Hallucination Mitigation or Conservative Decoding? Rethinking Faithfulness Evaluation in LVLMs

Anonymous ACL submission

Abstract

Recent inference-time hallucination mitigation methods for large vision-language models (LVLMs) report strong gains on hallucination benchmarks. However, it remains unclear whether lower hallucination scores reflect improved multimodal grounding or more conservative generation. We evaluate six mitigation methods across three LVLMs and four benchmarks, including hallucination-focused evaluation and the diverse capability benchmark MMSTAR. Our analysis reveals two consistent patterns. First, hallucination reduction is often coupled with reduced informativeness: methods that lower hallucination rates also reduce object recall, visual coverage, or response detailedness. Second, improvements on hallucination benchmarks do not reliably transfer to broader multimodal capabilities, with methods showing inconsistent or degraded performance on fine-grained perception and reasoning tasks. Our findings suggest that current evaluation protocols may overestimate progress by rewarding conservative generation. We argue that hallucination mitigation should be evaluated as a faithfulness–informativeness–capability trade-off rather than through hallucination scores alone. Data and code are available at https://anonymous.4open.science/r/AssessHalVLM-48FB/

1 Introduction

Large vision-language models (LVLMs) achieve strong performance across multimodal tasks but remain prone to visual hallucination, including *object hallucination*: mentioning objects that are unsupported by the image. To address this, a growing body of work proposes training-free, decoding-time mitigation methods that adjust the model's output distribution without retraining (Leng et al., 2024; Huang et al., 2024; Favero et al., 2024; Woo et al., 2024). These methods report substantial gains on standard hallucination benchmarks such as CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023), and AMBER (Wang et al., 2023), and have become standard baselines for inference-time hallucination mitigation.

We argue that these gains may be misleading. Many widely used hallucination benchmarks emphasize object-level hallucination in COCO-style natural images, where a method can lower hallucination scores by generating shorter, non-committal responses that mention fewer visual entities, minimizing opportunities to err without improving visual grounding. We call this behavior *risk suppression*: reducing unsupported visual claims by reducing visual commitment, rather than by improving multimodal grounding. Additionally, a method may score well on hallucination benchmarks while failing to improve, or even compromising, broader multimodal capabilities such as reasoning, and fine-grained perception. We call this disconnect the *faithfulness–capability gap*: improvements in narrow faithfulness metrics do not reliably transfer to broader multimodal competence.

We conduct a diagnostic evaluation of six inference-time mitigation methods across three LVLMs and three benchmarks. We evaluate risk suppression on CHAIR and AMBER, and faithfulness–capability gap on MMSTAR to test whether mitigation gains transfer to perception, reasoning, science, and mathematics tasks. To detect risk suppression, we pair hallucination metrics with informativeness metrics drawn from the same benchmark: object recall for CHAIR and visual coverage (Cover) for AMBER. Our analysis yields two findings:

- **Hallucination reduction often reflects risk suppression.** Across both CHAIR and AMBER, methods that reduce hallucination tend to reduce informativeness in tandem: lower CHAIR correlates with lower object recall, and lower AMBER hallucination correlates

Evaluation Axis	Component Class	Target Instantiations / Metrics
Base Models	Open-Source LVLMs	LLaVA-1.5, LLaVA-NeXT, InstructBLIP
Mitigation Methods	Contrastive Decoding	VCD (Leng et al., 2024), M3ID (Favero et al., 2024)
	Attention Calibration	AGLA (An et al., 2025), CAAC (Fazli et al., 2026a)
	Hidden-State Modification	CEI (Fazli et al., 2026b), AFTER (Wang et al., 2026)
Hallucination Benchmarks	faithfulness-informativeness	CHAIR (Rohrbach et al., 2018) AMBER (Wang et al., 2023)
Capability Benchmark	vision-indispensable benchmark	MMSTAR (Chen et al., 2024b)

Table 1: Systematic diagnostic evaluation matrix spanning our 54 experimental configurations, decoupling error mitigation from information retention and broad reasoning capability.

with lower Cover. The methods that score best on hallucination are often the most conservative, rather than the most informative.

- **Hallucination gains do not reliably transfer to broader capabilities.** On MMSTAR, mitigation methods show inconsistent or degraded performance, with effects that flip in direction across models and categories. Gains on object-hallucination benchmarks frequently fail to translate to fine-grained perception and reasoning tasks such as math, science, and logical reasoning.

These results suggest that current evaluation protocols may overestimate progress on LVLM faithfulness by rewarding conservative decoding behavior. We argue that hallucination mitigation should be evaluated as a *faithfulness–informativeness–capability trade-off* rather than by hallucination scores alone.

2 Diagnostic Evaluation Setup

To rigorously expose the trade-offs of inference-time hallucination mitigation, we conduct an evaluation spanning 3 LVLMs $\times$ 6 methods $\times$ 3 benchmarks, totaling 54 distinct experimental configurations (Table 1).

Models & Mitigation Methods. We evaluate the 7B-parameter variants of three widely adopted LVLMs (Wang et al., 2026; Fazli et al., 2026a) in the hallucination literature: LLaVA-1.5 (Liu et al., 2024b), LLaVA-NeXT (Liu et al., 2024c), and InstructBLIP (Dai et al., 2023). For each architecture, we benchmark the vanilla decoding baseline against six representative inference-time hallucination mitigation methods spanning the three dominant families: *contrastive decoding* (VCD (Leng et al., 2024), M3ID (Favero et al., 2024)), *attention calibration* (AGLA (An et al., 2025), CAAC (Fazli et al., 2026a)), and *hidden-state modification* (CEI (Fazli et al., 2026b), AFTER (Wang et al., 2026)). Furthermore, this selection deliberately balances foundational, well-respected approaches (VCD, M3ID, AGLA) with recent state-of-the-art advancements (CAAC, CEI, AFTER) in hallucination mitigation.

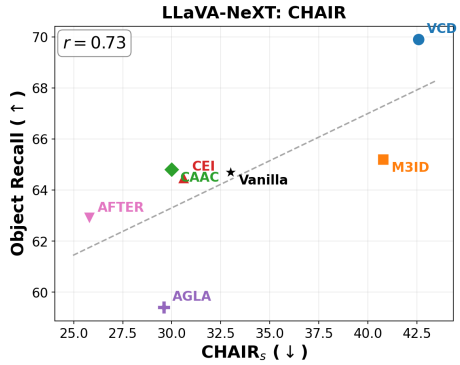
Benchmark Selection & Decoupled Metrics. To systematically decouple hallucination reduction from information loss and general intelligence degradation, we partition our evaluation suite into two distinct categories:

Hallucination-Informativeness Suites: We select two established benchmarks that specifically provide paired metrics for both hallucination rate and informativeness. CHAIR (Rohrbach et al., 2018) evaluates object-existence errors in open-ended MS-COCO captioning; we report CHAIR_s/CHAIR_i (lower is better) paired with *Object Recall* (higher is better) as an informativeness proxy. AMBER (Wang et al., 2023) evaluates generative responses; we report its object hallucination rate metrics (CHAIR and Hal) alongside *Cover*, the fraction of ground-truth visual entities successfully mentioned.

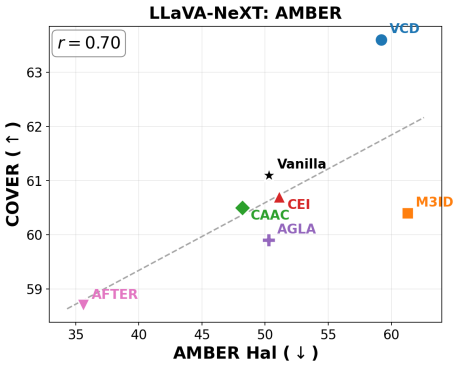
Broader Multimodal Capability Benchmark: To evaluate how hallucination mitigation methods influence broader model capabilities, we leverage MMSTAR (Wang et al., 2023; Chen et al., 2024b). Crucially, MMSTAR distills multimodal examples from existing benchmarks by purging samples solvable via textual cues or shallow visual shortcuts, enabling a rigorous evaluation of six core multimodal capabilities: coarse perception, fine-grained perception, instance reasoning, logical reasoning, science and technology, and mathematics. We report overall and category-specific accuracy.

3 Hallucination Reduction as Risk Suppression

Genuine improvements in multimodal faithfulness require minimizing hallucination rates without sacrificing visual information density. Under the *risk*



(a) CHAIR: CHAIR$_s$ vs. Object Recall



(b) AMBER: Hallucination Rate vs. Cover

Figure 1: The faithfulness–informativeness trade-off on LLaVA-NeXT-7B. Both (a) CHAIR and (b) AMBER exhibit a strict positive correlation between error rates and informativeness. The target optimal quadrant is the upper-left of the vanilla decoding (minimized errors, maximized recall).

suppression hypothesis, however, localized drops in hallucination errors often reflect a reduction in detailedness and visual commitment. An optimal model should navigate toward the upper-left quadrant of the original model in a hallucination–informativeness space, simultaneously achieving low error and high informativeness. As demonstrated below, current mitigation methods instead shift models along a risk-hedging diagonal.

Figure 1 illustrates these operational trade-offs for LLaVA-NeXT (7B), with full cross-model distributions for LLaVA-1.5 and InstructBLIP provided in Appendix B.

Information density trade-offs in captioning. Across both evaluation suites, hallucination rates and informativeness exhibit a powerful positive correlation. On CHAIR, sentence-level hallucination (CHAIR$_s$) and object recall are strongly correlated (Pearson $r = 0.73$). Only CAAC and CEI preserved the recall while mitigating hallucination, whereas other methods like AFTER and AGLA dropped the recall as they reduced the hallucinations. A similar pattern is manifested on AMBER, where instance-level hallucination (Hal) and object coverage (Cover) correlate at $r = 0.70$. Here, AFTER achieves the lowest nominal hallucination rate but achieves this at the cost of reduced coverage. Interestingly, contrastive decoding frameworks (e.g., VCD and M3ID) actively amplify vision-conditioned tokens over language priors. This mechanism shifts models toward a state of hyper-loquacity—inducing an over-descriptiveness that simultaneously inflates both hallucination rates and informativeness. None of the evaluated methods could penetrate the upper-left quadrant and most followed the hallucination–informativeness diagonal.

4 Hallucination Gains Do Not Reliably Transfer to General Capabilities

To verify if a decoding-time intervention genuinely improves visual grounding, we must evaluate whether it preserves or enhances broader visual understanding and reasoning capabilities. Under the RISK SUPPRESSION framework, we hypothesize that methods enforcing a more conservative, less detailed output distribution will suffer performance regressions on benchmarks requiring explicit, non-trivial visual commitments. We evaluate this premise using MMSTAR (Chen et al., 2024b), testing across six distinct multimodal categories where models cannot rely on simple textual cues or risk-averse generation to succeed.

Figure 2 displays accuracy changes (Δ) relative to vanilla decoding across all 54 configurations, across six multimodal tasks from MMSTAR. Our matrix analysis reveals two key observations:

Impairing fine-grained perception and reasoning tasks. decoding modifications systematically damage fine-grained perception (FP) across the board. Out of all 18 evaluated FP configurations, only two settings managed to yield isolated gains of more than 1%, while the vast majority across all three models suffered performance degradation. Similarly, the mitigation methods overwhelmingly degrade the instance reasoning and logical reasoning particularly in InstructBLIP and LLaVA-1.5. Figure 3 showcases an example in LR where every single mitigation strategy erodes the base model's accurate response. More examples are provided in Appendix C.

Most effects are modest and inconsistent. No method consistently improves MMSTAR over

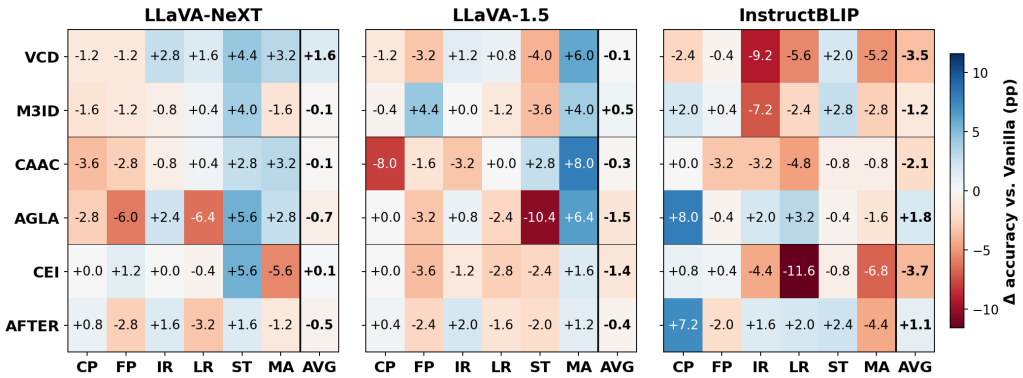


Figure 2: Per-category accuracy change (Δ vs. vanilla baseline, in percentage points) on MMSTAR across three distinct LVLM architectures. Columns track specific MMSTAR task categories: Coarse Perception (CP), Fine-grained Perception (FP), Instance Reasoning (IR), Logical Reasoning (LR), Science & Technology (ST), and Mathematics (MA). Red cells indicate performance degradation, while blue cells mark performance improvements.

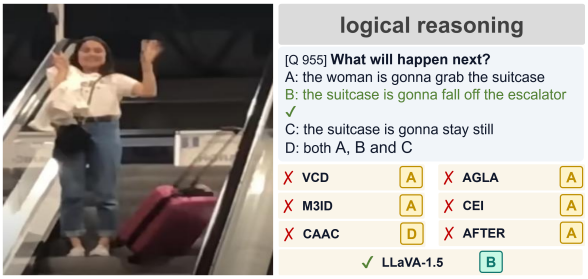


Figure 3: An MMSTAR example demonstrating logical reasoning collapse. While the vanilla LLaVA-1.5 accurately reasons about the image, all six mitigation methods lead to incorrect options.

vanilla across models or categories; in fact, every tested mitigation method fails to improve macro averages on more than one out of the three base architectures. Additionally, gains in one category are typically offset by losses elsewhere. For instance, AFTER boosts coarse perception (CP) by 7.2% on InstructBLIP, yet simultaneously undermines mathematical reasoning (MA) by 4.4% on the same architecture. Similarly, while AGLA on LLaVA-1.5 yields a 6.4% increase in mathematics, it compromises performance on scientific and technical tasks, dropping *Science & Technology* (ST) by a substantial 10.4 percentage points.

5 Related Work and Discussion

Inference-time hallucination mitigation. Training-free inference interventions have become a dominant paradigm for hallucination mitigation (Bai et al., 2025; Liu et al., 2024a). Contrastive decoding strategies bias token selection toward image-supported representations using alternative distributions (Chuang et al., 2024; Leng et al., 2024; Zhu et al., 2024; Chen et al., 2024a). Concurrently, attention-based methods intervene directly on visual grounding by amplifying image attention (Liu et al., 2024d), calibrating attention patterns (Zhu et al., 2025; Zhang et al., 2024), or penalizing uninformative and over-trusted tokens (Huang et al., 2024; Woo et al., 2024).

Hallucination benchmarks. Hallucination benchmarks are commonly discriminative, such as POPE (Li et al., 2023), MME (Fu et al., 2024), and HallusionBench (Guan et al., 2024), or generative, such as CHAIR (Rohrbach et al., 2018), and AMBER (Wang et al., 2023). Discriminative benchmarks provide objective yes/no-style evaluation, while generative benchmarks evaluate free-form responses but remain largely object-centric. Most hallucination mitigation studies primarily report hallucination rates, with limited evaluation of informativeness or broader multimodal capabilities. In contrast, our work jointly evaluates hallucination, informativeness, and transfer to general LVLM capability benchmarks.

6 Conclusion

We presented a diagnostic study of six inference-time hallucination mitigation methods across three LVLMs and three benchmarks. Our results show that hallucination reductions on CHAIR and AMBER often coincide with reduced informativeness, and these gains do not reliably transfer to broader multimodal capabilities on MMSTAR. We argue that hallucination mitigation must be evaluated as a faithfulness–informativeness–capability trade-off, as lower hallucination rates alone provide insufficient evidence of better visual grounding.

Limitations

Benchmark coverage. Our evaluation covers three benchmarks (CHAIR, AMBER, MMSTAR). While these span object-level hallucination, generative coverage, LLM-judge scoring, and broad multimodal capability, they do not cover all relevant axes of LVLM faithfulness, such as attribute-level or relational hallucination, video understanding, or long-form instruction following. The patterns reported here may not extend uniformly to those settings.

Scope of mitigation methods. We focus on representative inference-time hallucination mitigation methods, including decoding-time and attention-based interventions. These methods are widely used because they are training-free and can be applied to existing LVLMs, but they do not cover all possible mitigation strategies like training-based approaches, preference optimization, and retrieval-augmented grounding. Our conclusions therefore apply most directly to inference-time distribution-shaping methods.

Model scale. We evaluate three 7B-scale LVLMs. Larger or more recent models may respond differently to decoding-time interventions, and the trade-offs we observe may vary with model scale and architecture.

Hyperparameter Scoping. We implement each mitigation method using the hyperparameter configurations specified by their respective authors. When applying a method to an LVLM not included in its original study, we maintain these default settings rather than conducting exhaustive per-model optimization. While custom tuning might alter individual performance profiles, our approach reflects standard, out-of-the-box deployment conditions.

Causal interpretation. Our analysis demonstrates that hallucination metrics are entangled with informativeness and do not reliably transfer to broader capabilities. We do not directly measure "grounding" as a latent property; RISK SUPPRESSION describes the observed evaluation pattern, not a mechanistic claim about why each method behaves as it does.

Potential Risks. While our work primarily focuses on the critical analysis and evaluation of existing hallucination mitigation methods, a potential risk is that highlighting the systematic trade-offs of these techniques could inadvertently guide the development of generation strategies that look superficially accurate while masking underlying unreliability. Additionally, understanding how to bypass or exploit conservative decoding behaviors could theoretically be used to manipulate large vision-language models (LVLMs) in safety-critical applications.

AI Assistance Disclosure AI language models were utilized for text editing, grammatical refinement, and coding assistance. The authors have fully reviewed and verified all technical content, experimental results, and core claims.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of Multimodal Large Language Models: A Survey, April 2025. URL http://arxiv.org/abs/2404.18930. arXiv:2404.18930 [cs].

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Hallucination in Large Vision-Language Models, November 2024a. URL http://arxiv.org/abs/2411.15268. arXiv:2411.15268 [cs].

Lin Chen, Jinsong Li, Xiaoyi 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? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models, March 2024. URL http://arxiv.org/abs/2309.03883. arXiv:2309.03883 [cs].

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14303–14312, 2024.

Mehrdad Fazli, Bowen Wei, Ahmet Sari, and Ziwei Zhu. Caac: Confidence-aware attention calibration to reduce hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1242–1251, 2026a.

Mehrdad Fazli, Bowen Wei, and Ziwei Zhu. Context-aware decoding for faithful vision-language generation. In *arXiv preprint arXiv:2601.05939*, 2026b.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models, March 2024. URL http://arxiv.org/abs/2306.13394. arXiv:2306.13394 [cs].

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, March 2024. URL http://arxiv.org/abs/2310.14566. arXiv:2310.14566 [cs].

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 (CVPR)*, pages 13418–13427, 2024.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 12888–12900, 2022.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A Survey on Hallucination in Large Vision-Language Models, May 2024a. URL http://arxiv.org/abs/2402.00253. arXiv:2402.00253 [cs].

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, January 2024c.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying More Attention to Image: A Training-Free

Method for Alleviating Hallucination in LVLMs, July 2024d. URL http://arxiv.org/abs/2407.21771. arXiv:2407.21771 [cs].

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 (EMNLP)*, pages 4035–4045, 2018.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Tianbo Wang, Yuqing Ma, Kewei Liao, Zhange Zhang, Simin Li, Jinyang Guo, and Xianglong Liu. AFTER: Mitigating the Object Hallucination of LVLM via Adaptive Factual-Guided Activation Editing. In *The Fourteenth International Conference on Learning Representations*, 2026.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *arXiv preprint arXiv:2405.17820*, 2024.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing Clearly by Layer Two: Enhancing Attention Heads to Alleviate Hallucination in LVLMs, November 2024. URL http://arxiv.org/abs/2411.09968. arXiv:2411.09968 [cs].

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. IBD: Alleviating Hallucinations in Large Vision-Language Models via Image-Biased Decoding, February 2024. URL http://arxiv.org/abs/2402.18476. arXiv:2402.18476 [cs].

Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. Mitigating Object Hallucinations in Large Vision-Language Models via Attention Calibration, February 2025. URL http://arxiv.org/abs/2502.01969. arXiv:2502.01969 [cs].

A Full Experimental Details

A.1 Hyperparameters

We organize hyperparameters by method family. Within each family we follow the values reported in the original publication; deviations are flagged explicitly. The full per-method, per-benchmark configuration is also provided in the released codebase at: https://anonymous.4open.science/r/AssessHalVLM-48FB/.

Contrastive decoding methods. VCD (Leng et al., 2024) and M3ID (Favero et al., 2024) both compute a contrast between the model's logits on the original image and its logits on a perturbed logit distribution, then subtract the reference's distribution from the conditional one to lower the hallucinatory token's probabilities. Both methods retain only tokens that exceed an adaptive plausibility threshold (controlled by β) to avoid amplifying tail noise. Table 2 lists the values used for both methods.

Method	Parameter	Value
VCD (Leng et al., 2024)	α (contrast strength)	1.0
	β (plausibility threshold)	0.1
	DDPM noise step	500
M3ID (Favero et al., 2024)	λ (amplification coef.)	0.2
	β (plausibility threshold)	0.1

Table 2: Hyperparameters for contrastive-decoding methods. The M3ID λ=0.2 matches the runner-level default of the CAAC baselines codebase (Fazli et al., 2026a); the original M3ID paper reports a slightly different default that we did not retune for fairness across cells.

Attention-based methods. CAAC (Fazli et al., 2026a) calibrates the attention weights placed on visual tokens at decoding time, downweighting tokens whose attention concentration is inconsistent with overall visual saliency. AGLA (An et al., 2025) assembles two attention paths: a *global* path attending to the full image and a *local* path attending to a masked, augmented image; the per-token logits are mixed additively as $\text{logits} + \alpha \cdot \text{logits}_{\text{cd}}$, with an adaptive masking ratio derived from a BLIP image–text-matching score (Li et al., 2022). We use the released default configurations for both methods, summarised in Table 3.

Method	Parameter	Value
CAAC (Fazli et al., 2026a)	Calibration scope	All decoder layers
	Confidence aggregation	Token-level (default)
	Inference budget	Matches vanilla
AGLA (An et al., 2025)	α (mixing ratio)	2.0
	β (plausibility threshold)	0.5
	Augmentation backbone	BLIP-ITM (large), Grad-CAM block 6
	Masking ratio	$1 - \mathrm{ITC}/2$
	Inference budget	Matches vanilla

Table 3: Hyperparameters for attention-based methods. Both methods are applied with the unmodified released configurations from their official codebases. For LLaVA-NeXT, AGLA's augmented images are precomputed in a separate environment with `transformers` 4.34 (required by LAVIS-BLIP) and consumed by a token-by-token contrastive-decoding loop in `transformers` 4.47.

Method	Parameter	Value
CEI (Fazli et al., 2026b)	Injection layer	10
	Schedule	Cosine on Top-K mass
	Word-start gating δ	0.5
	Per-model scalars	See prose above
AFTER (Wang et al., 2026)	Top-K heads	64
	Intervention α	7
	Edit location	Last token
	Source $\rightarrow$ target	Image+Q $\rightarrow$ Text+Q

Table 4: Architecture-level hyperparameters for hidden-state methods. CEI's per-model scalars $(\alpha_{\max}, \beta, K_{\mathrm{mass}})$ are listed in the preceding paragraph since they vary by LVLM. Decoding-length budgets are pooled in the Decoding Settings section below.

Hidden-state methods. CEI (Fazli et al., 2026b) injects a contextual embedding into a fixed decoder layer at inference time, with the injection strength scheduled by a cosine over Top-K probability-mass statistics. AFTER (Wang et al., 2026) transfers attention activations from a source pattern (image+query) onto a target pattern (text+query) at the top-K heads selected from AMBER training activations, then continues decoding from the modified state. Both methods leave model parameters untouched.

The architecture-level hyperparameters used for both methods are summarized in Table 4. CEI uses per-model values released by its authors: $(\alpha_{\max}, \beta, K_{\mathrm{mass}}) = (0.40, 0.70, 40)$ for InstructBLIP, $(0.25, 0.55, 40)$ for LLaVA-1.5, and $(0.17, 0.35, 80)$ for LLaVA-NeXT, with injection always at decoder layer 10. AFTER uses K=64 heads with intervention strength α=7, applied at the last token, with the QAO offset MLP pretrained on AMBER for 10 epochs at learning rate 10^{-3}.

Vanilla baseline. The vanilla decoding baseline for each model utilizes its standard, native generation configuration stripped of external modifications—meaning no contrastive terms, no attention reweighting, and no hidden-state calibration. Crucially, to eliminate stochastic decoding variability and isolate the exact impacts of each inference intervention, the vanilla baseline is uniformly restricted to deterministic greedy decoding across all evaluation suites, strictly mirroring the prompt constraints and token length budgets enforced during mitigation testing.

Implementation notes. VCD and M3ID are implemented from the CAAC baselines codebase, with patches limited to (i) adding LLaVA-NeXT model-loading support and (ii) filtering unsupported sampling kwargs in the decoding utilities; the method logic is unmodified. AGLA is implemented from its official repository; we adapt the original POPE evaluation scripts to CHAIR, AMBER, and MMSTAR by removing the "answer with one word" prompt suffix. The evaluation benchmarks (e.g., MMStar) and LVLM weights used in this study are publicly available under their respective open-source licenses

A.2 Decoding Settings

All runs use `seed=42` (Python, NumPy, PyTorch) with greedy decoding for the vanilla baseline. The `max_new_tokens` budget is 512 for CHAIR and AMBER and 32 for MMSTAR. Minor per-codebase deviations (e.g., the AGLA repo's default of multinomial sampling at T=1.0 for its target models) are documented in the released code.

A.3 Benchmark Descriptions

We use three benchmarks chosen to separate object-level hallucination from broader visual informativeness and from general multimodal capability.

The hallucination-focused benchmarks (CHAIR, AMBER) each pair an error metric with an informativeness metric drawn from the *same* response; the capability benchmark (MMSTAR) covers reasoning, science, and mathematics categories that hallucination scores do not measure.

CHAIR (Rohrbach et al., 2018). The Caption Hallucination Assessment with Image Relevance (CHAIR) protocol measures object-existence errors in open-ended captioning on MSCOCO images. Given a generated caption, an automatic pipeline extracts mentioned object nouns, compares them against MSCOCO's ground-truth object annotations for the same image, and reports two complementary rates: CHAIR_s (the fraction of *sentences* containing at least one hallucinated object) and CHAIR_i (the fraction of all mentioned *instances* that are hallucinated). We pair these error metrics with *object recall*, the fraction of ground-truth objects that the caption successfully mentions, which serves as an informativeness proxy. We evaluate on 500 images sampled with `seed=42` from the COCO 2014 validation split; the generation prompt is "*Please describe this image in detail.*".

AMBER (Wang et al., 2023). The Adversarial Multimodal Benchmark for hallucination Evaluation in LVLMs uses a curated set of 1004 generative items spanning object existence, attributes, and inter-object relations, with dense human-annotated object lists per image. We use the full generative split with prompt "*Describe this image.*" AMBER reports two error metrics over the generated response — CHAIR (object-existence hallucination) and Hal (instance-level hallucination rate including attribute and relation errors) — and one informativeness metric, Cover, the fraction of the image's ground-truth objects mentioned. Throughout the paper we use Hal paired with Cover, since this pairing is the most directly comparable to CHAIR's (CHAIR_s, Recall).

MMSTAR (Chen et al., 2024b). MMSTAR is a vision-indispensable multimodal benchmark constructed by filtering existing evaluation suites to remove items that are solvable from text priors alone. The result is 1500 carefully audited multiple-choice questions (250 per category) covering six capability axes: *coarse perception*, *fine-grained perception*, *instance reasoning*, *logical reasoning*, *science & technology*, and *mathematics*. Each item presents the image and a four-option MCQ; we append the VLMEvalKit suffix "*Please select the correct answer from the options above.*" to elicit a single-letter response. We report per-category accuracy and the macro average across the six categories. MMSTAR serves as our general-capability probe: a mitigation method that genuinely improves multimodal grounding should at least preserve performance on these reasoning-heavy tasks, not just on object-existence benchmarks like CHAIR or AMBER.

A.4 Hardware and Software

All generation runs use a single NVIDIA A100-SXM4-80GB GPU. Judge runs are CPU-only. LLaVA-1.5 and InstructBLIP use PyTorch 2.0.1, and `transformers` 4.31. LLaVA-NeXT requires `transformers` 4.47 and uses PyTorch 2.5.1 with bitsandbytes 0.45 for 8-bit quantization. HuggingFace checkpoints used for the three LVLMs are: `llava-hf/llava-v1.6-vicuna-7b-hf`, `llava-hf/llava-1.5-7b-hf`, `Salesforce/instructblip-vicuna-7b`

B Full Results

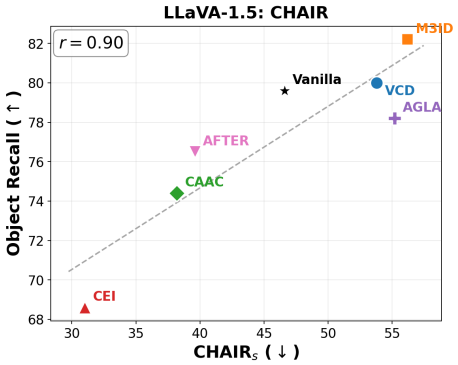
This appendix reports the complete numerical results, per-model scatter plots complementing Figure 1, an alternative bar-plot view of MMSTAR deltas, and full CHAIR, AMBER, and MMSTAR result tables.

B.1 Per-Model Trade-off Scatters

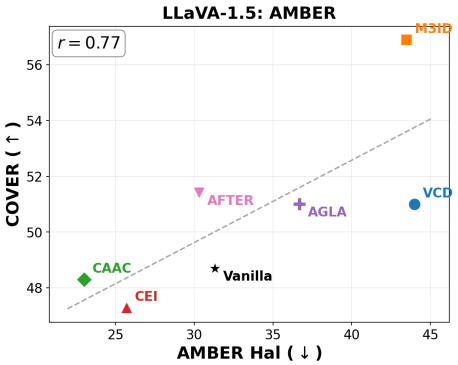
Figures 4 and 5 display the decoupled hallucination–informativeness trade-off scatter plots for LLaVA-1.5 and InstructBLIP, providing cross-architecture support for the faithfulness-informativeness trade-off. Across both additional foundation models: localized drops in hallucination error rates are fundamentally entangled with a loss of detailedness. Perhaps the only exception is LLaVA-1.5 with AMBER (Figure 4b), where AFTER manages to reduce hallucination while increasing the Cover. On the other hand, on CHAIR, the positive association between hallucination error rate and object recall is exceptionally pronounced both for LLaVA-1.5 (Figure 4a) and InstructBLIP (Figure 5a). As illustrated in Figure 4, top-performing mitigation methods like CEI and CAAC minimize nominal error rates entirely via *risk suppression*, forcing the model into the lowest-left margins where object

Method	$C_s \downarrow$	$C_i \downarrow$	Recall $\uparrow$	Len
LLaVA-NeXT	33.0	9.1	64.7	173.0
+ VCD	42.6	10.7	**69.9**	170.6
+ M3ID	40.8	12.3	65.2	189.6
+ CAAC	30.0	8.7	64.8	151.7
+ CEI	30.6	9.8	64.5	140.3
+ AGLA	29.6	**8.2**	59.4	173.2
+ AFTER	**25.8**	9.3	62.9	216.3
LLaVA-1.5	46.6	14.8	79.6	88.1
+ VCD	53.8	16.0	80.0	102.0
+ M3ID	56.2	15.7	**82.2**	94.4
+ CAAC	38.2	12.1	74.4	78.4
+ CEI	**31.0**	**9.5**	68.6	75.3
+ AGLA	55.2	14.7	78.2	101.3
+ AFTER	39.6	10.5	76.5	96.6
InstructBLIP	50.8	15.9	**75.7**	102.1
+ VCD	59.2	17.1	74.8	104.1
+ M3ID	69.8	21.3	74.0	102.0
+ CAAC	38.6	11.6	72.8	94.0
+ CEI	**30.4**	8.0	69.2	103.0
+ AGLA	51.0	13.5	72.4	106.1
+ AFTER	30.6	**7.6**	69.4	97.7

Table 5: Full CHAIR results. Len denotes average response length in tokens. The best and second-best values for each performance metric are boldfaced and underscored respectively.

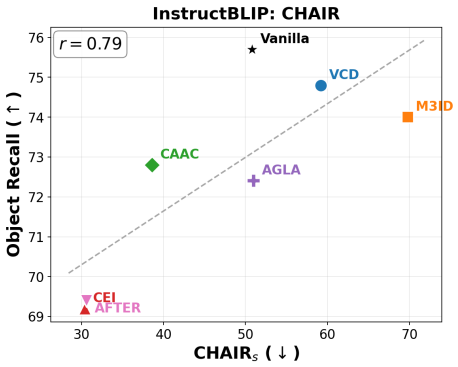


(a) CHAIR: CHAIRs vs Recall

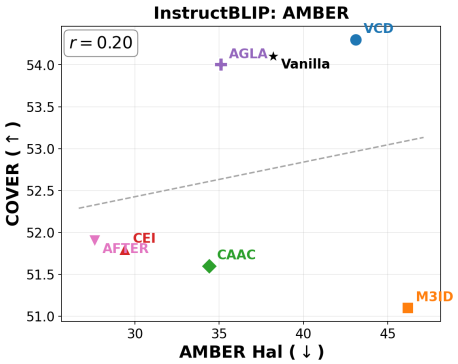


(b) AMBER: Hal vs Cover

Figure 4: Hallucination–informativeness trade-off on LLaVA-1.5.



(a) CHAIR: CHAIRs vs Recall



(b) AMBER: Hal vs Cover

Figure 5: Hallucination–informativeness trade-off on InstructBLIP.

recall sinks below 75% and visual coverage settles under 49%.

B.2 CHAIR

Table 5 reports the complete CHAIR results across the three LVLMs and seven decoding configurations (vanilla plus six mitigation methods). For each cell we report sentence-level hallucination ($CHAIR_s$, lower is better), instance-level hallucination ($CHAIR_i$, lower is better), object recall (higher is better), and the average response length in tokens. Crucially, a key observation from these results is that no single method is capable of simultaneously minimizing hallucination rates while maximizing object coverage. Instead, the configurations exhibit distinct trade-offs; for instance, contrastive decoding methods generally excel at maintaining high coverage, though this often comes at the cost of higher hallucination rates.

B.3 AMBER

Table 6 reports the complete AMBER results. For each (model, method) cell we report AMBER's CHAIR (object-existence hallucination, lower is better), Cover (fraction of ground-truth objects

Method	CHAIR ↓	Cover ↑	Hal ↓	Cog
LLaVA-NeXT	9.2	61.1	50.3	5.0
+ VCD	11.1	**63.6**	59.2	5.3
+ M3ID	13.4	60.4	61.3	5.5
+ CAAC	8.6	60.5	48.2	4.7
+ CEI	10.2	60.7	51.1	5.5
+ AGLA	8.8	59.9	50.3	4.5
+ AFTER	**7.1**	58.7	**35.6**	2.4
LLaVA-1.5	7.4	48.7	31.3	3.4
+ VCD	10.2	51.0	44.0	4.3
+ M3ID	8.0	**56.9**	43.5	3.2
+ CAAC	**5.4**	48.3	**23.0**	2.1
+ CEI	**5.4**	47.3	25.7	1.8
+ AGLA	7.6	51.0	36.7	3.9
+ AFTER	6.4	51.4	30.3	3.6
InstructBLIP	8.4	54.1	38.2	4.1
+ VCD	9.3	**54.3**	43.1	4.3
+ M3ID	10.3	51.1	46.2	4.7
+ CAAC	7.7	51.6	34.4	3.5
+ CEI	6.3	51.8	29.4	2.6
+ AGLA	7.5	54.0	35.1	3.8
+ AFTER	**5.7**	51.9	**27.6**	2.8

Table 6: Full AMBER results. Cog denotes the cognitive metric reported by AMBER. The best and second-best values for each performance metric are boldfaced and underscored respectively.

mentioned, higher is better), and Hal (instance-level hallucination rate including attributes and relations, lower is better). Throughout the main paper we pair Hal with Cover as the hallucination–informativeness signal on AMBER. Mirroring the trade-offs observed on CHAIR, methods that score exceptionally well on reducing hallucination metrics (CHAIR and Hal) often lag significantly behind on Cover. This consistent drop in coverage reinforces our primary claim that **most inference-time mitigation methods rely heavily on a risk-averse strategy—suppressing hallucinations by simply reducing visual commitment** rather than genuinely enhancing multimodal grounding.

B.4 MMStar

Table 7 reports per-category and overall accuracy on MMSTAR for each (model, method) cell. The six category columns are Coarse Perception (CP), Fine-grained Perception (FP), Instance Reasoning (IR), Logical Reasoning (LR), Science & Technology (ST), and Mathematics (MA); the final column is the macro average over the six categories. These absolute accuracies are the source data underlying the per-category change (Δ vs. vanilla) heatmap in Figure 2. Two critical patterns emerge from these absolute scores. First, no single mitigation method consistently performs well across all capability axes or across the three baseline LVLMs, with the granular results exhibiting high volatility. Second, the vanilla decoding configuration proves remarkably competitive, securing the second-highest overall accuracy once (on InstructBLIP) and the third-highest twice (on LLaVA-NeXT and LLaVA-1.5). Taken together, these trends demonstrate that current inference-time mitigation methods fail to reliably improve—and frequently degrade—the holistic multimodal capabilities of the base models.

C Qualitative Examples

To ground our macro findings, Figure 6 presents four verbatim task logs from MMSTAR illustrating the transition from *risk suppression* to *capability cannibalization*. Across all core tasks, the unmitigated base models output natively correct choices (green checkmarks), whereas the mitigation methods consistently produce systematic errors. In Figure 6a, the mitigation methods over-compensate for background density to incorrectly over-count crowd targets (**C**). Similarly, Figure 6b demonstrates that *risk suppression* alters confidence thresholds such that fine-grained details are filtered out, defaulting to understated group counts (**A** or **B**).

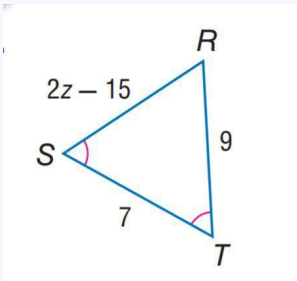
This performance degradation is most acute during complex reasoning. In Figure 6c, the mitigation methods squash vital semantic pathways, causing all six options to collapse onto a false generic attribute (**C**). Additionally, Figure 6d shows that flattening the token distribution space breaks the fragile sequential token commitments vital for exact arithmetic, causing every evaluated framework to deviate into an incorrect ratio fraction (**D**).

Two additional cases that show the failure mode of contrastive decoding methods are shown below where amplifying the visual-dependent tokens can force a wrong prediction.

Model	Method	CP	FP	IR	LR	ST	MA	Overall
LLaVA-NeXT	Vanilla	0.568	0.316	0.408	0.292	0.176	0.252	0.336
	VCD	0.560	0.304	**0.436**	**0.308**	0.220	**0.284**	**0.352**
	M3ID	0.556	0.304	0.400	0.296	0.216	0.236	0.335
	CAAC	0.536	0.288	0.400	0.296	0.204	**0.284**	0.335
	CEI	0.572	**0.328**	0.408	0.288	**0.232**	0.196	0.337
	AGLA	0.544	0.256	0.432	0.228	**0.232**	0.280	0.329
	AFTER	**0.580**	0.288	0.424	0.260	0.192	0.240	0.331
LLaVA-1.5	Vanilla	0.568	0.272	0.388	0.292	0.252	0.196	0.328
	VCD	0.556	0.240	0.400	**0.300**	0.212	0.256	0.327
	M3ID	0.564	**0.316**	0.388	0.280	0.216	0.236	**0.333**
	CAAC	0.488	0.256	0.356	0.292	**0.280**	**0.276**	0.325
	CEI	0.568	0.236	0.376	0.264	0.228	0.212	0.314
	AGLA	0.568	0.240	0.396	0.268	0.148	0.260	0.313
	AFTER	**0.572**	0.248	**0.408**	0.276	0.232	0.208	0.324
InstructBLIP	Vanilla	0.340	0.300	0.388	0.348	0.236	**0.308**	0.320
	VCD	0.316	0.296	0.296	0.292	0.256	0.256	0.285
	M3ID	0.360	**0.304**	0.316	0.324	**0.264**	0.280	0.308
	CAAC	0.340	0.268	0.356	0.300	0.228	0.300	0.299
	CEI	0.348	**0.304**	0.344	0.232	0.228	0.240	0.283
	AGLA	**0.420**	0.296	**0.408**	**0.380**	0.232	0.292	**0.338**
	AFTER	0.412	0.280	0.404	0.368	0.260	0.264	0.331

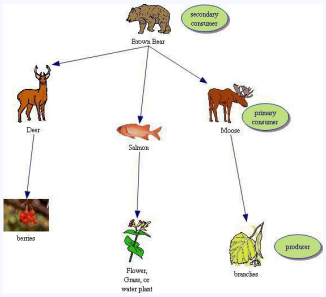
Table 7: Full MMSTAR results across the 3×7 model×method matrix.

Confidently wrong on a math question

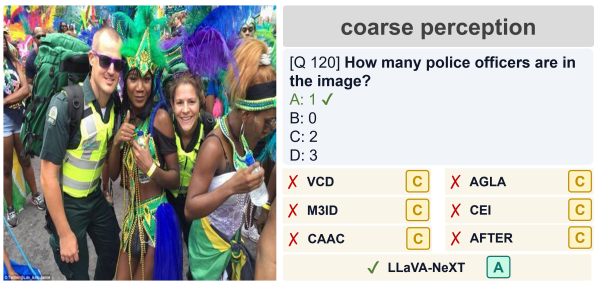


Benchmark: MMSTAR
Model: LLaVA-NeXT
Method: VCD
Category: mathematics
Question: Find z.
Options: (A) 7 (B) 12 (C) 9 (D) 15
Ground truth: B
Model response: *"The correct answer is (D) 15."*
Likely explanation: Wrong (predicted D, correct B). The method commits confidently to a fluent but incorrect option which appears in the image. This could be due to amplifying the image-dependent tokens in VCD.

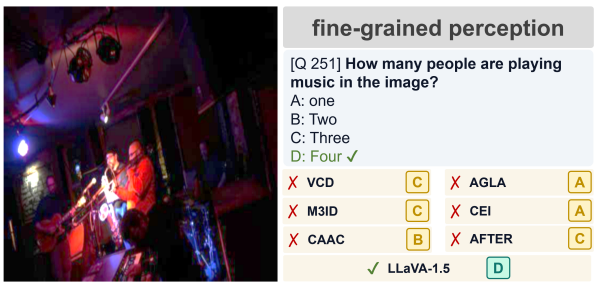
Wrong commitment on a food-web question



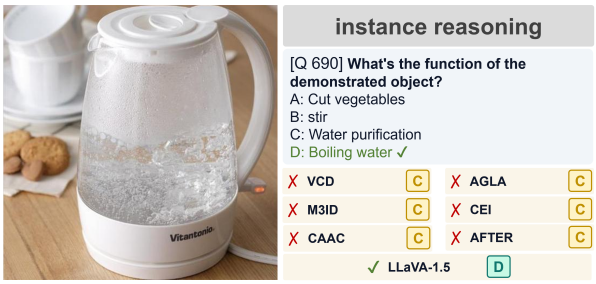
Benchmark: MMSTAR
Model: LLaVA-1.5
Method: M3ID
Category: logical reasoning
Question: According to the given food web, which is the animal that eats other animals and also becomes a prey to other animals?
Options: (A) salmon (B) deer (C) moose (D) none of the above
Ground truth: D
Model response: *"A: salmon."*
Likely explanation: Wrong (predicted A, correct D). The method commits to a specific incorrect option which appears in the image rather than reasoning about the question, a side effect of amplifying image-dependent tokens.



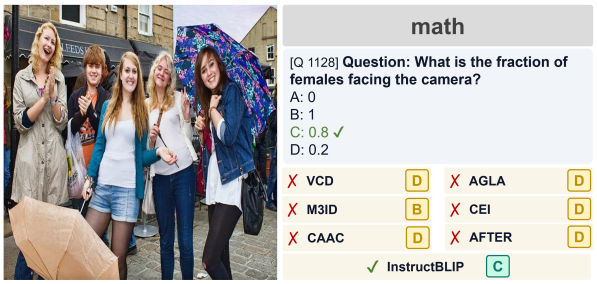
(a) Coarse Perception Failure



(b) Fine-Grained Perception Failure



(c) Instance Reasoning Failure



(d) Mathematical Ratio Calculation Failure

Figure 6: Extended comparative case studies on MMSTAR. Across perception, instance tracking, and fractional calculation, the base models' vanilla decoding output correct choices (green checkmarks). In contrast, the six inference-time mitigation strategies consistently induce systematic prediction failures across all evaluation tasks.