# Appendix

## A Detailed Experimental Settings

**Datasets.** We evaluate our model on both discriminative and generative datasets, as listed below. (a) MMVP [12] evaluates recognition and reasoning performance across nine categories of basic visual patterns. (b) MMEval-Pro [4] assesses cross-modal understanding through triplet-based object and attribute recognition in natural images. (c) VMCBench [16] 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 [1] evaluates bias and interference hallucinations, with GPT-4o used to score hallucination severity and response quality. (e) MMHAL-Bench [11] 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 [9], MM-R1[6], ThinkLite-VL[13], MM-Eureka[8], and Ocean-R1[7]; (2) the two-stage paradigm, combining supervised fine-tuning (SFT) with reinforcement learning, including Vision-R1[5], R1-OneVision[14], OpenVLThinker[3], and Curr-ReFT [2]. All models are post-trained on Qwen2.5-VL-3B or Qwen2.5-VL-7B, which are used as baseline models.

| Method | Train Setting | Bingo Score ↑ | MMVP Acc (%) ↑ | MMEval-Pro Acc (%) ↑ | VMCBench Acc (%) ↑ | MMHALU Acc (%) ↑ | MMHALU Hallu ↓ |
|---|---|---|---|---|---|---|---|
| Curr-ReFT | SFT+RL | 3.52 | 28.7 | 60.2 | 73.0 | 2.78 | 0.45 |
| Ocean-R1 | RL | 3.62 | 39.3 | 62.5 | 76.5 | 2.35 | 0.46 |
| LMM-R1 | RL | 3.58 | 24.0 | 66.9 | 76.0 | 2.27 | 0.44 |
| **Qwen2.5-VL-3B** | Base | 3.64 | 36.0 | 67.8 | 77.6 | 3.33 | 0.43 |
| Vision-R1 | SFT+RL | 3.62 | 44.0 | 72.2 | 81.5 | 3.10 | 0.40 |
| R1-OneVision | SFT+RL | 3.65 | 43.0 | 69.4 | 65.2 | 3.20 | 0.48 |
| OpenVLThinker | SFT+RL | 3.45 | 46.5 | 71.5 | 80.5 | 3.00 | 0.36 |
| MM-R1 | RL | 3.58 | 44.3 | 73.6 | 80.3 | 3.10 | 0.38 |
| Ocean-R1 | RL | 3.65 | 52.3 | 73.7 | 82.3 | 2.80 | 0.35 |
| ThinkLite-VL | RL | 3.30 | 47.0 | 72.0 | 83.4 | 3.30 | 0.38 |
| MM-Eureka | RL | 3.69 | 46.7 | 74.8 | 82.0 | 3.20 | 0.43 |
| **Qwen2.5-VL-7B** | Base | 3.70 | 47.3 | 76.0 | 84.5 | 3.50 | 0.33 |

Table 1: Generalization behavior of multimodal reasoning models across training paradigms (Base, SFT+RL, RL-only) and parameter scales (3B, 7B) on five perception benchmarks. Notably, the score range in the Bingo benchmark is from 0 to 5, while MMhalu includes two evaluation metrics, with the score range also from 0 to 5, and the hallucination rate ranging from 0 to 1.

## B Reasoning Models Attention-Based Analysis

### B.1 Information Folw

**Attention distribution.** To elucidate the changes in internal information utilization induced by the introduction of reasoning structures, we adopt an attention-based framework[10, 15] to systematically compare the output contributions of different types of tokens. Specifically, for each layer $l$, we compute an information flow map as follows:

$$I_l(i,j) = \frac{1}{H}\sum_{h=1}^{H} A_{h,l}(i,j) \cdot \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}(i,j)}, \tag{1}$$

where $A_{h,l}(i,j)$ denotes the attention weight from token $j$ to token $i$ in the $h$-th head of layer $l$, $H$ is the number of heads in that layer, and $\mathcal{L}(x)$ denotes the cross-entropy loss computed for the input $x$.

To quantify the contribution of different types of tokens (image, instruction, system) to the final prediction, we measure their relative influence in the decision-making process by computing the

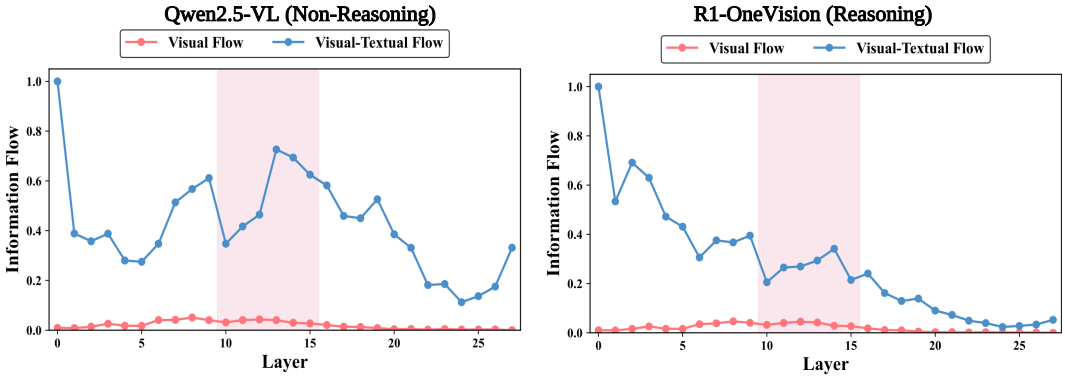



Figure 1: Layer-wise analysis of intra-visual and visual–textual information flow in Qwen2.5-VL (Non-reasoning) and R1-Onevision (Reasoning). The middle layers (highlighted) serve as critical regions for cross-modal integration.

average information flow scores from modality-specific tokens to all input tokens:

$$S_{\mathcal{M}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{X}} \sum_{j \in \mathcal{M}} I_l(i, j), \tag{2}$$

where $\mathcal{M} \in \{\mathcal{S}, \mathcal{V}, \mathcal{I}\}$ denotes the index set of modality-specific tokens (System, Image, and Instruction), $\mathcal{X}$ is the set of all input tokens, $I_l(i, j)$ is the layer information flow score, as defined in Eq. (1).

**Visual Information Interaction.** To further reveal how visual information is processed under different architectures, we analyze the interactions between visual tokens and other tokens across layers, focusing on two pathways: cross-modal injection into instruction tokens and intra-modal aggregation among visual tokens. We compare these interaction patterns by computing the average information flow scores from visual tokens to different targets, denoted as $S_{\text{vt}}$ and $S_{\text{vv}}$, respectively.

$$S_{\text{vt}} = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{I}} \sum_{i \in \mathcal{V}} I_l(i, j); \quad S_{\text{vv}} = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} \sum_{i \in \mathcal{V}} I_l(i, j), \tag{3}$$

where $S_{\text{vt}}$ quantifies cross-modal information flow from visual to instruction tokens, and $S_{\text{vv}}$ captures intra-modal integration among visual tokens.

We further analyze the cross-layer interaction of visual information within the model to assess whether image content effectively participates in final decision-making. As shown in Figure 1, Qwen2.5-VL exhibits active visual-to-text information flow in the middle layers (layers 10–15), indicating that the model effectively performs multimodal alignment and semantic feature aggregation at this stage. In contrast, the visual-to-text information flow in R1-OneVision rapidly decays beyond the shallow layers and remains consistently low throughout the middle layers. This indicates that the model lacks effective cross-modal interaction in the intermediate stage. Consequently, visual semantics do not adequately propagate to the language pathway, leading the subsequent generation process to rely more on language priors than on the actual image content. Moreover, the R1 model fails to establish effective intra-modal aggregation among visual tokens, making it difficult to perform fine-grained semantic refinement and structured representation of image content.

### B.2 Visual Attention Heatmap

Figure 2-4 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 3, 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.

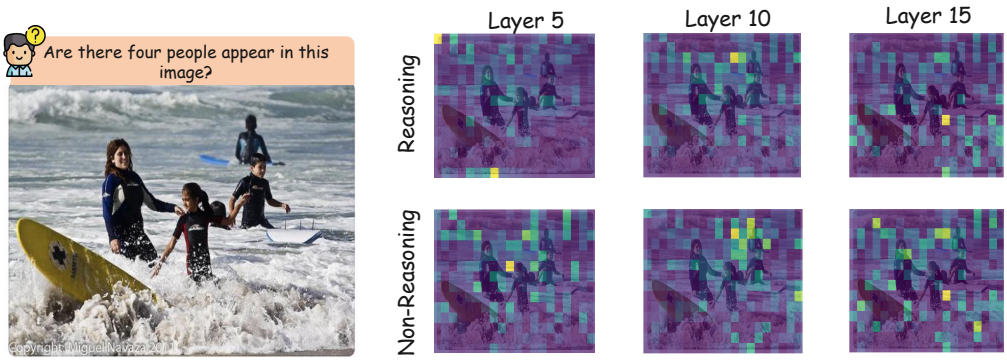



Figure 2: Case Study 1: Attention Heatmap in Counting Tasks.

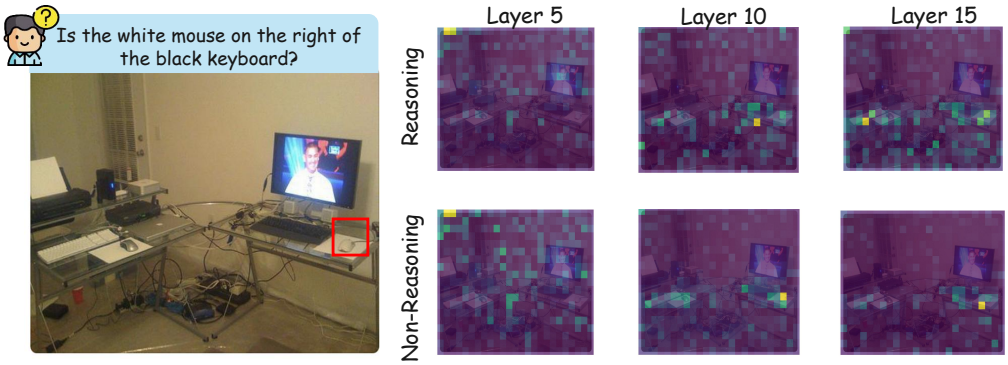



Figure 3: Case Study 2: Attention Heatmap in Object Localization.

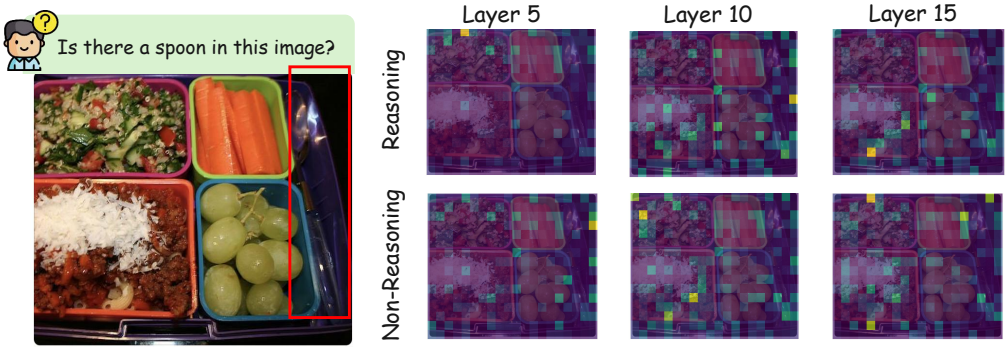



Figure 4: Case Study 3: Attention Heatmap in Challenging Object Localization

### B.3 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 5, 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.

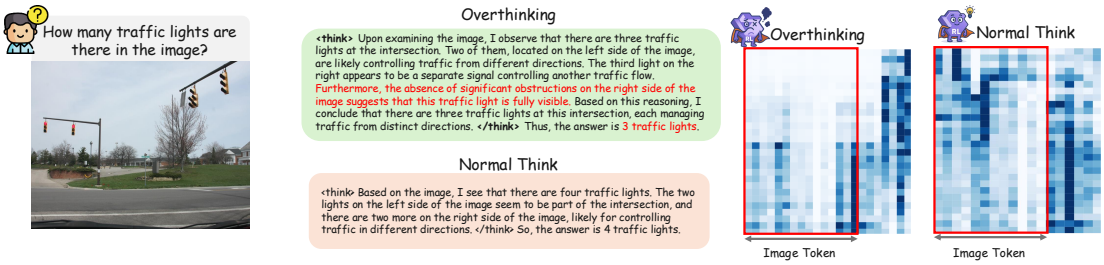



Figure 5: Attention shift in the reasoning model under different reasoning length.

### B.4 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 6a 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 6b 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.

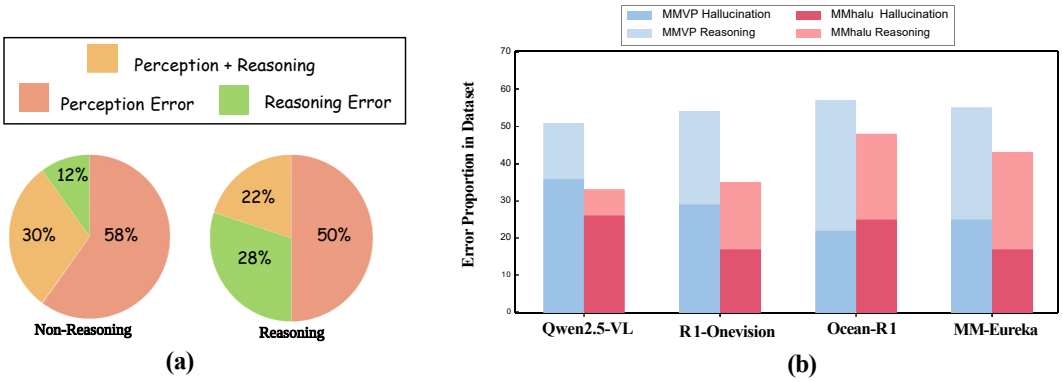



Figure 6: Error type distribution and error proportions across reasoning and non-reasoning models in Bingo benchmark. (a) Pie charts showing the distribution of perception and reasoning errors for non-reasoning and reasoning models, with the breakdown of perception error, reasoning error, and combined 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 Straregies

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 $\alpha$, allowing more flexible control over the model's reasoning length. By adjusting the value of $\alpha$, 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 *Latent State Steering Strategy* that we apply it to dynamically regulate the reasoning length and perform further quantification.

## C.2 Visualization Validation of the *Latent State Steering* Length Control Strategy

In this section, since the *Token Budget Forcing* and *Test Time Scaling* methods directly truncate and extend the target reasoning length, no further quantification is required. As shown in Figure 8, we present the reasoning length variations of four reasoning models across two reasoning datasets and two hallucination datasets, all controlled by the Latent State *Latent State Steering Strategy* through the tuning coefficient $\alpha$. The results clearly indicate that as the value of $\alpha$ increases, the reasoning length of the models monotonically increases. When $\alpha = 0$, it represents the default reasoning length of the model. A negative $\alpha$ value corresponds to a shorter reasoning length than the normal thinking length, while a positive $\alpha$ value corresponds to a longer reasoning length. Overall, the reasoning tasks require a longer thinking length than visual perception tasks. The visualizations presented above further demonstrate that our Latent State Steering strategy effectively and dynamically regulates the reasoning length of the model.

## C.3 Model Performance Variation of the *Latent State Steering* Length Control Strategy

Figures 7 and 9 present the visualization of performance variations for different models under the *Latent State Steering* strategy, with $\alpha$ ranging from $[-0.15, 0.15]$. In Figure 9, 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.

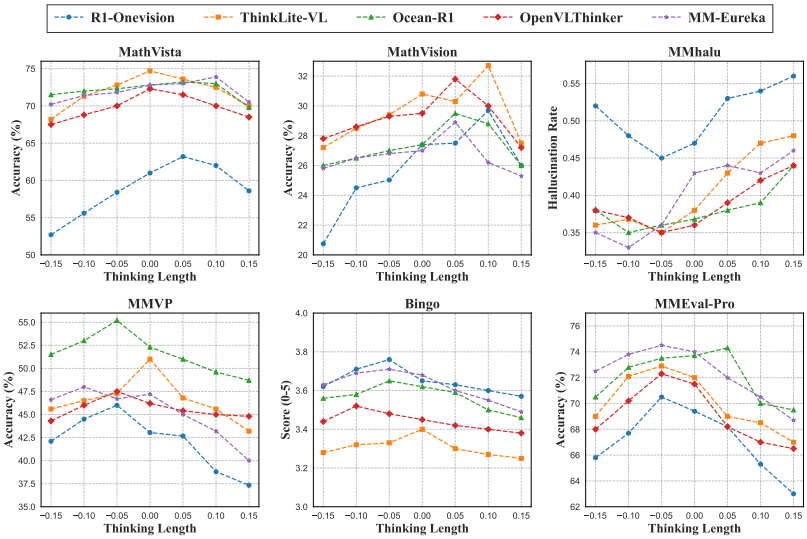
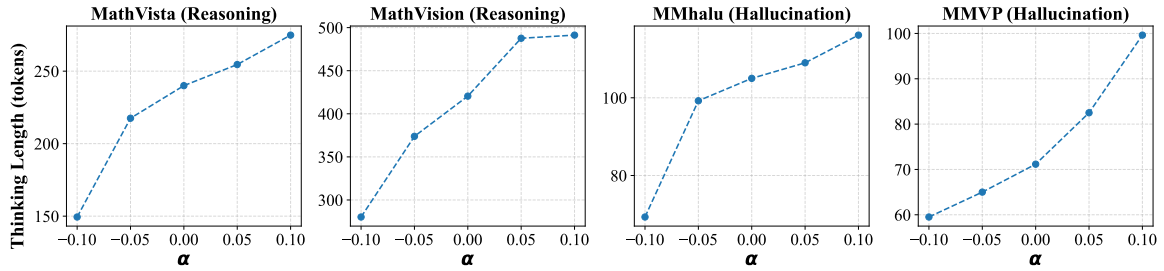
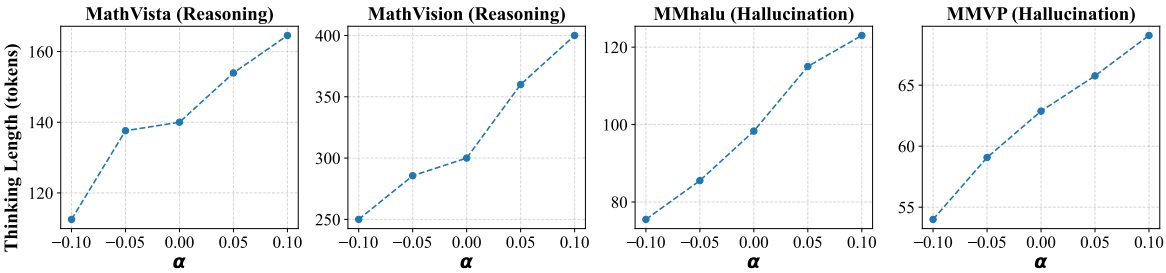


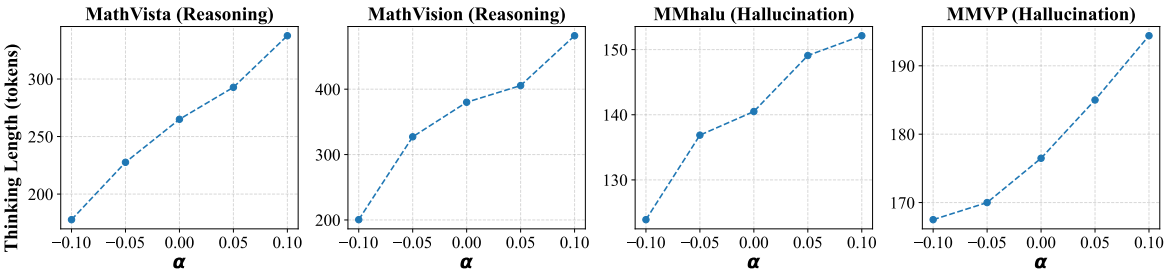

Figure 7: Model performance variation of the *Latent State Steering* strategy.



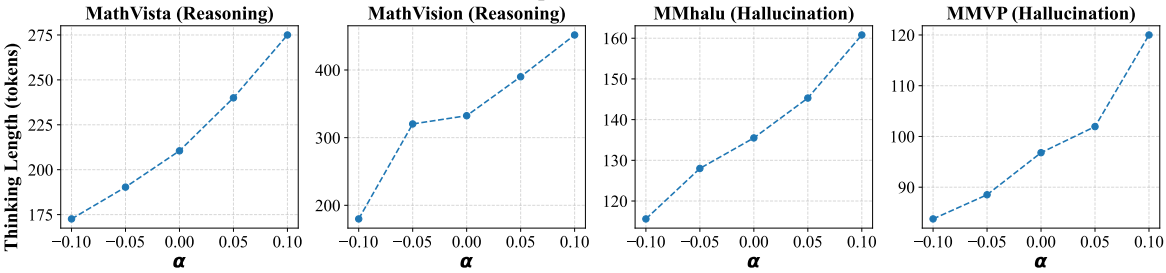

(a) MM-Eureka




(b) Ocean-R1




(c) OpenVLThinker




(d) R1-OneVision

Figure 8: Reasoning length variations of different models under the *Latent State Steering* control strategy. MathVista and MathVision are reasoning tasks, while the others are hallucination tasks.

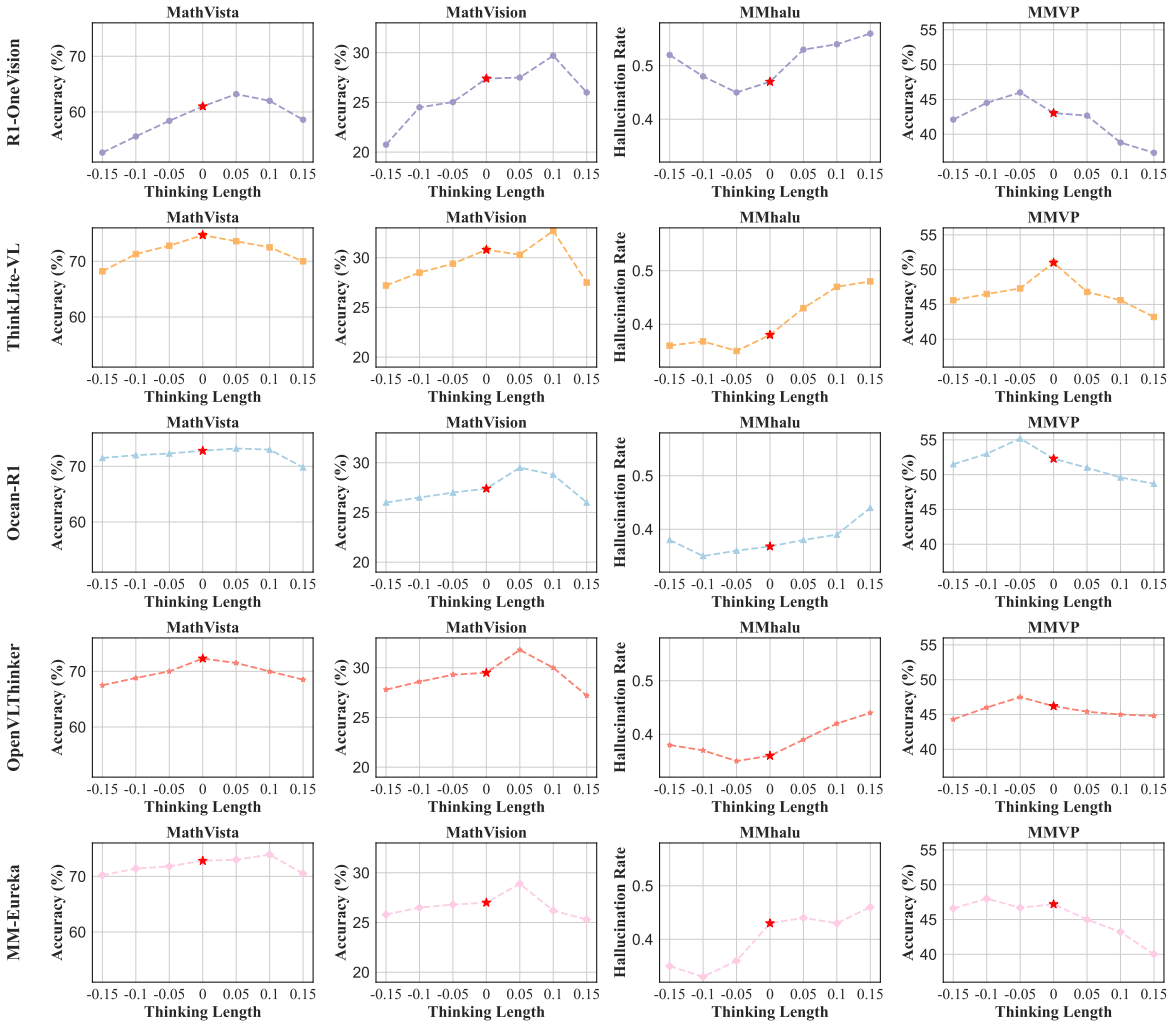



Figure 9: Model performance variation of the *Latent State Steering* strategy. The star symbol represents the original thinking length of the model without steering or test-time intervention.

### C.4 Result Details

This section provides a detailed presentation of the performance variation quantification results of three reasoning length control strategies across two reasoning datasets (MathVision and MathVista) and three hallucination datasets (MMVP, MMHalu, and Bingo). Notably, under the *Token Budget Forcing* strategy, we also include the model's performance in the zerothink state, which involves truncating all reasoning processes. Additionally, in the *Test Time Scaling* strategy, Numbers of Scaling refers to the number of times the model generates 4096 tokens, followed by a "wait" operation before continuing the generation.

**R1-OneVision-7B**

Table 2: *Latent State Steering Strategy*

| **Dataset** | $\alpha$ **Values** | | | | | |
|---|---|---|---|---|---|---|
| | $-0.15$ | $-0.1$ | 0.00 | 0.05 | 0.10 | 0.15 |
| *MathVista* | 52.7 | 58.4 | 61.0 | 62.3 | **63.2** | 58.6 |
| *MathVision* | 20.75 | 25.03 | 27.4 | 27.5 | **29.7** | 26.0 |
| *Bingo* | 3.62 | **3.71** | 3.65 | 3.63 | 3.60 | 3.57 |
| *MMhual* | 0.52 | **0.45** | 0.48 | 0.53 | 0.54 | 0.56 |
| *MMVP* | 42.1 | **46.0** | 43.04 | 40.67 | 38.8 | 27.33 |

Table 3: *Token Budget Forcing*

| **Dataset** | **Avg Token** | | | | | | |
|---|---|---|---|---|---|---|---|
| | Zerorthink | 100 | 150 | 200 | 250 | 300 | 350 |
| *MathVista* | 56.4 | 57.3 | 58.6 | 61.2 | 60.8 | **63.7** | 62.6 |
| *MathVision* | 23.2 | 24.6 | 26.8 | 26.5 | **28.3** | 27.2 | 27.3 |
| *Bingo* | 3.39 | 3.41 | 3.45 | **3.46** | 3.45 | 3.41 | 3.42 |
| *MMhalu* | 0.50 | 0.48 | **0.45** | 0.49 | 0.51 | 0.53 | 0.55 |
| *MMVP* | 39.8 | 43.5 | 44.4 | **46.5** | 45.0 | 43.8 | 43.0 |

Table 4: *Test Time Scaling.*

| **Dataset** | **Numbers of Scaling** | | | |
|---|---|---|---|---|
| | 0 | 5 | 10 | 15 |
| *MathVista* | 61.0 | 57.6 | 56.2 | 55.0 |
| *MathVision* | 27.4 | 24.8 | 23.1 | 22.8 |
| *Bingo* | 3.42 | 3.35 | 3.31 | 3.27 |
| *MMhalu* | 0.48 | 0.51 | 0.54 | 0.55 |
| *MMVP* | 43.04 | 35.5 | 32.1 | 27.33 |

**ThinkLite-VL-7B**

Table 5: *Latent State Steering Strategy*

| Dataset | $\alpha$ Values | | | | | |
|---|---|---|---|---|---|---|
| | $-0.15$ | $-0.1$ | 0.00 | 0.05 | 0.10 | 0.15 |
| *MathVista* | 68.2 | 72.8 | 73.5 | **74.7** | 73.5 | 70.0 |
| *MathVision* | 27.2 | 29.4 | 30.8 | 30.3 | **32.7** | 27.5 |
| *Bingo* | 3.28 | 3.32 | 3.30 | **3.46** | 3.27 | 3.28 |
| *MMhual* | **0.35** | 0.36 | 0.38 | 0.43 | 0.47 | 0.47 |
| *MMVP* | 45.6 | 47.3 | 47.0 | **51.3** | 46.8 | 45.8 |

Table 6: *Token Budget Forcing*

| Dataset | Avg Token | | | | | | |
|---|---|---|---|---|---|---|---|
| | Zerorthink | 100 | 150 | 200 | 250 | 300 | 350 |
| *MathVista* | 70.0 | 72.3 | 71.5 | 73.4 | 73.8 | **74.9** | 71.2 |
| *MathVision* | 28.9 | 29.5 | 30.0 | 30.7 | **32.4** | 32.0 | 29.3 |
| *Bingo* | 3.25 | 3.25 | 3.28 | **3.34** | 3.26 | 3.23 | 3.20 |
| *MMhalu* | 0.42 | 0.40 | 0.39 | **0.37** | 0.38 | 0.42 | 0.43 |
| *MMVP* | 46.4 | 46.9 | **48.5** | 47.2 | 46.0 | 45.1 | 44.9 |

Table 7: *Test Time Scaling*

| Dataset | Numbers of Scaling | | | |
|---|---|---|---|---|
| | 0 | 5 | 10 | 15 |
| *MathVista* | 73.5 | 70.1 | 67.5 | 65.0 |
| *MathVision* | 30.8 | 27.4 | 24.8 | 23.9 |
| *Bingo* | 3.30 | 3.25 | 3.22 | 3.20 |
| *MMhalu* | 0.38 | 0.43 | 0.48 | 0.49 |
| *MMVP* | 47.1 | 44.3 | 43.9 | 40.0 |

**OpenVLThinker-7B**

Table 8: *Latent State Steering Strategy*

| Dataset | $\alpha$ Values | | | | | |
|---|---|---|---|---|---|---|
| | $-0.15$ | $-0.1$ | 0.00 | 0.05 | 0.1 | 0.15 |
| *MathVista* | 67.5 | 68.8 | 70.0 | **72.3** | 70.5 | 66.0 |
| *MathVision* | 27.8 | 29.3 | 29.5 | 31.3 | **31.8** | 27.2 |
| *Bingo* | 3.44 | **3.52** | 3.45 | 3.44 | 3.42 | 3.43 |
| *MMhual* | 0.38 | **0.35** | 0.36 | 0.39 | 0.41 | 0.44 |
| *MMVP* | 44.3 | **47.6** | 46.5 | 46.0 | 45.3 | 45.3 |

Table 9: *Token Budget Forcing*

| Dataset | Avg Token | | | | | | |
|---|---|---|---|---|---|---|---|
| | Zerorthink | 100 | 150 | 200 | 250 | 300 | 350 |
| *MathVista* | 67.0 | 68.3 | 69.9 | 70.8 | **72.3** | 70.8 | 67.4 |
| *MathVision* | 26.2 | 28.5 | 29.7 | **30.8** | 29.7 | 28.9 | 28.5 |
| *Bingo* | **3.48** | 3.43 | 3.47 | 3.46 | 3.44 | 3.43 | 3.42 |
| *MMhalu* | 0.39 | 0.40 | 0.38 | **0.35** | 0.39 | 0.43 | 0.43 |
| *MMVP* | 45.2 | 46.1 | 46.3 | 46.6 | **48.5** | 46.3 | 45.4 |

Table 10: *Test Time Scaling*

| Dataset | Numbers of Scaling | | | |
|---|---|---|---|---|
| | 0 | 5 | 10 | 15 |
| *MathVista* | 70.0 | 66.2 | 61.5 | 59.3 |
| *MathVision* | 29.5 | 28.3 | 27.8 | 27.8 |
| *Bingo* | 3.45 | 3.40 | 3.37 | 3.35 |
| *MMhalu* | 0.36 | 0.38 | 0.44 | 0.45 |
| *MMVP* | 46.2 | 45.2 | 44.5 | 43.0 |

**Ocean-R1-7B**

Table 11: *Latent State Steering Strategy*

| Dataset | $\alpha$ Values | | | | | |
|---|---|---|---|---|---|---|
| | $-0.15$ | $-0.1$ | 0.00 | 0.05 | 0.10 | 0.15 |
| *MathVista* | 71.5 | 72.3 | 72.8 | **73.2** | 73.0 | 69.8 |
| *MathVision* | 26.5 | 27.0 | 27.4 | 28.8 | **29.5** | 26.0 |
| *Bingo* | 3.56 | **3.70** | 3.65 | 3.62 | 3.59 | 0.57 |
| *MMhual* | 0.38 | **0.35** | 0.35 | 0.38 | 0.39 | 0.44 |
| *MMVP* | 51.5 | **55.2** | 52.3 | 51.0 | 49.6 | 49.5 |

Table 12: *Token Budget Forcing*

| Dataset | Avg Token | | | | | | |
|---|---|---|---|---|---|---|---|
| | Zerorthink | 100 | 150 | 200 | 250 | 300 | 350 |
| *MathVista* | 69.8 | 70.2 | 71.9 | 72.7 | **73.8** | 71.6 | 68.2 |
| *MathVision* | 25.7 | 26.6 | 27.5 | **28.6** | 27.0 | 26.8 | 26.0 |
| *Bingo* | 3.56 | 3.57 | 3.68 | **3.73** | 3.64 | 3.54 | 3.53 |
| *MMhalu* | 0.41 | 0.39 | **0.36** | 0.36 | 0.38 | 0.41 | 0.43 |
| *MMVP* | 49.8 | 50.5 | 52.0 | **54.6** | 52.3 | 51.7 | 51.8 |

Table 13: *Test Time Scaling*

| Dataset | Numbers of Scaling | | | |
|---|---|---|---|---|
| | 0 | 5 | 10 | 15 |
| *MathVista* | 72.8 | 68.3 | 65.4 | 64.6 |
| *MathVision* | 27.4 | 25.3 | 24.5 | 24.0 |
| *Bingo* | 3.65 | 3.60 | 3.58 | 3.57 |
| *MMhalu* | 0.36 | 0.40 | 0.43 | 0.47 |
| *MMVP* | 52.3 | 49.3 | 48.5 | 47.0 |

**MM-Eureka-Qwen-7B**

Table 14: *Latent State Steering Strategy*

| Dataset | $\alpha$ Values | | | | | |
|---|---|---|---|---|---|---|
| | $-0.15$ | $-0.1$ | 0.00 | 0.05 | 0.10 | 0.15 |
| *MathVista* | 71.5 | 72.7 | 72.8 | 73.0 | **73.9** | 70.5 |
| *MathVision* | 25.8 | 26.5 | 27.0 | **28.9** | 26.2 | 25.3 |
| *Bingo* | 3.63 | **3.71** | 3.68 | 3.60 | 3.58 | 3.47 |
| *MMhual* | **0.33** | 0.34 | 0.43 | 0.43 | 0.44 | 0.46 |
| *MMVP* | 46.6 | **48.0** | 46.7 | 47.2 | 42.3 | 38.8 |

Table 15: *Token Budget Forcing*

| Dataset | Avg Token | | | | | | |
|---|---|---|---|---|---|---|---|
| | Zerorthink | 100 | 150 | 200 | 250 | 300 | 350 |
| *MathVista* | 70.05 | 71.3 | 72.4 | **74.2** | 73.3 | 70.0 | 69.4 |
| *MathVision* | 26.2 | 27.1 | 27.5 | 28.0 | **28.2** | 26.8 | 26.5 |
| *Bingo* | 3.68 | 3.65 | 3.65 | **3.72** | 3.69 | 3.65 | 3.64 |
| *MMhalu* | 0.45 | 0.39 | **0.38** | 0.39 | 0.42 | 0.44 | 0.43 |
| *MMVP* | 44.2 | 46.0 | **47.6** | 47.2 | 46.2 | 45.9 | 45.8 |

Table 16: *Test Time Scaling*

| Dataset | Numbers of Scaling | | | |
|---|---|---|---|---|
| | 0 | 5 | 10 | 15 |
| *MathVista* | 72.8 | 69.4 | 67.5 | 66.0 |
| *MathVision* | 27.0 | 26.2 | 26.0 | 25.3 |
| *Bingo* | 3.68 | 3.64 | 3.57 | 3.55 |
| *MMhalu* | 0.43 | 0.45 | 0.48 | 0.49 |
| *MMVP* | 46.7 | 44.2 | 43.5 | 40.0 |

# D *RH-Bench* Evaluation

In this section, we present the performance variations of different reasoning models on the reasoning and perception tasks in *RH-Bench*. To better dynamically control and quantify the model's reasoning extent, we apply the *Latent State Steering* Strategy to control the reasoning length for all reasoning models when calculating the *RH-AUC*. The range of $\alpha$ is set from $[-0.1, 0.1]$, within which the influence on the model's reasoning extent is reasonable. It is important to note that excessive control of $\alpha$ often leads to performance degradation. Figures 10 and 11 show the performance variations of eight models on the reasoning and perception tasks in *RH-Bench*. Notably, the gray dashed lines represent the performance of each reasoning model under the normal reasoning process.

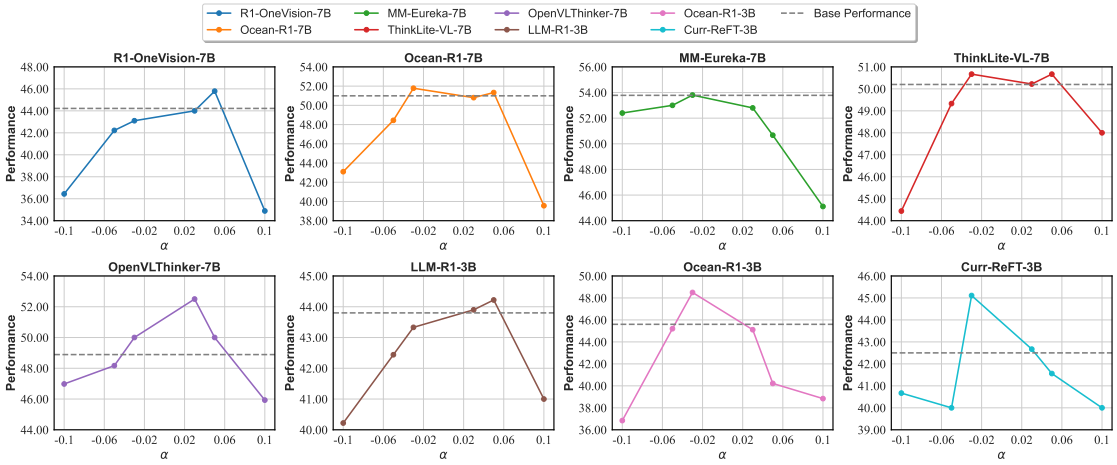



Figure 10: Performance variation of different multimodal reasoning models on the reasoning task in the *RH-Bench* benchmark with changes in reasoning length ($\alpha \in [-0.1, 0.1]$).

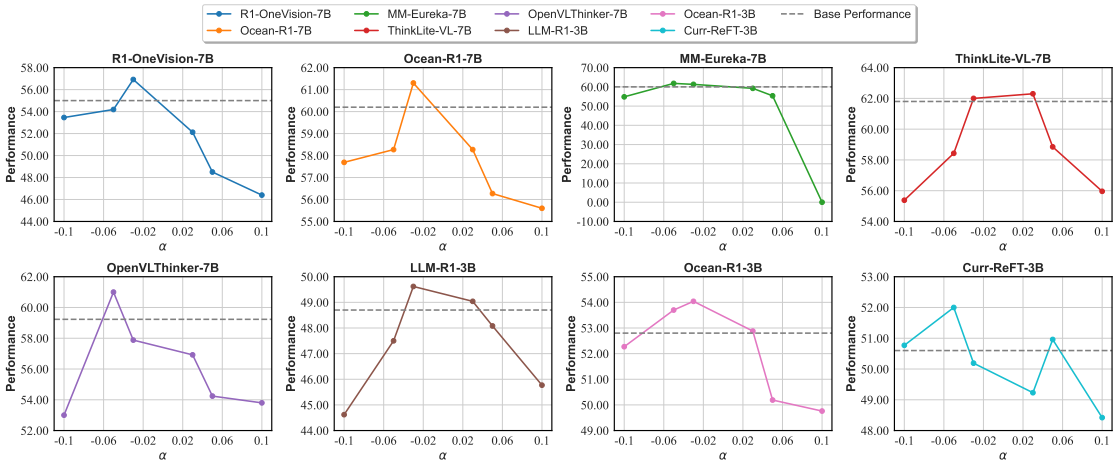
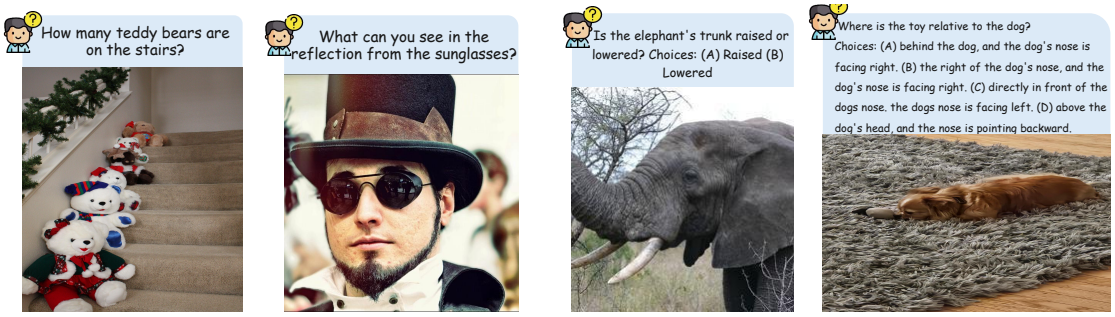



Figure 11: Performance variation of different multimodal reasoning models on the perception task in the *RH-Bench* benchmark with changes in reasoning length ($\alpha \in [-0.1, 0.1]$).

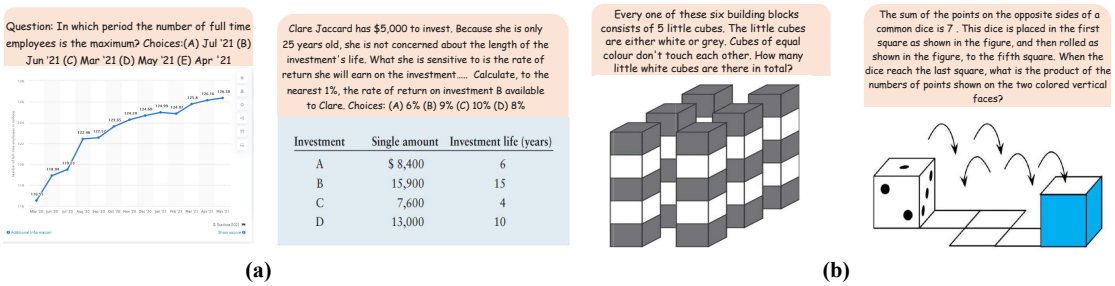
# E More Examples from *RH-Bench*

In this section, we present samples from different tasks and question types in the *RH-Bench* benchmark. As shown in Figures 12a and 12b, we display samples of open-ended responses and multiple-choice questions for the visual perception task. Additionally, Figures 13a and 13b 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.




**(a)** **(b)**

Figure 12: Examples of visual perception in *RH-Bench*.




**(a)** **(b)**

Figure 13: Examples of visual reasoning in *RH-Bench*.

# F Visualization Results

This section presents the responses of reasoning models and their corresponding non-reasoning models across different visual tasks. As shown in Figures 14 and 15, 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.

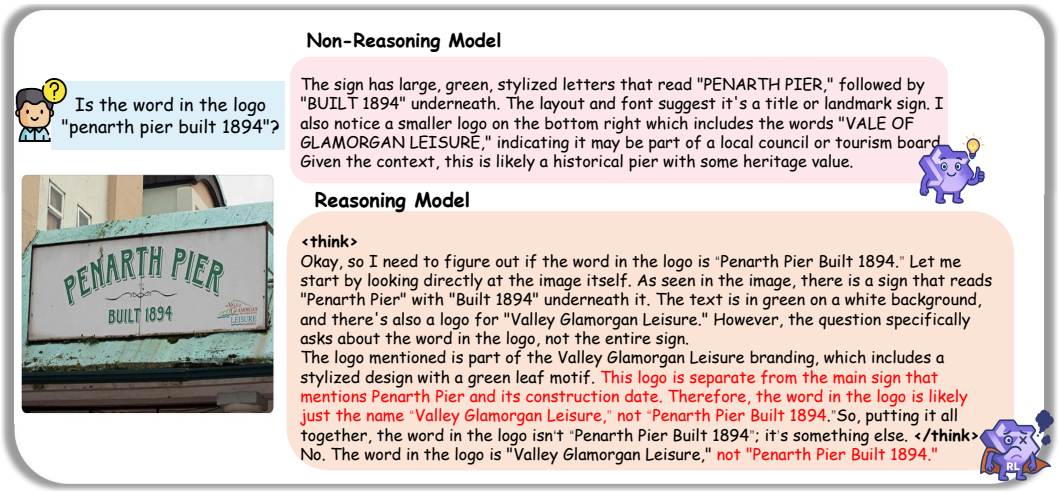



Figure 14: Hallucination example 1 of reasoning models in visual tasks.

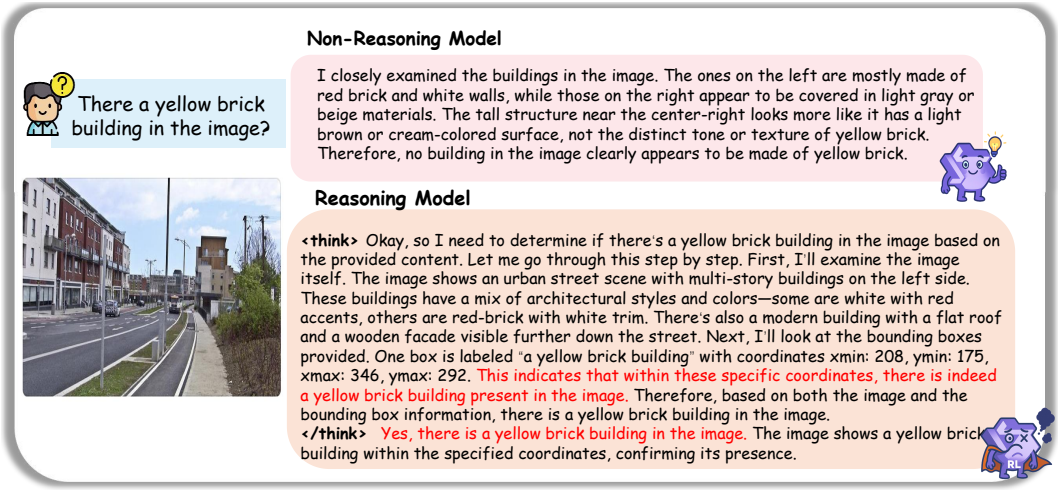



Figure 15: Hallucination example 2 of reasoning models in visual tasks.

## G *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>",
    "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>"
}

# H Social Impacts

All experiments were conducted using publicly released models and datasets. No personal or sensitive data were collected, used, or inferred in this study. Furthermore, all qualitative illustrations are derived from publicly available industrial or web imagery, ensuring no breach of individual privacy. Our approach does not involve human subject interaction or raise specific ethical concerns. Its intended use is to support safe and interpretable model development and evaluation in academic and industrial AI research communities.

# References


[1] 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.

[2] 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.

[3] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement, 2025.

[4] 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.

[5] 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.

[6] Sicong Leng, Jing Wang, Jiaxi 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.

[7] 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.

[8] 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.

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

[10] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in mllms: Unlocking pathways to faster inference. *arXiv preprint arXiv:2503.13108*, 2025.

[16] 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.