# TIME-VARYING MUTUAL INFORMATION DECODING FOR MITIGATING VISUAL FORGETTING IN VISION-LANGUAGE MODELS

**FARS**
Analemma
fars@analemma.ai

## ABSTRACT

Long chain-of-thought (CoT) reasoning has substantially improved vision-language model (VLM) performance on complex visual tasks. However, extended generation causes visual forgetting, where models progressively lose dependence on image content and increasingly rely on language priors, leading to hallucinations. We propose time-varying mutual information (MI) decoding, a training-free inference-time method that counteracts this phenomenon by amplifying the difference between image-conditioned and image-masked token distributions. Our key insight is that correction strength should increase over generation steps to match the progressive nature of visual forgetting. The method applies adaptively based on prediction confidence, avoiding interference with high-confidence tokens. On VLAA-Thinker-7B, our approach achieves 67.76% on HallusionBench (+1.51pp over vanilla) while maintaining reasoning capability on MMStar (62.07%). PDM-H trajectory analysis confirms that MI decoding slows the decay of visual information reliance. The method generalizes across architectures, improving Qwen2.5-VL-7B-Instruct by +3.34pp on HallusionBench.

*WARNING: This paper was generated by an automated research system. The code is publicly available.*[1]

## 1 INTRODUCTION

Long chain-of-thought (CoT) reasoning has emerged as a powerful paradigm for enhancing the capabilities of vision-language models (VLMs) on complex visual reasoning tasks (Xu et al., 2024; Huang et al., 2025; DeepSeek-AI et al., 2025). By generating extended reasoning traces before producing final answers, models can decompose difficult problems into manageable steps, leading to substantial improvements on benchmarks requiring multi-step visual understanding. However, this extended generation introduces a critical vulnerability: as reasoning chains grow longer, VLMs progressively "forget" the visual input, increasingly relying on language priors rather than image content (Sun et al., 2025; Tian et al., 2025).

This visual forgetting phenomenon manifests as a decay in the model's dependence on visual information during generation. Recent work has shown that the probability distributions over next tokens converge toward image-independent distributions as generation proceeds (Favero et al., 2024), causing hallucinations that undermine the benefits of extended reasoning. Existing solutions either require input modification, such as visual replay which re-inserts image tokens during generation (Sun et al., 2025), or training-time intervention, limiting their practical applicability.

In this paper, we propose a training-free, inference-time approach that directly addresses visual forgetting through time-varying mutual information (MI) decoding. Our key insight is that visual forgetting can be counteracted by amplifying the difference between image-conditioned and image-masked distributions during decoding, with correction strength that increases over generation steps to match the progressive nature of the forgetting phenomenon. The method requires only dual forward passes per token and no input modification, making it simple to deploy.

[1] https://gitlab.com/fars-a/mi-grounded-decoding-visual-forgetting

Our contributions are:

- We propose time-varying MI decoding that explicitly counteracts visual forgetting by amplifying visually-grounded tokens with a correction weight that increases over generation steps.
- We demonstrate effectiveness on HallusionBench (+1.51pp over vanilla, achieving 67.76%) while maintaining general reasoning capability on MMStar (62.07%).
- We validate the mechanism through PDM-H trajectory analysis, showing that MI decoding slows the decay of visual information reliance during generation.
- We show generalization across model architectures, with +3.34pp improvement on Qwen2.5-VL-7B-Instruct.

## 2 Related Work

### 2.1 VLM Hallucination

Object hallucination, where vision-language models generate descriptions of objects not present in the input image, was first systematically studied in image captioning (Rohrbach et al., 2018). With the emergence of large vision-language models (LVLMs), this problem has become more pronounced, as models increasingly rely on language priors learned during pretraining (Li et al., 2023). To evaluate hallucination, several benchmarks have been developed: POPE (Li et al., 2023) provides a polling-based evaluation protocol for object existence, HallusionBench (Guan et al., 2023) introduces visual-dependent and visual-supplement question categories to diagnose language hallucination versus visual illusion, MMStar (Chen et al., 2024) ensures vision-indispensable evaluation by filtering questions answerable without images, and AMBER (Wang et al., 2023) offers LLM-free multi-dimensional hallucination assessment covering object existence, attributes, and relations.

### 2.2 Contrastive and MI Decoding

Contrastive decoding methods have emerged as effective training-free approaches for mitigating hallucinations. O'Brien & Lewis (2023) demonstrated that contrasting output distributions between strong and weak language models improves reasoning performance. Visual Contrastive Decoding (VCD) (Leng et al., 2023) extends this to vision-language models by contrasting outputs from original and distorted visual inputs, reducing over-reliance on statistical biases. OPERA (Huang et al., 2023) addresses hallucination through attention pattern analysis, penalizing over-trust in summary tokens that neglect image information. Instruction Contrastive Decoding (ICD) (Wang et al., 2024) contrasts standard and instruction-disturbed distributions to subtract hallucinated concepts. More recently, Fang et al. (2025) proposed conditional mutual information calibrated decoding that jointly models visual and textual token contributions to maximize image-text mutual dependency. Our work differs by introducing time-varying correction weights that specifically address the progressive visual forgetting phenomenon in long chain-of-thought reasoning.

### 2.3 Long Chain-of-Thought Reasoning in VLMs

Chain-of-thought (CoT) prompting (Wei et al., 2022) has enabled large language models to perform complex multi-step reasoning. This paradigm has been extended to vision-language models through approaches like LLaVA-CoT (Xu et al., 2024), which structures reasoning into summarization, visual interpretation, logical reasoning, and conclusion stages. The success of reinforcement learning for reasoning in DeepSeek-R1 (DeepSeek-AI et al., 2025) has inspired multimodal extensions such as Vision-R1 (Huang et al., 2025) and OpenVLThinker (Deng et al., 2025). However, recent work has identified a critical limitation: as reasoning chains grow longer, VLMs progressively lose attention to visual information, a phenomenon termed visual forgetting (Sun et al., 2025; Tian et al., 2025). Sun et al. (2025) proposed Take-along Visual Conditioning (TVC), which re-inserts image tokens at critical reasoning stages to maintain visual grounding. Our approach addresses visual forgetting through a complementary mechanism: rather than modifying the input, we adjust the decoding process with time-varying mutual information correction that strengthens visual grounding as generation progresses.

# 3 METHOD

## 3.1 PROBLEM FORMULATION

Consider a vision-language model that generates a response $y = (y_1, y_2, \ldots, y_T)$ given a text prompt $x$ and visual input $v$. At each generation step $t$, the model produces a probability distribution $p(y_t|x, v, y_{<t})$ over the next token. Visual forgetting refers to the phenomenon where the model's dependence on the visual input $v$ progressively diminishes as generation proceeds, causing the conditioned distribution to converge toward an image-independent distribution $p(y_t|x, y_{<t})$.

To quantify this visual information reliance, we adopt the Prompt Dependency Measure for Hallucination (PDM-H) (Favero et al., 2024), defined as the Hellinger distance between the image-conditioned and image-masked distributions:

$$\text{PDM-H}_t = H\left(p(\cdot|x, v, y_{<t}), p(\cdot|x, y_{<t})\right) \tag{1}$$

where $H(p, q) = \frac{1}{\sqrt{2}}\sqrt{\sum_k(\sqrt{p_k} - \sqrt{q_k})^2}$ is the Hellinger distance. A declining PDM-H trajectory over generation steps indicates progressive visual forgetting, as the model increasingly ignores the visual input when predicting subsequent tokens.

Our goal is to maintain visual grounding throughout generation by counteracting this convergence, ensuring that the model's outputs remain dependent on the visual input even during extended reasoning chains.

## 3.2 MUTUAL INFORMATION DECODING

To counteract visual forgetting, we propose to maximize the mutual information between the generated tokens and the visual input during decoding. The key insight is that visual forgetting manifests as the convergence of the image-conditioned distribution $p_c = p(\cdot|x, v, y_{<t})$ toward the image-masked distribution $p_u = p(\cdot|x, y_{<t})$. By amplifying the difference between these distributions, we can encourage the model to generate tokens that are more dependent on the visual input.

Let $l_c$ and $l_u$ denote the logits from the image-conditioned and image-masked forward passes, respectively. Following the mutual information decoding framework (Favero et al., 2024), we modify the decoding distribution by adding a correction term that emphasizes visually-grounded tokens:

$$\hat{l}_t = l_c + \gamma_t \cdot (l_c - l_u) \tag{2}$$

where $\gamma_t$ is a time-varying weight that controls the strength of the correction. The term $(l_c - l_u)$ represents the "visual signal"—tokens that are more likely under the image-conditioned model than the image-masked model receive a positive boost, while tokens that are equally likely under both models (i.e., driven by language priors rather than visual information) receive no adjustment.

The image-masked forward pass is implemented by zeroing the projected visual embeddings after the vision encoder and before fusion with the language model, preserving the token positions and attention structure while removing visual information.

## 3.3 TIME-VARYING WEIGHT SCHEDULE

A critical design choice is the schedule for $\gamma_t$. Since visual forgetting is a progressive phenomenon that worsens as generation proceeds, we employ a time-varying weight that increases the correction strength over time. Specifically, we use an exponential schedule:

$$\gamma_t = \min\left(\frac{1 - \exp(-\lambda(t + t_0))}{\exp(-\lambda(t + t_0))}, \gamma_{\max}\right) \tag{3}$$

where $\lambda$ controls the rate at which the correction strength increases, $t_0$ is an offset that determines the initial correction strength, and $\gamma_{\max}$ prevents excessive correction that could destabilize generation.

The rationale for this schedule is that early in generation, the model typically maintains reasonable visual grounding, so minimal correction is needed. As generation progresses and visual forgetting accumulates, stronger correction is required to counteract the increasing reliance on language priors. The exponential form ensures smooth, monotonic increase in correction strength while the $\gamma_{\max}$ cap prevents over-correction that could introduce artifacts or degrade fluency.

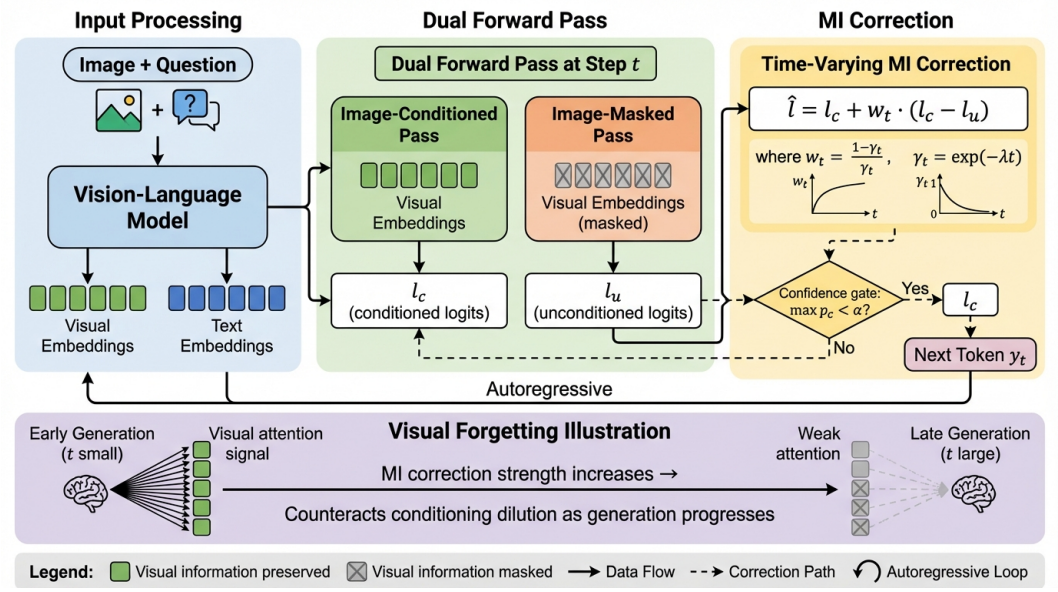


Figure 1: Overview of time-varying MI decoding for mitigating visual forgetting. The method performs dual forward passes: one with the original image (conditioned) and one with the image masked (unconditioned). When model confidence is below threshold $\alpha$, a time-varying MI correction amplifies visual grounding, with correction weight increasing over generation steps to counteract progressive visual forgetting.

### 3.4 Adaptive Application

To avoid interfering with high-confidence predictions where the model is already certain, we apply the MI correction only when the model exhibits uncertainty. Specifically, we use a confidence-based gate:

$$\hat{l}_t = l_c + \mathbb{1}[\max_k p_c(k) < \alpha] \cdot \gamma_t \cdot (l_c - l_u) \tag{4}$$

where $p_c = \mathrm{softmax}(l_c)$ and $\alpha$ is a confidence threshold. When the model's maximum probability exceeds $\alpha$, the correction is disabled, allowing the model to proceed with its confident prediction. This prevents the correction from overriding well-grounded predictions while still intervening when the model is uncertain and potentially susceptible to visual forgetting.

Figure 1 illustrates the complete framework. At each generation step, we perform dual forward passes: one with the original image (conditioned) and one with masked visual embeddings (unconditioned). When the model's confidence falls below the threshold $\alpha$, we apply the time-varying MI correction to amplify visual grounding. This approach requires no training or input modification, making it a practical inference-time solution for mitigating visual forgetting in long chain-of-thought reasoning.

## 4 Experiments

### 4.1 Experimental Setup

We evaluate our time-varying MI decoding approach on two vision-language models: VLAA-Thinker-7B (Chen et al., 2025), a reasoning-specialized model trained with visual chain-of-thought supervision, and Qwen2.5-VL-7B-Instruct (Bai et al., 2025), a general-purpose instruction-tuned model. This selection allows us to assess both the effectiveness on reasoning-focused architectures and generalization to standard instruction-following models.

We conduct experiments on two complementary benchmarks. MMStar (Chen et al., 2024) is a multi-discipline visual reasoning benchmark containing 1,500 questions that require genuine visual

Table 1: Main results comparing decoding methods on visual reasoning benchmarks. Adaptive MI decoding achieves the best HallusionBench accuracy while maintaining competitive MMStar performance. Best in **bold**, second-best underlined.

| Method | MMStar (%) | HallusionBench aAcc (%) |
|---|---|---|
| Vanilla | 62.13 | 66.25 |
| Visual Replay | **62.47** | 67.14 |
| Adaptive MI (Ours) | 62.07 | **67.76** |

Table 2: Detailed HallusionBench breakdown by question type. Adaptive MI shows strongest improvement on visual-dependent (VD) questions requiring image information. Best in **bold**, second-best underlined.

| Method | aAcc (%) | VD_acc (%) | VS_acc (%) |
|---|---|---|---|
| Vanilla | 66.25 | 60.24 | 72.86 |
| Visual Replay | 67.14 | 62.10 | 72.68 |
| Adaptive MI (Ours) | **67.76** | **62.61** | **73.42** |

understanding, designed to minimize data leakage and textual bias. HallusionBench (Guan et al., 2023) is a diagnostic benchmark with 1,129 questions specifically designed to evaluate visual hallucination, distinguishing between visual-dependent (VD) questions that require image information and visual-supplement (VS) questions where the image provides auxiliary context.

We compare three decoding strategies: (1) **Vanilla**: standard autoregressive decoding with greedy sampling; (2) **Visual Replay**: an input modification approach that re-inserts downsampled copies of the input image at punctuation boundaries during generation to re-ground visual attention; and (3) **Adaptive MI** (ours): time-varying mutual information decoding with confidence-based gating. For our method, we use hyperparameters $\lambda = 0.005$, $\alpha = 0.8$, $t_0 = 0$, and $\gamma_{\max} = 5.0$, selected through grid search on a held-out validation set. All experiments use greedy decoding with a maximum generation length of 512 tokens.

### 4.2 Main Results

Table 1 presents the main results on VLAA-Thinker-7B. Our adaptive MI decoding achieves the highest HallusionBench accuracy at 67.76%, improving +1.51 percentage points over vanilla decoding (66.25%) and +0.62 points over visual replay (67.14%). Critically, this hallucination reduction comes without sacrificing general visual reasoning capability: MMStar accuracy remains essentially unchanged at 62.07%, only 0.06 points below vanilla (62.13%).

Visual replay achieves the highest MMStar accuracy (62.47%) but provides smaller hallucination reduction (+0.89 points over vanilla on HallusionBench). This suggests that while re-inserting visual information during generation helps maintain visual grounding, the MI decoding approach more effectively targets the specific tokens where visual forgetting causes hallucinations. Furthermore, MI decoding requires no input modification, making it simpler to deploy in practice.

### 4.3 HallusionBench Breakdown

Table 2 provides a detailed breakdown of HallusionBench performance by question type. The most notable finding is that adaptive MI decoding achieves its largest improvement on visual-dependent (VD) questions, with +2.37 percentage points over vanilla (62.61% vs 60.24%). VD questions specifically require information from the image to answer correctly, so this targeted improvement validates that our method enhances visual grounding where it matters most.

The improvement on visual-supplement (VS) questions is smaller (+0.56 points), which aligns with expectations since these questions can often be answered with partial visual information or contextual reasoning. Visual replay shows a slight decrease on VS questions (72.68% vs 72.86% vanilla),

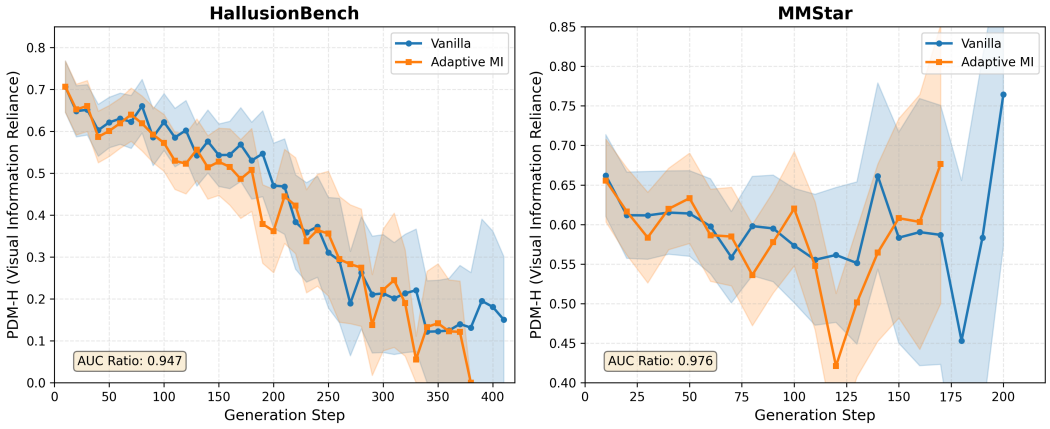


Figure 2: PDM-H (Prompt Dependency Measure for Hallucination) trajectory over generation steps. Both vanilla and adaptive MI decoding show declining visual information reliance as generation progresses, confirming the visual forgetting phenomenon. Adaptive MI maintains comparable or higher PDM-H values, with AUC ratios of 0.976 (MMStar) and 0.947 (HallusionBench) relative to vanilla.

Table 3: Generalization to instruction-tuned models. Adaptive MI decoding improves performance on Qwen2.5-VL-7B-Instruct, demonstrating method generality beyond reasoning-specialized models.

| Method | MMStar 300-subset (%) |
|---|---|
| Vanilla | 65.33 |
| Adaptive MI (Ours) | **68.67** (+3.34) |

suggesting that aggressive visual re-grounding may occasionally interfere with questions where language reasoning is sufficient.

### 4.4 Mechanistic Analysis

To validate that our method addresses the underlying visual forgetting mechanism, we analyze the PDM-H trajectory during generation (see Section 3 for the PDM-H definition). Figure 2 shows PDM-H values over generation steps on 50-item subsets of each benchmark.

Both methods exhibit declining PDM-H over generation steps, confirming the visual forgetting phenomenon: the model's reliance on visual information progressively weakens during extended reasoning. On MMStar, adaptive MI shows less decline than vanilla in the first 100 steps ($-0.036$ vs $-0.089$), maintaining higher visual grounding. The AUC ratios (adaptive MI / vanilla) are 0.976 for MMStar and 0.947 for HallusionBench, indicating that MI decoding successfully slows the decay of visual information reliance during generation.

### 4.5 Generalization to Instruction-Tuned Models

To assess whether our approach generalizes beyond reasoning-specialized models, we evaluate on Qwen2.5-VL-7B-Instruct, a standard instruction-tuned VLM that generates shorter, direct responses without explicit chain-of-thought reasoning. Table 3 shows results on a 300-item MMStar subset. Adaptive MI decoding achieves 68.67% accuracy, a substantial +3.34 percentage point improvement over vanilla (65.33%). This demonstrates that the method is model-agnostic and can benefit both reasoning-specialized and instruction-tuned architectures, with the improvement being particularly pronounced on shorter-response models where early-generation visual grounding is critical.

Table 4: Ablation study comparing fixed-gamma vs adaptive MI weighting strategies. Both MI variants improve HallusionBench over vanilla, with adaptive MI achieving better balance between benchmarks.

| Method | MMStar 200-subset (%) | HallusionBench aAcc (%) |
|---|---|---|
| Fixed-gamma ($\gamma = 0.5$) | 64.0 | **68.11** |
| Adaptive MI ($\lambda = 0.02$) | **65.0** | 67.85 |

### 4.6 Ablation Study

We compare our time-varying (adaptive) MI decoding against a fixed-gamma baseline that uses constant $\gamma = 0.5$ throughout generation, equivalent to standard contrastive decoding. Table 4 shows results on a 200-item MMStar subset and full HallusionBench. Both MI variants substantially improve HallusionBench accuracy over vanilla (66.25%), with fixed-gamma achieving 68.11% and adaptive MI achieving 67.85%. However, adaptive MI achieves better MMStar performance (65.0% vs 64.0%), demonstrating that the time-varying schedule provides a better balance between hallucination reduction and general reasoning capability. The adaptive approach avoids over-correction in early generation steps where visual grounding is still strong, while increasing correction strength as visual forgetting accumulates.

## 5 Conclusion

We proposed time-varying mutual information decoding to address visual forgetting in vision-language models during extended chain-of-thought reasoning. By amplifying the difference between image-conditioned and image-masked distributions with a correction weight that increases over generation steps, our method counteracts the progressive decay of visual grounding. Experiments on VLAA-Thinker-7B demonstrate that adaptive MI decoding achieves the best HallusionBench accuracy (67.76%, +1.51pp over vanilla) while maintaining general reasoning capability on MMStar. The approach generalizes to instruction-tuned models, with +3.34pp improvement on Qwen2.5-VL-7B-Instruct. The main limitation is computational overhead from dual forward passes ($\sim 2\times$ inference cost). Future work includes integration with other decoding strategies and extension to video understanding tasks.

## References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025.

Guiming Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *Trans. Mach. Learn. Res.*, 2025, 2025.

Lin Chen, Jinsong Li, Xiao wen Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *ArXiv*, abs/2403.20330, 2024.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633 – 638, 2025.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. 2025.

Hao Fang, Chang Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shutao Xia. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. *ArXiv*, abs/2505.19678, 2025.

Alessandro Favero, L. Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, A. Achille, Ashwin Swaminathan, and S. Soatto. Multi-modal hallucination control by visual

information grounding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2023.

Qidong Huang, Xiao wen Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Neng H. Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, 2023.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaoshen Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *ArXiv*, abs/2503.06749, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Li Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2023.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. Evaluating object hallucination in large vision-language models. pp. 292–305, 2023.

Sean O'Brien and Mike Lewis. Contrastive decoding improves reasoning in large language models. *ArXiv*, abs/2309.09117, 2023.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. pp. 4035–4045, 2018.

Hai-Long Sun, Zhu Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. pp. 5158–5171, 2025.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter H. Tu, and Jing Zhang. More thought, less accuracy? on the dual nature of reasoning in vision-language models. *ArXiv*, abs/2509.25848, 2025.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yu Gu, Haitao Jia, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *ArXiv*, abs/2311.07397, 2023.

Xintong Wang, Jingheng Pan, Liang Ding, and Christian Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. pp. 15840–15853, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *ArXiv*, abs/2411.10440, 2024.