
More Thinking, Less Seeing? Assessing Amplified Hallucination in Multimodal Reasoning Models

Zhongxing Xu^{1,3*} Chengzhi Liu^{2*} Qingyue Wei¹ Juncheng Wu³
James Zou¹ Xin Eric Wang² Yuyin Zhou³ Sheng Liu¹
¹Stanford University ²UC Santa Barbara ³UC Santa Cruz

chengzhi@ucsb.edu, zhongxingxu1@gmail.com, shengl@stanford.edu

Abstract

Test-time compute has empowered multimodal large language models to generate extended reasoning chains, yielding strong performance on tasks such as multimodal math reasoning. However, we observe that this improved reasoning ability often comes with increased hallucination: as generations become longer, models tend to drift away from image-grounded content and rely more on language priors. Attention analysis reveals that longer reasoning chains reduce focus on visual inputs, contributing to hallucination. To systematically study this phenomenon, we introduce *RH-AUC*, a metric that quantifies how a model’s perception accuracy changes with reasoning length, enabling evaluation of whether the model preserves visual grounding while reasoning. We also release *RH-Bench*, a diagnostic benchmark covering diverse multimodal tasks, designed to jointly assess the balance of reasoning ability and hallucination. We find that (i) larger models generally exhibit a better balance between reasoning and perception; (ii) reasoning and perception balance depends more on the types and domains of the training data than its volume. Our findings highlight the need for evaluation frameworks that account for both reasoning quality and perceptual reliability.

1 Introduction

Large reasoning models scale test-time computation to improve complex reasoning. These models [6, 7, 32, 2] generate longer outputs and engage in deeper reasoning before producing final answers, resulting in more comprehensive solutions for complex mathematical and scientific problems. This paradigm has been extended to multimodal large language models: non-reasoning base models are supervised finetuned (SFT), or finetuned with reinforcement learning (RL) to obtain strong reasoning ability [36, 56, 10, 49, 54, 21], demonstrating exceptional capabilities in multimodal reasoning tasks, particularly in domains like mathematical problem solving.

Most existing studies on multimodal reasoning models focus on enhancing reasoning performance, with limited attention paid to perception-focused tasks. As illustrated in Figure 1a, although the reasoning model generates an extended reasoning chain in visual question answering, its answer is largely driven by language priors rather than visual evidence, leading to hallucination. Our empirical study reveals a consistent and significant finding: although reasoning models can generate more detailed reasoning chains, they introduce more hallucinations in perception-focused tasks than the non-reasoning counterparts, as shown in Figure 1b.

Through attention analysis, we investigate the decrease of attention on visual tokens in multimodal reasoning models, which exacerbates visual hallucinations. The reasoning model allocates significantly less attention to visual tokens compared to its non-reasoning counterpart, while directing more attention to the instruction tokens. This bias increases reliance on language priors and amplifies

¹<https://mlrm-halu.github.io/>. Work was partially done while ZX was visiting Stanford.

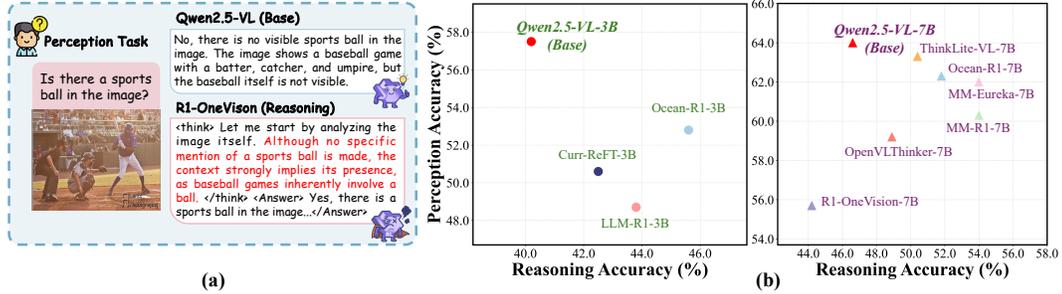


Figure 1: (a) Example of outputs from a reasoning model and a non-reasoning model on a perception task. Red highlights indicate visual hallucination. *Multimodal reasoning models are generally more prone to amplifying hallucinations during the reasoning process compared to their non-reasoning counterparts.* (b) Performance of different models on reasoning and perception tasks in the *RH-Bench* dataset. Better performing models are positioned in the upper right corner. *Baseline non-reasoning models of varying scales typically exhibit weaker reasoning capabilities and fewer hallucination, whereas reasoning models display the opposite trend.*

hallucination risk. Moreover, the extension of the reasoning chain further weakens the visual attention allocation, leading to an increase in hallucinations, as the model becomes more dependent on language-based reasoning rather than visual evidence.

Based on these findings, we further investigate the impact of reasoning chain length on model reasoning and hallucination. The results indicate that the influence of reasoning chain length on reasoning-hallucination exhibits a non-monotonic relationship. Additionally, the optimal reasoning range differs across tasks, while traditional evaluation metrics, such as accuracy and hallucination rate, are inadequate for capturing the dynamic balance between reasoning and visual grounding.

To address this, we introduce *RH-AUC*, a new metric designed to assess the balance between reasoning and hallucination in multimodal reasoning models. This metric is computed by calculating the area under the curve formed by reasoning performance and hallucination performance at different reasoning lengths, with higher values indicating better balance. Alongside this metric, we release *RH-Bench*, a diagnostic benchmark containing 1,000 samples across various reasoning and perception tasks, with each task featuring both multiple-choice questions and open-ended questions. Through the evaluation of *RH-Bench*, we observe three key findings: (i) Larger models typically demonstrate better reasoning and hallucination balance. (ii) RL-only training models promote more adaptive reasoning, resulting in a better balance between reasoning and hallucination compared to SFT+RL. (iii) Reasoning-Hallucination balance is more influenced by the types and domains of the training data than by its volume. To sum up, our contributions are listed as follows:

- We observe that multimodal reasoning models are more prone to hallucinations than their non-reasoning counterparts in perception tasks, which can be attributed to a decline in visual attention allocation. Longer reasoning chains further diminish visual attention.
- We reveal that the relationship between reasoning chain length and the model’s reasoning and perception performance is non-monotonic, with the optimal length varying across tasks.
- We introduce the new *RH-AUC* metric and the *RH-Bench* diagnostic dataset to systematically evaluate the balance between reasoning and hallucination across varying reasoning lengths in multimodal reasoning models.

2 Multimodal Reasoning Can Amplify Visual Hallucination

In this section, we begin by investigating whether multimodal reasoning models introduce more hallucination in perception-focused tasks. Specifically, we compare 8 recent multimodal reasoning models against their backbone non-reasoning-based counterparts across multiple hallucination benchmarks, including MMVP [42], MMEval-Pro [13],VMCBench [61],Bingo [5],MMHAL [39].

2.1 Hallucination Increases Consistently Compared to Base Models

To systematically assess the impact of multimodal reasoning on visual grounding, we evaluated eight reasoning-augmented models against their non-reasoning Qwen2.5-VL backbones on five

hallucination datasets. As shown in Figure 2, all reasoning models trace markedly smaller radar areas than their baselines, indicating uniformly higher hallucination rates on perception-focused tasks. This deficit remains consistent at both the 3 B and 7 B scales, demonstrating that the elevated hallucination rate stems from the reasoning paradigm itself rather than model size.

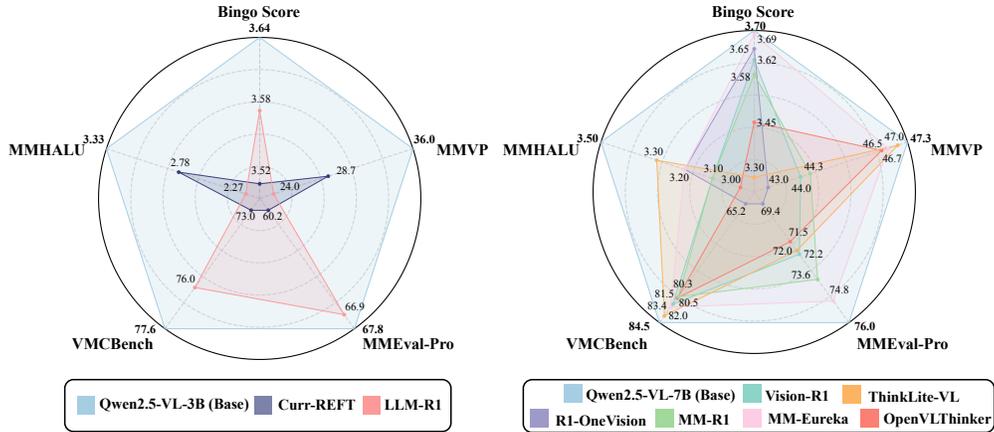


Figure 2: Comparison of reasoning and non-reasoning models on five perception benchmarks. Results are shown for 3B models (left) and 7B models (right). Higher scores indicate lower hallucination.

2.2 Does Training Paradigm Matter? Comparison Between RL and SFT+RL

Current multimodal reasoning models typically adopt one of two training regimes: (1) pure reinforcement learning (RL-only) or (2) supervised fine-tuning followed by reinforcement learning (SFT+RL). Figure 3 shows a consistent performance hierarchy across four perception benchmarks: The Qwen2.5-VL baseline achieves the highest scores, followed by RL-only fine-tuning, with the SFT+RL pipeline performing the worst. This pattern highlights the robustness of baseline model in visual grounding and indicates that subsequent RL or hybrid fine-tuning weakens this robustness, with the supervised-preceded RL strategy leading to the most significant performance degradation.

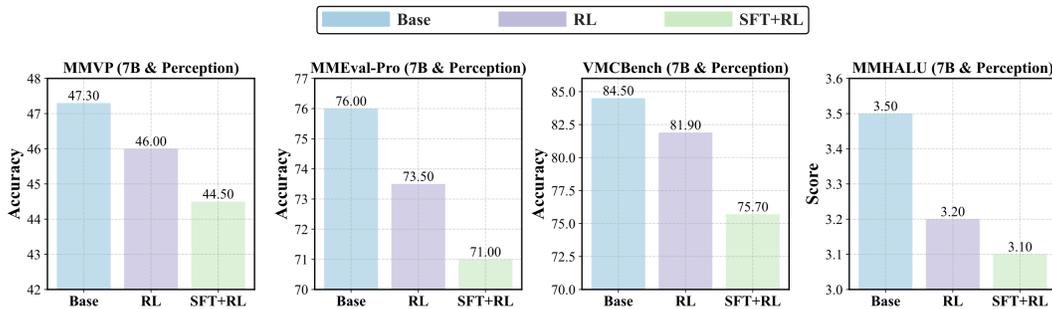


Figure 3: Performance across four perception benchmarks comparing Base, RL, and SFT+RL.

2.3 Case Study of Hallucinations in Reasoning Models

Figure 4 presents two representative hallucination patterns observed in multimodal reasoning models, arising from visual misrecognition and reasoning bias, respectively. In Figure 4a, the reasoning model fails to identify fine-grained visual cues and miscounts four individuals as three, reflecting a localized deficiency in visual perception. In Figure 4b, the reasoning model increasingly relies on linguistic priors during the reasoning process while overlooking early visual evidence, ultimately generating an incorrect response. In contrast, the baseline model exhibits a lower hallucination rate under identical inputs. These observations raise a crucial question: why do multimodal reasoning models, despite their strong reasoning performance, exhibit weakened visual grounding? In the next section, we provides an in-depth analysis based on the internal attention mechanisms of the reasoning models.



Figure 4: Two common types of hallucination patterns observed in multimodal reasoning models. (a) corresponds to hallucinations caused by visual misrecognition, while (b) reflects hallucinations arising from reasoning biases. Hallucinated spans are highlighted in red.

Takeaway 1: Reasoning Models Amplify Visual Hallucinations

Across training paradigms and model scales, multi-modal reasoning models exhibit a consistent drop in accuracy and rise in hallucination rates on general visual benchmarks.

3 Why Reasoning Models Amplify Hallucinations?

Many previous studies have investigated the role of attention mechanisms in hallucination, identifying insufficient attention allocation as a potential key factor contributing to hallucinations [14, 16, 53]. In this section, we conduct an attention based analysis to explore the underlying causes of hallucination amplification in multimodal reasoning models. Section 3.1 indicates that hallucinations may result from limited attention allocated to visual inputs, while Section 3.2 shows that longer reasoning chains further weaken the model’s visual focus.

3.1 Hallucination Resulting from Weak Visual Attention

We conduct a comparative analysis of the attention distributions over visual, instruction, and system tokens across all layers in the reasoning and non-reasoning models. As shown in Figure 5a, the reasoning model consistently assigns low attention to visual tokens, with a further decrease observed in deeper layers, indicating a limited ability to integrate visual evidence. Meanwhile, more attention is shifted to instruction tokens, reflecting a heightened reliance on linguistic priors. In contrast, the non-reasoning model maintains a relatively high and stable level of visual attention from shallow to intermediate layers. The visual attention heatmap in Figure 5b further supports this observation: while the non-reasoning model progressively focuses on semantically salient regions, the reasoning model exhibits sparse and dispersed attention, failing to consistently engage with key visual areas. This phenomenon indicates that the weakening of visual attention undermines the reasoning model’s ability to achieve effective visual grounding, exacerbating the occurrence of hallucinations.

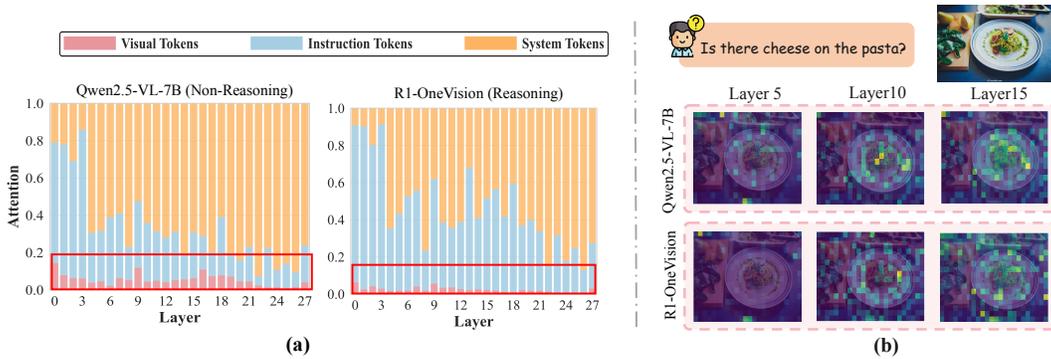


Figure 5: Attention allocation and visual grounding between reasoning and non reasoning models. *The reduction of visual attention in reasoning models amplifies visual hallucinations.*

3.2 Visual Focus Declines with Longer Reasoning Chains

As shown in Figure 6, we visualize the attention distributions of the reasoning model under two reasoning modes: normal thinking and overthinking. As the reasoning chain length increases, the heatmaps clearly reveal a systematic shift in the model’s attention focus: under the overthinking mode, attention to visual tokens significantly decreases, while attention to instruction tokens intensifies. This pattern indicates that longer reasoning chains cause the model to increasingly rely on linguistic cues rather than grounded visual evidence. For instance, when asked whether a gray wall is present, the model under normal thinking correctly identifies the gray wall and provides a correct response. In contrast, under over-reasoning conditions, the model exhibits further diminished attention to visual tokens, with increased focus directed toward the end of the user instruction. This suggests that longer reasoning chains tend to further exacerbate the degradation of the model’s visual grounding, potentially leading to an increase in hallucinations.

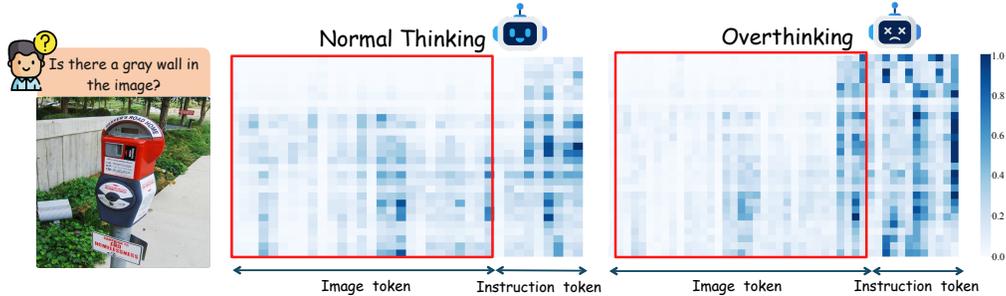


Figure 6: Attention shift in the reasoning model under different reasoning length. In normal thinking, the model generates outputs as typically expected, while in overthinking, the reasoning length is adjusted using Latent State Steering (Section 4.1). *Longer reasoning chains further exacerbate the degradation of attention to visual information and focus toward linguistic priors.*

4 Effects of Reasoning Length on Reasoning-Hallucination Balance

In this section, we explore the impact of reasoning length on the balance between hallucination and reasoning. Section 4.1 provides an overview of the proposed control strategy: latent state steering as well as techniques that are previously used in the literature[31]: budget forcing, and test time scaling. In Section 4.2, we explore the optimal generation length for various benchmarks and analyze the trade-off between hallucination and reasoning performance as reasoning length varies.

4.1 Overview of Reasoning Length Control Strategies

To systematically control the reasoning length in reasoning models, we adopt three strategies:

- (1) *Token Budget Forcing*: A hard constraint on reasoning length is enforced by predefining a generation budget at decoding time, directly limiting the number of tokens allocated for the reasoning.
- (2) *Test Time Scaling*: Reasoning is incrementally extended during inference through staged generation. The model first produces partial reasoning under a 4096-token constraint and halts midway. It is then prompted to continue by appending a simple token ("Wait"), enabling soft extension of reasoning while preserving contextual coherence.
- (3) *Latent State Steering*: Inspired by recent works on latent space steering for behavior control in large language models [23, 22, 1, 29], we propose a method to steer the model toward generating reasoning traces of varying lengths. Specifically, we extract steering directions from the post-attention hidden states by computing the difference of latent states between long and short reasoning trajectories. These direction vectors are obtained and applied across all layers of the text decoder, with a scaling factor controlling both the magnitude of guidance on the reasoning length. Specifically, we collect responses from the test benchmark and categorize them into long reasoning traces $\mathcal{R}_{\text{long}}$ and short reasoning traces $\mathcal{R}_{\text{short}}$ based on token length. The query and reasoning steps for each sample are input into the model, from which hidden representations S^ℓ are extracted at each layer. $S^\ell(q, t)$ denotes the hidden representation at layer ℓ for token position t in the response to query q . We

compute the average hidden representation over reasoning tokens, where \mathcal{H}_i represents the set of token positions within the reasoning span. The average representation is then calculated across the long and short reasoning traces to obtain layerwise embeddings:

$$\mathcal{S}_{\text{long}}^\ell = \frac{1}{|\mathcal{R}_{\text{long}}|} \sum_{q \in \mathcal{R}_{\text{long}}} \frac{1}{|\mathcal{H}_i|} \sum_{t \in \mathcal{H}_i} \mathcal{S}^\ell(q, t), \quad \mathcal{S}_{\text{short}}^\ell = \frac{1}{|\mathcal{R}_{\text{short}}|} \sum_{q \in \mathcal{R}_{\text{short}}} \frac{1}{|\mathcal{H}_i|} \sum_{t \in \mathcal{H}_i} \mathcal{S}^\ell(q, t) \quad (1)$$

The reasoning length direction at layer ℓ is defined as the difference between the long and short embeddings, denoted as d^ℓ , which captures the variation in the model’s representation resulting from different reasoning chain lengths. To adjust the hidden representation based on this direction, we introduce a parameter $\alpha \in [-0.15, 0.15]$ to dynamically control the reasoning length and its magnitude. As α increases, the length of the reasoning chain extends, as shown below:

$$d^\ell = \mathcal{S}_{\text{long}}^\ell - \mathcal{S}_{\text{short}}^\ell, \quad \mathcal{S}_{\text{steering}}^\ell = \mathcal{S}^\ell + \alpha d^\ell. \quad (2)$$

These strategies are applied to five representative multimodal reasoning models and evaluated on six benchmark datasets, covering both reasoning and perception tasks. In Figure 7, we present two benchmarks for both tasks. All implementation details and results are provided in Appendix C.

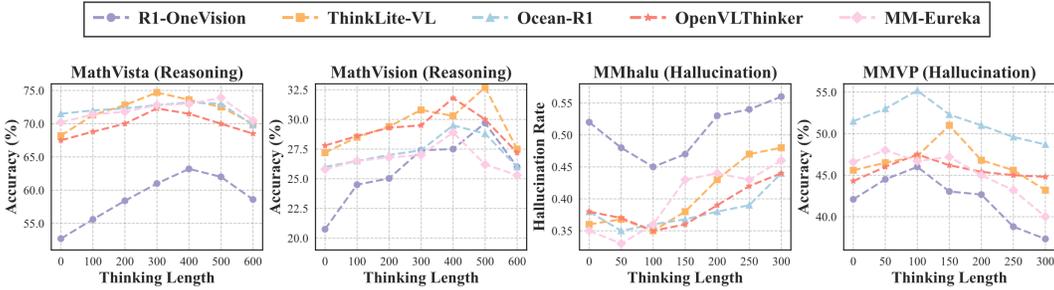


Figure 7: Reasoning-Hallucination balance of multimodal reasoning models under varying reasoning lengths. Thinking lengths are controlled within [0–600] tokens for reasoning and [0–300] for hallucination, corresponding to the longer chains required for reasoning and shorter for hallucination.

4.2 Dynamic Balance Between Reasoning and Hallucination

Non-monotonic Effect of Reasoning Length on Reasoning and Perception Performance. The relationship between reasoning length and model performance typically exhibits a non-monotonic pattern under reasoning and perception tasks. Across various length control strategies, a consistent trend emerges: *moderate reasoning depth tends to yield optimal performance, whereas overly short or excessively long reasoning chains often lead to a decline in accuracy.* As shown in Figure 7, we employ the *Latent State Steering* method adjusts the thinking step for reasoning and perception tasks. It is evident that as the thinking length increases, the model’s performance across tasks generally follows a rising-then-falling trajectory. This indicates that enhanced reasoning does not linearly improve model performance, but instead follows a dynamic trade-off pattern.

Task-Specific Variability of Optimal Reasoning Intervals. While most tasks exhibit non-monotonic relationships between reasoning length and performance, we further observe that the optimal reasoning length varies significantly across tasks. Figure 7 reveals that reasoning benchmarks such as MathVista [27] tend to benefit from longer reasoning chains, whereas perception and hallucination-oriented tasks such as MMHalu achieve their best performance at shorter or moderate lengths. This indicates that the balance between reasoning depth and performance is task-specific, and unified length control strategies are unlikely to be effective across all task types.

Impact of the Zerothink Condition. Zerothink retains the reasoning structure but lacks substantive content. As shown in Figure 7, this setting leads to a consistent drop in model performance on both reasoning and perception benchmarks, notably lower than results under normal reasoning lengths. These results indicate that the absence of reasoning content diminishes the reasoning model’s performance in both perception and reasoning.

Limitation of Conventional Metric. Conventional metrics like reasoning accuracy and hallucination rate, when computed at a fixed generation length, fail to capture the dynamic balance between deeper

reasoning and perception. Figure 7 shows that reasoning and perception often peak at different reasoning lengths, making it misleading to evaluate models using single-point metrics or simple averages between reasoning and hallucination performance. For instance, a short reasoning trace may yield a lower hallucination rate but poor reasoning depth, while a longer trace may improve reasoning at the cost of increased hallucination, yet both scenarios could yield the same average score.

To capture this evolving balance, in the next section, we propose an AUC-style metric that summarizes the balance curve between reasoning and perception fidelity across various reasoning lengths. This provides a more faithful and holistic measure of performance, revealing both the model’s optimal balance and its stability across varying generation lengths.

Takeaway 2: Moderate Reasoning Length Strikes the Best Reasoning-Hallucination Balance

Reasoning length exerts a non-monotonic effect on model performance: both insufficient and excessive reasoning degrade accuracy, and the optimal length is task-dependent.

5 Evaluation on the Reasoning-Hallucination Balance

To comprehensively quantify the balance between reasoning and hallucination in multimodal large reasoning models at different reasoning depths, we introduce a new metric *RH-AUC*. This metric captures how hallucination risk evolves with reasoning depth while also reflecting the cumulative effects of reasoning and perception. Additionally, we present *RH-Bench*, a new diagnostic dataset of 1000 samples, designed for the integrated evaluation of reasoning and perception tasks, offering a robust basis for analyzing reasoning ability and perceptual hallucinations.

5.1 Setup

Benchmark Overview. *RH-Bench* consists of two types of tasks: reasoning and perception, with each task including two types of questions: multiple-choice and open-ended. The reasoning task includes 500 samples sourced from MathVision [44], MathVista [27], MMMU [55], and ScienceQA [28], while the visual perception task includes 500 samples from MMhalu, MMVP, HallusionBench, and VMCBench. Both task types use accuracy as the evaluation metric. For multiple-choice questions, evaluation is based on matching the final options. For open-ended questions, both tasks are evaluated using GPT-4o. The reasoning task determines whether the generated response is consistent with the correct answer, whereas the visual task evaluates the generated response against the correct answer, assigning a score within the range of 0 to 6. Responses with a score below 3 are classified as hallucinations. All sample ground-truth and evaluation answers have undergone manual inspection.

Method	Paradigms	Perception		Reasoning		Training Data		<i>RH-AUC</i>
		Acc.(%) ↑	Length	Acc.(%) ↑	Length	Perc.	Reas.	
LLM-R1-3B	RL	48.7	121.9	43.8	391.8	65k	40k	0.46
Curr-ReFT-3B	SFT+RL	50.6	133.7	42.5	472.61	6k	3k	0.47
Ocean-R1-3B	RL	52.8	131.2	45.6	414.5	20k	63k	0.53
R1-OneVision-7B	SFT+RL	55.7	162.9	44.2	457.3	80k	77k	0.46
ThinkLite-VL-7B	RL	63.3	110.4	50.4	435.4	62k	8k	0.52
OpenVLThinker-7B	SFT+RL	59.2	187.7	48.9	460.1	25k	25k	0.54
MM-Eureka-7B	RL	62.0	139.6	54.0	450.5	-	15k	0.55
MM-R1-7B	RL	60.3	139.6	54.0	430.0	-	6k	0.57
Ocean-R1-7B	RL	62.3	90.4	51.8	262.2	20k	63k	0.63

Table 1: Comparison of model performance on *RH-Bench*, including task-specific accuracy and *RH-AUC* scores. **Perc.** and **Reas.** respectively denote training data for visual perception and reasoning.

RH-AUC We define reasoning length as T , which controls the extent of the model’s generated reasoning trace. For each length T , we compute R_T , which represents the reasoning performance at length T , and H_T , representing performance on hallucination at the same length.

By evaluating the model at multiple lengths on the *RH-bench* benchmark, we obtain a series of (R_T, H_T) pairs that form a balance curve between reasoning and perceptual hallucination. To compute the area under this curve, we first sort the pairs in ascending order of reasoning performance R_T . Let

the sorted indices be denoted as $T^{(0)}, T^{(1)}, \dots, T^{(n-1)}$, such that $R_{T^{(0)}} \leq R_{T^{(1)}} \leq \dots \leq R_{T^{(n-1)}}$. To ensure comparability across models, both R_T and H_T are min-max normalized to the range $[0, 1]$. The $RH-AUC$ is then computed using the trapezoidal rule as:

$$RH-AUC = \sum_{i=0}^{n-2} \frac{R_{T^{(i+1)}} - R_{T^{(i)}}}{2} \cdot (H_{T^{(i+1)}} + H_{T^{(i)}}), \quad (3)$$

where n is the number of evaluated reasoning lengths. A higher $RH-AUC$ indicates a model that better balances reasoning and hallucination across different reasoning lengths.

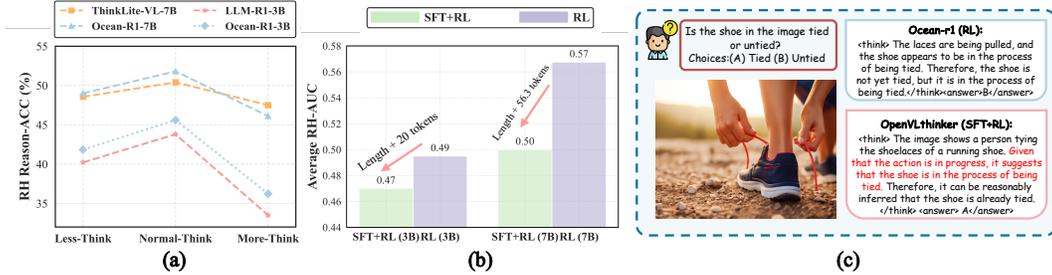


Figure 8: (a) Accuracy trends on the $RH-Bench$ reasoning task across different reasoning lengths for 3B and 7B models. Larger models typically exhibit more stable performance across varying reasoning lengths. (b) Comparison of SFT+RL and RL-only training paradigms in terms of $RH-AUC$, with arrow directions indicating the increase in reasoning length for SFT+RL relative to RL-only. RL-only training tends to generate more concise reasoning chains, leading to a better perception hallucination balance. (c) Case study comparing RL-only and SFT+RL models. SFT+RL models often introduce rigid imitation reasoning paths, which limit the flexibility of visual reasoning.

5.2 Result Diagnosis

We conduct an in-depth analysis of model performance based on the evaluation results from the $RH-Bench$ diagnostic dataset, investigating the influence of three key factors: model scale, training paradigm, and training dataset on the reasoning-hallucination balance.

Model Scaling. As shown in Table 1, the 7B model generally outperforms the 3B model in $RH-Bench$, demonstrating higher $RH-AUC$, primarily due to its larger parameter size and stronger representational capacity. As illustrated in Figure 8a, the larger model maintains higher stability, especially under longer reasoning chains, while the smaller models show a noticeable decline in performance. **This suggests that larger models typically exhibit better robustness and adaptability.**

Training Paradigms. A comparison between the two-stage SFT+RL-trained model and the RL-only trained model reveals that RL demonstrates a stronger balance between reasoning and perception. For example, in Figure 8c, although the OpenVLThinker model maintains a longer reasoning chain, the introduction of redundant reasoning interferes with visual perception, leading to an incorrect inference that the shoe is already tied. In contrast, the RL-only model, Ocean-R1, uses a shorter reasoning chain, enabling it to more efficiently capture key visual features and avoid unnecessary complex reasoning steps. This advantage is particularly evident at different reasoning length, as shown in the average $RH-AUC$ in Figure 8b, which is significantly higher for the RL-only model compared to SFT+RL. This phenomenon suggests that **although SFT helps the model learn reasoning formats, it may introduce rigid imitation reasoning paths, limiting the model’s adaptability to dynamic tasks and ultimately resulting in redundant reasoning.** In contrast, RL encourages the model to generate more adaptive reasoning behaviors, enhancing the integration of reasoning and perception.

Training Dataset. The diversity and quality of training data play a crucial role in the reasoning-hallucination balance of models. Through a statistical analysis of the multimodel reasoning models training data and a comparison with the results in Table 1, we have observed some interesting phenomena: **(1) More visual perception data does not necessarily improve the balance between reasoning and perception.** Increasing the training samples of visual perception data can enhance the balance of reasoning models to a certain degree. For example, the ThinkLite-VL model, supported by large scale visual perception data, demonstrates strong hallucination and reasoning balance. Similarly,

the Ocean-R1 model adopts a two-stage training strategy, first enhancing reasoning ability and then strengthening visual perception, achieving the highest *RH-AUC* on *RH-bench*. However, this phenomenon is not consistent. For example, despite the R1-OneVision model utilizing a large amount of visual perception data, it demonstrates a weaker balance between reasoning and perception, which may be attributed to the limitations of its training paradigm design. **(2) Perception and Reasoning balance can be achieved through training on domain-specific data.** Training on domain-specific data helps enhance the balance of the reasoning model. For example, the MM-Eureka model, trained on a larger mathematical dataset, shows a higher *RH-AUC*, proving its effectiveness in balancing reasoning and perception. Similarly, despite being trained on only 6k mathematical data, the MM-R1 model still performs well on *RH-bench*. This highlights the potential of domain-specific data to stimulate the balance capabilities of reasoning models, even with smaller datasets. **(3) The size of the training data is not always a guarantee for the reasoning-perception balance.** The training data size does not always directly correlate with the model’s balance capability. For example, both the LLM-R1, trained on over 60k visual perception samples, and the R1-OneVision, with a dataset of 150k samples, exhibit inadequate reasoning-hallucination balance, with the *RH-AUC* of only 0.46.

6 Related Work

Multimodal Reasoning Tasks. Multimodal reasoning requires integrating information across modalities to solve complex problems. It is generally categorized into general reasoning and domain-specific reasoning. General reasoning typically occurs in natural image scenarios, where models must combine visual perception with knowledge and commonsense. Representative benchmarks include multiple-choice datasets such as MMMU [55], MMVP [42], MMBench [24], MMStar [3], MMEval-Pro [13], and VMCBench [61], as well as open-ended evaluations like Bingo [5], MMHAL-Bench [13], POPE [19], CHAIR [35], and HallusionBench [12]. Domain-specific reasoning focuses on technical tasks within particular domains. For mathematical reasoning, benchmarks such as MathVista [27], MATH-Vision [44], MM-Math [37], WeMath [34] evaluate models’ ability to solve math problems grounded in visual contexts. For physical reasoning, datasets like PhysBench [4] and CRAVE [38] test understanding of physics and commonsense reasoning from visual inputs.

Reinforcement Learning in MLLMs. Recent approaches enhance the reasoning capabilities of multimodal large models by incorporating chain-of-thought supervision during supervised fine-tuning or reinforcement learning [62, 58, 48, 41, 50, 52]. Methods like RLHF-V [54], LLaVA-Reasoner [60], and Insight-V [11] leverage large-scale CoT-style datasets and preference optimization to improve model reasoning. Following DeepSeek-R1, the GRPO (Group Relative Policy Optimization) algorithm has become a standard paradigm in training multimodal large reasoning models [26, 57, 47, 43, 18, 45, 25]. Some models, such as R1-OneVision [51], Reason-RFT [40], and R1-VL [59], follow a two-stage SFT + RL pipeline, while others like Ocean-R1 [20], ThinkLite-VL [46], and MM-Eureka [30] apply rule-based reinforcement learning directly at scale.

7 Conclusion

In conclusion, this paper investigates the balance between reasoning and hallucination in multimodal reasoning models, with a focus on how reasoning chain length and visual attention allocation impact performance. While longer reasoning chains enhance performance on complex tasks, they also exacerbate hallucinations by diminishing visual attention and increasing reliance on language priors. To address these challenges, the paper introduces the *RH-AUC* metric and the *RH-Bench* benchmark, which provide a systematic method to evaluate the balance between reasoning ability and hallucination risk. The findings reveal that reasoning-augmented models are more prone to hallucinations, highlighting the importance of developing evaluation frameworks that assess both the quality of reasoning and the accuracy of perception.

Limitation. Although our study provides a comprehensive analysis of visual hallucinations in multimodal reasoning models, it also has several limitations. First, our evaluation is limited to models built on the Qwen2.5-VL backbone, which may constrain the generalizability of our findings to architectures with different modalities or pretraining objectives. Second, our analysis of the influence of training data is based solely on technical reports and publicly available documentation of existing models, without conducting controlled retraining experiments. Therefore, our conclusions are observational and may not fully capture causal effects.

References

- [1] Zouying Cao, Yifei Yang, and Hai Zhao. Nothing in excess: Mitigating the exaggerated safety for llms via safety-conscious activation steering. *arXiv preprint arXiv:2408.11491*, 2024.
- [2] Edward Y. Chang, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *CoRR*, abs/2502.03373, 2025.
- [3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.
- [4] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.
- [5] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.
- [6] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.
- [7] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding Zeng. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024.
- [8] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning, 2025.
- [9] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Open-ylthinker: An early exploration to complex vision-language reasoning via iterative self-improvement, 2025.

- [10] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.
- [11] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.
- [12] 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.
- [13] Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, et al. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. *arXiv preprint arXiv:2407.00468*, 2024.
- [14] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.
- [15] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- [16] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [17] Sicong Leng, Jing Wang, Jiayi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Hang Zhang, Yuming Jiang, Xin Li, Deli Zhao, Fan Wang, Yu Rong, Aixin Sun, and Shijian Lu. Mmr1: Advancing the frontiers of multimodal reasoning. <https://github.com/LengSicong/MMR1>, 2025.
- [18] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025.
- [19] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, December 2023. Association for Computational Linguistics.
- [20] Ming Lingfeng, Li Yadong, Chen Song, Xu Jianhua, Zhou Zenan, and Chen Weipeng. Ocean-r1: An open and generalizable large vision-language model enhanced by reinforcement learning. <https://github.com/VLM-RL/Ocean-R1>, 2025. Accessed: 2025-04-03.
- [21] Chengzhi Liu, Yuzhe Yang, Kaiwen Zhou, Zhen Zhang, Yue Fan, Yannan Xie, Peng Qi, and Xin Eric Wang. Presenting a paper is an art: Self-improvement aesthetic agents for academic presentations, 2025.
- [22] Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*, 2023.
- [23] Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*.

- [24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.
- [25] Yue Liu, Shengfang Zhai, Mingzhe Du, Yulin Chen, Tri Cao, Hongcheng Gao, Cheng Wang, Xinfeng Li, Kun Wang, Junfeng Fang, Jiaheng Zhang, and Bryan Hooi. Guardreasoner-vl: Safeguarding vlms via reinforced reasoning. *arXiv preprint arXiv:2505.11049*, 2025.
- [26] 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.
- [27] 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.
- [28] 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, 2022.
- [29] Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Darshan Thaker, Aditya Chattopadhyay, Chris Callison-Burch, and René Vidal. Pace: Parsimonious concept engineering for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [30] 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.
- [31] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025.
- [32] OpenAI. Learning to reason with LLMs. 2024.
- [33] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b llms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.
- [34] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma Gongque, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *CoRR*, abs/2407.01284, 2024.
- [35] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.
- [36] Haozhan Shen, Zilun Zhang, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. 2025. Accessed: 2025-02-15.
- [37] Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. MM-MATH: Advancing multimodal math evaluation with process evaluation and fine-grained classification. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1358–1375, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [38] Shangkun Sun, Xiaoyu Liang, Bowen Qu, and Wei Gao. Content-rich aigc video quality assessment via intricate text alignment and motion-aware consistency. *arXiv preprint arXiv:2502.04076*, 2025.

- [39] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.
- [40] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.
- [41] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.
- [42] 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.
- [43] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. VI-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.
- [44] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- [45] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.
- [46] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025.
- [47] Wenyi Xiao, Leilei Gan, Weilong Dai, Wangui He, Ziwei Huang, Haoyuan Li, Fangxun Shu, Zhelun Yu, Peng Zhang, Hao Jiang, et al. Fast-slow thinking for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*, 2025.
- [48] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.
- [49] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024.
- [50] Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, et al. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*, 2025.
- [51] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.
- [52] Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. Twinmarket: A scalable behavioral and social simulation for financial markets, 2025.
- [53] Hao Yin, Guangzong Si, and Zilei Wang. Clearlight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2503.13107*, 2025.
- [54] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwan 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.

- [55] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [56] Liu Yuqi, Peng Bohao, Zhong Zhisheng, Yue Zihao, Lu Fanbin, Yu Bei, and Jia Jiaya. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. 2025.
- [57] Liu Yuqi, Peng Bohao, Zhong Zhisheng, Yue Zihao, Lu Fanbin, Yu Bei, and Jia Jiaya. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement, 2025.
- [58] 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.
- [59] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.
- [60] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024.
- [61] Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, et al. Automated generation of challenging multiple-choice questions for vision language model evaluation. *arXiv preprint arXiv:2501.03225*, 2025.
- [62] Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*, 2024.

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: Our contributions are summarized in the Abstract and Introduction sections.

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: Please see Conclusion.

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: This paper does not include theoretical results.

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: Please see the Setup in sec.5.1 and Detailed Experimental Settings in appendix A

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: As mentioned in the Abstract, our code and data will be made publicly available.

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: The detailed results can be found in 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: The detailed results can be found in 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: The detailed results can be found in 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 the NeurIPS Code of Ethics and confirm that our research complies with these ethical standards.

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: Please see the Appendix G.

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: This 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: All the thing in this paper credited and are the license and terms of use explicitly mentioned and properly respected.

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: This 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: This 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: This 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: [NA]

Justification: The method development in this paper does not involve LLMs as any components.

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.

Appendix

A Detailed Experimental Settings

Datasets. We evaluate our model on both discriminative and generative datasets, as listed below. (a) MMVP [42] evaluates recognition and reasoning performance across nine categories of basic visual patterns. (b) MMEval-Pro [13] assesses cross-modal understanding through triplet-based object and attribute recognition in natural images. (c) VMCBench [61] use adversarial distractors to test fine-grained discriminative ability across diverse tasks such as commonsense reasoning, image-text matching. All three datasets (a–c) adopt accuracy as the evaluation metric. (d) Bingo [5] evaluates bias and interference hallucinations, with GPT-4o used to score hallucination severity and response quality. (e) MMHAL-Bench [39] evaluates model capabilities beyond object hallucination, with GPT-4o used to assess hallucination rate and response informativeness.

Implementation Details. We select nine representative multimodal reasoning models to evaluate their hallucination performance on general vision tasks. We categorize these models into two major training paradigms: (1) the RL-only paradigm, where models are trained solely via reinforcement learning, including LMM-R1 [33], MM-R1[17], ThinkLite-VL[46], MM-Eureka[30], and Ocean-R1[20]; (2) the two-stage paradigm, combining supervised fine-tuning (SFT) with reinforcement learning, including Vision-R1[15], R1-OneVision[51], OpenVLThinker[9], and Curr-ReFT [8]. All models are post-trained on Qwen2.5-VL-3B or Qwen2.5-VL-7B, which are used as baseline models.

B Reasoning Models Attention-Based Analysis

B.1 Visual Attention Heatmap

Figure 9-10 compares the visual attention distribution between multimodal reasoning models and their corresponding non-reasoning models. The results indicate that, compared to non-reasoning models, reasoning models exhibit weaker focus on key image details, with attention more dispersed across other regions of the image. Specifically, reasoning models display a greater degree of attention dispersion at lower layers, and their attention is not concentrated on critical areas of the image. In contrast, non-reasoning models demonstrate more precise visual grounding. For instance, as shown in Figure 10, the attention maps of the non-reasoning model at layers 10 and 15 consistently focus on the target object, the white mouse, highlighting its sustained attention on the target.

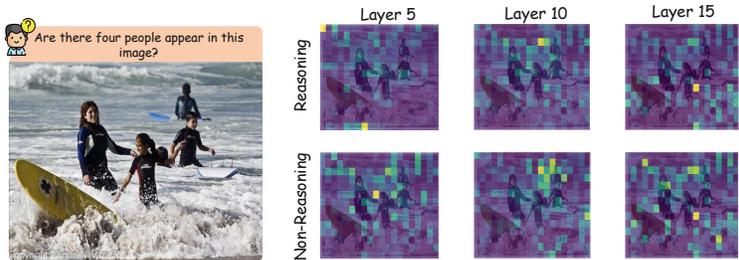


Figure 9: Case Study 1: Attention Heatmap in Counting Tasks.

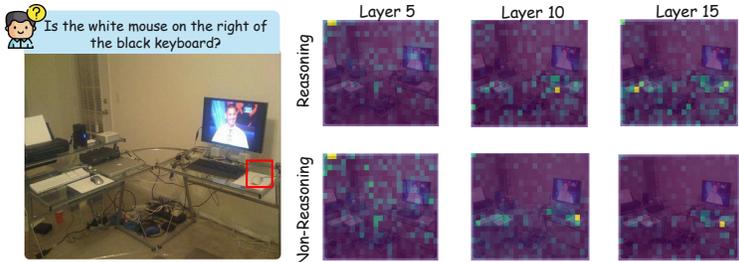


Figure 10: Case Study 2: Attention Heatmap in Object Localization.

B.2 More Examples of the Impact of Reasoning Length on Visual Perception Degradation

This section presents an additional example, including a visual task involving counting, comparing the results of reasoning models and their corresponding attention maps under different reasoning lengths. It is evident that an excessively lengthy reasoning process causes the model to disregard the visual information inherent in the image, instead relying more heavily on prior linguistic knowledge. In Figure 11, the attention maps clearly show that, under over reasoning conditions, the model’s attention shifts more towards the instruction tokens following the image tokens, particularly towards the latter part of the instruction. This suggests that prolonged reasoning reduces the model’s focus on the visual information, leading it to depend more on the guidance provided by the linguistic instructions.

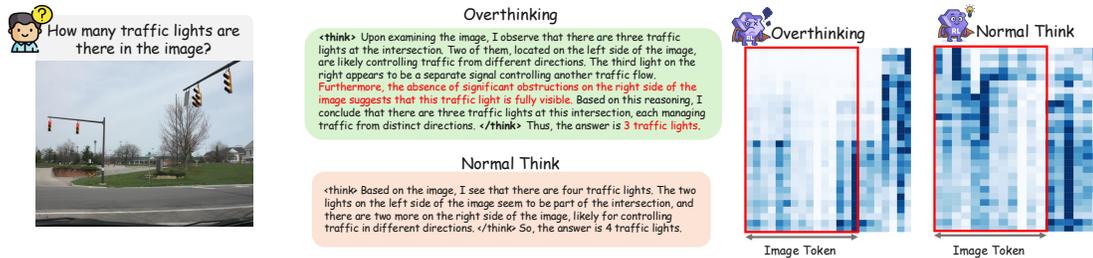


Figure 11: Attention shift in the reasoning model under different reasoning length.

B.3 Error Analysis

In this section, we further focus on the error rates of multimodal reasoning models and non-reasoning models across different problem types, conducting a statistical analysis to compare the differences between the two. Figure 12a presents the error type statistics for the Bingo benchmark samples. By combining GPT-4o evaluations with manual inspection, we analyze the reasoning process and final answers of the reasoning model to determine whether the errors stem from reasoning or perception. If the model’s errors arise from both reasoning and perception, we classify them as "perception and reasoning" errors. The statistical results indicate that the proportion of perception errors in the reasoning model decreases, with more errors originating from the reasoning process. This suggests that the reasoning model does not completely fail to interpret the image information, but rather diminishes its focus on perceptual information during reasoning. The evaluation results in Figure 12b further confirm this phenomenon: the overall error rate of the reasoning model is higher than that of the non-reasoning model, with a more prominent proportion of errors coming from reasoning.

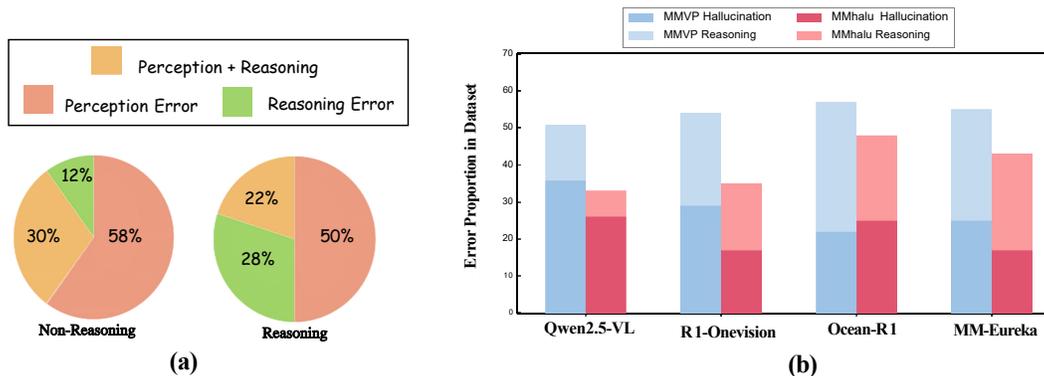


Figure 12: Error type distribution and error proportions across reasoning and non-reasoning models in Bingo benchmark. (a) Pie charts showing the distribution of perception and reasoning errors for non-reasoning and reasoning models, with the breakdown of perception error, reasoning error, and combined perception & reasoning errors. (b) Bar chart illustrating error proportions in the MMVP and MMhalu benchmarks, comparing hallucination and reasoning errors across reasoning models.

C Reasoning Length Control

C.1 Comparison of Three Reasoning Length Control Strategies

In the manuscript, we have thoroughly explored three methods: *Token Budget Forcing*, *Test Time Scaling*, and *Latent State Steering*. The first two methods directly control the model’s reasoning length by using fixed-length truncation or soft expansion of the reasoning length, ensuring dynamic expansion within a predefined thinking length range. However, the limitations of *Token Budget Forcing* and *Test Time Scaling* are that they can only control the model’s reasoning length to shorten or lengthen, lacking flexibility for more nuanced adjustments. In contrast, *Latent State Steering* introduces a tuning coefficient α , allowing more flexible control over the model’s reasoning length. By adjusting the value of α , we can effectively quantify changes in the extent of reasoning. All of our *Latent State Steering* experiments are dynamically adjusted within the range of $\alpha \in [-0.15, 0.15]$. Furthermore, in the subsequent *RH-bench* calculation of *RH-AUC*, it is precisely due to the flexibility of the *Latent State Steering Strategy* that we apply it to dynamically regulate the reasoning length and perform further quantification.

C.2 Model Performance Variation of the *Latent State Steering* Length Control Strategy

Figure 13 presents the visualization of performance variations for different models under the *Latent State Steering* strategy, with α ranging from $[-0.15, 0.15]$. In Figure 13, the star symbol represents the performance under the base condition. It is clearly observed that the variation in reasoning length shows that the optimal intervals for reasoning models differ between reasoning and hallucination tasks, with both exhibiting non-monotonicity.

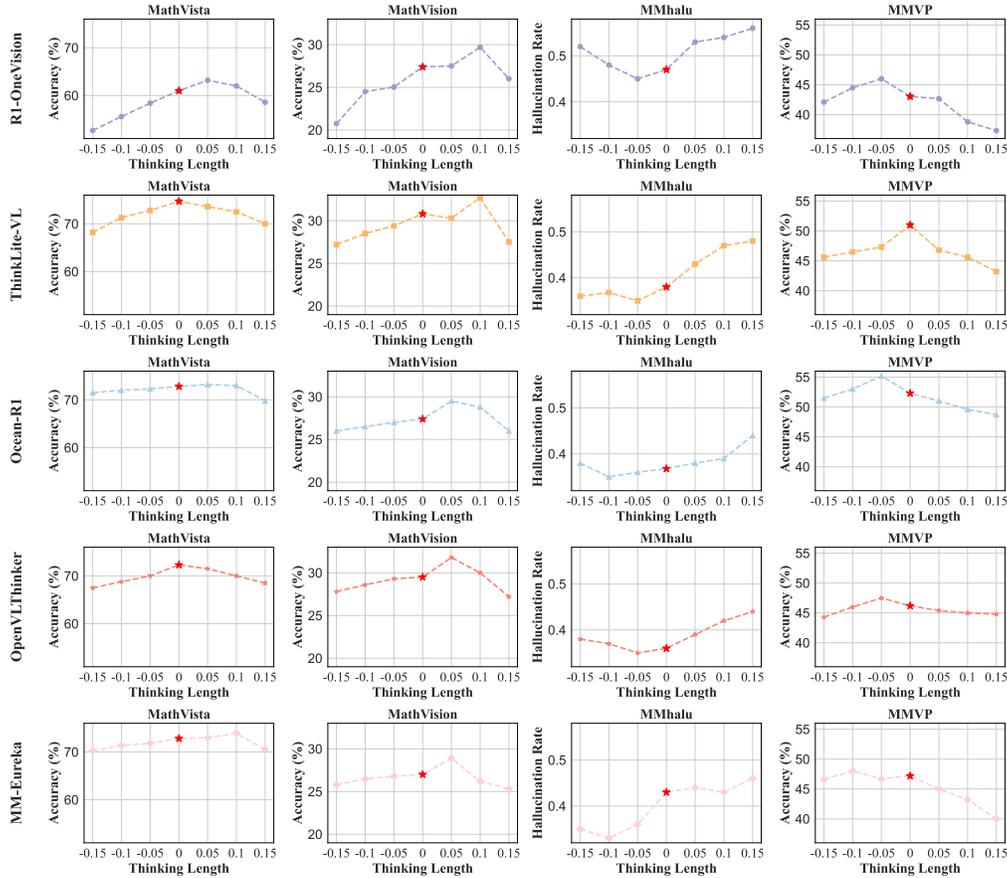


Figure 13: Model performance variation of the *Latent State Steering* strategy. The star symbol represents the original thinking length of the model without steering or test-time intervention.

D More Examples from *RH-Bench*

In this section, we present samples from different tasks and question types in the *RH-Bench* benchmark. As shown in Figures 14a and 14b, we display samples of open-ended responses and multiple-choice questions for the visual perception task. Additionally, Figures 15a and 15b showcase samples of multiple-choice questions and open-ended responses for the visual reasoning task. The focus of the questions differs across tasks. For instance, the visual perception task typically emphasizes image content recognition and understanding, whereas the visual reasoning task places more focus on the ability to draw conclusions from the image. To ensure the accuracy of the evaluation, we select samples that maintain as much independence between tasks as possible, minimizing interference between different tasks. This independence allows each task to reflect the model’s performance in a specific task type, thereby providing a more accurate assessment of the model’s performance across various tasks and its ability to balance performance between different task types.



Figure 14: Examples of visual perception in *RH-Bench*.

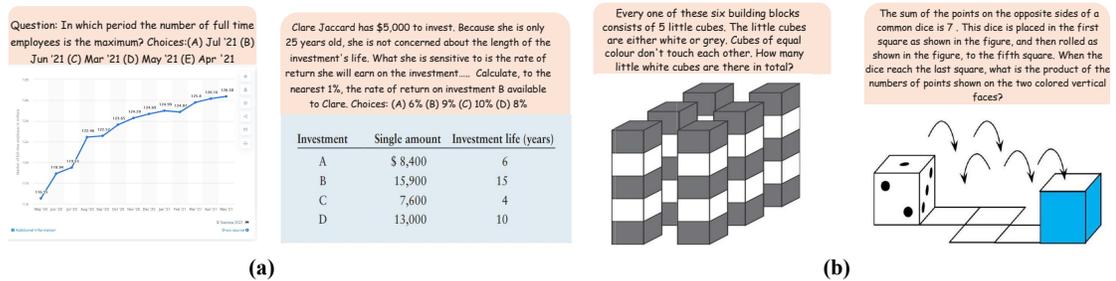


Figure 15: Examples of visual reasoning in *RH-Bench*.

E Visualization Results

This section presents the responses of reasoning models and their corresponding non-reasoning models across different visual tasks. As shown in Figures 16 and 17, in these error examples, it is evident that reasoning models typically exhibit errors during the reasoning process, leading to inaccurate reasoning results and, consequently, the emergence of hallucinations in the final answers.

 Is the word in the logo "penarth pier built 1894"?



Non-Reasoning Model

The sign has large, green, stylized letters that read "PENARTH PIER," followed by "BUILT 1894" underneath. The layout and font suggest it's a title or landmark sign. I also notice a smaller logo on the bottom right which includes the words "VALE OF GLAMORGAN LEISURE," indicating it may be part of a local council or tourism board. Given the context, this is likely a historical pier with some heritage value.

Reasoning Model

<think> Okay, so I need to figure out if the word in the logo is "Penarth Pier Built 1894." Let me start by looking directly at the image itself. As seen in the image, there is a sign that reads "Penarth Pier" with "Built 1894" underneath it. The text is in green on a white background, and there's also a logo for "Valley Glamorgan Leisure." However, the question specifically asks about the word in the logo, not the entire sign. The logo mentioned is part of the Valley Glamorgan Leisure branding, which includes a stylized design with a green leaf motif. **This logo is separate from the main sign that mentions Penarth Pier and its construction date. Therefore, the word in the logo is likely just the name "Valley Glamorgan Leisure," not "Penarth Pier Built 1894."** So, putting it all together, the word in the logo isn't "Penarth Pier Built 1894"; it's something else. **</think>** No. The word in the logo is "Valley Glamorgan Leisure," not "Penarth Pier Built 1894."

Figure 16: Hallucination example 1 of reasoning models in visual tasks.

 There a yellow brick building in the image?



Non-Reasoning Model

I closely examined the buildings in the image. The ones on the left are mostly made of red brick and white walls, while those on the right appear to be covered in light gray or beige materials. The tall structure near the center-right looks more like it has a light brown or cream-colored surface, not the distinct tone or texture of yellow brick. Therefore, no building in the image clearly appears to be made of yellow brick.

Reasoning Model

<think> Okay, so I need to determine if there's a yellow brick building in the image based on the provided content. Let me go through this step by step. First, I'll examine the image itself. The image shows an urban street scene with multi-story buildings on the left side. These buildings have a mix of architectural styles and colors—some are white with red accents, others are red-brick with white trim. There's also a modern building with a flat roof and a wooden facade visible further down the street. Next, I'll look at the bounding boxes provided. One box is labeled "a yellow brick building" with coordinates xmin: 208, ymin: 175, xmax: 346, ymax: 292. **This indicates that within these specific coordinates, there is indeed a yellow brick building present in the image.** Therefore, based on both the image and the bounding box information, there is a yellow brick building in the image. **</think>** Yes, there is a yellow brick building in the image. The image shows a yellow brick building within the specified coordinates, confirming its presence.

Figure 17: Hallucination example 2 of reasoning models in visual tasks.

F *RH-Bench* Evaluation Prompts

RH-Bench Visual Reasoning MCQ & Open-Ended Prompt

MCQ: You are an impartial evaluator assessing the correctness of a model's answer to a multiple-choice question.

Question: {question}
Choices: {choices}
Model's Answer: {model answer}
Correct Answer: {ground truth}

Please evaluate whether the model's answer is correct by considering:

1. Whether the model's answer matches the correct answer exactly (e.g., same option letter).
2. If the model's answer is a value, whether it matches the value of the correct option.
3. Whether the model's reasoning (if provided) supports its answer.

Your response should be a JSON object with the following structure:

```
{  
  "is_correct": <true or false>,  
  "reason": "<brief explanation of your evaluation>"  
}
```

Open-Ended: You are an impartial evaluator assessing the correctness of a model's answer to a multiple-choice question.

Question: {question}
Model's Answer: {model answer}
Correct Answer: {ground truth}

Please evaluate whether the model's answer is correct by considering:

1. Whether the model's answer matches the correct answer exactly (e.g., same option letter).
2. If the model's answer is a value, whether it matches the value of the correct option.
3. Whether the model's reasoning (if provided) supports its answer.

RH-Bench Visual Perception MCQ Prompt

Please evaluate whether the model's answer to the multiple-choice question is correct by considering: 1. Whether the model's answer matches the correct answer exactly (same option letter). 2. If the model's answer is a value, whether it matches the value of the correct option. 3. Whether the model's reasoning (if provided) supports its answer.

Question: {}
Options: {}
Correct Answer: {}
Model's Answer: {}

Your response should be a JSON object with the following structure:

```
{  
  "is_correct": <boolean>,  
  "reason": "<explanation of your evaluation>",&br/>  "model_answer_extracted": "<the extracted answer from the model's response>"  
}
```

RH-Bench Visual Perception Open-Ended Prompt

Please act as an impartial and objective judge to evaluate the presence and severity of hallucination in the response provided by a Large Multimodal Model (LMM) to the user question. Hallucination, in this context, refers to a situation where the LMM generates a response that includes information not present or implied in the image or previous conversation. A hallucination could be a false claim about an object, action, emotion, or any other detail not grounded in the image.

Your task is to determine whether hallucination exists and, if present, to categorize its severity based on the extent and impact of the false information. Use the provided image contents, question, standard human-generated answer, and LMM response to make your judgment.

Evaluation Guidelines:

Compare the LMM response with the standard human-generated answer. Identify any claims in the LMM response that are not supported by the image or question. Assess the severity of hallucination based on the nature and extent of the false information. The standard human-generated answer may not describe all objects or details comprehensively, so focus on whether the LMM introduces false claims rather than omitting details. Detailed reasoning or analysis by the LMM that is consistent with the image contents is acceptable and should not be considered hallucination.

Task: Evaluate the LMM response for hallucination based on the provided image contents, question, and standard human-generated answer. Provide a brief explanation of your analysis, identifying any false claims and their severity. Then, assign a rating from the following six levels:

Rating: 5: No hallucination (response is fully consistent with image contents).

Rating: 4: Minimal hallucination (minor, negligible false detail that does not significantly alter the response's accuracy, e.g., a slight misdescription of color or background).

Rating: 3: Mild hallucination (small false claim that adds unsupported but non-critical detail, e.g., mentioning a minor object or attribute not present).

Rating: 2: Moderate hallucination (clear false claim that affects the response's accuracy, e.g., incorrect object count or unsupported environmental detail).

Rating: 1: Severe hallucination (significant false claim that misrepresents key elements, e.g., entirely wrong objects or actions).

Rating: 0: Extreme hallucination (response is largely or entirely detached from the image contents, with multiple or critical false claims).

Question { }

Ground-Truth Answer { }

LMM Response to Evaluate { }

Your response should be a JSON object with the following structure:

```
{
  "hallucination_score": <integer from 0 to 5>,
  "reason": "<brief explanation of your evaluation>"
}
```