compromising training efficiency, we integrate offline data filtration [74] into our approach, forming Online Reward Filtration (ORF). In each iteration, for given question \mathbf{x} with G sampled responses $\mathbf{r}_1, \dots, \mathbf{r}_G$, Logos first computes the corresponding rewards r_1, \dots, r_G using the predefined reward function \mathcal{R} . If all responses share the same reward ($r_1 = \dots = r_G$), the question is discarded for that iteration, ensuring only diverse, meaningful samples contribute to optimization.

Table 2: Comparison of recent state-of-the-art MLLMs on MIRAGE. Best results are **bolded**.

| Model | Type | Accuracy \uparrow | Factuality | | LHS \uparrow |
|----------------------------------------|-----------|---------------------|----------------------------|-----------------------------|-------------------------------------|
| | | | $F_{\text{step}} \uparrow$ | $F_{\text{claim}} \uparrow$ | |
| Black-Box MLLMs | | | | | |
| Gemini-2-Flash-Thinking [56] | Reasoning | 47.6 | 51.5 | 50.7 | 0.7517\pm0.0168 |
| O1 [24] | Reasoning | 49.7 | 41.3 | 42.7 | 0.6193 \pm 0.0091 |
| Gemini-2-Flash [56] | General | 44.1 | 47.8 | 47.4 | 0.6882 \pm 0.0496 |
| GPT-4o [23] | General | 35.0 | 39.2 | 40.6 | 0.6332 \pm 0.0111 |
| Open-sourced \sim 72B MLLMs | | | | | |
| Qwen2.5-VL-72B-Instruct [2] | General | 38.8 | 47.4 | 44.6 | 0.7223\pm0.0339 |
| InternVL-2.5-78B [5] | General | 29.6 | 39.0 | 36.6 | 0.6377 \pm 0.0325 |
| Qwen2-VL-72B-Instruct [61] | General | 24.5 | 29.7 | 26.2 | 0.4928 \pm 0.0332 |
| QvQ-72B-Preview [57] | Reasoning | 31.0 | 46.1 | 45.3 | 0.5717 \pm 0.0597 |
| Virgo-72B [12] | Reasoning | 37.4 | 47.1 | 45.0 | 0.6328 \pm 0.0251 |
| Open-sourced \sim 7B MLLMs | | | | | |
| Qwen2.5-VL-7B-Instruct [2] | General | 28.8 | 34.7 | 31.7 | 0.5996 \pm 0.0123 |
| Qwen2-VL-7B-Instruct [2] | General | 19.5 | 21.9 | 18.6 | 0.3633 \pm 0.0106 |
| Qwen2.5-VL-7B-Instruct+VIC [86] | Reasoning | 26.9 | 22.8 | 25.2 | 0.4478 \pm 0.0177 |
| Qwen2.5-VL-7B-Instruct+Reflection [16] | Reasoning | 26.7 | 40.1 | 33.4 | 0.5826 \pm 0.0124 |
| R1-OneVision-7B [5] | Reasoning | 22.9 | 30.7 | 30.2 | 0.5098 \pm 0.0099 |
| Mulberry-Qwen2-VL-7B [72] | Reasoning | 22.6 | 29.2 | 24.4 | 0.4740 \pm 0.0147 |
| InternVL-2.5-8B [5] | General | 20.8 | 31.9 | 26.4 | 0.4838 \pm 0.0156 |
| Llama-3.2-Vision-11B [18] | General | 18.7 | 26.9 | 22.3 | 0.4265 \pm 0.0141 |
| Llava-CoT-11B [67] | Reasoning | 17.4 | 26.9 | 22.4 | 0.4267 \pm 0.0140 |
| Logos-7B (Ours) | Reasoning | 37.1 | 43.3 | 38.3 | 0.6568\pm0.0179 |
| Open-sourced \sim 3B MLLMs | | | | | |
| Qwen2.5-VL-3B-Instruct [2] | General | 18.8 | 23.1 | 18.8 | 0.3422 \pm 0.0244 |
| Phi-3.5-Instruct [1] | General | 12.9 | 16.6 | 13.8 | 0.3181 \pm 0.0161 |
| Logos-3B (Ours) | Reasoning | 29.4 | 38.9 | 34.5 | 0.5840\pm0.0216 |

Hallucination Types Distribution Across Question Types

| Question Types | Algebraic | Arithmetic | Geometry | Logical | Scientific | Spatial | Statistical |
|----------------|-----------|------------|----------|---------|------------|---------|-------------|
| Logical | 34.7% | 32.2% | 32.3% | 35.4% | 36.0% | 25.5% | 26.0% |
| Spatial | 5.6% | 15.0% | 20.2% | 22.1% | 9.0% | 24.9% | 14.3% |
| Factuality | 24.9% | 23.1% | 20.1% | 15.0% | 26.6% | 22.4% | 33.6% |
| Context | 19.7% | 16.1% | 16.2% | 16.0% | 17.4% | 14.0% | 14.3% |
| Fabrication | 15.2% | 13.6% | 11.2% | 11.4% | 11.0% | 13.2% | 11.7% |

Pearson Correlation Between Hallucination Types

| | Logical | Spatial | Factuality | Context | Fabrication |
|-------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Logical | 1.000*** p=0.000 | 0.635 p=0.125 | 0.886** p=0.008 | 0.986*** p=0.000 | 0.913** p=0.004 |
| Spatial | 0.635 p=0.125 | 1.000*** p=0.000 | 0.541 p=0.210 | 0.595 p=0.159 | 0.631 p=0.128 |
| Factuality | 0.886** p=0.008 | 0.541 p=0.210 | 1.000*** p=0.000 | 0.927** p=0.003 | 0.917** p=0.004 |
| Context | 0.986*** p=0.000 | 0.595 p=0.159 | 0.927** p=0.003 | 1.000*** p=0.000 | 0.960*** p=0.001 |
| Fabrication | 0.913** p=0.004 | 0.631 p=0.128 | 0.917** p=0.004 | 0.960*** p=0.001 | 1.000*** p=0.000 |

*p<0.05, **p<0.01, ***p<0.001

Figure 4: Distribution between question types and reasoning hallucination types.

Figure 5: Pearson correlation among reasoning hallucination types.

5.4 Collaborative Hint Inference

To further reduce reasoning hallucinations beyond training, we introduce Collaborative Hint Inference (CHI), which leverages an auxiliary LLM ϕ to provide context-specific guidance. Given a question \mathbf{x} , CHI first uses a predefined question classification prompt to guide ϕ in predicting the question type \mathbf{c} . Based on this type, CHI generates two structured hints: a topic-specific hint $\mathbf{h}_{\text{topic}}$, reflecting the general approach for the given type \mathbf{c} , and a question-specific hint $\mathbf{h}_{\text{question}}$, tailored to the particular content of \mathbf{x} . During inference, we generate response by $\mathbf{r} = \pi([\mathbf{h}_{\text{topic}}, \mathbf{x}, \mathbf{h}_{\text{question}}])$. The optimized MLLM can benefit from CHI and generate more accurate chains than vanilla MLLMs.

6 Experiments

We state the full experimental setup of MIRAGE evaluation and training of Logos in App. B. And in the following, we illustrate the insightful findings on MIRAGE and the effectiveness of Logos.

6.1 Empirical Analysis

Overall results. As shown in Table 2, O1 achieves the highest accuracy at 49.7, outperforming Gemini-2-flash-thinking by 2.1. However, O1 scores lower in factuality, likely due to generating

Table 3: Probability of different types of reasoning hallucinations for each model.

| Model | Logical | Spatial | Factuality | Context | Fabrication |
|-------------------------|---------|---------|------------|---------|-------------|
| Gemini-2-flash | 54.66% | 29.33% | 39.03% | 28.48% | 22.32% |
| Qwen2.5-VL-7B | 68.94% | 35.35% | 51.07% | 30.98% | 23.91% |
| Gemini-2-flash-thinking | 47.88% | 25.44% | 32.85% | 22.51% | 18.04% |
| Virgo-72B | 63.97% | 29.18% | 40.65% | 32.26% | 21.32% |
| QvQ-72B-Preview | 73.37% | 37.93% | 47.91% | 47.52% | 30.19% |

Table 4: Manually fixing reasoning chains experimental results on 10% sampled questions with reasoning hallucination.

| Model | Fix Reasoning | Accuracy |
|----------------|---------------|----------|
| GPT-4o | - | 12.1 |
| GPT-4o | ✓ | 68.5 |
| Qwen2.5-VL-72B | - | 10.4 |
| Qwen2.5-VL-72B | ✓ | 72.4 |

Table 5: Correlation matrix of all metrics. All correlations are significant at $p < 0.001$ (***).

| Metric | Accuracy | F_{step} | F_{claim} | LHS |
|--------------------|----------|-------------------|--------------------|----------|
| Accuracy | 1.000 | 0.864*** | 0.918*** | 0.889*** |
| F_{step} | 0.864*** | 1.000 | 0.975*** | 0.915*** |
| F_{claim} | 0.918*** | 0.975*** | 1.000 | 0.933*** |
| LHS | 0.889*** | 0.915*** | 0.933*** | 1.000 |

shorter, more summarized reasoning chains that reduce step coverage and recall. For most models except GPT-4o and O1, step scores exceed claim scores, suggesting that generating coarse reasoning steps is easier than detailed calculation steps. The LHS scores align with step scores, confirming the reliability of MIRAGE’s metrics. Focusing on open-source Qwen-VL models, increasing parameters from 3B to 72B raises accuracy from 18.8 to 38.8. Moreover, better pretraining in Qwen2.5-VL improves both accuracy and factuality/LHS, indicating that enhanced pretraining reduces hallucinations.

Consistency among evaluation metrics. As shown in Table 5, we compute pearson correlation coefficients between *Accuracy*, F_{step} , F_{claim} , and LHS across all models. All pairs of metrics exhibit very strong positive correlations ($r = 0.86\text{--}0.98$), with all correlations being highly significant ($p < 0.001$). These results indicate that the hallucination rate in the reasoning chains has large correlation with final answers, and inspire our hallucination mitigation method.

Correlation between reasoning hallucination and final accuracy. Given correlations between reasoning chains and answer accuracy, we conduct preliminary study to show the impact of hallucination mitigation. We manually corrected the reasoning chains for about 10% of commonly misanswered questions by GPT-4o and Qwen2.5-VL-72B, then prompt the models to reconsider their final answers. As shown in Table 4, this correction significantly improves the answer accuracy to around 70%, confirming that reducing reasoning hallucinations directly enhances overall model performance.

Relation between pretraining data and hallucination types. We also explore relations between pretraining data and hallucination types. Specifically, we keep use Qwen-VL [61, 2] with different pretraining data (*i.e.*, Qwen2-VL and Qwen2.5-VL) and compare the hallucination rates of each hallucination type. As shown in Table 6, Qwen2.5-VL models have less logical, factual, and fabrication hallucination rates than those of Qwen2-VL models. A possible explanation is that pretraining data with higher quality provides more accurate factual knowledge and reasoning chains to models, such that models can avoid logical and factuality hallucinations during inference. Nevertheless, the spatial hallucination does not significantly reduced, which indicates that current MLLMs still show weak visual reasoning capabilities.

Relation between model size and hallucination types. We also explore relations between pretraining data and hallucination types. Specifically, we keep use Qwen2.5-VL [2] with different model sizes (*i.e.*, 3B/7B/72B) and compare the hallucination rates of each hallucination type. As shown in Table 7, Larger Qwen2.5-VL models have less logical, factual, and fabrication hallucination rates than those of smaller models. A possible explanation is that models owning more model parameters have more capabilities for accurate factual knowledge and reasoning chains to models, such that models can avoid logical and factuality hallucinations during inference. Nevertheless, the spatial hallucination does not significantly reduced, which indicates that current MLLMs still show weak visual reasoning capabilities.

Correlation between question and hallucination types. We also analyze the relationship between question types and hallucination patterns, as shown in Fig. 4. Results indicate that logical hallucinations are widespread across various question types, while certain hallucination types are more closely associated with specific question types. Specifically, logical and spatial hallucinations are particularly common in logical questions, reflecting the high demands for complex reasoning and visual transformations that current MLLMs struggle with [4]. In contrast, statistical and scientific questions tend

Table 6: Hallucination type rates in MIRAGE questions of Qwen-7/72B with different pertaining data. Pretraining with higher quality data leads to less logical, fabrication, and factual hallucinations.

| Model | Logical | Factuality | Spatial | Context | Fabrication |
|----------------|---------|------------|---------|---------|-------------|
| Qwen2.5-VL-72B | 47.7% | 33.7% | 29.2% | 21.6% | 16.5% |
| Qwen2-VL-72B | 59.3% | 45.4% | 32.7% | 32.6% | 26.5% |
| Qwen2.5-VL-7B | 64.7% | 45.7% | 33.4% | 29.3% | 25.5% |
| Qwen2-VL-7B | 74.0% | 60.6% | 35.6% | 42.7% | 35.4% |

Table 7: Hallucination type rates in MIRAGE benchmark questions of Qwen-2.5-VL. Larger Models lead to less logical, fabrication, and factual hallucinations.

| Model | Logical | Factuality | Spatial | Context | Fabrication |
|----------------|---------|------------|---------|---------|-------------|
| Qwen2.5-VL-72B | 47.7% | 33.7% | 29.2% | 21.6% | 16.5% |
| Qwen2.5-VL-7B | 64.7% | 45.7% | 33.4% | 29.3% | 25.5% |
| Qwen2.5-VL-3B | 78.9% | 60.1% | 36.7% | 37.9% | 38.1% |

to exhibit more factuality hallucinations, likely due to their reliance on precise knowledge retrieval. These findings highlight the specific vulnerabilities of MLLMs in handling diverse reasoning tasks.

Correlation among hallucination types. We further analyze correlations among hallucination types using pearson coefficients. As shown in Fig. 5, logical hallucinations strongly correlate with factuality, context, and fabrication errors, likely because flawed logic often leads to context inconsistency and factual errors. Notably, spatial hallucinations, which arise from complex visual operations, show relatively low correlation with other hallucinations, suggesting they are more independent and unique to multimodal models rather than text-based LLMs. These findings highlight the need for targeted mitigation strategies for hallucination types, particularly for challenging spatial reasoning errors.

Hallucination rate comparison across models. To quantitatively assess the impact of model design and training on reasoning hallucinations, we analyzed five representative MLLMs, as shown in Table 3. QvQ-72B-Preview exhibits the highest overall hallucination rates, especially in Logical (73.37%) and Context (47.52%) categories, significantly higher than Virgo-72B, which shares the same base model but benefits from more effective fine-tuning. In contrast, Gemini-2-flash-thinking consistently shows the lowest hallucination rates, particularly in Logical (47.88%), Spatial (25.44%), and Fabrication (18.04%) categories, indicating superior robustness.

Existing solutions are not sufficient to mitigate hallucination. Training-free methods like self-reflection [16] and visual inference chain [86] generally degrade both accuracy and LHS on base models without sufficient reasoning capabilities (Table 2), highlighting their limitations. Similarly, SFT-based methods can improve hallucination mitigation on larger models (e.g., Virgo-72B) but often fail to enhance smaller models, suggesting that model capacity plays a critical role in the effectiveness of external supervision. More detailed analysis can be found in App. D. These findings motivate our integration of CHI into CRFT-optimized MLLMs rather than directly applying it to base models.

6.2 Empirical Analysis of Logos

To quantitatively investigate the component of our baseline method Logos, we use Logos-7B as an example and conduct an ablation study on both MIRAGE and a standard benchmark, MathVista [41]. More in-depth analysis can be found in App. D.

Comparison with previous methods. We compare Logos-7B with other 7B models to validate its effectiveness. Compared to the base model Qwen2.5-VL-7B[2], Logos-7B achieves an 8.3 gain in accuracy, and outperforms the base by 8.6 on F_{step} and 6.6 on F_{claim} , approaching the performance of the larger Virgo-72B[12]. These results, consistent with LHS scores, indicate that Logos effectively reduces reasoning hallucinations, improving reliability across reasoning chains. Similar gains are also observed for Logos-3B, showing the compatibility of our framework across different model scales.

Whether Logos reduces reasoning hallucination or not. Finally we investigate the hallucination mitigate effect on each hallucination type. As shown in Table 9, Logos-7B reduces logical hallucination by 15.4% and fabrication hallucination by 10%. Similar results can also be found in Logos-3B. Nevertheless, we do not find significant hallucination mitigation on spatial and factuality hallucination on both Logos models. A possible reason is that reinforcement learning does not introduce new knowledge and only refines the logic of reasoning chains.

Table 8: Ablation study of Logos-7B, where CHI means collaborative hint inference.

| Method | GRPO | CRFT | h_{topic} | h_{question} | Accuracy | F_{step} | F_{claim} | LHS | MathVista |
|--------------------------|------|------|--------------------|-----------------------|-------------|-------------------|--------------------|---------------|-------------|
| Qwen2.5-VL-7B | ✗ | ✗ | ✗ | ✗ | 28.8 | 34.7 | 31.7 | 0.5996 | 68.2 |
| +GRPO | ✓ | ✗ | ✗ | ✗ | 33.7 | 41.0 | 35.9 | 0.6180 | 70.7 |
| +CRFT | ✓ | ✓ | ✗ | ✗ | 35.7 | 41.8 | 37.3 | 0.6193 | 71.9 |
| + h_{topic} | ✓ | ✓ | ✓ | ✗ | 36.2 | 42.6 | 37.6 | 0.6335 | 72.2 |
| + h_{question} | ✓ | ✓ | ✗ | ✓ | 36.5 | 42.2 | 37.6 | 0.6224 | 72.2 |
| +full CHI | ✗ | ✗ | ✓ | ✓ | 37.1 | 43.3 | 38.3 | 0.6568 | 72.3 |
| Qwen2.5-VL-7B + full CHI | ✓ | ✓ | ✓ | ✓ | 29.0 | 34.9 | 32.1 | 0.6011 | 68.3 |

Table 9: Hallucination type rates in MIRAGE benchmark questions of Qwen2.5-VL-3B/7B and corresponding Logos-3B/7B. Our proposed method leads to less logical and fabrication hallucinations.

| Model | Logical | Factual | Spatial | Context | Fabrication |
|---------------|---------|---------|---------|---------|-------------|
| Qwen2.5-VL-7B | 64.7% | 45.7% | 33.4% | 29.3% | 25.5% |
| Logos-7B | 49.3% | 39.7% | 29.9% | 23.8% | 15.6% |
| Qwen2.5-VL-3B | 78.9% | 60.1% | 36.7% | 37.9% | 38.1% |
| Logos-3B | 57.1% | 47.4% | 36.7% | 31.8% | 24.0% |

Ablation Study of Each Component of Logos. We first investigate the effect of each key component in Logos. The experimental results are shown in Table 8. After adopting reinforcement learning on the base model, the accuracy on MIRAGE and MathVista achieves 33.7 and 70.7 respectively. Benefiting from RL, the step score and claim score also increase to 41.0 and 35.9. After adopting CRFT, both accuracy and F_{claim} further increases to 35.7 and 37.3 respectively. By further integrating CHI, Logos-7B achieves 37.1 on MIRAGE and 72.3 on Mathvista. Note that directly adopt CHI on base model does not lead to performance improvement, which further proves the findings in Sec. 6.1.

Effect of different hints in CHI. The primary role of h_{question} is to guide the model to identify key entities and starting conditions within the question. It excels at grounding the model’s initial focus but does not provide a complete, step-by-step problem-solving template or structure. In contrast, h_{topic} provides a high-level problem-solving schema or template. It offers a clear, structured formulation for the entire reasoning process, ensuring logical completeness from start to finish. The difference leads to performance gap in different metrics. For CRFT + h_{question} , it has better final accuracy (+0.3) than CRFT + h_{topic} , but the metrics regarding intermediate steps (*i.e.*, F_{step} and LHS) are slightly lower. Logos enhanced by CRFT is already optimized to generate complete and precise reasoning chains. When this strong base model is augmented with h_{topic} , it directly benefits from the clear structural guidance, making it easier to construct a perfectly formed reasoning chain. This synergy explains why + h_{topic} achieves slightly better scores on metrics like F_{step} and LHS compared to + h_{question} .

How Logos mitigates reasoning hallucination? We evaluated the impact of CRFT by comparing the accuracy of Logos-7B on 8-round sampling across training dataset, before and after training. The accuracy increases from 24.8% to 68.3%, indicating that CRFT effectively guides the model to generate correct reasoning chains. Meanwhile, we calculate the “Logical” hallucination rate in Qwen2.5-VL-7B and Logos-7B, which reduces from 57.1 to 49.3. This result shows that CRFT benefits to encourage model learning logic-consistent reasoning chains to mitigate hallucination.

7 Conclusion

We propose MIRAGE, which isolates reasoning hallucinations by questions where inputs are correctly perceived but reasoning errors persist. For analysis of reasoning hallucination, MIRAGE proposes multi-level evaluation metrics, covering different levels of the reasoning chains. Our findings reveal that the model scale, data scale, and training stages of MLLMs: (1) significantly influence the degree of logical, fabrication, and factual hallucinations; (2) show no effective improvement on spatial hallucinations caused by misinterpretations of spatial relationships, suggesting that current MLLMs exhibit weak visual reasoning capabilities and struggle to benefit from simple scaling of training resources; and 3) correlations between question types and specific reasoning hallucination patterns, highlighting critical challenges and mitigation for specific types. These findings will provide insights for future MLLM development. To address this, we propose Logos, a method using curriculum reinforcement fine-tuning and collaborative hint inference to reduce logical hallucination for higher accuracy. Logos provides a baseline and offers insights for reducing hallucinations.

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.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibor Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mocha: Multi-objective reinforcement mitigating caption hallucinations. *arXiv preprint arXiv:2312.03631*, 2, 2023.
- [4] Huanqia Cai, Yijun Yang, and Winston Hu. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models. *arXiv preprint arXiv:2502.00698*, 2025.
- [5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.
- [6] Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023.
- [7] Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16259–16273, 2024.
- [8] Jieren Deng, Haojian Zhang, Kun Ding, Jianhua Hu, Xingxuan Zhang, and Yunkuan Wang. Zero-shot generalizable incremental learning for vision-language object detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [9] Bowen Dong, Zitong Huang, Guanglei Yang, Lei Zhang, and Wangmeng Zuo. Mr-gdino: efficient open-world continual object detection. *arXiv preprint arXiv:2412.15979*, 2024.
- [10] Bowen Dong, Pan Zhou, YAN Shuicheng, and Wangmeng Zuo. Lpt: Long-tailed prompt tuning for image classification. In *The Eleventh International Conference on Learning Representations*, 2023.
- [11] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.
- [12] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*, 2025.
- [13] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.
- [14] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- [15] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

- [16] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, and Kamile Lukošiusite. The capacity for moral self-correction in large language models. *Parameters*, 109(1010):1011.
- [17] Deepanway Ghosal, Vernon Toh Yan Han, Yew Ken Chia, , and Soujanya Poria. Are language models puzzle prodigies? algorithmic puzzles unveil serious challenges in multimodal reasoning. *arXiv preprint arXiv:2403.03864*, 2024.
- [18] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [19] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.
- [20] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
- [21] 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.
- [22] Yukang Huo and Hao Tang. When continue learning meets multimodal large language model: A survey. *arXiv preprint arXiv:2503.01887*, 2025.
- [23] 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.
- [24] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [25] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM hallucination via self reflection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore, December 2023. Association for Computational Linguistics.
- [26] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024.
- [27] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.
- [28] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*.
- [29] 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 *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- [30] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

- [31] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.
- [32] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3923–3954, 2024.
- [33] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6929–6938, 2022.
- [34] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- [35] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024.
- [36] 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.
- [37] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.
- [38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- [39] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.
- [40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [41] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.
- [42] Pan Lu, Swaroop Mishra, Tony 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. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- [43] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*, 2025.
- [44] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.
- [45] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, 2022.
- [46] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

- [47] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024.
- [48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32, pages 8024–8035. Curran Associates, Inc., 2019.
- [49] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-rl: Empowering 3b llms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.
- [50] Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. Receval: Evaluating reasoning chains via correctness and informativeness. *arXiv preprint arXiv:2304.10703*, 2023.
- [51] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan O Arik, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In *The Thirteenth International Conference on Learning Representations*.
- [52] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [53] 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.
- [54] Taiwei Shi, Yiyang Wu, Linxin Song, Tianyi Zhou, and Jieyu Zhao. Efficient reinforcement finetuning via adaptive curriculum learning. *arXiv preprint arXiv:2504.05520*, 2025.
- [55] Maciej Świechowski, Konrad Godlewski, Bartosz Sawicki, and Jacek Mańdziuk. Monte carlo tree search: A review of recent modifications and applications. *Artificial Intelligence Review*, 56(3):2497–2562, 2023.
- [56] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [57] Qwen Team. QVQ: To See the World with Wisdom. <https://qwenlm.github.io/blog/qvq-72b-preview/>, 2024.
- [58] Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. Uncertainty-based abstention in llms improves safety and reduces hallucinations. *arXiv preprint arXiv:2404.10960*, 2024.
- [59] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.
- [60] Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- [61] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [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*, 35:24824–24837, 2022.

- [63] Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. *arXiv preprint arXiv:2405.17821*, 2024.
- [64] Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Boyd-Graber, et al. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8395–8419, 2024.
- [65] xAI. Grok 3 Beta — The Age of Reasoning Agents. <https://x.ai/grok>, 2025.
- [66] Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. *Advances in Neural Information Processing Systems*, 37:92012–92035, 2024.
- [67] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024.
- [68] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [69] Jun Cheng Yang, Zuchao Li, Shuai Xie, Wei Yu, Shijun Li, and Bo Du. Soft-prompting with graph-of-thought for multi-modal representation learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15024–15036, 2024.
- [70] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [71] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.
- [72] Huanjin Yao, Jiaying Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.
- [73] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.
- [74] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.
- [75] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024.
- [76] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.
- [77] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.
- [78] Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*, 2024.

- [79] Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pages 196–213. Springer, 2024.
- [80] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.
- [81] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. V1-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.
- [82] Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025.
- [83] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.
- [84] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.
- [85] Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 419–428, 2024.
- [86] Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. *arXiv preprint arXiv:2411.12591*, 2024.
- [87] Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models’ reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, 2024.
- [88] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We have claimed the paper's contributions and scope in both the abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We have discussed the limitation in the appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: The paper does not include theoretical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We have provided all the details (including prompts) in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: <https://github.com/DongSky/MIRAGE>

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have stated the experimental details in both main body of the paper and the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We have noted the statistical significance and state the calculation method in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We list the detailed experiments compute resources in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have read and confirm the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We have stated the broader impacts in the appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: A few samples in the proposed dataset come from existing benchmarks (e.g., MathVista). We have cited related works and we have listed the license of related works in the appendix.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are used as a data preprocessing tool and an evaluation tool. And we also use LLM to revise the grammar of the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Table 10: Categories of multimodal reasoning hallucination investigated in MIRAGE.

| Hallucination Type | Hallucination Description |
|---------------------------|------------------------------------------------------------------------------------------------|
| Spatial Hallucination | Errors in reasoning about spatial relationships, shapes, or complex visual operations. |
| Logical Hallucination | Errors in logical consistency or reasoning, even when surface-level facts are correct. |
| Factuality Hallucination | Factually incorrect claims about scientific principles or established knowledge in input data. |
| Context Hallucination | Inconsistencies between intermediate reasoning steps and final predictions. |
| Fabrication Hallucination | Entirely invented values, entities, or relationships not in input data or real world. |

This appendix mainly contains:

- Hallucination type definition in Section A
- Detailed experimental settings in Section B
- Additional quantitative results in Section C
- Additional in-depth analysis in Section D
- Dataset examples in Section E
- More qualitative results in Section F
- Statement of limitations in Section G
- Statement of broader impact in Section H

A Hallucination Type Definition

In this section, we summarize the hallucination types mentioned in Sec. 4.2, which are listed in Table 10. Specifically, the multimodal reasoning hallucinations can be categorized into five distinct types, *i.e.*, spatial hallucination, logical hallucination, factuality hallucination, context hallucination, and fabrication hallucination. The detailed descriptions are summarized in Table 10.

B Experimental Details of MIRAGE and Logos

B.1 Experimental Setup

Implementation Details. During MIRAGE evaluation, we leverage GPT-4o [23] to judge the accuracy of final answers. And for factuality and LLMs hallucination score metrics, to reduce the cost while keeping comparable evaluation accuracy, we utilize DeepSeek-V3 [36] for both metrics, and utilize Qwen2.5-72B-Instruct [68] as well as Llama-3.1-70B-Instruct [18] for LLMs hallucination score. For Logos-7B, the pretrained model is Qwen-2.5-VL-7B-Instruct; and for Logos-3B, the pretrained model is Qwen-2.5-VL-3B-Instruct. The visual encoder is frozen to avoid catastrophic forgetting of visual perception ability [22, 8, 9]. During training, we collect 13K mathematical questions with K12-level difficulty and ~ 1 K text-only math questions from LIMO [74] as training data. The batch size is 128. For each training sample, the rollout samples G is 8 by default. The initial learning rate is 1×10^{-6} , both warmup strategy and cosine learning rate scheduler are adopted to stabilize training. We optimize Logos by 10 epochs using AdamW [40] during each stage. The number of CRFT stages is set to 1, and we will discuss this choice in Sec. 6.2. Benefiting from the filtration mechanism in CRFT and ORF, the total training time is less than 24 hours, Logos-3B is faster. All programs are constructed by PyTorch [48] toolkit and vLLM [29] framework. All the experiments are conducted on 8 NVIDIA RTX A6000 GPUs.

B.2 Detailed Breakdown of the Initial Data Sources

Our selection of the seven reasoning taxonomies was a principled decision aimed at ensuring comprehensive coverage of both broad, general-purpose reasoning and deep, domain-specific skills. For general reasoning capabilities, to evaluate skills applicable to everyday scenarios, we included tasks requiring spatial reasoning (object locations, attributes) and statistical reasoning (interpretation of charts and plots). For domain-specific reasoning capabilities, to probe more complex, multi-step

logical deductions, we incorporated challenging problems from mathematics (geometry, algebra), science (physics, chemistry, biology), and formal logic (IQ-test style problems). During data collection, our strategy is to curate a diverse dataset from high-quality, specialized sources, each chosen to target a specific reasoning skill:

- **Mathematical Reasoning:** We sourced questions from MathVista and MathVision, which are renowned for their high-quality, complex geometric diagrams and mathematical problems.
- **Logical Reasoning:** To assess formal logic, we drew from the official question banks of national civil service examinations, which feature rigorously designed IQ-style test questions.
- **Scientific Reasoning:** We utilized questions from prestigious science exams (*e.g.*, USA Biology Olympiad), which require domain-specific reasoning in science subjects.
- **Spatial Reasoning:** Our data comes from established traffic-related datasets and the synthetic SuperCLEVR dataset, which emphasize the reasoning of complex spatial relationships.

We ensure the inherent difficulty of initial data by combining empirical validation with source selection. For public benchmarks (*e.g.*, MathVision), we confirm their challenge by observing that even powerful MLLMs (*e.g.*, Qwen2.5-VL-7B) achieve relatively low accuracy (25.4%), which provides a strong empirical indicator of their difficulty. For crawled data, we curate questions from sources renowned for their complexity, *e.g.*, Physics / Biology Olympiads and IQ tests, which are by design created to probe the limits of logical deduction. This approach ensures the initial data poses a meaningful challenge to current MLLMs.

B.3 Baseline Models

We evaluate various reasoning and CoT-enhanced general MLLMs, including black-box MLLMs [24, 56, 23] and open-sourced MLLMs [2, 61, 5, 18, 1]. We also analyze reasoning-enhanced methods [72, 67, 57, 12, 86, 16] to explore the hallucination mitigation effectiveness.

Black-box MLLMs include GPT-4o [23] and O1 [24] from OpenAI, as well as Gemini-2-flash and Gemini-2-flash-thinking [56] from Google. These models have shown state-of-the-art reasoning or chain-of-thought thinking capabilities in various tasks.

Open-sourced MLLMs cover both specifically-designed reasoning MLLMs (*e.g.*, QvQ-72B [57] and Virgo-72B [12]), and General MLLMs including Qwen2-VL [61], Qwen2.5-VL [2], InternVL-2.5 [5], Llama-3.2-Vision [18] and Phi-3.5-Instruct [1]. All these models have shown competitive inherent or chain-of-thought reasoning capabilities. Note that the parameter numbers of selected models are largely varied from 3B to 72B, ensuring that models with different scales can be analyzed in our experiments.

Reasoning-enhanced General MLLMs. Besides, to comprehensively evaluate the capabilities of reducing reasoning hallucination, we also assess multiple training-free and training-based hallucination mitigation methods, including self-reflection [16], question decomposition [86], and supervised fine-tuning [71]. All these methods have shown effectiveness in improving reasoning capabilities.

B.4 Prompts Used in Construction and Evaluation

For the reproducibility of our work, we release the critical prompts used in MIRAGE construction and evaluation. Fig. 8 shows the prompt for extracting intermediate results (*i.e.*, steps and claims). Fig. 9 shows the prompt of intermediate results matching results (*i.e.*, factuality evaluation prompt). Fig. 10 shows the prompt used to detect specific reasoning hallucination types. And Fig. 11 shows the prompt used to calculate the LHS score.

B.5 Computational Resources and Time

During the first stage of C-RFT, Logos-7B uses 8 NVIDIA A6000 GPUs to train a 7B model, and the total training time is 16 hours. During the second stage (since the optimal stage number k is 1, as discussed in App. D), Logos-7B uses the same 8 A6000 GPUs and requires 6 hours to complete

training. During Inference, benefiting from the optimization of the vLLM [29] framework, Logos-7B only requires one A6000 GPU for inference reasoning chains. And both training and inference of Logos-3B model require fewer computational resources.

B.6 Significance Computation

To calculate the correlation and the corresponding significance value, we leverage scipy package and call `pearsonr` function to calculate pearson correlation coefficient with corresponding significance (*i.e.*, p-value).

Table 11: Effect of reinforcement learning algorithm. We remove CHI and directly assess the original output of each model.

| RL | MIRAGE | MathVista |
|------|--------|-----------|
| PPO | 29.9 | 69.3 |
| DAPO | 30.0 | 69.6 |
| GRPO | 35.7 | 71.9 |

Table 12: Effect of online reward filtration.

| ORF | MIRAGE | MathVista |
|-----|--------|-----------|
| ✗ | 34.2 | 69.6 |
| ✓ | 37.1 | 72.3 |

Table 13: Effect of the KL-divergence in Logos.

| KL-Div | MIRAGE | MathVista |
|-----------|--------|-----------|
| 1e-2 | 31.0 | 67.0 |
| 1e-3 | 35.0 | 70.0 |
| 1e-4 | 36.7 | 71.1 |
| 0 (Logos) | 37.1 | 72.3 |

Table 14: Comparison between CRFT and vanilla RL with longer training epochs.

| Method | Total Epochs | MIRAGE | MathVista |
|------------|--------------|--------|-----------|
| Vanilla RL | 20 | 35.5 | 71.4 |
| CRFT | 10+10 | 37.1 | 72.3 |

Table 15: Effect of curriculum learning stage k .

| k | MIRAGE | MathVista |
|-----|--------|-----------|
| 0 | 35.0 | 70.7 |
| 1 | 37.1 | 72.3 |
| 2 | 37.2 | 72.3 |
| 3 | 37.2 | 72.3 |

C Detailed Quantitative Results

In addition to reporting average accuracy and overall LHS score for each model, we also report accuracy in each question topic and LHS score in each dimension. The detailed per-topic accuracy comparison results are shown in Table 16, and the detailed per-dimension accuracy comparison results are shown in Table 18. Generally, all models suffer from unpromising accuracy on logical and spatial questions, which indicate that existing models still do not obtain sufficient visual pattern and relation reasoning abilities. Meanwhile, mathematic reasoning and statistical question reasoning are usually perform well on state-of-the-art MLLMs. As for the LHS score, the score of logical consistency and reasoning completeness of some unpromising models are relatively low, which indicates that previous models still struggle in insufficient reasoning capabilities and result in reasoning hallucination. These results reveal the the vulnerability in reasoning MLLMs.

D More Analysis

D.1 Qualitative Results of Manually Fixing Examples

As stated in Sec. 6.2, manually fixing hallucinations in the reasoning chains enhances overall model performance. In addition to the quantitative results, we also illustrate qualitative results from GPT-4o [23], which is shown in Fig. 6. The corrected reasoning chain (right) can guide MLLMs to predict correct answers, and the original reasoning chain with hallucination (left) still results in wrong answers.

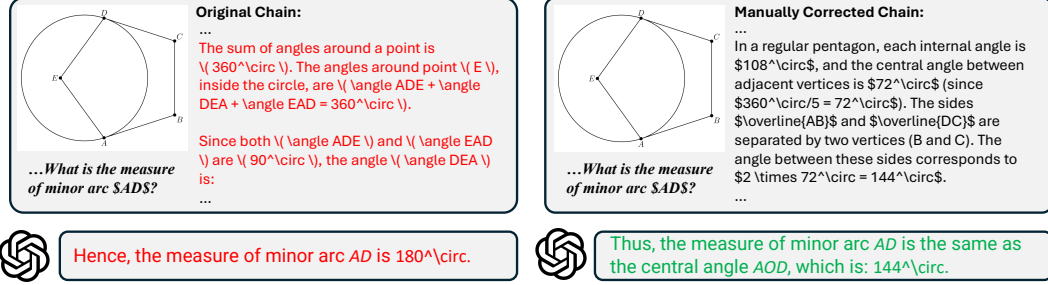


Figure 6: Qualitative results of manually fixing the reasoning hallucination in the reasoning chain and inference the refined answers. The corrected reasoning chain (right) can guide MLLMs to predict correct answers, and the original reasoning chain with hallucination (left) still results in wrong answers.

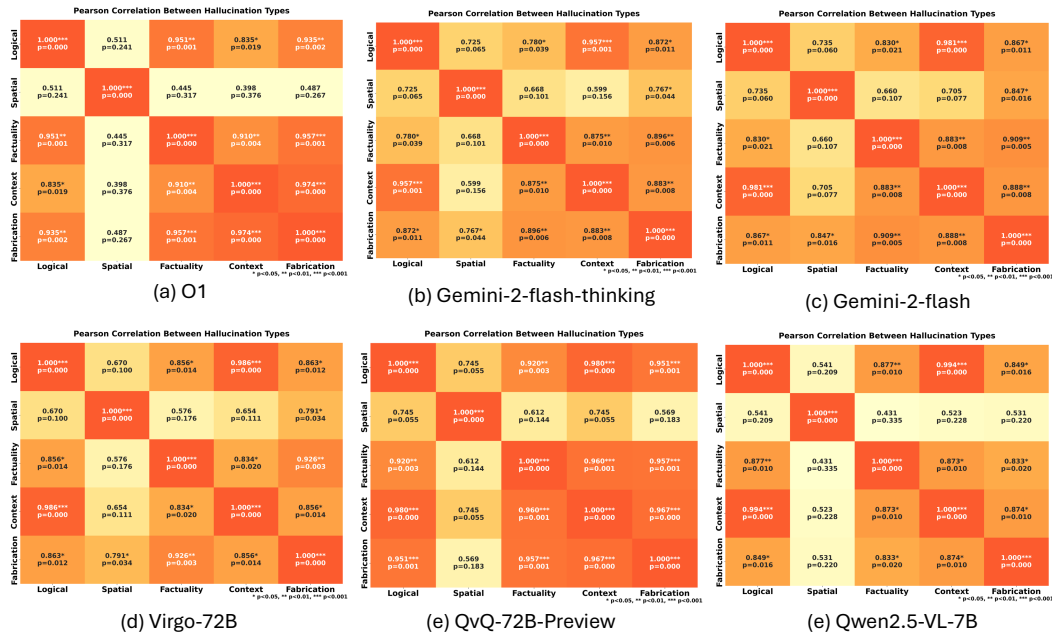


Figure 7: Pearson correlation regarding hallucination types from six most representative MLLMs. All models tend to represent a similar pattern.

D.2 Pearson Correlation of Hallucination Type Among Single Model

In addition to stating the overall pearson correlation coefficient in Fig. 5, we also illustrate corresponding correlation from six most representative models, *i.e.*, O1 [24], Gemini-2-flash-thinking [56], Gemini-2-flash [56], Virgo-72B [12], QvQ-72B-Preview [57], and Qwen2.5-VL-7B [2]. As shown in Fig 7, all models show a similar correlation pattern, which is consistent with Fig. 5. These results indicate the shared vulnerability in reasoning MLLMs.

D.3 Training-free Methods in MIRAGE.

We also explore some training-free methods to verify the hallucination mitigation capabilities. Therefore, we evaluate prompt-based self-reflection [16] and visual inference chain [86]. As shown in Table 2, compared to Qwen2.5-VL-7B, both methods suffer performance degradation on both accuracy and LHS. These results indicate that for models without insufficient reasoning capabilities and hallucination-defending abilities, introducing training-free methods does not help to mitigate

reasoning hallucination. This observation inspires us to integrate CHI into CRFT-enhanced MLLMs rather than the base model.

D.4 Supervised fine-tuning (SFT) Methods in MIRAGE

Intuitively, reasoning hallucination could be solved by further supervised fine-tuning to integrate the correct Hence, we select several SFT-based methods on 7B-level [72, 67, 71] and 72B-level [57, 12], and evaluate on MIRAGE. As shown in Table 2, on 72B MLLMs, introducing SFT can lead to better accuracy and partly mitigate reasoning hallucinations. Nevertheless, on 7B MLLMs, only Mulberry surpasses the base model by 3.1 on accuracy, while other methods do not lead to performance improvement and hallucination mitigation. This contradiction may come from the model’s capacity for model sizes. Larger models with more inherent knowledge may be easier to mitigate hallucination via external supervision, while smaller models usually struggle with SFT.

Table 16: Accuracy comparison of each question topic in MIRAGE.

| Model | Algebra | Arithmetic | Geometry | Logical | Scientific | Spatial | Statistical | Overall |
|----------------------------------------|---------|------------|----------|---------|------------|---------|-------------|---------|
| Black-Box MLLMs | | | | | | | | |
| Gemini-2-Flash-Thinking [56] | 56.1 | 66.3 | 53.7 | 33.3 | 41.4 | 26.1 | 55.0 | 47.6 |
| O1 [24] | 50.9 | 64.1 | 60.0 | 37.7 | 42.5 | 37.8 | 51.9 | 49.7 |
| Gemini-2-Flash [56] | 51.2 | 57.6 | 50.2 | 34.1 | 32.3 | 26.7 | 55.8 | 44.1 |
| GPT-4o [23] | 38.8 | 40.2 | 29.5 | 24.6 | 28.0 | 42.2 | 47.2 | 35.0 |
| Open-sourced ~72B MLLMs | | | | | | | | |
| Qwen2.5-VL-72B-Instruct [2] | 44.9 | 50.0 | 37.8 | 29.0 | 24.7 | 32.2 | 49.6 | 38.8 |
| InternVL-2.5-78B [5] | 31.1 | 38.0 | 31.4 | 21.7 | 21.0 | 27.2 | 40.3 | 29.6 |
| Qwen2-VL-72B-Instruct [61] | 20.1 | 30.4 | 26.0 | 18.8 | 19.9 | 25.0 | 38.0 | 24.5 |
| QvQ-72B-Preview [57] | 30.1 | 44.6 | 32.1 | 23.9 | 28.0 | 25.0 | 41.1 | 31.0 |
| Virgo-72B [12] | 44.6 | 47.8 | 37.5 | 29.0 | 23.1 | 38.9 | 41.1 | 37.4 |
| Open-sourced ~7B MLLMs | | | | | | | | |
| Qwen2.5-VL-7B-Instruct [2] | 28.0 | 34.8 | 28.6 | 26.1 | 17.2 | 28.9 | 46.5 | 28.8 |
| Qwen2-VL-7B-Instruct [2] | 16.3 | 13.0 | 14.6 | 22.5 | 15.6 | 32.8 | 27.1 | 19.5 |
| Qwen2.5-VL-7B-Instruct+VIC [86] | 30.1 | 31.5 | 26.0 | 23.2 | 17.7 | 24.4 | 39.5 | 26.9 |
| Qwen2.5-VL-7B-Instruct+Reflection [16] | 23.9 | 31.5 | 24.8 | 2.3 | 17.2 | 31.7 | 39.5 | 26.7 |
| R1-OneVision-7B [5] | 20.1 | 22.8 | 21.0 | 26.1 | 16.1 | 28.3 | 33.3 | 22.9 |
| Mulberry-Qwen2-VL-7B [72] | 19.7 | 26.1 | 27.0 | 22.5 | 19.9 | 19.4 | 24.8 | 22.6 |
| InternVL-2.5-8B [5] | 11.4 | 26.1 | 22.5 | 16.7 | 17.2 | 30.6 | 30.2 | 20.8 |
| Llama-3.2-Vision-11B [18] | 12.5 | 21.7 | 20.0 | 14.5 | 16.1 | 17.2 | 38.0 | 18.7 |
| Llava-CoT-11B [67] | 12.1 | 14.1 | 19.7 | 10.9 | 16.1 | 20.0 | 31.0 | 17.4 |
| Logos-7B (Ours) | 39.1 | 39.1 | 38.7 | 32.6 | 20.4 | 34.4 | 59.7 | 37.1 |
| Open-sourced ~3B MLLMs | | | | | | | | |
| Qwen2.5-VL-3B-Instruct [2] | 10.7 | 27.2 | 20.6 | 25.4 | 16.1 | 16.1 | 27.1 | 18.8 |
| Phi-3.5-Instruct [1] | 4.1 | 16.3 | 14.0 | 15.2 | 14.5 | 13.9 | 21.7 | 12.9 |
| Logos-3B (Ours) | 27.3 | 38.0 | 31.4 | 24.6 | 17.7 | 27.2 | 48.0 | 29.4 |

D.5 In-depth Analysis of Logos in Mitigating Spatial Hallucinations

We conducted experiments to compare Logos-7B against other methods specifically on spatial hallucination. The results in Table 17 are striking. While mitigating spatial deficits remains a significant challenge, Logos is the only method that shows a slight improvement, whereas all other methods, including training-free and SFT-based approaches, actually degrade performance. The reasons are two-fold. Firstly, training-free methods (*e.g.*, Self-Reflection) These methods cannot introduce new knowledge. By forcing the base model, which already lacks spatial reasoning ability, to adhere to rigid prompting rules, they can amplify existing errors. Secondly, SFT-based methods (*e.g.*, R1-OneVision) are often trained on datasets dominated by other domains (*e.g.*, mathematics). This can cause the model to overfit to mathematical reasoning patterns at the expense of its already weak spatial capabilities. In contrast, our Logos avoids these pitfalls. The RL approach enhances reasoning without the overfitting risk of SFT, while our CHI component provides targeted hints that guide the model on difficult spatial problems. Therefore, Logos is the only method avoiding performance degradation.

Table 17: Spatial hallucination error rate (%) comparison across methods. Lower is better.

| Model | Spatial Hallucination (%) |
|-----------------|---------------------------|
| Qwen2.5-VL-7B | 33.4 |
| +Reflection | 37.4 |
| +VIC | 36.7 |
| R1-OneVision-7B | 35.6 |
| Logos-7B | 29.9 |

D.6 Effect of RL algorithms

As mentioned in Sec. 5, the multiple sampling pipeline in the GRPO algorithm is naturally aligned with our hallucination mitigation proposal, *i.e.*, encouraging models to predict along the correct chain for correct answers. To verify the effect of RL algorithms, we compare Logos-7B using GRPO [53] and that using PPO [52]. We remove all CHI stages and directly assess the effect of RL. As shown in Table 11, Logos-7B using GRPO surpasses that using PPO by 5.8 and 2.6 on MIRAGE and MathVista, respectively. These results verify our motivation and provide strong support for the Logos framework design. We also evaluate the effect of recently proposed DAPO [76], a specially-designed GRPO variant. Nevertheless, it does not lead to better performance on both benchmarks. A possible explanation is that the newly introduced constraints in DAPO lead to overfitting in the model and restrict the final performance.

Table 18: LLMs Hallucination Score (LHS) comparison of each dimension in MIRAGE.

| Model | Factual | Logical | Reasoning | Conceptual | Appropriateness | Overall |
|----------------------------------------|---------|---------|-----------|------------|-----------------|---------|
| Black-Box MLLMs | | | | | | |
| Gemini-2-Flash-Thinking [56] | 0.7182 | 0.7558 | 0.7689 | 0.7349 | 0.7372 | 0.7517 |
| O1 [24] | 0.6054 | 0.6427 | 0.5793 | 0.6384 | 0.6306 | 0.6193 |
| Gemini-2-Flash [56] | 0.6640 | 0.7053 | 0.7265 | 0.6862 | 0.7007 | 0.6882 |
| GPT-4o [23] | 0.5811 | 0.6466 | 0.6777 | 0.6198 | 0.6404 | 0.6332 |
| Open-sourced ~72B MLLMs | | | | | | |
| Qwen2.5-VL-72B-Instruct [2] | 0.6330 | 0.7464 | 0.7912 | 0.6793 | 0.7321 | 0.7233 |
| InternVL-2.5-78B [5] | 0.5830 | 0.6441 | 0.6700 | 0.6088 | 0.6313 | 0.6377 |
| Qwen2-VL-72B-Instruct [61] | 0.4665 | 0.5115 | 0.5339 | 0.4746 | 0.4774 | 0.4928 |
| QvQ-72B-Preview [57] | 0.5024 | 0.5495 | 0.5698 | 0.5368 | 0.5168 | 0.5717 |
| Virgo-72B [12] | 0.6094 | 0.6185 | 0.6437 | 0.6252 | 0.6187 | 0.6328 |
| Open-sourced ~7B MLLMs | | | | | | |
| Qwen2.5-VL-7B-Instruct [2] | 0.5333 | 0.6201 | 0.6765 | 0.5786 | 0.6130 | 0.5996 |
| Qwen2-VL-7B-Instruct [2] | 0.3512 | 0.3960 | 0.4120 | 0.3573 | 0.3519 | 0.3633 |
| Qwen2.5-VL-7B-Instruct+VIC [86] | 0.4600 | 0.4746 | 0.4336 | 0.4449 | 0.4261 | 0.4478 |
| Qwen2.5-VL-7B-Instruct+Reflection [16] | 0.5658 | 0.6242 | 0.6008 | 0.5806 | 0.6117 | 0.5826 |
| R1-OneVision-7B [5] | 0.4565 | 0.5227 | 0.5809 | 0.4822 | 0.5070 | 0.5098 |
| Mulberry-Qwen2-VL-7B [72] | 0.4545 | 0.4819 | 0.5070 | 0.4605 | 0.4660 | 0.4740 |
| InternVL-2.5-8B [5] | 0.4515 | 0.4967 | 0.5317 | 0.4636 | 0.4757 | 0.4838 |
| Llama-3.2-Vision-11B [18] | 0.4014 | 0.4473 | 0.4741 | 0.4030 | 0.4066 | 0.4265 |
| Llava-CoT-11B [67] | 0.4050 | 0.4417 | 0.4735 | 0.4116 | 0.4267 | 0.4267 |
| Logos-7B (Ours) | 0.5841 | 0.6533 | 0.7052 | 0.6233 | 0.6566 | 0.6568 |
| Open-sourced ~3B MLLMs | | | | | | |
| Qwen2.5-VL-3B-Instruct [2] | 0.3282 | 0.3593 | 0.3712 | 0.3279 | 0.3242 | 0.3422 |
| Phi-3.5-Instruct [1] | 0.2983 | 0.3443 | 0.3459 | 0.3049 | 0.2968 | 0.3181 |
| Logos-3B (Ours) | 0.5486 | 0.5947 | 0.6411 | 0.5600 | 0.5757 | 0.5840 |

D.7 Effect of KL Divergence

Since we remove the KL-divergence term in Logos training, to analyze the effect, we conduct an ablation study on the KL-divergence weight. As shown in Table 13, when gradually increasing the weight of the KL-divergence term, the accuracy on both datasets is gradually decreased. When the KL-divergence term is relatively large (*e.g.*, $1e-2$), the accuracy on MathVista is even slightly lower than the base model (68.2). A possible explanation is that the distribution of reasoning MLLM has a non-negligible gap with the corresponding base models. To mitigate original reasoning hallucination

and bring inherent reasoning capabilities, one should disable the KL-divergence term to tolerate the distribution gap between two models.

D.8 Effect of online reward filtration.

Next, we explore the effect of online reward filtration and report experimental results in Table 12. After integrating ORF into training, Logos-7B surpasses the counterpart by 2.9 on MIRAGE and 2.7 on MathVista. This improvement proves the effectiveness of ORF and points the future direction for more effective RL algorithms.

D.9 Effect of curriculum learning.

We also conduct experiments to verify the necessity of the CRFT stage. Specifically, we conduct vanilla RL training on all training data and ensure the number of training epochs is equal to the total CRFT. As shown in Table 14, even using longer training, one-stage RL still falls behind Logos-7B with CRFT by 1.6 on MIRAGE accuracy and 0.9 on MathVista accuracy. These results indicate that, benefiting from multi-stage difficulty filtration, the learning efficiency of Logos is highly improved. Meanwhile, we explore the effect of the curriculum learning stage k in Logos. As shown in Table 15, Logos-7B can easily obtain 37.1 on MIRAGE and 72.3 on MathVista. Further increasing k only introduces a marginal improvement. These results prove the effectiveness and efficiency of CRFT. Therefore, we select $k = 1$ to optimize Logos.

D.10 Performance on Multimodal Reasoning Datasets

To show the generalizability of Logos beyond MIRAGE, we evaluate Logos-7B on three general-purpose benchmarks: MathVista, MathVision, and MathVerse. As shown in Table 19, Logos-7B not only surpasses its strong base model but also outperforms concurrent state-of-the-art methods, achieving the highest average score. This shows that Logos to mitigating hallucinations also enhances reasoning abilities, leading to comparable performance on academic benchmarks.

Table 19: Performance comparison on multimodal reasoning datasets.

| Method | MathVista | MathVision | MathVerse | Avg |
|------------------|-----------|------------|-----------|------|
| Qwen2.5-VL-7B | 68.2 | 25.4 | 47.9 | 47.1 |
| R1-OneVision-7B | 64.1 | 19.7 | 46.4 | 46.8 |
| OpenVLThinker-7B | 70.2 | 25.3 | 47.9 | 47.8 |
| MM-EUREKA | 73.0 | 26.9 | 50.3 | 50.0 |
| Logos-7B (Ours) | 72.3 | 29.8 | 52.5 | 51.5 |

D.11 Comparision with DAPO and ADARFT

Comparison with DAPO [76]: The primary distinction lies in the policy gradient loss design. DAPO introduces several strategies (e.g., clip-higher, token-level loss) optimized for text-only math problems, which often feature very long reasoning chains. Our experiments found these strong regularizations were less effective and could lead to overfitting in the multimodal context, where reasoning chains are typically more concise (see Appendix D.5). Logos employs a simpler, more robust sample-level policy gradient loss that proved more effective for multimodal training, as shown in Table 8.

Comparison with ADARFT [54]: The key difference is the curriculum learning mechanism. ADARFT uses a static offline difficulty annotation and greedy strategy to select training batches. This approach does not guarantee that all data is utilized effectively. Logos employs a more dynamic hybrid online-offline framework. ORF ensures each batch is filled with high-value samples, while offline difficulty thresholding stages the curriculum effectively. This hybrid approach maximizes data utility and leads to better performance. To provide direct empirical evidence, we compare Logos (w/o CHI) against these methods on MIRAGE and MathVista. Logos demonstrates superior performance on both.

Table 20: Performance comparison between DAPO, ADARFT, and Logos (w/o CHI) on MIRAGE and MathVista.

| Method | MIRAGE | MathVista |
|-----------------|--------|-----------|
| DAPO | 30.0 | 69.6 |
| ADARFT | 34.0 | 70.1 |
| Logos (w/o CHI) | 35.7 | 71.9 |

D.12 The Quality of Automatic Annotation

We also assess the accuracy of automatically annotated reasoning chains in different phase, which is shown in Table 21. The accuracy of O3-mini initialized reasoning chains achieves 43.9. By incorporating DeepSeek-R1 and guided by answers, the accuracy of refined reasoning chains achieves 73.7. The relative high accuracy of refined reasoning chains ensures that one can reduce the human labor to correct the reasoning chains with reasoning hallucinations.

Annotation cost. Finally we also concern the detailed annotation cost. By using the annotation method proposed in Sec. 3.2, the total cost is $\sim 22\$$. And the total human working hour is 36 hours*person. We also estimate the annotation cost using O1, which is nearly 200\$. And if all the questions are annotated by human experts, the total working hour is nearly 200 hours*person. These results show the efficiency of our annotation method.

Table 21: The reasoning annotation accuracy in each phase.

| Annotation Phase | Reasoning Chain Accuracy |
|---------------------------------|--------------------------|
| O3-mini (init) | 43.9 |
| +DeepSeek-R1 (refine w/ answer) | 73.7 |

D.13 Human Double Checking for Metric Reliability and Consistency

To verify the reliability and consistency of evaluation metrics, the human double checking procedure is built on three pillars.

First is mitigating circularity. To prevent self-enhancement bias, we intentionally use an older, isolated model version (gpt-4o-2024-08-06) as our primary evaluator, ensuring it has not been trained on the models or data it is evaluating. For open-source evaluators, we designed highly structured, specialized prompts (see Appendix) to constrain the evaluation and minimize potential bias.

Second is human-in-the-loop verification to double-check the validity of LLM-as-a-judge. We do not rely solely on automated metrics. We conducted manual double-checking for our key metrics (factuality and LHS) on representative models. The results in Table 22 demonstrate a strong alignment between our automated scores and human expert judgment. Specifically, for factuality scores in the following table, the automated scores closely match the manually verified scores, confirming their accuracy.

Table 22: Comparison of automated and manual factuality scores for step and claim metrics.

| Model | Step | Claim |
|-------------------------|------|-------|
| Gemini-2-flash | 47.8 | 47.4 |
| Gemini-2-flash (manual) | 48.5 | 47.9 |
| Qwen2.5-VL-7B | 34.7 | 31.7 |
| Qwen2.5-VL-7B (manual) | 33.9 | 31.3 |

And for LHS, we analyze the absolute difference between automated and manual LHS scores. The results in Table 23 show that the metric is highly stable, with over 75% of samples having a negligible difference of less than 0.1.

Besides, we further validate our framework by calculating the correlation between all metrics in Table 5. The high correlation scores across the board indicate strong internal consistency and robustness.

Table 23: Absolute difference between automated and manual LHS scores at different percentiles. Over 75% of samples have a negligible difference (<0.1).

| Model | 25% | 50% | 75% | 90% |
|----------------|------|------|------|------|
| Gemini-2-flash | 0.02 | 0.04 | 0.07 | 0.11 |
| Qwen2.5-VL-7B | 0.04 | 0.05 | 0.09 | 0.13 |

E Dataset Examples

To clearly show the structure of MIRAGE, we provide detailed examples of MIRAGE. Fig. 12 shows an example of geometry questions. Fig. 13 shows an example of algebraic questions. Fig. 14 shows an example of arithmetic questions. Fig. 15 shows an example of scientific questions. Fig. 16 shows an example of spatial questions. Fig. 17 shows an example of logical questions. Fig. 18 shows an example of statistical questions.

Summary of Topic-specific Hint. We also release the topic-specific hints used in MIRAGE and Logos. As shown in Fig. 19. The topic-specific hints include key concepts and basic rules regarding the question topics. Meanwhile, the classical reasoning process of corresponding question topics is also included in the hints.

F More Qualitative Results

We also illustrate a couple of raw outputs from some representative models, *i.e.*, Qwen2.5-VL-7B-Instruct, Gemini-2-flash-thinking, and our Logos-7B. As shown in the Fig. 20 and 21. We find that in this example, Qwen2.5-VL-7B has consistent logic in the reasoning chain, but suffers from factual hallucination (only two 90 degree angles should be calculated). In contrast, Gemini-2-flash-thinking and Logos-7B correctly solve the question.

G Limitation

The limitation of this paper is two-fold. First, MIRAGE does not include multiple images or video question-answering problems. Hence, the hallucination from the temporal dimension and the hallucination regarding cross-image relations are not fully explored. Our primary goal with MIRAGE was to first establish a foundational and high-quality benchmark for reasoning chain hallucinations in the prevalent single-image context. This deliberate focus allows for a deep and controlled analysis of a core challenge in MLLMs. For scenarios that can be adapted, we propose a practical method of concatenating multiple images into a single composite image and annotating object indices, effectively converting a multi-image problem into a format compatible with the current benchmark. And second, the theoretical analysis of why MLLMs suffer from reasoning hallucination is still insufficient. These limitations motivate us to conduct more in-depth exploration in the future.

Limitation in the Logos’s effectiveness across all identified hallucination types. The behavior stems from two distinct challenges:

- **Inherent Gaps in Base Models (spatial hallucination):** Our analysis in Table 1 and 2 reveals that even top-performing base models have a fundamental weakness in spatial reasoning. Scaling model or data size does not significantly reduce this type of error. Since reinforcement learning in CRFT works by rewarding and reinforcing correct behaviors that the model can already sample, it cannot effectively teach a capability that is fundamentally absent. If the base model rarely generates a correct spatial reasoning path, there are few positive examples for Logos to amplify.
- **The Nature of RL vs. factual knowledge (factuality hallucination):** For strong base models, the remaining factual errors are often subtle and stem from inherent biases in their pre-training data. The goal of Logos’s RL framework is to optimize the policy for generating correct reasoning chains, not to inject new factual knowledge. If all sampled reasoning paths from the base model contain the same ingrained factual error (e.g., a slightly incorrect constant), Logos lacks a ground-truth signal to correct it.

****Task****

Given a reasoning chain, extract the key ****steps**** (critical reasoning or calculation steps) and ****claims**** (correct intermediate results or derived facts). Follow these guidelines for each problem type:

1. ****Geometry****: Angles, lengths, area/volume formulas, congruence/similarity, geometric theorems (e.g., Pythagorean theorem).
2. ****Arithmetic****: Critical numerical calculations (e.g., division, fractions) or multi-step operations.
3. ****Algebra****: Equations, inequalities, variable substitutions, function definitions, or computed variable values.
4. ****Spatial Reasoning****: Spatial transformations (rotation, reflection), 3D object properties, coordinate systems.
5. ****Logical Patterns (Rule Identification)****: Explicit rules (e.g., "add 3"), sequence relationships, or logical operators.
6. ****Science (Physics/Chemistry/Biology)****: Scientific laws, formulas (e.g., $F=ma$), experimental steps, chemical reactions, biological processes.
7. ****Statistics (Chart/Graph Problems)****: Data trends, statistical measures (mean, median), chart interpretations, probability rules.

****If the reasoning chain is incomplete****, treat **"the answer is [answer in the input]"** as both a step and a claim.

****If the reasoning chain is too long, output AT MOST 10 most important (i.e., critical to solve the problem) steps and AT MOST 10 most important (i.e., critical to solve the problem) claims in the chain.**

****Formatting:****

- Wrap each step with `<step>...</step>`.
- Wrap each claim with `<claim>...</claim>`.

****Examples (omitted to fit one-page style):****

Now, process the following input:

Input: {cot}

Output:

Figure 8: The evaluation prompt used to extract intermediate results.

Hence, these findings show a crucial insight: tackling spatial and factual hallucinations may require architectural innovations or new data paradigms beyond the scope of post-training methods. These results also suggest that future work might explore better pertaining data, spatial-enhanced network design and advanced framework to mitigate such hallucinations.

H Broader Impact

The broader impact of this paper lies in advancing the reliability and accuracy of multimodal large language models (MLLMs) by systematically isolating and evaluating reasoning hallucinations. MIRAGE offers a targeted benchmark for diagnosing and mitigating reasoning errors, which is essential for applications in fields like autonomous systems, medical imaging, and scientific discovery, where accurate multimodal reasoning is critical. By revealing key weaknesses in current MLLMs, such as their struggles with complex spatial reasoning, our work encourages the development of more robust, transparent, and context-aware AI systems, ultimately promoting safer and more trustworthy AI deployment.

```

**Task:**
THINK STEP BY STEP and Analyze three components:
1. **Question** (provide the basic judge background information)
2. **Model-generated steps** (between `<steps>` tags)
3. **Ground-truth steps** (between `<steps>` tags)
**Output the answer, then provide additional analysis to enhance the accuracy**
**Important Notes:**
1. The number of decision in <pred_match>...</pred_match> should be the exactly same with
steps in the prediction.
2. The number of decision in <gt_match>...</gt_match> should be the exactly same with steps in
the ground-truth.
... (more detailed rules omitted to fit one-page style) ...
---
### Output Format:

```

<<<xml
<pred_match>MATCH,REASONABLE,CONFLICT,...</pred_match>
<gt_match>MATCH,CONFLICT,...</gt_match>
<precision>0.XX</precision>
<recall>0.XX</recall>
<f1>0.XX</f1>
>>>

```


---
### Input:
Question:
{question}

Prediction from Model:
<steps>
{pred_steps}
</steps>

Ground-Truth:
<steps>
{gt_steps}
</steps>

**Output:**

```

Figure 9: The evaluation prompt used for factuality assessment (e.g., F_{step}).

****Task Requirement:****

You are given the output of a multi-modal reasoning model, including a chain-of-thought (CoT) and an input consisting of a question and an accompanying image (represented by the placeholder <image>). You are also provided with a human-verified reference chain-of-thought (Ref-CoT) that correctly answers the question. Your task is to analyze the predicted CoT and determine whether it contains any hallucinations based on comparison with the Ref-CoT and the input.

****Types of Hallucinations:****

1. **Spatial Hallucination**: Misunderstanding the visual structure, shape, angle, or spatial configuration based on the image.
2. **Logical Hallucination**: Flawed or missing reasoning steps, incorrect logical deductions, or invalid inferences.
3. **Factuality Hallucination**: Statements that contradict known math facts or the image data.
4. **Context Hallucination**: Inconsistencies between steps or contradictions with the predicted answer.
5. **Fabrication Hallucination**: Reasoning steps that are not derived from earlier steps, image data, or question context.

Please output a JSON result with the following fields, enclosed within `<result>` and `</result>` tags:

- `result` : Either `CONFIRM` or `HALLUCINATION`
- `reason` : A brief justification for your judgment

****Few-shot Examples for Reference:****

----- Examples are omitted to fit one-page style -----

****Now, for the following input, analyze the reasoning chain and provide your final evaluation according to the schema above.****

****Input:****

- **Image**: <image>
- **Question**: {question}
- **Predicted CoT**: {pred_cot}
- **Reference CoT**: {ref_cot}

****Output format:****

```
<result>
{ "Overall": {
  "result": "CONFIRM" or "HALLUCINATION",
  "reason": "Overall explanation here..." },
  "Spatial": { "result": "CONFIRM" or "HALLUCINATION",
    "reason": "Detailed explanation for Spatial hallucination..." },
  "Logical": { "result": "CONFIRM" or "HALLUCINATION",
    "reason": "Detailed explanation for Logical hallucination..." },
  "Factuality": { "result": "CONFIRM" or "HALLUCINATION",
    "reason": "Detailed explanation for Factuality hallucination..." },
  "Context": { "result": "CONFIRM" or "HALLUCINATION",
    "reason": "Detailed explanation for Context hallucination..." },
  "Fabrication": { "result": "CONFIRM" or "HALLUCINATION",
    "reason": "Detailed explanation for Fabrication hallucination..." }
}
</result>
```

****Output****

Figure 10: The evaluation prompt used to detect hallucination types in reasoning chains.

Task Requirement:

You are an expert reasoning chain evaluator. Your job is to score a model-generated Chain of Thought (CoT) based on six key dimensions that assess its correctness, coherence, and hallucination risk. You are provided with:

1. A problem statement.
2. A model-generated Chain of Thought (CoT).
3. One or more reference CoTs (these are valid, possibly diverse ground-truth reasoning paths).

You must evaluate the model's CoT based on the following 6 dimensions. For each dimension, assign a score between 1 and 5, where 5 is best. Also provide a concise explanation for each score.

Evaluation Dimensions and Scoring Criteria:****1. Factual Accuracy****

- 5 = All stated facts, definitions, and formulas are correct.
- 4 = Minor factual inaccuracies not critical to the reasoning.
- 3 = At least one notable factual mistake, but the main conclusion still follows.
- 2 = Several factual inaccuracies that impact the reasoning chain.
- 1 = Major factual errors or invented knowledge that invalidate the reasoning.

****2. Logical Consistency****

- 5 = Every step logically follows from the previous one; no contradictions.
- 4 = Mostly consistent with a minor leap or imprecision.
- 3 = Some unclear or unjustified transitions between steps.
- 2 = Multiple reasoning steps are logically invalid.
- 1 = Reasoning is incoherent or self-contradictory.

****3. Reasoning Completeness****

- 5 = Fully complete; all key steps are included.
- 4 = Minor omissions; conclusion still follows.
- 3 = Missing one important step; slightly weakens the argument.
- 2 = Missing multiple key steps; reasoning is hard to follow.
- 1 = Severely incomplete; the reasoning fails to connect to the answer.

****4. Conceptual Reasoning Accuracy****

- 5 = Correct use of all key mathematical, spatial, or logical concepts.
- 4 = Small misunderstanding or imprecision.
- 3 = Misuse of one core concept that weakens reasoning.
- 2 = Multiple concept-level misapplications.
- 1 = Misuse or hallucination of core principles or definitions.

****5. Strategy Appropriateness****

- 5 = The chosen strategy is valid and optimal or close to optimal.
- 4 = The strategy is valid but suboptimal.
- 3 = Strategy is unconventional but valid; minor concerns.
- 2 = The strategy is logically flawed or poorly suited.
- 1 = The approach is fundamentally invalid for the problem.

Output Format:

Return your evaluation as a valid JSON object with the following format (omitted to fit onepage style):

Few-shot Evaluation Example (omitted to fit onepage style):

Now please evaluate the following CoT:

****Problem:****

{prompt}

****Model CoT:****

{model_cot}

****Reference CoT:****

{reference_cot}

Important Notes:

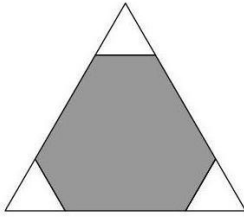
1. Be strict and honest during scoring - high scores should be reserved for flawless or near-flawless reasoning.
2. Pay special attention when final answers differ between model and reference CoTs.
3. Use the negative example as a reference for identifying serious flaws in reasoning.
4. DO NOT give high scores unless the CoT fully deserves it across all dimensions.

Return your scores and explanations in the JSON format as shown above.

****Output****

Figure 11: The evaluation prompt used for LLMs hallucination score extraction.

Geometry



Question:

Three small equilateral triangles of the same size are cut from the corners of a larger equilateral triangle with sides 6 cm as shown. The sum of the perimeters of the three small triangles is equal to the perimeter of the remaining hexagon. What is the side-length of one of the small triangles?

Choices:

A: 1 cm

B: 1.2 cm

C: 1.25 cm

D: 1.5 cm

E: 2 cm

Reasoning Chains:

- Understanding the Problem:**
 - A large equilateral triangle (side length 6 cm) has three smaller equilateral triangles cut from its corners, forming a hexagon.
 - The sum of the perimeters of the three small triangles equals the perimeter of the hexagon.
 - Goal: Find the side length x of the small triangles.
- Perimeter of the Small Triangles:**
 - Each small triangle has a perimeter of $3x$.
 - Total for three triangles: $3 \times 3x = 9x$.
- Perimeter of the Hexagon:**
 - Each original side of the large triangle contributes a middle segment of $(6 - 2x)$ (after removing x from both ends).
 - Each corner cut adds one new side of the small triangle (length x) to the hexagon.
 - Total perimeter of the hexagon: $3 \times (6 - 2x) + 3 \times x = 18 - 6x + 3x = 18 - 3x$.
- Equating Perimeters:**
 - Set the total perimeter of the small triangles equal to the hexagon perimeter: $9x = 18 - 3x$.
 - Solving for x : $12x = 18 \Rightarrow x = \frac{18}{12} = 1.5 \text{ cm}$.

Intermediate Results:

Understand the problem: large equilateral triangle with three smaller equilateral triangles forming a hexagon.

Sum of perimeters of small triangles equals perimeter of hexagon

Calculate perimeter of each small triangle: $3x$

Total perimeter for three small triangles: $9x$

Determine perimeter of hexagon: original side segments $(6 - 2x)$ and new sides (x)

Hexagon perimeter: $3 \times (6 - 2x) + 3 \times x = 18 - 3x$

Set perimeters equal: $9x = 18 - 3x$

Equation: $9x = 18 - 3x$

Solve for x: $12x = 18 \Rightarrow x = 1.5 \text{ cm}$

the answer is 1.5 cm

Description:

The image shows a 3D green graphic of a right circular cone with a sphere inscribed inside it. The sphere is approximately at the center of the cone, touching its curved surface at a single point. This model is essential for understanding the spatial relationship between the sphere and the cone, particularly focusing on their radii. The top and bottom of the truncation (where the large base meets the curved surface) should be the key focal points in solving the problem, as the question is about the ratio of these radii given certain volume relationships. The sphere being inscribed within the cone suggests that its diameter is equal to the height of the truncated part of the cone, which could help in deriving specific relationships. There are no labels or measurements provided, which reinforces the need to rely on visual proportions rather than exact figures. However, the vivid illustration aids in visualizing and solving the problem by providing clear geometric shapes necessary for solving the question.

Hints:

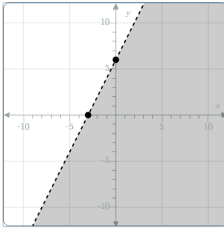
topic_specific_hint: To solve the geometry problem, systematically integrate visual and textual data: cross-verify labels, lengths, angles, and relationships, prioritizing explicit textual details over visual estimates. Identify key components (objective, given data) and apply relevant theorems (e.g., Pythagorean theorem, circle properties, coordinate formulas). Break computations into logical steps, using auxiliary lines or algebraic variables as needed. Verify unit consistency, answer reasonableness, and cross-check via alternative methods. For multiple-choice questions, employ elimination or backsolving. Avoid assumptions about scale or formula misuse; prioritize textual precision and rigorous validation to mitigate errors.

Question_specific_hint: Focus on calculating the volumes of the sphere and the truncated cone. Use the volume formula for a sphere $(V = \frac{4}{3} \pi r^3)$ and for a truncated cone $(V = \frac{1}{3} \pi h (R_1^2 + R_1 R_2 + R_2^2))$. Find the relationship between these volumes based on the given condition to solve for the ratio of the radii (R_1) and (R_2) .

Answer: D

Figure 12: The example of geometry question in MIRAGE.

Algebraic



Question:

State the inequality that describes the region drawn in the number plane.

Intermediate Results:

Identify the boundary line as a linear inequality due to the dashed line
Boundary line suggests a linear inequality (e.g., $y < mx + c$ or $y > mx + c$)

Verify the boundary line equation $y = 2x + 6$ using intercepts
At $x = 0$, $y = 6$ (y-intercept)
At $y = 0$, $x = -3$ (x-intercept)

Check the shading direction of the inequality
Inequality $y < 2x + 6$ implies shading below the line

Test given points $(-3, -10)$ and $(5, 10)$ in the inequality
For $(-3, -10)$: $-10 < 2(-3) + 6 \rightarrow -10 < 0$
For $(5, 10)$: $10 < 2(5) + 6 \rightarrow 10 < 16$

Conclude the correct inequality
Inequality $y < 2x + 6$ correctly describes the shaded region

the answer is $y < 2x + 6$
the answer is $y < 2x + 6$

Hints:

topic_specific_hint:
 To solve the algebraic problem, focus on extracting and interpreting textual information to define variables, expressions, and equations. Identify the objective (e.g., solve for a variable, simplify an expression, evaluate an equation) and distinguish known values, unknowns, and their relationships. Translate word problems or symbolic descriptions into formal algebraic expressions. If an image is present, assess whether it provides meaningful cues or serves as a placeholder; prioritize textual clarity when visual content is ambiguous or abstract. Apply appropriate algebraic techniques (e.g., substitution, factoring, combining like terms, solving systems) in clear logical steps. Clearly define each variable and explain the transformations at each stage. Check your final answer for consistency by substituting it back into the original equation or constraints. For multiple-choice questions, use elimination, estimation, or back-substitution. Avoid misinterpreting irrelevant visual elements or oversimplifying complex expressions prioritizing rigorous algebraic reasoning grounded in the problem statement.

Question_specific_hint: Identify the slope and y-intercept of the dashed line using the points where it intersects the axes or notable grid points. Formulate the linear equation ($y = mx + c$). Determine which side of the line the shaded region falls into to establish the inequality, considering the direction indicated by shading.

Reasoning Chains:

Problem Analysis

The question involves interpreting a graphical region defined by an inequality. The image description specifies a shaded area bounded by a dashed line (indicating a strict inequality) and references points $(-3, -10)$ and $(5, 10)$. The goal is to determine the inequality describing this region.

Key Observations

- Boundary Line:** The dashed boundary suggests a linear inequality (e.g., $y < mx + c$ or $y > mx + c$).
- Critical Points:** The points $(-3, -10)$ and $(5, 10)$ likely relate to the boundary line or the shaded region.
- Axes Interaction:** The region is bounded by the line and the axes, narrowing the inequality's form.

Deriving the Boundary Line

The provided answer $y < 2x + 6$ implies the boundary line is $y = 2x + 6$. To verify:

- Intercepts:**
 - At $x = 0$, $y = 6$ (y-intercept).
 - At $y = 0$, $x = -3$ (x-intercept).
- Shading Direction:** Since the inequality is $y < 2x + 6$, the region below the line is shaded.

Reconciling with Given Points

The points $(-3, -10)$ and $(5, 10)$ are not on the line $y = 2x + 6$ (e.g., substituting $x = 5$ gives $y = 16$), not 10). However, they may lie within the shaded region:

- For $(-3, -10)$: $-10 < 2(-3) + 6 \rightarrow -10 < 0$
- For $(5, 10)$: $10 < 2(5) + 6 \rightarrow 10 < 16$

Conclusion

The inequality $y < 2x + 6$ correctly describes the region below the dashed line $y = 2x + 6$, consistent with the graph description. The initial assumption of a circle was incorrect due to misinterpreting the boundary as curved rather than linear.

Final Answer

$y < 2x + 6$

Description:

The image shows a graph with axes labeled as x and y , along with a shaded region defined by a dashed boundary. The graph features two points plotted at coordinates $(-3, -10)$ and $(5, 10)$. This demonstrates that we are looking at a region involving both positive and negative values on the x and y axes, bounded on the left side by the line segment and on the axes themselves. The relevant points $(-3, -10)$ and $(5, 10)$ should be noted for understanding the area being described. The boundaries help define the inequality that describes this region.

Answer: $y < 2x + 6$

Figure 13: The example of algebraic question in MIRAGE.

Arithmetic

Number of Stamps by Decade

| Country | 50s | 60s | 70s | 80s |
|---------|-----|-----|-----|-----|
| Brazil | 4 | 7 | 12 | 8 |
| France | 8 | 4 | 12 | 15 |
| Peru | 6 | 4 | 6 | 10 |
| Spain | 3 | 9 | 13 | 9 |

Juan's Stamp Collection

Question:

Juan organizes the stamps in his collection by country and by the decade in which they were issued. The prices he paid for them at a stamp shop were: Brazil and France, 6 cents each, Peru 4 cents each, and Spain 5 cents each. (Brazil and Peru are South American countries and France and Spain are in Europe.)

In dollars and cents, how much did his South American stamps issued before the '70s cost him?

- A: \$0.40 B: \$1.06
C: \$1.80 D: \$2.38
E: \$2.64



Reasoning Chains:

Step 1: Identify Relevant Countries and Decades
South American countries are Brazil (6 ¢ /stamp) and Peru (4 ¢ /stamp). "Before the '70s" includes stamps from the 1950s and 1960s.

Step 2: Extract Stamp Counts from the Table
Assuming the table provided the following counts (based on the correct answer):

- Brazil: 11 stamps (1950s + 1960s)
- Peru: 10 stamps (1950s + 1960s)

Step 3: Calculate Total Cost per Country
- Brazil: $(11 \times 6 \text{ ¢}) = 66 \text{ ¢}$
- Peru: $(10 \times 4 \text{ ¢}) = 40 \text{ ¢}$

Step 4: Sum Costs and Convert to Dollars
Total cost: $(66 \text{ ¢} + 40 \text{ ¢}) = 106 \text{ ¢} = \1.06



Intermediate Results:

Identify relevant countries and decades: Brazil (6 ¢ /stamp) and Peru (4 ¢ /stamp), focusing on 1950s and 1960s.
Relevant countries: Brazil and Peru;
Decades: 1950s and 1960s

Extract stamp counts from the table: Brazil (11 stamps) and Peru (10 stamps) for the specified decades.
Brazil: 11 stamps; Peru: 10 stamps

Calculate total cost per country: Brazil $(11 \times 6 \text{ ¢})$ and Peru $(10 \times 4 \text{ ¢})$.
Brazil: 66 cents; Peru: 40 cents

Sum costs and convert to dollars: 66 cents + 40 cents = 106 cents = \$1.06.
Total cost: \$1.06

the answer is B
the answer is B



Description:

The image shows a table titled "Number of Stamps by Decade" that outlines the number of stamps Juan has by country and by the decade in which they were issued. The columns represent the decades from the 1950s to the 1980s. The rows list the countries deemed relevant to the problem: Brazil, France, Peru, and Spain. The table is essential for solving the problem as it provides the necessary data about the number of stamps Juan has for each decade and country. The relevant part of the table corresponds to the South American countries (Brazil, Peru) issued before the '70s, which are the '50s, '60s, and '70s decades, and their respective countries.

Each cell in the table represents specific information: the count of stamps issued before the '70s for a particular country. Juan's cost for the stamps can be computed by multiplying the number of stamps from South American countries (Brazil and Peru) purchased before the '70s by their respective prices (6 cents for Brazil and 4 cents for Peru).



Hints:

topic_specific_hint:

To solve the arithmetic problem, systematically extract quantitative information from both visual and textual sources. Identify objects, patterns, counts, and spatial relationships that can be translated into numerical values. Determine the objective (e.g., total value, missing term, comparison) and map visual elements to algebraic representations or equations. Cross-check for background elements (e.g., overlapping objects, symmetry, groupings, or hidden features) that may influence the count or operation. Apply basic arithmetic operations (addition, subtraction, multiplication, division) in clear, logical steps, and ensure consistency with visual evidence. Clearly state the inferred numerical value of each symbol or object and how they relate through equations. For multiple-choice questions, use elimination or plug-in strategies. Avoid visual misinterpretation or overreliance on superficial similarity; prioritize logical modeling, numerical precision, and verification through reverse calculation or consistency checks.

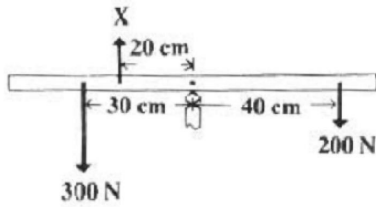
Question_specific_hint: Focus on the "Brazil" and "Peru" rows for stamps from the 50s and 60s. Multiply the number of stamps by their respective costs (Brazil: 6 cents, Peru: 4 cents), sum the total cost, and convert to dollars and cents.



Answer: B

Figure 14: The example of arithmetic question in MIRAGE.

Scientific



Question:

A uniform meterstick is balanced at its midpoint with several forces applied as shown below. If the stick is in equilibrium, the magnitude of the force X in newtons (N) is

- | | | |
|----------|----------|----------|
| A. 250 N | B. 150 N | C. 300 N |
| D. 200 N | E. 400 N | F. 100 N |
| G. 500 N | H. 600 N | I. 50 N |
| J. 350 N | | |

Reasoning Chains:

To solve for X, we need to use the principle of moments (torque equilibrium). The sum of moments (torques) about any point should be zero for equilibrium.

We'll take moments about the midpoint of the stick. The distances of the forces from the midpoint are:

- The 300 N force is 30 cm to the left of the midpoint.
- The 200 N force is 40 cm to the right of the midpoint.
- The X force is 20 cm to the right of the midpoint.

The moment caused by a force is given by the force multiplied by the distance from the pivot. In this case, we will set up the equation for the clockwise and counterclockwise torques.

The torque due to the 300 N force will be:

$$\text{Torque}_1 = 300 \times 30 = 9000 \text{ N} \cdot \text{cm}$$

This is a counterclockwise moment.

The torque due to the 200 N force will be:

$$\text{Torque}_2 = 200 \times 40 = 8000 \text{ N} \cdot \text{cm}$$

This is a clockwise moment.

The torque due to the X force will be:

$$\text{Torque}_3 = X \times 20 = 20X \text{ N} \cdot \text{cm}$$

This is also a clockwise moment.

For equilibrium, the sum of torques in both directions must be equal:

$$\text{Torque}_1 = \text{Torque}_2 + \text{Torque}_3$$

Substitute the values:

$$9000 = 8000 + 20X$$

Solving for X:

$$9000 - 8000 = 20X$$

$$1000 = 20X$$

$$X = 1000 / 20 = 50 \text{ N}$$

Thus, the magnitude of force X is 50 N.

Intermediate Results:

Identify forces and their distances from the midpoint: 300 N (30 cm left), 200 N (40 cm right), X (20 cm right)

Forces and distances: 300 N (30 cm left), 200 N (40 cm right), X (20 cm right)

Calculate torque due to 300 N force: $300 \times 30 = 9000 \text{ N} \cdot \text{cm}$ (counterclockwise)

Torque₁ = 9000 N · cm

Calculate torque due to 200 N force: $200 \times 40 = 8000 \text{ N} \cdot \text{cm}$ (clockwise)

Torque₂ = 8000 N · cm

Calculate torque due to X force: $X \times 20 = 20X \text{ N} \cdot \text{cm}$ (clockwise)

Torque₃ = 20X N · cm

Set up torque equilibrium equation: $9000 = 8000 + 20X$

Equation: $9000 = 8000 + 20X$

Solve for X: $9000 - 8000 = 20X \rightarrow 1000 = 20X \rightarrow X = 50 \text{ N}$

X = 50 N

the answer is 50 N

the answer is 50 N

Hints:

Topic-specific hint: To solve a scientific problem, identify objectives, key variables, and relationships. Distinguish knowns from unknowns, use diagrams and formulas effectively, and apply relevant principles accurately. Break complex steps into parts, ensure unit consistency, and check result reasonableness. For multiple-choice questions, use strategies like elimination or back-checking. Avoid common errors and maintain clarity and rigor throughout.

Question-specific hint: Determine the torques around the midpoint on each side of the meterstick. Use the formula torque = force × distance (from the pivot), ensuring that counterclockwise and clockwise torques are balanced. Consider the distances 20 cm, 30 cm, and 40 cm for each relevant force.

Description

The diagram shows a uniform meterstick balanced at its midpoint (50 cm mark) with three forces acting on it. A 300 N downward force is applied at the 20 cm mark (30 cm to the left of the fulcrum), and a 200 N downward force is applied at the 90 cm mark (40 cm to the right of the fulcrum). An unknown upward force, labeled X, acts at the 30 cm mark (20 cm to the left of the fulcrum). The system is in equilibrium, meaning the sum of clockwise and counterclockwise torques about the fulcrum is zero.

Answer: I

Figure 15: The example of scientific question in MIRAGE.

Spatial



Reasoning Chains:

In the image, the train is clearly closer to the camera and situated along the track, while the bench is further back on the platform. Therefore, the bench is not positioned in front of the train. It is located behind the train relative to the viewpoint of the image. This suggests that the spatial relationship described in the caption is incorrect.



Hints:

topic_specific_hint: To solve the spatial reasoning problem, carefully interpret both textual and visual inputs to understand object attributes (e.g., shape, color, size) and spatial relations (e.g., left/right, behind/in front, inside/above/below). Systematically extract the objective (e.g., identify, compare, count, color) and all given relational facts. Construct a mental or symbolic spatial map to track object positions and interrelations. Use consistent reference points (e.g., camera view, object centers) and verify directional logic step-by-step. Pay attention to subtle qualifiers like "exactly," "same," or "nearest." For complex scenes, decompose multi-object relationships into pairwise comparisons. Avoid assumptions about 3D depth without explicit cues. Rigorously validate inferred positions or relations using all available evidence. In multiple-choice scenarios, use elimination and contradiction checks to isolate the most consistent answer.

Question_specific_hint: Identify the position of the bench in relation to the train. Check if the bench is positioned in front of the train's direction of movement or the front end.



Question:

`<image>` Please determine whether the spatial relationships described in the captions below are correct based on the image.

Caption: The bench is in front of the train.

A: Correct

B: Error



Intermediate Results:

`<step>`Analyze the spatial relationship between the train and the bench in the image`</step>`

`<claim>`The train is closer to the camera and situated along the track`</claim>`

`<step>`Determine the position of the bench relative to the train`</step>`

`<claim>`The bench is further back on the platform, behind the train relative to the viewpoint`</claim>`

`<step>`Evaluate the accuracy of the caption's claim`</step>`

`<claim>`The caption's claim that the bench is in front of the train is incorrect`</claim>`

`<step>`the answer is B`</step>`

`<claim>`the answer is B`</claim>`



Description

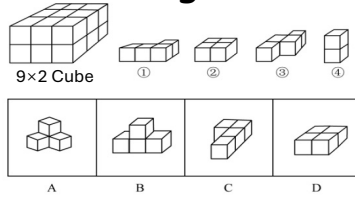
This image shows a train station with a train passing through. The foreground shows a portion of the platform with a track, and there are some benches visible along the platform. The train is moving in the direction of the camera. The caption claims that the bench is in front of the train. We need to evaluate whether this spatial relationship is accurate based on the image.



Answer: B

Figure 16: The example of spatial question in MIRAGE.

Logical



Question:

From the given four options, select the one that can form a 9x2 block combination as shown in the figure with the given parts ①, ②, ③, and ④.

Reasoning Chains:

The problem involves finding the length of side BC in trapezoid $ABCD$ with given area, altitude, and lengths of sides AB and CD . Here's the step-by-step solution:

1. Trapezoid Area Formula:

The area of a trapezoid is given by:

$$A = \frac{(b_1 + b_2)}{2} \times \text{height}$$

Here, b_1 and b_2 are the lengths of the two parallel bases. The given area is 164 , and the height (altitude) is 8 .

$$164 = \frac{(b_1 + b_2)}{2} \times 8 \implies b_1 + b_2 = \frac{164 \times 2}{8} = 41$$

2. Identify the Bases:

$AB = 10$ and $CD = 17$ are the non-parallel sides (legs). Let AD and BC be the two bases. Thus:

$$AD + BC = 41$$

3. Horizontal Projections of Legs:

When dropping perpendiculars from A and D to BC , forming right triangles:

- For AB :

$$\text{Horizontal segment} = \sqrt{AB^2 - \text{height}^2} = \sqrt{10^2 - 8^2} = \sqrt{36} = 6$$

- For CD :

$$\text{Horizontal segment} = \sqrt{CD^2 - \text{height}^2} = \sqrt{17^2 - 8^2} = \sqrt{225} = 15$$

4. Relate Bases:

The total horizontal difference between the bases is $6 + 15 = 21$.

$$BC = AD + 21$$

Substituting $AD = 41 - BC$ into the equation:

$$BC = (41 - BC) + 21 \implies 2BC = 62 \implies BC = 31$$

Final Answer: 31

Intermediate Results:

Apply trapezoid area formula:
 $164 = \frac{(b_1 + b_2)}{2} \times 8 \rightarrow b_1 + b_2 = 41$

Identify non-parallel sides AB and CD :
 $AB = 10$ cm, $CD = 17$ cm

Calculate horizontal projections of legs AB and CD :
 Horizontal segment for $AB = \sqrt{10^2 - 8^2} = 6$ cm
 Horizontal segment for $CD = \sqrt{17^2 - 8^2} = 15$ cm

Compute total horizontal difference: $6 + 15 = 21$
 $BC = AD + 21$

Substitute $AD = 41 - BC$ into $BC = AD + 21$:
 $BC = (41 - BC) + 21 \rightarrow 2BC = 62 \rightarrow BC = 31$ cm

the answer is C
 the answer is C

Description

The diagram shows a uniform meterstick balanced at its midpoint (50 cm mark) with three forces acting on it. A 300 N downward force is applied at the 20 cm mark (30 cm to the left of the fulcrum), and a 200 N downward force is applied at the 90 cm mark (40 cm to the right of the fulcrum). An unknown upward force, labeled X, acts at the 30 cm mark (20 cm to the left of the fulcrum). The system is in equilibrium, meaning the sum of clockwise and counterclockwise torques about the fulcrum is zero.



Answer: C

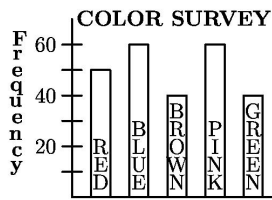
Hints:

Topic-specific hint: To solve the visual logic problem, analyze the image grid systematically by comparing visual components across rows, columns, or diagonals. Identify changes in shape, position, count, color, size, or orientation. Look for consistent rules or transformations such as rotation, reflection, addition, subtraction, XOR, or pattern shifts. Describe each pattern evolution step-by-step using symbolic or relational expressions (e.g., "Image C = Image A XOR Image B", or "Right = Left rotated 90° clockwise"). Clearly state the inferred rule and apply it to missing or target positions. When multiple rules are possible, list and compare them, selecting the one that best fits all visible patterns. For multiple-choice questions, eliminate inconsistent options by checking each against the rule. Avoid relying on visual intuition alone—prioritize rule-based reasoning, pattern consistency, and explicit justification for your answer.

Question-specific hint: Identify the number and arrangement of cubes needed to recreate the 9x2 block from the smaller parts ①, ②, ③, and ④. Compare each option (A, B, C, D) with the target shape to find a matching configuration.

Figure 17: The example of logical question in MIRAGE.

Statistical



Question:

<image> The bar graph shows the results of a survey on color preferences. What percent preferred blue?
 A: \$20\%\$ B: \$24\%\$
 C: \$30\%\$ D: \$36\%\$
 E: \$42\%\$



Reasoning Chains:

To determine the percentage of people who preferred blue in the survey, we need to analyze the bar graph and calculate the proportion of the frequency for blue relative to the total frequency.

Step 1: Identify the frequencies from the bar graph
 The bar graph shows the following frequencies for each color:
 - Red: ~50 - Blue: ~60 - Brown: ~60 - Pink: ~40 - Green: ~40

Step 2: Calculate the total frequency
 Add up the frequencies of all the colors:
 $[50 + 60 + 60 + 40 + 40 = 250]$
 So, the total number of responses is 250.

Step 3: Find the frequency for blue
 From the graph, the frequency for blue is approximately 60.

Step 4: Calculate the percentage for blue
 The percentage is calculated as:

$$\left[\frac{\text{Frequency of blue}}{\text{Total frequency}} \right] \times 100$$

 Substitute the values:

$$\left[\frac{60}{250} \right] \times 100 = 0.24 \times 100 = 24\%$$

Step 5: Match with the given options
 The calculated percentage is 24%, which matches option B.



Intermediate Results:

<step>Identify the frequencies from the bar graph</step>
 <claim>Red: ~50, Blue: ~60, Brown: ~60, Pink: ~40, Green: ~40</claim>

<step>Calculate the total frequency</step>
 <claim> $50 + 60 + 60 + 40 + 40 = 250$ </claim>

<step>Find the frequency for blue</step>
 <claim>Frequency for blue is approximately 60</claim>

<step>Calculate the percentage for blue</step>
 <claim>Percentage for blue = $(60/250) \times 100 = 24\%$ </claim>

<step>Match with the given options</step>
 <claim>The calculated percentage is 24%, which matches option B</claim>

<step>the answer is B: 24%</step>
 <claim>the answer is B: 24%</claim>



Hints:

topic_specific_hint:

To solve the statistical problem, carefully examine all visual representations (e.g., bar charts, line graphs, tables, pie charts) and extract precise numerical values and labels. Identify the objective (e.g., compute, compare, interpret, predict) and determine the relevant data segments. Cross-check axes, scales, units, and legends to avoid misreading quantities. Look for patterns, trends, distributions, and outliers across categories or time. Apply appropriate statistical operations (e.g., mean, median, mode, range, percentage change, probability) based strictly on the data provided. For compound questions, decompose tasks and perform step-by-step computations. Use estimation, elimination, or reverse-checking in multiple-choice settings. Avoid assumptions not grounded in the visual/textual evidence. Prioritize accuracy in data extraction, clarity in method, and logical coherence in interpreting results.

Question_specific_hint: Identify the height of the bar labeled "BLUE" to find the frequency. Then calculate the percentage based on the total frequency sum of all colors.



Description

The image features a bar graph titled "COLOR SURVEY." The horizontal axis lists different color options (Green, Red, Brown, Pink, Blue), and the vertical axis represents frequency, labeled as "Frequency." The height of each bar represents the frequency of each color mentioned. For instance, the Red color has a bar reaching up to 45 on the frequency scale, and the Blue color has a bar that reaches 60 on the same scale. This bar graph is necessary to solve the problem because it visually represents the preferences for each color, and these preferences are used to calculate the percentage that preferred Blue.

To further break down the bar graph, each color has an equivalent label at the base of its corresponding bar on the horizontal axis. The color preferences are ordered from left to right as Red (approx. 45), Brown (approx. 55), Blue (approx. 60), Pink (approx. 25), and Green (approx. 30). The frequencies are annotated with individual bars above each color.



Answer: I

Figure 18: The example of statistical question in MIRAGE.

Algebraic: To solve the algebraic problem, focus on extracting and interpreting textual information to define variables, expressions, and equations. Identify the objective (e.g., solve for a variable, simplify an expression, evaluate an equation) and distinguish known values, unknowns, and their relationships. Translate word problems or symbolic descriptions into formal algebraic expressions. If an image is present, assess whether it provides meaningful cues or serves as a placeholder; prioritize textual clarity when visual content is ambiguous or abstract. Apply appropriate algebraic techniques (e.g., substitution, factoring, combining like terms, solving systems) in clear logical steps. Clearly define each variable and explain the transformations at each stage.

Geometry: To solve the geometry problem, systematically integrate visual and textual data: cross-verify labels, lengths, angles, and relationships, prioritizing explicit textual details over visual estimates. Identify key components (objective, given data) and apply relevant theorems (e.g., Pythagorean theorem, circle properties, coordinate formulas). Break computations into logical steps, using auxiliary lines or algebraic variables as needed. Verify unit consistency, answer reasonableness, and cross-check via alternative methods.

Arithmetic: To solve the arithmetic problem, systematically extract quantitative information from both visual and textual sources. Identify objects, patterns, counts, and spatial relationships that can be translated into numerical values. Determine the objective and map visual elements to algebraic representations or equations. Cross-check for background elements (e.g., overlapping objects, symmetry, groupings, or hidden features) that may influence the count or operation. Apply basic arithmetic operations (addition, subtraction, multiplication, division) in clear, logical steps, and ensure consistency with visual evidence. Clearly state the inferred numerical value of each symbol or object and how they relate through equations.

Scientific: To solve the scientific problem, carefully extract the objective and identify key entities, variables, and relationships described in the prompt. Distinguish between known quantities, assumptions, and what needs to be derived. Integrate textual data with diagrams or formulas where provided, and prioritize explicit numerical or symbolic information over vague intuitions. Apply relevant scientific principles (e.g., Newton's laws, conservation laws, reaction equations, biological processes) precisely and in context. Break down complex processes into sequential steps—use dimensional analysis, algebraic manipulation, or system modeling as needed. Ensure unit consistency and check for reasonable magnitudes in your answer.

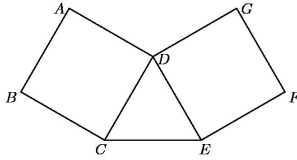
Logical: To solve the visual logic problem, analyze the image grid systematically by comparing visual components across rows, columns, or diagonals. Identify changes in shape, position, count, color, size, or orientation. Look for consistent rules or transformations such as rotation, reflection, addition, subtraction, XOR, or pattern shifts. Describe each pattern evolution step-by-step using symbolic or relational expressions (e.g., "Image C = Image A XOR Image B", or "Right = Left rotated 90° clockwise"). Clearly state the inferred rule and apply it to missing or target positions. When multiple rules are possible, list and compare them, selecting the one that best fits all visible patterns.

Spatial: To solve the spatial reasoning problem, carefully interpret both textual and visual inputs to understand object attributes (e.g., shape, color, size) and spatial relations (e.g., left/right, behind/in front, inside/above/below). Systematically extract the objective (e.g., identify, compare, count, color) and all given relational facts. Construct a mental or symbolic spatial map to track object positions and interrelations. Use consistent reference points (e.g., camera view, object centers) and verify directional logic step-by-step. Pay attention to subtle qualifiers like "exactly," "same," or "nearest." For complex scenes, decompose multi-object relationships into pairwise comparisons. Avoid assumptions about 3D depth without explicit cues.

Statistical: To solve the statistical problem, carefully examine all visual representations (e.g., bar charts, line graphs, tables, pie charts) and extract precise numerical values and labels. Identify the objective (e.g., compute, compare, interpret, predict) and determine the relevant data segments. Cross-check axes, scales, units, and legends to avoid misreading quantities. Look for patterns, trends, distributions, and outliers across categories or time. Apply appropriate statistical operations (e.g., mean, median, mode, range, percentage change, probability) based strictly on the data provided.

Figure 19: The topic-specific hints used in MIRAGE.

Geometry



Question:

In the adjoining figure, CDE is an equilateral triangle and ABCD and DEFG are squares.

<image1>

The measure of $\angle GDA$ is

** Choices: **

A: 90°

B: 105°

C: 120°

D: 135°

E: 150°

Qwen2.5-VL-7B-Instruct

Let's analyze the problem step-by-step to find the measure of $\angle GDA$.

Step-by-Step Analysis

1. **Identify the Geometric Shapes and Their Properties:**

- $\square ABCD$ and $\square DEFG$ are squares.
- $\triangle CDE$ is an equilateral triangle.
- All sides of squares are equal, and each angle in a square is 90° .
- Each angle in an equilateral triangle is 60° .

2. **Consider the Angles at Point D:**

- The angles at point D from the squares $\square ABCD$ and $\square DEFG$ are:
 $\angle ADC = 90^\circ$, $\angle CDE = 90^\circ$, $\angle EDF = 90^\circ$
- The angle at point D from the equilateral triangle $\triangle CDE$ is:
 $\angle CDE = 60^\circ$

3. **Sum the Angles at Point D:**

- The total sum of the angles at point D from the squares is:
 $90^\circ + 90^\circ + 90^\circ = 270^\circ$ [Factual Hallucination Here, only two 90° angles]
- Adding the angle from the equilateral triangle:
 $270^\circ + 60^\circ = 330^\circ$ [wrong answers from previous factual hallucination, but logic in this step is correct.]

4. **Determine the Angle $\angle GDA$:

- The total angle sum around any point is 360° . Therefore, the measure of $\angle GDA$ is the difference between 360° and the sum of the other angles at D:
 $360^\circ - 330^\circ = 30^\circ$

Given the options provided:

A: 90°

B: 105°

C: 120°

D: 135°

E: 150°

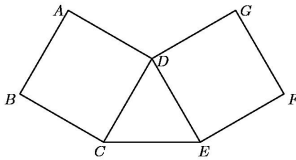
The correct angle measure for $\angle GDA$ is 30° .

Final Answer

$\boxed{30^\circ}$

Figure 20: Response example from Qwen2.5-VL-7B-Instruct. Red font means reasoning hallucination and corresponding judgement.

Geometry



Question:

In the adjoining figure, CDE is an equilateral triangle and ABCD and DEFG are squares.

The measure of $\angle GDA$ is

** Choices: **

A: 90°

B: 105°

C: 120°

D: 135°

E: 150°

Logos-7B

think Given that $\triangle CDE$ is an equilateral triangle, we know that all its interior angles are 60° . Therefore, $\angle ECD = \angle CDE = \angle DEC = 60^\circ$.

Since $ABCD$ and $DEFG$ are squares, we know that all their interior angles are 90° . Therefore, $\angle ACD = 90^\circ$ and $\angle EDG = 90^\circ$.

We need to find $\angle GDA$. To do this, we can determine the measure of $\angle GDC$. Since $\angle ECD = 60^\circ$ and $\angle ACD = 90^\circ$, the angle $\angle GDC$ is:

$\angle GDC = 90^\circ - 60^\circ = 30^\circ$.

Since $\angle EDG = 90^\circ$, the angle $\angle GDA$ is:

$\angle GDA = 90^\circ + 30^\circ = 120^\circ$.

Thus, the measure of $\angle GDA$ is 120° .

Therefore, the answer is:

think

answer C

Gemini-2-flash-thinking

Rationale:

We are given that ABCD is a square, CDE is an equilateral triangle, and DEFG is a square. We need to find the measure of $\angle GDA$.

Since ABCD is a square, $\angle ADC = 90^\circ$.

Since CDE is an equilateral triangle, $\angle CDE = 60^\circ$.

Since DEFG is a square, $\angle GDE = 90^\circ$.

The angles $\angle ADC$, $\angle CDE$, $\angle GDE$ are consecutive angles around point D.

The sum of the angles $\angle ADC$, $\angle CDE$, $\angle GDE$ is $\angle ADG$ (reflex angle).

$\angle ADG_{\text{reflex}} = \angle ADC + \angle CDE + \angle GDE = 90^\circ + 60^\circ + 90^\circ = 240^\circ$.

The angle $\angle GDA$ is the other angle between the lines DA and DG.

The sum of the angles around a point is 360° . Therefore, the sum of $\angle ADG_{\text{reflex}}$ and $\angle GDA$ is 360° .

$\angle GDA = 360^\circ - \angle ADG_{\text{reflex}} = 360^\circ - 240^\circ = 120^\circ$.

Thus, the measure of $\angle GDA$ is 120° .

Final Answer: The final answer is \boxed{C}

Figure 21: Response examples from Logos-7B and Gemini-2-flash-thinking.