CausalLens: Sensitivity-Guided Multi-Head Causal Intervention for Hallucination Mitigation in Large Vision-Language Models

Supplementary Material

1. More Details about Benchmarks

We evaluate our model on five representative benchmarks that jointly assess grounding faithfulness (i.e., resistance to hallucination), perceptual sensitivity, and general multimodal reasoning capabilities in large vision-language models (LVLMs).

POPE [4] (Polling-based Object Probing Evaluation) is a discriminative benchmark designed to quantitatively evaluate object hallucination. To assess robustness across diverse image distributions, POPE is constructed using validation images from three distinct datasets: **MSCOCO** [5], **A-OKVQA** [7], and **GQA** [2]. For each dataset, 500 images are sampled, and binary "Yes/No" questions (e.g., "Is there a {object} in the image?") are generated. The evaluation comprises three splits based on the negative sampling strategy: (1) *Random*: negative objects are randomly sampled from the dataset vocabulary; (2) *Popular*: negative objects are selected from the most frequent categories in the dataset; (3) *Adversarial*: negative objects are chosen based on co-occurrence frequency with ground-truth objects, specifically targeting the model's reliance on language priors. In total, the benchmark consists of 9,000 questions per dataset (3,000 per split), balanced between positive and negative samples. We report Accuracy, Precision, Recall, F1-score, and the "Yes" ratio to measure both performance and the tendency for over-affirmation.

MMHal-Bench [8](Multimodal Hallucination Benchmark) is a challenging open-ended VQA benchmark specifically crafted to trigger and penalize hallucinations in LVLMs. It consists of 96 carefully designed image-question pairs sourced from the OpenImages validation set, covering 12 diverse object categories and eight hallucination-sensitive question types: object existence, attributes, relations, counting, position, environment, adversarial objects, and complex reasoning. Responses are scored automatically (for short answers) and with GPT-4o (for informativeness and groundedness), yielding a hallucination rate (lower is better) and an informativeness score. This benchmark emphasizes fine-grained misalignment between vision and language, making it highly sensitive to subtle hallucinations.

CHAIR [6] (Caption Hallucination Assessment with Image Relevance) is a classic metric for quantifying object hallucination in open-ended image captioning. It is computed on the MSCOCO captioning subset (5,000 images from val2014, often referred to as the "Karpathy test split" in hallucination studies). Using the fixed prompt "Please describe this image in detail," generated captions are parsed for MSCOCO object mentions (via WordNet synonyms). Hallucinated objects are those mentioned but absent according to ground-truth instance segmentations. Two variants are reported: $\mathbf{CHAIR}_i$ (C_i, the average ratio of hallucinated object instances per caption) and $\mathbf{CHAIR}_s$ (C_s, the percentage of captions containing at least one hallucinated object). Lower values indicate better grounding faithfulness.

MME [1] (Multimodal Evaluation) is a comprehensive perception+cognition benchmark with 2,800 yes/no questions across 14 subtasks, all manually annotated to avoid data contamination. It comprises two major components: (i) *Perception* (full score 2,000) with 10 subtasks testing basic visual abilities (existence, count, position, color, OCR, poster, celebrity, scene, landmark, artwork); and (ii) *Cognition* (full score 800) with 4 subtasks (commonsense reasoning, numerical calculation, text translation, code reasoning). Each subtask has a maximum score of 200, and the overall score is the sum. MME emphasizes short-answer robustness and is widely used to measure both low-level perception and higher-order reasoning without hallucination bias in yes/no responses.

LLaVA-Bench (In-the-Wild) is a challenging open-ended benchmark targeting complex real-world multimodal understanding. It contains 24 diverse images (indoor/outdoor scenes, memes, artwork, sketches, abstract images, etc.) paired with 60 expert-curated questions of three types: conversational (simple QA), detailed description, and complex reasoning. Each image is accompanied by a high-quality ground-truth description. Evaluation is performed by GPT-4 (text-only), which scores model responses (1–10 scale) against reference answers generated from detailed annotations, considering helpfulness, relevance, accuracy, and level of detail. Relative GPT-4 scores are reported, with higher values indicating stronger instruction-following and reasoning in unconstrained, "in-the-wild" scenarios.

2. Implementation and Reproducibility Details

In VCD [3] method, the decoding probability is computed as follows:

$$p_{\text{vcd}}(y \mid v, v', x) = \text{softmax}\Big[(1+\alpha_{vcd})\,\text{logit}_\theta(y \mid v, x) - \alpha_{vcd}\,\text{logit}_\theta(y \mid v', x)\Big], \quad (1)$$

where α_{vcd} is a tunable coefficient that controls the strength of contrastive decoding. Here, v denotes the original image, and v' represents a corrupted version of the image obtained

GPT-4o Prompt for MMHAL-Bench Evaluation

You are a precise evaluator for vision-language model outputs. Given an image, a user question, the model's answer, and the ground-truth object labels, your task is to assess the model output based on:

1. **Hallucination**: Does the response mention any object, attribute, or detail that is not present in the image? Use the ground-truth labels as reference.
2. **Informativeness**: How informative is the answer in addressing the question, on a scale of 1 (uninformative or vague) to 5 (specific, accurate, and helpful)?

Please respond in the following JSON format:

```
{
  "hallucination": "Yes" or "No",
  "informativeness": 1 to 5,
  "justification": "Brief explanation (1-2 sentences)"
}
```

Table 1. GPT-4o Prompt for MMHAL-Bench Evaluation

Description:
An AI model designed to evaluate and score the accuracy and detailedness of image descriptions.

Instructions:
You are tasked with evaluating the performance of two AI assistants in describing a given image. Your evaluation will focus on two criteria: accuracy and detailedness. Accuracy is determined by identifying any hallucinations—i.e., parts of the description that are inconsistent with the image. Detailedness refers to how comprehensive the description is, excluding hallucinated content. For each assistant, you will assign a score between 1 and 10 for both accuracy and detailedness. After scoring, provide a rationale for your assessment, ensuring that your explanation is unbiased and not influenced by the order in which the responses are presented.

Input format:
[Assistant 1[
{Response 1}
[End of Assistant 1[

[Assistant 2[
{Response 2}
[End of Assistant 2[

Output format:
Accuracy:
Scores for both responses:
Reasoning:

Detailedness:
Scores for both responses:
Reasoning:

Table 2. **GPT-4V-Aided Evaluation Setup.** This table outlines the prompt used to instruct GPT-4V in evaluating LVLM responses based on accuracy and detailedness.

by applying diffusion-based noise. This contrastive formulation encourages the model to focus on visual features that are robust to noise, thereby improving grounding consistency during generation.

We implement VAF [9] as a baseline. VAF modifies the hidden representation at each attention layer by amplifying

the influence of visual tokens and suppressing that of system tokens:

$$\widehat{Z}_{l,h} = Z_{l,h} + \alpha_{vaf} \cdot M^{\text{enh}}_{l,h} \circ Z_{l,h} - \beta_{vaf} \cdot M^{\text{sup}}_{l,h} \circ Z_{l,h}, \tag{2}$$

where $Z_{l,h}$ is the original output from layer l and head h, and $\circ$ denotes element-wise multiplication. The coefficient $\alpha_{vaf} > 0$ controls the amplification strength of visual features, while $\beta_{vaf} \in (0, 1)$ determines the level of suppression applied to system tokens. The enhancement and suppression masks are defined as:

$$\begin{aligned} M^{\text{enh}}_{l,h}(i,j) &= \mathbb{I}(i \in \mathcal{T},\ j \in \mathcal{V}), \\ M^{\text{sup}}_{l,h}(i,j) &= \mathbb{I}(i \in \mathcal{T},\ j \in \mathcal{S}), \end{aligned} \tag{3}$$

where $\mathcal{T}$, $\mathcal{V}$, and $\mathcal{S}$ denote the sets of query tokens, visual tokens, and system prompt tokens, respectively.

For the baselines VCD and VAF, we maintain uniform experimental hyperparameters, which are detailed in the tables below.

Hyperparameters	**Value**
Amplification Factor α_{vcd}	1
Diffusion Noise Step	999

Table 3. VCD Hyperparameters Settings

Hyperparameters	**Value**
Enhancement Coefficient α_{vaf}	0.15
Suppression Coefficient β_{vaf}	0.1
Layer l	8-15

Table 4. VAF Hyperparameters Settings

We implement a self-correcting decoding framework (DeGF) [10] as a baseline, which leverages generative feedback to refine token predictions and mitigate hallucinations. At each timestep t, two output distributions are generated: one conditioned on the original image v and another conditioned on a synthesized visual reference v'. The divergence between these distributions is measured using Jensen-Shannon (JS) divergence:

$$d_t(v, v') = \mathcal{D}_{\text{JS}}\left(p_\theta(y_t \mid v, \mathbf{x}, \mathbf{y}_{<t}) \parallel p_\theta(y_t \mid v', \mathbf{x}, \mathbf{y}_{<t})\right). \tag{4}$$

If the JS divergence falls below a threshold γ, the original prediction is confirmed. Otherwise, it is revised based on the feedback from the generated reference. Specifically, the final token is sampled from a weighted combination of the two distributions under the following conditions to determine the next token y_t during decoding.

Condition 1. If $d_t(v, v') < \gamma$, we confirm the original prediction by enhancing it with the generated visual reference:

$$y_t \sim p_\theta(y_t) = \text{Softmax}\begin{bmatrix} f_\theta(y_t \mid v, \mathbf{x}, \mathbf{y}_{<t}) \\ +\alpha_1 f_\theta(y_t \mid v', \mathbf{x}, \mathbf{y}_{<t}) \end{bmatrix}. \tag{5}$$

Condition 2. If $d_t(v, v') \geq \gamma$, we revise the original prediction by adjusting it based on the generated reference:

$$y_t \sim p_\theta(y_t) = \text{Softmax}\begin{bmatrix} (1+\alpha_2)\, f_\theta(y_t | v, \mathbf{x}, \mathbf{y}_{<t}) \\ -\alpha_2 f_\theta(y_t | v', \mathbf{x}, \mathbf{y}_{<t}) \end{bmatrix}. \tag{6}$$

In our experiments, we set $\alpha_1 = 3$, $\alpha_2 = 1$, and $\gamma = 0.1$ for the baseline decoding process.

Mechanistic Intervention and Terminology. Throughout this paper, the term "causal" explicitly refers to a *mechanistic intervention* on the computational graph—specifically, restoring the $V \rightarrow H \rightarrow Y$ causal pathway. While attention variance is utilized as a statistical heuristic to select mediating heads, their mechanistic necessity is strictly validated through the ablation studies presented in the main text. Furthermore, during intervention, we deliberately subtract the hidden states of system tokens $H^{(sys)}$ rather than user instruction tokens $H^{(text)}$. This design choice is driven by our observation that system prompt tokens disproportionately act as "attention sinks" (occupying over 60% of the attention mass), whereas $H^{(text)}$ contains essential user commands that, if suppressed, would damage the model's core instruction-following capabilities. Additionally, please note that the visualizations provided in the main text (e.g., attention distributions and variance plots) are derived using the LLaVA-v1.5-7B model. The qualitative generation examples depict snapshots of the autoregressive generation process at a specific timestep (i.e., next-token prediction state), rather than finalized sentences.

3. More Experimental Results and Analysis

3.1. Full Results on LLaVA-Bench

The results presented in Table 5 demonstrate the effectiveness of our approach in improving both accuracy and detailedness on LLaVA-Bench. Our method consistently outperforms other hallucination mitigation strategies, such as VCD and VAF, on the LLaVA-v1.5-7B model. Specifically, we observe significant gains in accuracy and detailedness, with our method achieving a notable increase over regular methods and other baselines. These improvements highlight the robustness of our approach in generating more accurate and detailed image descriptions, suggesting its potential for enhancing large vision-language models in real-world tasks.

Method	LLaVA-v1.5-7B	
	Acc. ↑	Det. ↑
Regular	3.76	4.19
Ours	**4.39**	**5.24**
VCD	3.92	4.26
Ours	**4.24**	**4.78**
VAF	3.96	4.38
Ours	**4.18**	**4.62**

Table 5. **GPT-4V-aided evaluation on LLaVA-Bench.**

3.2. Ablation and Domain Generalization Analysis

Effect of the mid-layer. We study the effect of the mid-layer intervention window on model performance by varying the range of layers to which CausalLens is applied. The results on the POPE (MSCOCO Popular) benchmark using the LLaVA-v1.5-7B model are reported in Table 6. The original fixed window L10–L20 (used in all main experiments) already achieves strong performance (POPE Accuracy 88.20%, F1 87.66%). Slightly broader or shifted windows (L10–L22, L12–L20, L12–L22) yield identical or marginally lower scores, demonstrating that the method is robust to the exact choice of mid-layer range.

Mid-Layer Window	POPE Acc↑	POPE F1↑
L8–L16	87.76	87.47
L8–L18	87.83	87.58
L8–L20	87.83	87.59
L10–L18	87.90	87.52
L10–L20	88.20	87.66
L10–L22	88.20	87.50
L12–L20	88.20	87.50
L12–L22	88.20	87.50
L12–L24	88.17	87.46

Table 6. Ablation study of mid-layer window on the POPE (MSCOCO Popular) benchmark, using the LLaVA-v1.5-7B model.

To further ensure that the structural preference for mid-layers generalizes, we evaluated applying our intervention to the front 30%, middle 30%, and last 30% of transformer layers across different models. As demonstrated in Table 7, the chosen middle layers consistently achieve the best or highly competitive performance. This confirms our hypothesis that autoregressive transformers undergo a critical "vision-to-text" semantic transition in the middle layers, making it the optimal intervention window.

Model	L_{10}–L_{20}	Front 30%	Mid 30%	Last 30%
LLaVA-v1.5-7B	**85.19**	84.74	**85.19**	84.86
LLaVA-v1.5-13B	**85.94**	85.36	85.66	85.32
Qwen2-VL-7B	**86.66**	86.49	**86.66**	86.34

Table 7. POPE COCO Adversarial F1 score comparison across different structural layer groups.

Sensitivity of Intervention Strength (λ). We also verified the robustness of the hyperparameter λ using Qwen2-VL-7B-Instruct. We evaluated its performance across three distinct benchmarks: MME Color, POPE-COCO (Adversarial), and POPE-GQA (Random). As shown in Table 8, the model's performance consistently peaks around $\lambda = 0.15$ and remains generally superior to the VAF baseline across a broad range of values.

Dataset	VAF	λ				
		0.05	**0.10**	**0.15**	**0.20**	**0.25**
MME Color	180	**185**	**185**	**185**	180	180
MSCOCO	86.87	86.87	87.07	**87.10**	86.93	86.93
GQA	89.87	90.13	90.27	**90.60**	90.57	90.53

Table 8. λ sensitivity analysis on Qwen2-VL-7B-Instruct compared against VAF.

Robustness to System Prompt Length. To determine if our method is sensitive to the token length of the system prompt, we evaluated Qwen2-VL-7B-Instruct on the POPE (COCO-Popular) split using short (26 tokens), regular (42 tokens), and long (67 tokens) system prompts. Table 9 demonstrates that our approach maintains a stable and consistent performance gain over the baseline, independent of the length of the provided system instructions.

Method	Short (26 token)		Reg. (42 token)		Long (67 token)	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**
VAF	87.93	86.87	88.10	87.06	88.30	87.29
Ours	**88.73**	**88.03**	**88.90**	**88.30**	**89.03**	**88.45**

Table 9. Robustness on POPE evaluating under different system-prompt lengths.

Domain Generalization Analysis. To further evaluate the effectiveness of our method in mitigating hallucination beyond natural images, we conduct experiments on two challenging subsets of the MME benchmark: *OCR* and *Artwork*. These subsets are known to pose difficulties

for vision-language models due to low-level visual distortions and domain shift. As shown in Table 10, our method achieves the best performance on both tasks, surpassing the baseline, VCD, DeGF and VAF variants. This demonstrates the robustness and generalization capability of our approach in handling non-natural and artistic image domains.

Method	OCR ↑	Artwork ↑
Baseline	120.00	118.00
VCD	127.50	120.75
DeGF	132.50	120.75
VAF	132.50	121.50
Ours	**137.50**	**126.75**

Table 10. Evaluation on the OCR and Artwork subsets of the MME benchmark using LLaVA-v1.5-7B. Our method shows improved hallucination mitigation on non-natural and artistic images.

Impact on Global Context and Language Priors. One might intuitively assume that enhancing visual sensitivity (attention variance) predominantly favors object-centric tasks at the expense of global contextual reasoning. However, high visual sensitivity essentially denotes *selectivity* rather than just object-centricity. Even for global tasks, the model must accurately focus on relevant image patches rather than attending diffusely. Empirically, CausalLens outperforms baselines on MME subtasks requiring global structural reasoning (e.g., *Position*, *Count*), proving that suppressing diffuse heads clarifies rather than harms holistic understanding.

Moreover, our method is highly self-damping: ambiguous visual inputs yield lower sensitivity scores ($s_{l,i}$), which automatically scales down the intervention to avoid amplifying noise. Because our intervention precisely targets the system-token attention sink without dampening useful textual reasoning pathways, it successfully preserves helpful language priors. This allows our model to achieve sustained improvements even in categories heavily reliant on complex linguistic reasoning, such as the "Reasoning" and "Holistic" categories in MMHal-Bench.

4. Limitations and Future Directions

Despite its strong performance and efficiency, CausalLens has several limitations. The number of intervened layers remains fixed, and the intervention is currently applied only to mid-layers to preserve fluency, leaving strong early-layer (L0–L3) and occasional late-layer visual signals unused. All evaluations focus on short single-turn tasks; long-form generation and multi-turn conversations, where hallucinations accumulate severely, have not been tested. Finally, although no harm is observed on general reasoning benchmarks, stronger visual grounding may hurt tasks that legitimately rely on linguistic priors.

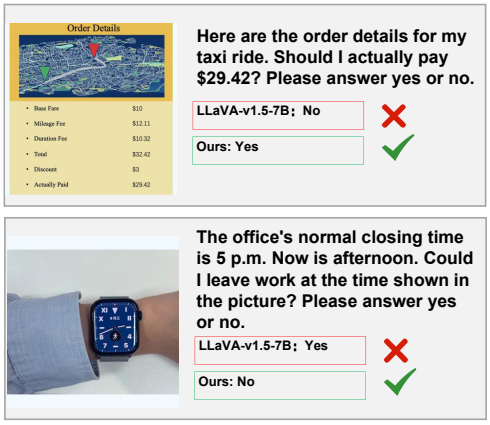

Figure 1. Some examples of complex visual understanding.

Future work will explore fully adaptive layer count and weighting, safe extension to early/late layers, systematic long-sequence and conversational evaluation, and lightweight hybrid training-interventions to push hallucination rates even lower while retaining real-time efficiency.

References

[1] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1

[2] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1

[3] 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*, pages 13872–13882, 2024. 1

[4] 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*, pages 292–305, 2023. 1

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 1

[6] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image cap-

tioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. 1

[7] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 1

[8] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13088–13110, 2024. 1

[9] Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14625–14634, 2025. 2

[10] Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia P Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3